Auton Agent Multi-Agent Syst
DOI 10.1007/s10458-009-9092-y

A probabilistic multimodal approach for predicting listener backchannels

Louis-Philippe Morency · Iwan de Kok · Jonathan Gratch

Abstract During face-to-face interactions, listeners use backchannel feedback such as head nods as a signal to the speaker that the communication is working and that they should continue speaking. Predicting these backchannel opportunities is an important milestone for building engaging and natural virtual humans. In this paper we show how sequential probabilistic models (e.g., Hidden Markov Model or Conditional Random Fields) can automatically learn from a database of human-to-human interactions to predict listener backchannels using the speaker multimodal output features (e.g., prosody, spoken words and eye gaze). The main challenges addressed in this paper are automatic selection of the relevant features and optimal feature representation for probabilistic models. For prediction of visual backchannel cues (i.e., head nods), our prediction model shows a statistically significant improvement over a previously published approach based on hand-crafted rules.

Keywords Listener backchannel feedback · Nonverbal behavior prediction · Sequential probabilistic model · Conditional random field · Head nod · Multimodal

1 Introduction

Natural conversation is fluid and highly interactive. Participants seem tightly enmeshed in something like a dance, rapidly detecting and responding, not only to each other's words, but

L.-P. Morency (✉) · J. Gratch
Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Rey, CA 90292, USA
e-mail: morency@ict.usc.edu

J. Gratch
e-mail: gratch@ict.usc.edu

I. de Kok
Human Media Interaction Group, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands
e-mail: i.a.dekok@student.utwente.nl

Published online: 10 May 2009

to speech prosody, gesture, gaze, posture, and facial expression movements. These "extra-linguistic" signals play a powerful role in determining the nature of a social exchange. When these signals are positive, coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in such diverse areas as negotiations and conflict resolution [9,13], psychotherapeutic effectiveness [37], improved test performance in classrooms [10] and improved quality of child care [4].

Not surprisingly, supporting such fluid interactions has become an important topic of virtual human research. Most research has focused on individual behaviors such as rapidly synthesizing the gestures and facial expressions that co-occur with speech [5,25,22,35] or real-time recognition the speech and gesture of a human speaker [30,8]. But as these techniques have matured, virtual human research has increasingly focused on dyadic factors such as the feedback a listener provides in the midst of the other participants speech [16,23]. These include recognizing and generating backchannel or jump-in points [39] turn-taking and floor control signals, postural mimicry [14] and emotional feedback [19,1]. In particular, backchannel feedback (the nods and paravebals such as "uh-huh" and "mm-hmm" that listeners

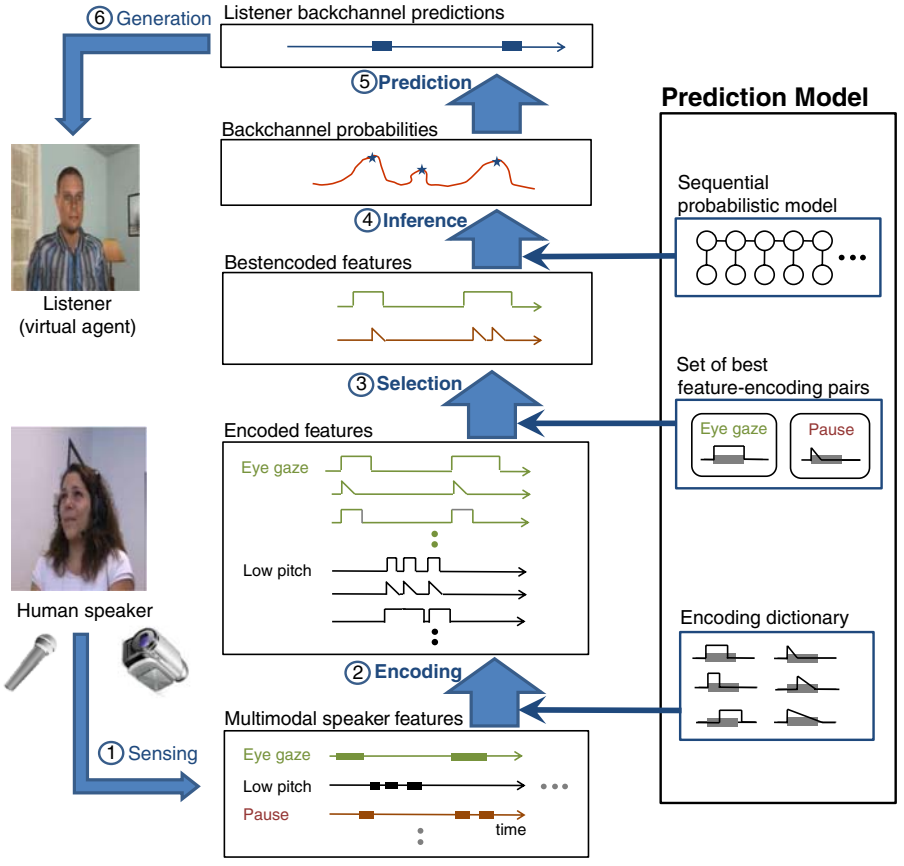

Fig. 1 Our prediction model is designed for generating in real-time backchannel feedback for a listener virtual agent. It uses speaker multimodal features such as eye gaze and prosody to make predictions. The timing of the backchannel predictions and the optimal subset of features is learned automatically using a sequential probabilistic model

produce as some is speaking) has received considerable interest due to its pervasiveness across languages and conversational contexts and this paper addresses the problem of how to predict and generate this important class of dyadic nonverbal behavior.

Generating appropriate backchannels is a notoriously difficult problem. Listener backchannels are generated rapidly, in the midst of speech, and seem elicited by a variety of speaker verbal, prosodic and nonverbal cues. Backchannels are considered as a signal to the speaker that the communication is working and that they should continue speaking [40]. There is evidence that people can generate such feedback without necessarily attending to the content of speech [3], and this has motivated a host of approaches that generate backchannels based solely on surface features (e.g., lexical and prosodic) that are available in real-time.

This paper describes a general probabilistic framework for learning to predict and generate dyadic conversational behavior from multimodal conversational data, and applies this framework to listener backchanneling behavior. As shown in Fig. 1, our approach is designed to generate real-time backchannel feedback for virtual agents. The paper provides several advances over prior art. Unlike prior approaches that use a single modality (e.g., speech), we incorporate multi-modal features (e.g., speech and gesture). We present a machine learning method that automatically selects appropriate features from multimodal data and produces sequential probabilistic models with greater predictive accuracy than prior approaches. The general probabilistic framework presented in this paper was originally published at the IVA 2008 conference [29].

The following section describes previous work in backchannel generation and explains the differences between our prediction model and other predictive models. Section 3 describes the details of our prediction model including the encoding dictionary and our feature selection algorithm. Section 4 presents the way we collected the data used for training and evaluating our model as well as the methodology used to evaluate the performance of our prediction model. In Sect. 5 we discuss our results and conclude in Sect. 6.

2 Previous work

Several researchers have developed models to predict when backchannel should happen. In general, these results are difficult to compare as they utilize different corpora and present varying evaluation metrics. In fact, we are not aware of a paper that makes a direct comparison between alternative methods.

Ward and Tsukahara [39] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110 ms during speech. Models were produced manually through an analysis of English and Japanese conversational data.

Nishimura et al. [31] present a unimodal decision-tree approach for producing backchannels based on prosodic features. The system analyzes speech in 100 ms intervals and generates backchannels as well as other paralinguistic cues (e.g., turn taking) as a function of pitch and power contours. They report a subjective evaluation of the system where subjects were asked to rate the timing, naturalness and overall impression of the generated behaviors but no rigorous evaluation of predictive accuracy.

Cathcart et al. [6] propose a unimodal model based on pause duration and trigram part-of-speech frequency. The model was constructed by identifying, from the HCRC Map Task Corpus [2], trigrams ending with a backchannel. For example, the trigram most likely to predict a backchannel was (⟨NNS⟩ ⟨pau⟩ ⟨bc⟩), meaning a plural noun followed by a pause of at least 600 ms. The algorithm was formally evaluated on the HCRC data set, though there was no direct comparison to other methods. As part-of-speech tagging is a challenging

requirement for a real-time system, this approach is of questionable utility to the design of interactive virtual humans.

Fujie et al. [11] used Hidden Markov Models to perform head nod recognition. In their paper, they combined head gesture detection with prosodic low-level features from the same person to determine strongly positive, weak positive and negative responses to yes/no type utterances.

Maatman et al. [27] present a multimodal approach where Ward and Tsukhara's prosodic algorithm is combined with a simple method of mimicking head nods. No formal evaluation of the predictive accuracy of the approach was provided but subsequent evaluations have demonstrated that generated behaviors do improve subjective feelings of rapport [20] and speech fluency [14].

To our knowledge, no system has demonstrated how to automatically learn a predictive model of backchannel feedback from multi-modal conversational data nor have there been definitive head-to-head comparisons between alternative methods.

3 Prediction model

The goal of our prediction model is to create real-time predictions of listener backchannel based on multimodal features from the human speaker. Our prediction model learns automatically which speaker feature is important and how they affect the timing of listener backchannel. For example, in our experiments described in Sect. 4 we learn a prediction model of when a listener is likely to head nod based on the speaker actions (prosody, pause, visual gestures and spoken words). We achieve this goal by using a machine learning approach: we train a sequential probabilistic model from a database of human–human interactions and use this trained model in a real-time backchannel generator (as depicted in Fig. 1).

3.1 Sequential probabilistic model

A sequential probabilistic model takes as input a sequence of observation features (e.g., the speaker features) and returns a sequence of probabilities (i.e., probability of listener backchannel). Two of the most popular sequential models are Hidden Markov Model (HMM) [33] and Conditional Random Field (CRF) [24]. One of the main difference between these two models is that CRF is discriminative (i.e., tries to find the best way to differentiate cases where the listener gives backchannel to cases where it does not) while HMM is generative (i.e., tries to find the best way to generalize the samples from the cases where the listener gives backchannel without looking at the cases where the listener did not give backchannel). Our prediction model is designed to work with both types of sequential probabilistic models.

Sequential probabilistic models such as HMM and CRF have some constraints that need to be understood and addressed before using them:

- *Limited learning*: The more informative your features are, the better your sequential model will perform. If the input features are too noisy (e.g., direct signal from microphone), it will make it harder for the HMM or CRF to learn the important part of the signal. Also, because of the pre-processing your input features to highlight their influences on your label (e.g., listener backchannel) you improve your chance of success.
- *Over-fitting* The more complex your model is, the more training data it needs. Every input feature that you add increases its complexity and at the same time its need for a larger

training set. Since we usually have a limited set of training sequences, it is important to keep the number of input features low.

In our prediction model we directly addressed these issues by focusing on the feature representation and feature selection problems:

- *Encoding dictionary*: To address the limited learning constraint of sequential models, we suggest to use more than binary encoding to represent input features. Our encoding dictionary contains a series of encoding templates that were designed to model different relationship between a speaker feature (e.g., a speaker in not currently speaking) and listener backchannel. The encoding dictionary and its usage are described in Sect. 3.2.
- *Automatic feature and encoding selection*: Because of the over-fitting problem happening when too many uncorrelated features (i.e., features that do not influence listener backchannel) are used, we suggest two techniques for automatic feature and encoding selection based on co-occurrence statistics and performances evaluation on a validation dataset. Our feature selection algorithms are described in Sect. 3.3.2.

The following two sections describe our encoding dictionary and feature selection algorithm. Section 3.4 describes how the probabilities output from our sequential model are used to generate backchannel.

3.2 Encoding dictionary

The goal of the encoding dictionary is to propose a series of encoding templates that capture the coarse relationship between speaker features and listener backchannel. These encoding templates will help to represent long-range dependencies (when the influence of an input feature decay slowly, possibly with a delay) that are otherwise hard to learn using a sequential probabilistic model. An example of a long-range dependency will be the effect of low-pitch regions on backchannel feedback with an average delay of 0.7 s (observed by Ward and Tsukahara [39]). In our framework, the prediction model will pick an encoding template with a 0.5 s delay and the exact alignment will be learned by the sequential probabilistic model which will also take into account the influence of other input features.

The Fig. 2 shows the 13 encoding templates used in our experiments. These encoding templates were selected to represent a wide range of ways that a speaker feature can influence the listener backchannel. These encoding templates were also selected because they can easily be implemented in real-time since the only needed information is the start time of the speaker feature. Only the binary feature also uses the end time. In all cases, no knowledge of the future is needed.

The three main types of encoding templates are:

- *Binary encoding*: This encoding is designed for speaker features which influence on listener backchannel is constraint to the duration of the speaker feature.
- *Step function*: This encoding is a generalization of binary encoding by adding two parameters: width of the encoded feature and delay between the start of the feature and its encoded version. This encoding is useful if the feature influence on backchannel is constant but with a certain delay and duration.
- *Ramp function*: This encoding linearly decreases for a set period of time (i.e., width parameter). This encoding is useful if the feature influence on backchannel is changing over time.

It is important to note that a feature can have an *individual* influence on backchannel and/or a *joint* influence. An *individual* influence means the input feature directly influences

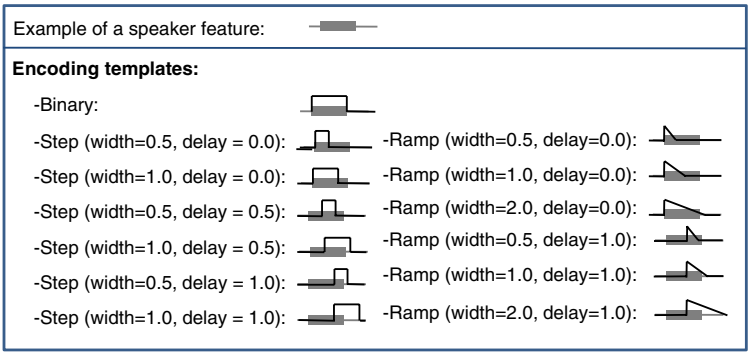

Fig. 2 *Encoding dictionary*: This figure shows the different encoding templates used by our prediction model. Each encoding templates were selected to model different relationships between speaker features (e.g., a pause or an intonation change) and listener backchannels. We included a delay parameter in our dictionary since listener backchannels can sometime happen later after speaker features (e.g., Ward and Tsukahara [39]). This encoding dictionary gives a more powerful set of input features to the sequential probabilistic model which improves the performance of our prediction model

listener backchannel. For example, a long pause can by itself trigger backchannel feedback from the listener. A *joint* influence means that more than one feature is involved in triggering the feedback. For example, saying the word "and" followed by a look back at the listener can trigger listener feedback. This also means that a feature may need to be encoded more than one way since it may have a *individual* influence as well as one or more *joint* influences.

One way to use the encoding dictionary with a small set of features is to encode each input feature with each encoding template. We tested this approach in our experiment with a set of 12 features (see Sect. 5) but because of the problem of over-fitting, a better approach is to select the optimal subset of input features and encoding templates. The following section describes our feature selection algorithm.

3.3 Automatic feature selection

We perform the feature selection based on the same concepts of *individual* and *joint* influences described in the previous section. Individual feature selection is designed to asses the individual performance of each speaker feature while the joint feature selection looks at how features can complement each other to improve performance.

3.3.1 Individual feature selection

Individual feature selection is designed to do a pre-selection based on (1) the statistical co-occurrence of speaker features and listener backchannel, and (2) the individual performance of each speaker feature when trained with any encoding template and evaluated on a validation set.

The first step of individual selection looks at statistics of co-occurrence between backchannel instances and speaker features. The number of co-occurrence is equal to the number of times a listener backchannel instance happened between the start time of the speaker feature and up to 2 s after it. This threshold was selected after analysis of the average co-occurrence histogram for all features. After this step the number of features is reduced to 50.

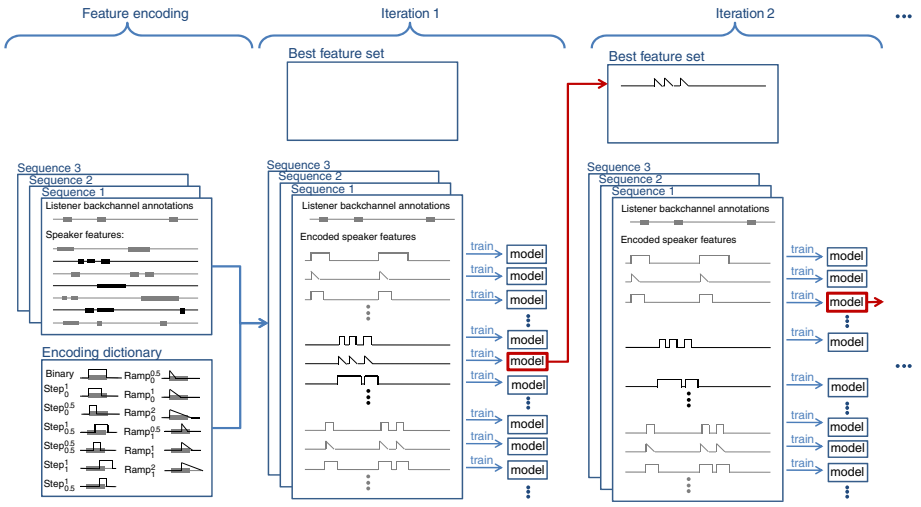

Fig. 3 *Joint Feature selection*: This figure illustrates the feature encoding process using our encoding dictionary as well as two iterations of our joint feature selection algorithm. The goal of joint selection is to find a subset of features that best complement each other for prediction of listener backchannel

The second step is to look at the best performance an individual feature can reach when trained with any of the encoding templates in our dictionary. For each top-50 feature, 13 probabilistic sequential model (e.g., HMM or CRF) are trained (one for each encoding template) and then evaluated. A ranking is made based on the best performance of each individual feature (out of all 13 encoding templates) and a subset of 12 best features is selected.

3.3.2 Joint feature selection

Given the subset of features that performed best when trained individually, we now build the complete set of feature hypothesis to be used by the joint feature selection process. This set represents each feature encoded with all possible encoding templates from our dictionary. The goal of joint selection is to find a subset of features that best complements each other for prediction of backchannel. Figure 3 shows the first two iterations of our algorithm.

The algorithm starts with the complete set of feature hypothesis and an empty set of *best* features. At each iteration, the best feature hypothesis is selected and added to the best feature set. For each feature hypothesis, a sequential model is trained and evaluated using the feature hypothesis and all features previously selected in the best feature set. While the first iteration of this process is really similar to the individual selection, every iteration afterward will select a feature that best complement the current best features set. Note that during the joint selection process, the same feature can be selected more than once with different encodings. The procedure stops when the performance starts decreasing.

3.4 Generating listener backchannel

The goal of the prediction step is to analyze the output from the sequential probabilistic model (see example in Fig. 1) and make discrete decision about when backchannel should happen. The output probabilities from HMM and CRF models are smooth over time since both

models have a transition model that insures no instantaneous transitions between labels. This smoothness of the output probabilities makes it possible to find distinct peaks. These peaks represent good backchannel opportunities. A peak can easily be detected in real-time since it is the point where the probability starts decreasing. For each peak we get a backchannel opportunity with associated probability.

Interestingly, Cathcart et al. [6] note that human listeners varied considerably in their backchannel behavior (some appear less expressive and pass up "backchannel opportunities") and their model produces greater precision for subjects that produced more frequent backchannels. The same observation was made by Ward and Tsukahara [39]. An important advantage of our prediction model over previous work is the fact that for each backchannel opportunity returned, we also have an associated probability. This makes it possible for our model to address the problem of expressiveness. By applying an expressiveness threshold on the backchannel opportunities, our prediction model can be used to create virtual agents with different levels of nonverbal expressiveness.

4 Experiments

For training and evaluation of our prediction model, we used a corpus of 50 human-to-human interactions. This corpus is described in Sect. 4.1. Section 4.2 describes the speaker features used in our experiments as well as our listener backchannel annotations. Finally Sect. 4.3 discusses our methodology for training the probabilistic model and evaluate it (see Fig. 4).

4.1 Data collection

One hundred and four participants (67 women, 37 men) were recruited through Craigslist.com from the greater Los Angeles area and compensated $20. Of the 52 subject pairs, two were excluded due to recording equipment failure, resulting in 50 valid sessions.

We elicited natural conversational behaviors from pairs of subjects via a face-to-face quasi-monologue storytelling task. In this task, one subject ('the Speaker') describes a

Fig. 4 Setup for training and evaluation corpus. This study of face-to-face narrative discourse ('quasi-monologic' storytelling) included 104 subjects. The speaker was instructed to retell the stories portrayed in two video clips to the listener

previously-watched movie clip to another subject ('the Listener') that has not seen the movie. The movie was selected to elicit a range of nonverbal feedback from listeners. Speakers watched and described a 2-min video taken from a sexual harassment awareness video by Edge Training Systems, Inc. The video dramatizes two incidents of workplace harassment: The first incident is about a woman at work who receives unwanted instant messages from a colleague at work and the second is about a man at work who is confronted by a female business associate, who asks him for a foot massage in return for her business. Subjects face each other across a table (approximately 8' apart) and could see see and respond to each other's nonverbal feedback.

Participants in groups of two entered the laboratory and were told they were participating in a study to evaluate communication technology. They completed a consent form and pre-experiment questionnaire contained questions about subject's demographic background, personality [18], self-monitoring [26], self-consciousness [34] and shyness [7]. Subjects were randomly assigned the role of listener or speaker. The listener was asked to wait outside the room while the speaker viewed the short video clip. The listener was then led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. Elicited stories were approximately 2 min in length on average.

Finally, the experimenter led the speaker to a separate side room. The speaker completed a post-questionnaire assessing their impressions of the interaction while the listener remained in the room and spoke to the camera what s/he had been told by the speaker. Participants were debriefed individually and dismissed.

We collected synchronized multimodal data from each participant including voice and upper-body movements. Both the speaker and listener wore a lightweight headset with microphone. Three camcorders were used to videotape the experiment: one was placed in front the speaker, one in front of the listener, and one was attached to the ceiling to record both speaker and listener.

4.2 Speaker features and listener backchannels

From the video and audio recordings several features were extracted. In our experiments the speaker features were sampled at a rate of 30 Hz so that visual and audio feature could easily be concatenated.

Pitch and intensity of the speech signal were automatically computed from the speaker audio recordings, and acoustic features were derived from these two measurements. The following prosodic features were used (based on Ward and Tsukahara [39]):

- Regions of pitch lower than the 26th percentile continuing for at least 110 ms (i.e., lowness)
- Downslopes in pitch (at least 1.5 percentile) continuing for at least 40 ms
- Utterances longer than 700 ms
- Gradual drop or rise in intensity of speech
- Fast drop or rise in intensity of speech
- Vowel volume (spoken softer than 80% of the average volume)

Human coders manually annotated the narratives with several relevant features from the audio recordings. All elicited narratives were transcribed, including pauses, filled pauses (e.g., "um"), incomplete and prolonged words. These transcriptions were double-checked by a second transcriber. This provided us with the following extra lexical and prosodic features:

– All individual words (i.e., unigrams)
– Pause (i.e., no speech)
– Filled pause (e.g. "um")
– Lengthened words (e.g., "I li::ke it")
– Emphasized or slowly uttered words (e.g., "ex_a_c_tly")
– Incomplete words (e.g., "jona-")
– Words spoken with continuing intonation
– Words spoken with falling intonation (e.g., end of an utterance)
– Words spoken with rising intonation (i.e., question mark)

From the speaker video the eye gaze of the speaker was annotated on whether he/she was looking at the listener. After a test on five sessions we decided not to have a second annotator go through all the sessions, since annotations were almost identical (less than 2 or 3 frames difference in segmentation). The feature we obtained from these annotations is:

– Speaker looking at the listener

Note that although some of the speaker features were manually annotated in this corpus, all of these features can be recognized automatically given the recent advances in real-time keyword spotting [17], eye gaze estimation and prosody analysis. Our feature *Speaker looking at the listener* can be computed from a coarse estimate of the eye gaze (e.g., intensity-based eye gaze systems [32,38]) since it only needs to do a binary decision: is the speaker looking at the listener or not?

Finally, the listener videos were annotated for visual backchannels (i.e., head nods) by two coders. These annotations form the labels used in our prediction model for training and evaluation.

4.3 Methodology

To train our prediction model we split the 50 session into three sets, a training set, a validation set and a test set. This is done by doing a 5-fold testing approach. This means that 10 sessions are left out for test purposes only and the other 40 are used for training and validation. This process is repeated five times in order to be able to test our model on each session. Validation is done by using the holdout cross-validation strategy. In this strategy a subset of 10 sessions is left out of the training set.

The performance is measured by using the F-measure since it is well established in Natural Language Processing (NLP) and speech recognition. This is the weighted harmonic mean of precision and recall. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model. We use the same weight for both precision and recall, so called F_1. During validation we find all the peaks in our probabilities. A backchannel is predicted correctly if a peak in our probabilities (see Sect. 3.4) happens during an actual listener backchannel.

As discussed in Sect. 3.4, the expressiveness level is the threshold on the output probabilities of our sequential probabilistic model. This level is used to generate the final backchannel opportunities. In our experiments we picked the expressiveness level which gave the best F_1 measurement on the validation set. This level is used to evaluate our prediction model in the testing phase. For space constraint reason, all the results presented in this paper are using Conditional Random Fields [24] as sequential probabilistic model. We performed the same series of experiments with Hidden Markov Models [33] but the results were constantly lower.

Algorithm 1 Rule based approach of Ward and Tsukahara [39]

Upon detection of
P1: a region of pitch less than the 26th percentile pitch level and
P2: continuing for at least 100 milliseconds
P3: coming after at least 700 milliseconds of speech,
P4: providing you have not output backchannel feedback within the preceding 800 milliseconds,
P5: after 700 milliseconds wait,
you should produce backchannel feedback.

The hCRF library was used for training the CRF model [15]. The regularization term for the CRF model was validated with values $10^k, k = -1 \ldots 3$.

5 Results and discussion

We compared our prediction model with the rule based approach of Ward and Tsukahara [39] since this method has been employed effectively in virtual human systems and demonstrates clear subjective and behavioral improvements for human/virtual human interaction [14]. We re-implemented their rule based approach summarized in Algorithm 1. The two main features used by this approach are *low pitch regions* and *utterances* (see Sect. 4.2). We also compared our model with a "random" backchannel generator as defined in Ward and Tsukahara [39]: randomly generate a backchannel cue every time conditions P3, P4 and P5 are true (see Algorithm 1). The frequency of the random predictions was set to 60% which provided the best performance for this predictor, although differences were small.

Table 1 shows a comparison of our prediction model with both approaches. We performed an one-tailed t-test comparing our prediction model to both random and Ward's approach over our 50 independent sessions. Our performance is significantly higher than both random and the hand-crafted rule based approaches with p-values below 0.05. A second result from this table is that the performance of the rule-based and random approaches on our multimodal dataset are similar to the results previously published by Ward and Tsukahara [39] on a unimodal dataset predicting audio back-channel feedback. The one-tailed t-test comparison between Ward's system and random shows that that difference is statistically significant, suggesting that Ward and Tsukahara rule-based approach does also apply to visual backchannel prediction. Our prediction model outperforms both the random approach and the rule-based approach of Ward and Tsukahara [39].

Table 1 Comparison of our prediction model with previously published rule-based system of Ward and Tsukahara [39]

	Result			t-Test (p-value)	
	F_1	Precision	Recall	Random	Rule-based
Our prediction model	0.2562	0.1862	0.4106	<0.0001	0.0210
Rule-based approach [12]	0.1824	0.1262	0.3290	0.0042	–
Random	0.1137	0.1042	0.1250	–	–

By integrating the strengths of a machine learning approach with multimodal speaker features and automatic feature selection, our prediction model shows a statistically significant improvement over the unimodal rule-based and random approaches

Table 2 Compares the performance of our prediction model before and after joint feature selection (see Sect. 2)

	Result			*t*-Test (*p*-value)
	F_1	Precision	Recall	
Joint and individual feature selections	0.2562	0.1862	0.4106	0.1312
Only individual feature selection	0.2210	0.1407	0.5145	

We can see that joint feature selection is an important part of our prediction model

Table 3 Compares the performance of our prediction model with and without the visual speaker feature (i.e., speaker looking at the listener)

	Result			*t*-Test (*p*-value)
	F_1	Precision	Recall	
Multimodal features	0.2210	0.1407	0.5145	0.1454
Unimodal features	0.2064	0.1398	0.3941	

We can see that the multimodal factor is an important part of our prediction model

Our prediction model uses two types of feature selections: individual feature selection and joint feature selection (see Sect. 3.3.2 for details). It is very interesting to look at the features and encoding selected after both processes:

- *Pause* using binary encoding
- *Speaker looking at the listener* using ramp encoding with a width of 2 s and a 1 s delay
- *'and'* using step encoding with a width 1 s and a delay of 0.5 s
- *Speaker looking at the listener* using binary encoding

The joint selection process stopped after four iterations, the optimal number of iterations on the validation set. Note that *Speaker looking at the listener* was selected twice with two different encodings. This reinforces the fact that having different encodings of the same feature reveals different information of a feature and is essential to get high performance with this approach. It is also interesting to see that our prediction algorithm outperform Ward and Tsukahara without using their feature corresponding of low pitch. In Table 2 we show that the addition joint feature selection improved performance over individual feature selection alone. In the second case the sequential model was trained with all the 12 features returned by the individual selection algorithm and every encoding templates from our dictionary. These speaker features were: pauses, energy fast edges, lowness, speaker looking at listener, "and", vowel volume, energy edge, utterances, downslope, "like", falling intonations, rising intonations.

In Table 3 the importance of multimodality is showed. Both of these models were trained with the same 12 features described earlier, except that the unimodal model did not include the *Speaker looking at the listener* feature. Even though we only added one visual feature between the two models, the performance of our prediction model increased by approximately 3%. This result shows that multimodal speaker features is an important concept.

The capability to automatically learn nonverbal prediction models has wide applicability beyond the generation of nonverbal behavior of interactive virtual humans. The outcome is notable because it demonstrates the potential of a very general approach to learning to extract patterns of nonverbal behavior that can potentially illuminate fundamental questions in social

psychology and communication studies. Predicting such nonverbal feedback also has broad applicability to a number of human–machine systems. We have recently demonstrated that such prediction models can improve the accuracy of gesture recognition techniques by setting prior expectations [28]. For example if a head nod is more likely than a look down, given what the speaker just said, knowing this information can help distinguish ambiguous gestures. Additionally, prediction models can potentially help diagnose normal from abnormal behavior (for example, by training two prediction models on normal and schizophrenic subjects, it may be possible to automatically distinguish individuals on the basis of their nonverbal behavior). Finally, these capabilities can be combined to inform the behavior of advance virtual humans that can both recognize and synthesize natural nonverbal behavior for a variety of training and entertainment applications [36,21,12].

6 Conclusion

In this paper we presented how sequential probabilistic models can be used to automatically learn from a database of human-to-human interactions to predict listener backchannel using the speaker multimodal output features (e.g., prosody, spoken words and eye gaze). The main challenges addressed in this paper were automatic selection of the relevant features and optimal feature representation for probabilistic models. For prediction of visual backchannel cues (i.e., head nods), our prediction model was showed a statistically significant improvement over a previously published approach based on hand-crafted rules. Although we applied the approach to generating backchannel behavior, the method is proposed as a general probabilistic framework for learning to recognize and generate meaningful multimodal behaviors from examples of face-to-face interactions including facial expressions, posture shifts, and other interactional signals. Thus, it has importance, not only as a means to improving the interactivity and expressiveness of virtual humans but as an fundamental tool for uncovering hidden patterns in human social behavior.

Acknowledgements The authors would like to thank Nigel Ward for his valuable feedback, Marco Levasseur and David Carre for helping to build the original Matlab prototype, Brooke Stankovic, Ning Wang and Jillian Gerten. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the National Science Foundation under Grant # HS-0713603. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Allwood, J. (2008). Dimensions of embodied communication—towards a typology of embodied communication. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied communication in humans and machines*. Oxford University Press.
2. Anderson, H., Bader, M., Bard, E. G., Doherty, G., Garrod, S. Isard, S., et al. (1991). The mcrc map task corpus. *Language and Speech, 34*(4), 351–366.
3. Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology, 79*(6), 941–952.
4. Burns, M. (1984). Rapport and relationships: The basis of child care. *Journal of Child Care, 2*, 47–57.
5. Cassell, J., Vilhjlmsson, H., & Bickmore, T. (2001). Beat: The behavior expressive animation toolkit. In *Proceedings of the SIGGRAPH*.
6. Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *European ACL*, pp. 51–58.

7. Cheek, J. M. (1983). *The Revised Cheek and Buss Shyness Scale (RCBS)*. Wellesley, MA: Wellesley College.
8. Demirdjian, D., & Darrell, T. (2002). 3-d articulated pose tracking for untethered deictic reference. In *International conference on multimodal interfaces*.
9. Drolet, A. L., & Morris, M. W. (2000). Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Experimental Social Psychology, 36*, 26–50.
10. Fuchs, D. (1987). Examiner familiarity effects on test performance: Implications for training and practice. *Topics in Early Childhood Special Education, 7*, 90–104.
11. Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., & Kobayashi, T. (2004). A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of the international symposium on robot and human interactive communication* (pp. 159–164).
12. Gandhe, S., DeVault, D., Roque, A., Martinovski, B., Artstein, R., Leuski, A., et al. (2008). From domain specification to virtual humans: An integrated approach to authoring tactical questioning characters. In *Proceedings of interspeech 2008*.
13. Goldberg, S. B. (2005). The secrets of successful mediators. *Negotiation Journal, 21*(3), 365–376.
14. Gratch, J., Wang, N., Gerten, J., & Fast, E. (2007). Creating rapport with virtual agents. In *Proceedings of intelligent virtual agents (IVA 2007)*.
15. *hCRF library*. http://sourceforge.net/projects/hcrf/. Accessed March 2008.
16. Heylen, D., Bevacqua, E., Tellier, M., & Pelachaud, C. (2007). Searching for prototypical facial feedback signals. In *Proceedings of 7th international conference on intelligent virtual agents* (pp. 147–153).
17. Igor, S., Petr, S., Pavel, M., Luk, B., Michal, F., Martin, K., et al. (2005). Comparison of keyword spotting approaches for informal continuous speech. In *Proceedings of the joint workshop on multimodal interaction and related machine learning algorithms*.
18. John O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). Guilford Press.
19. Jónsdóttir, G. R., Gratch, J., Fast, E., & Thórisson, K. R. (2007). Fluid semantic back-channel feedback in dialogue: Challenges and progress. In *Proceedings of 7th international conference on intelligent virtual agents*.
20. Kang, S.-H., Gratch, J., Wang, N., & Watt, J. (2008). Does the contingency of agents' nonverbal feedback affect users' social anxiety? In *Proceedings of the international joint conference on autonomous agents and multiagent systems*.
21. Kenny, P., Parsons, T., Gratch, J., & Rizzo, A. (2008). Evaluation of justina: A virtual patient with ptsd. In *Proceedings of 8th international conference on intelligent virtual agents*, Tokyo, Japan, September 2008.
22. Kipp, M., Neff, M., Kipp, K. H., & Albrecht, I. (2007). Toward natural gesture synthesis: Evaluating gesture units in a data-driven approach. In *Proceedings of 7th international conference on intelligent virtual agents* (pp. 15–28). Springer.
23. Kopp, S., Stocksmeier, T., & Gibbon, D. (2007). Incremental multimodal feedback for conversational agents. In *Proceedings of 7th international conference on intelligent virtual agents* (pp. 139–146).
24. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of the eighteenth international conference on machine learning*.
25. Lee, J., & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *Proceedings of 6th international conference on intelligent virtual agents* (pp. 243–255).
26. Lennox, R. D., & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social psychology, 46*, 1349–1364.
27. Maatman, M., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. In *Proceedings of intelligent virtual agent (IVA 2005)* (pp. 25–36).
28. Morency, L.-P., de Kok, I., & Gratch, J. (2008). Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *Proceedings of 10th international conference on multimodal interfaces (ICMI 2008)*, October 2008.
29. Morency, L.-P., de Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of intelligent virtual agents (IVA 2008)*, September 2008.
30. Morency, L.-P., Sidner, C., Lee, C., & Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the international conference on multimodal interfaces*, October 2005.
31. Nishimura, R., Kitaoka, N., & Nakagawa, S. (2007). A spoken dialog system for chat-like conversations considering response timing. *Lecture Notes in Computer Science, 4629*, 599–606.
32. *OKAO Vision library*. http://www.omron.com/r_d/coretech/vision/okao.html. Accessed Dec 2008.
33. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

34. Scheier, M. F., & Carver, C. S. (1985). The self-consciousness scale: A revised version for use with general populations. *Journal of Applied Social Psychology, 15*, 687–699.
35. Thiebaux, M., Marshall, A., Marsella, S., & Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the international joint conference on autonomous agents and multiagent systems*.
36. Traum, D., Gratch, J., Marsella, S., Lee, J., & Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of 8th international conference on intelligent virtual agents*, Tokyo, Japan, September 2008.
37. Tsui, P., & Schultz, G. L. (1985). Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients. *American Journal of Orthopsychiatry, 55*, 561–569.
38. Valenti, R., & Gevers, T. (2008). Accurate eye center location and tracking using isophote curvature. In *IEEE conference on computer vision and pattern recognition (CVPR 2008)*, June 2008.
39. Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 23*, 1177–1207.
40. Yngve, V. H. (1970). On getting a word in edgewise. In *Proceedings of the sixth regional meeting of the Chicago Linguistic Society*.