Learning-Based Modeling of Multimodal Behaviors for Humanlike Robots

Chien-Ming Huang and Bilge Mutlu
Department of Computer Sciences, University of Wisconsin–Madison
1210 West Dayton Street, Madison, WI 53706 USA
{cmhuang, bilge}@cs.wisc.edu

ABSTRACT

In order to communicate with their users in a natural and effective manner, humanlike robots must seamlessly integrate behaviors across multiple modalities, including speech, gaze, and gestures. While researchers and designers have successfully drawn on studies of human interactions to build models of humanlike behavior and to achieve such integration in robot behavior, the development of such models involves a laborious process of inspecting data to identify patterns within each modality or across modalities of behavior and to represent these patterns as "rules" or heuristics that can be used to control the behaviors of a robot, but provides little support for validation, extensibility, and learning. In this paper, we explore how a *learning-based* approach to modeling multimodal behaviors might address these limitations. We demonstrate the use of a *dynamic Bayesian network (DBN)* for modeling how humans coordinate speech, gaze, and gesture behaviors in narration and for achieving such coordination with robots. The evaluation of this approach in a human-robot interaction study shows that this learning-based approach is comparable to conventional modeling approaches in enabling effective robot behaviors while reducing the effort involved in identifying behavioral patterns and providing a probabilistic representation of the dynamics of human behavior. We discuss the implications of this approach for designing natural, effective multimodal robot behaviors.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors, software psychology*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*evaluation/ methodology, user-centered design*

General Terms

Design, Human Factors

1. INTRODUCTION

In communication, people draw on a rich repertoire of behaviors from multiple modalities, including speech, gaze, gestures, and so on. While this repertoire offers a rich design space for creating robot behaviors that achieve natural, effective communication, how behaviors across different modalities are integrated is critical to the success of such robot behaviors. For example, previous work in human-robot interaction has shown that robots that do not employ gaze behaviors that are contingent with their speech impair learning, collaboration, and user experience, while robots that better align their gaze with their speech improve such interaction outcomes [9].

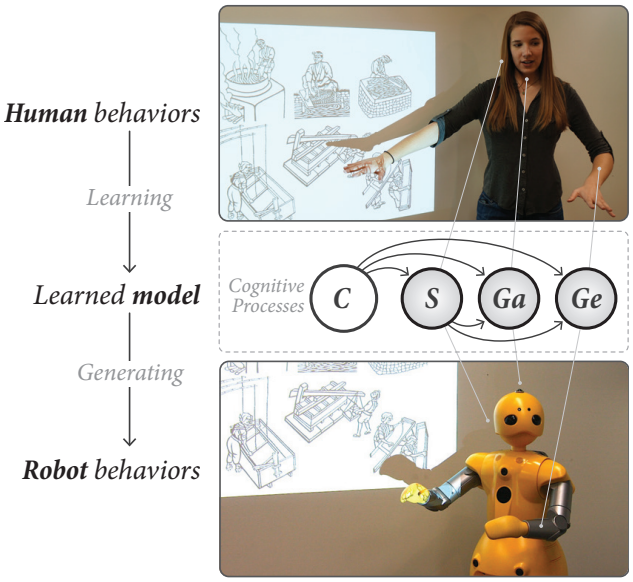


Figure 1: We used a learning-based approach to *model* how humans employ multimodal behaviors involving speech, gaze, and gestures during narration and *generate* multimodal behaviors for a humanlike robot to perform the same narration task.

How might robots employ behaviors across multiple modalities while ensuring proper alignment among them? To achieve such alignment, researchers and designers have developed *multimodal* models of behavior that specify patterns in which behaviors co-occur, such as how speakers coordinate their gaze and speech to manage conversational floor [30] or how gesture and speech are aligned in giving directions [32]. However, the development of such models involves a laborious process of sifting through data to identify and extract distribution and alignment parameters and provides limited support for validation, extensibility, and learning.

To facilitate this process and address its limitations, we propose a *learning-based* approach that involves the use of probabilistic graphical models (PGMs) to automatically learn distribution and alignment parameters from annotated data on human behavior. In this paper, we demonstrate the use of a dynamic Bayesian network (DBN) to model the alignment among speech, gaze, and gesture

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI'14, March 3–6, 2014, Bielefeld, Germany.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2658-2/14/03 ...$15.00.
http://dx.doi.org/10.1145/2559636.2559668.

behaviors in a narration task and to estimate distribution and alignment parameters for these behaviors to enable a humanlike robot to perform the same narration task (Figure 1). We also present an empirical evaluation of the effectiveness of this approach in generating robot behaviors against approaches that generate robot behaviors based on designer-specified parameters, randomly-generated parameters, and no behaviors.

2. BACKGROUND

2.1. Human Multimodal Behavior

Humans naturally use multimodal behaviors involving speech, gaze, and gestures during interaction. While *speech* serves as the main channel to convey information, *gestures* are used to illustrate imagery [21], draw attention from other participants [7], disambiguate unclear speech, and supplement speech with additional information [16]. *Gaze* also supplement speech, facilitating turn-taking in conversation [14], signaling conversational roles [5], and regulating intimacy between participants [1]. These three channels make up a rich repertoire of social behaviors that play a critical role in effective communication in settings such as narration [37] and reenactments of previous events or experiences [38].

Research in human communication has extensively studied the relationship between speech and gesture [15, 21], particularly the four common types of gestures: (1) *deictic* gestures, which involve pointing toward a shared reference, (2) *iconic* gestures that illustrate concrete objects or actions, (3) *metaphoric* gestures, which use concrete metaphors to represent abstract concepts such as time, and (4) *beat* gestures, which are rhythmic movements that mark the structure of the speaker's discourse. Gestures and speech are tightly linked at both semantic and structural levels. Deictic, iconic, and metaphoric gestures are closely related to the semantics in speech through *lexical affiliates*—the words or phrases with which gestures co-express semantic meaning [35]. Beat gestures are linked to speech at a structural level and indicate significant points in speech, such as connecting discontinuous parts in speech and introducing a new topic [21]. Research has also studied the relationship between speech and gaze [6, 14, 23], identifying a close coupling between these two channels, such as a tendency to gaze toward referents before they refer to them in speech [6, 23] and the use of gaze cues to signal opportunities for exchanging conversation floor [14].

2.2. Multimodal Behaviors in Robots

Previous research in human-robot interaction has explored the development of mechanisms for achieving natural and effective multimodal behaviors for robots, such as the development of models of gaze for displaying appropriate head movements at meaningful speech points for a museum guide robot [40], aligning gaze shifts with discourse structure for a storytelling robot [29], and synchronizing deictic verbal and gaze references for an instructional robot [9]. These examples highlight the importance of temporal alignment among different modalities of behavior to improve human-robot collaboration, perceptions of the robot, and overall user experience. Researchers have also developed models of gesture to improve human-robot interaction, including a probabilistic model to generate the four common types of gestures that are aligned with speech and achieve varying levels of expressiveness for robots [31] and a method for aligning speech and gesture strokes and smoothing transitions between gestures to produce more fluent behavior [34]. Robot that appropriately align their speech and gesture using such models were found to be more natural [32].

In our prior work, we explored how robots might display coherent multimodal behaviors involving speech, gaze, and gestures [10]. We explicitly specified the semantic link between speech and gestures and empirically obtained parameters to quantify the temporal speech-gesture and gaze-gesture alignments. Specifically, the speech-gesture associations were hand-coded by identifying the lexical affiliate for each gesture according to literature on human communication [21] and quantifying the alignment between them. We also modeled the link between gaze and gestures by obtaining distributions of where the speaker looked while performing different types of gestures. While this approach yielded acceptable behaviors, enabling us to study how gestures might shape interaction outcomes, it required a deliberate selection of alignment parameters.

Additionally, while such inspection-based approaches might be feasible for modeling a small number of behaviors from small datasets, this feasibility diminishes when a larger number of behaviors or large datasets are considered. The models built by these approaches are also highly sensitive to the decisions made and inspection methods used by the researcher or the designer in the modeling process. To address these limitations, we propose a learning-based approach to automatically learn these parameters from data.

2.3. Learning-based Modeling

Previous research in human-robot interaction has used learning-based approaches primarily to achieve autonomous human-robot interaction. Examples of such uses include unsupervised learning of associations between human gestural commands and robot actions using graphical models [24, 25] and using a multilayer Bayesian approach to realize active perception [4].

Learning-based approaches have commonly been used to build predictive models of human behavior and to control behaviors of embodied conversational agents (e.g., [19, 26, 33]). These approaches frequently use probabilistic graphical models (PGMs) for their support for modeling complex relationships under uncertainty.

Building on these approaches, the current work seeks to use PGMs to represent human multimodal behaviors, learn model parameters from annotated data on human behaviors, and draw on the learned model to achieve natural, humanlike robot behaviors.

3. DESIGNING MULTIMODAL BEHAVIOR

We conceptualized the problem of generating multimodal robot behaviors into two levels—the *feature* level and the *domain* level. Figure 2 illustrates this conceptualization for the process of generating speech, gaze, and gestures. The feature level represents high-level behavioral features of the target channel, such as "iconic gesture" for gesture type and "listener" for gaze target. At the domain level, behavioral features are associated with specific motions, such as specific arm motions for an iconic gesture and gaze shifts toward the listener. This separation allows us to modularize the problem space and develop and improve different components in isolation. In this work, we focus on learning and inference at the feature level and employ simple mechanisms to bridge the feature and domain levels. The following sections describe how human data was collected and annotated for learning and how the feature and domain levels were associated to produce robot behaviors.

3.1. Data Collection and Annotation

Our investigation of human multimodal behavior focused on three modalities—speech, gaze, and gestures—as widely observed behaviors in human interaction across various contexts and cultures [1, 21]. We contextualized our investigation in a narration task in which a narrator presents *the process of making paper* to a recipient with the aid of a projected illustration of the process (Figure 1). We chose narration to model the dynamics of human behavior, because narration elicits the use of a wide range of gestures [21], and narrative re-enactments involve a rich interplay between gaze and gestures [37]. Additionally, robots are envisioned to serve in roles similar to

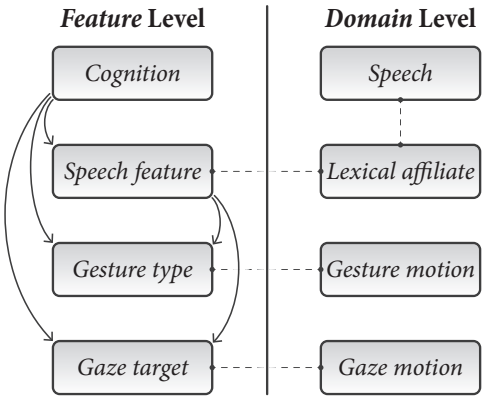


Figure 2: Our conceptualization of the process of generating speech, gaze, and gestures. Learning and inference for high-level features are performed at the *feature* level, while specific motions are defined at the *domain* level.

that of a narrator conveying information to users, such as museums tour guides, shopping assistants, or receptionists [2, 13, 20].

To understand how people concurrently use speech, gaze, and gestures, we conducted a human-human interaction study with 16 participants, recruited from the University of Wisconsin–Madison campus, whose ages ranged 19–26 ($M = 22.00, SD = 2.16$). Dyads were gender-balanced so that two dyads represented each gender combination in the study. Dyadic interactions yielded a total of 37.15 minutes ($M = 4.64, SD = 1.16$) of video data.

The video data was coded for speech, gaze, and gestures features. Four typical types of gestures—*deictics*, *iconics*, *metaphorics*, and *beats*—were coded according to the guidelines provided by McNeill [21]. Instances of no gestures were also coded. Four clusters of gaze targets—*reference*, *recipient*, *narrator's own gesture*, and *other places*—were observed in the collected data and used for coding to represent features of gaze behavior. Additionally, we coded speech for lexical affiliates for deictic, iconic, and metaphoric gestures and significant structural points for beat gestures. Used affinity diagramming, we categorized lexical affiliates and significant points in speech. The top three categories for each type of gesture served as speech features (Figure 3). A primary coder coded the eight interaction episodes, and a secondary coder coded 10% of the data to ensure inter-coder reliability. The reliability analysis showed almost perfect agreement for speech features (Cohen's κ = .870), gesture type (Cohen's κ = .845) and gaze target (Cohen's κ = .916) based on guidelines suggested by Landis and Koch [18].

3.2. Dynamic Bayesian Network

A dynamic Bayesian network (DBN) is a type of probabilistic graphical model (PGM) that provides compact representations of conditional independence among random variables [17]. DBNs generalize hidden Markov models (HMMs) to represent the hidden state and the observation as a set of random variables. In their basic form, DBNs are directed acyclic graphs in which nodes represent random variables and edges represent conditional dependencies (e.g., Figure 4). Semantically, an edge from a parent node A to a child node B means that node A has influence over node B. A dynamic Bayesian network extends static Bayesian nets (BNs) to incorporate the temporal dependencies among variables. These characteristics for dealing with *uncertain* and *temporal* relations among random variables make dynamic Bayesian networks particularly useful in modeling the dynamics of multimodal behaviors. Murphy [28] provides an extensive introduction to representation, learning, and inference in DBNs.

Gesture Type	Speech Features		
Deictic gestures	Concrete reference *"a big pot"*	Abstract reference *"the first step"*	Pronoun *"this person"*
Iconic gestures	Concrete object *"two boards"*	Descriptive verb *"peel it off"*	Non-descriptive action *"make it"*
Metaphoric gestures	Abstract concept *"for six hours"*	Abstract process *"how paper is made"*	Abstract object *"the water soluble elements"*
Beat gestures	Important information *"at least ten times of water"*	New information *"for example"*	Connector *"so that"*

Figure 3: Speech features for gestures. Features were identified through affinity diagramming of human data and by using the guideline suggested by McNeill [21]. An example of each type of feature is also provided.

3.3. Model Representation

Informed by literature in human communication [8, 11, 21], we propose a network structure, shown in Figure 4, to represent the relationships among speech, gaze, and gestures. In developing this network, we included a *hidden random variable* denoting a cognitive process (C) that directs how humans coordinate speech (S), gaze (Ga), and gestures (Ge), which were considered *observations*. We assumed the latent cognitive process was a discrete-time Markov process. This assumption is consistent with psycholinguistic models of speech production [8, 11]. Additionally, we assumed that speech influences gaze and gestures, as research suggests that nonverbal behaviors might be contingent on verbal utterances [21]. Based on our exploratory tests, we empirically determined there were three hidden states and that the discrete time window of the Markov process was 500 ms.

3.4. Learning and Inference

To learn the parameters of each conditional probability distribution (CPD) in the DBN (Figure 4), the expectation-maximization (EM) algorithm [3] was used. The eight coded episodes of interaction were used as the training data.

To control the robot's behaviors using the learned DBN, we assumed that the robot's speech features would be given and the most probable gesture type and gaze target would be inferred at any given time t. To this end, we used a junction tree algorithm to perform offline smoothing [28] and compute the most probable explanation (MPE)—the maximal posterior probability of a set of variables given observations of another set of variables—of gaze and gestures. We used the Bayes Net Toolbox [27] for learning and inference.

The latent cognitive process contained three states ($C = c_i, i = 1,2,3$). Gaze and gestures contained four ($Ga = a_i, i = 1,...,4$) and five ($Ge = e_i, i = 1,...,5$) values, respectively. Speech (S) was represented by 12 boolean variables, each of which corresponded to a speech feature (Figure 3). As a result, the model is characterized by a vector

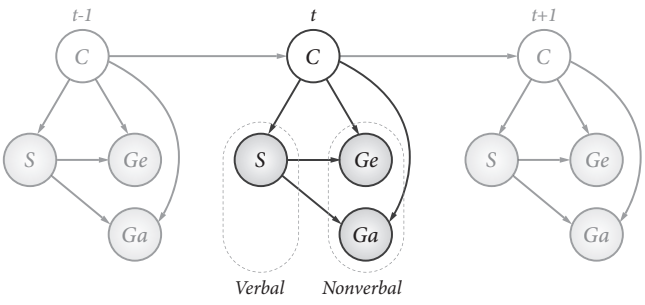


Figure 4: The proposed dynamic Bayesian network for modeling and generating multimodal behaviors. C denotes a latent cognitive process that directs verbal, involving speech (S), and nonverbal, involving gaze (Ga) and gestures (Ge), processes.

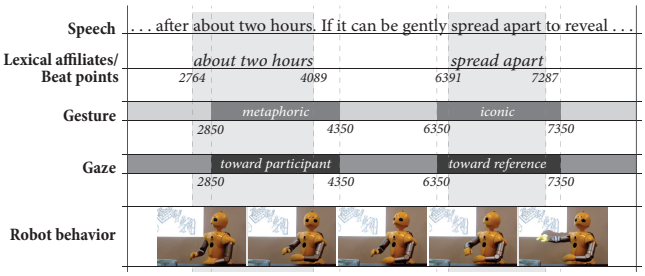


Figure 5: An example of the robot displaying speech, gaze, and gesture behaviors generated by the proposed learning-based approach.

of 15 discrete values at each time step. Given the speech features, the most probable latent cognitive state (C_t), gesture type (Ge_t), and gaze target (Ga_t) at any given time t over the duration of the speech ($S_{1:T}$) can therefore be computed using Equation 1. Here, X represents C, Ga, and Ge, and x represents c, a, and e.

$$\arg\max_i p(X_t = x_i | S_{1:T}) \qquad \forall t \in T \tag{1}$$

The features of the robot's speech were manually marked, and the annotated speech features were discretized into feature sequences at the rate of 500 ms per feature. During inference, the sequences of speech features were used as partial observations, and gaze and gesture were treated as missing values. The consecutive states inferred for each behavior were combined into a sequence of behavior that was considered to be one continuous instance of behavior. For example, six consecutive states of iconic gestures were combined into one iconic gesture lasting 3,000 ms.

3.5. Model Evaluation

An evaluation of the model sought to determine to what extent the DBN model, as illustrated in Figure 4, accurately predicted gaze targets and gesture types when given speech features by comparing gaze and gestures predicted by Equation 1 against the human data. The evaluation was conducted using eight-fold leave-one-out cross validation. We used gaze and gesture data from seven dyads to train the DBN and evaluated the performance of the trained model in predicting gaze targets and gesture types, given speech features from the eighth dyad. While the behaviors are continuous in time, we chose to discretize the data using a window 500 ms and to compare the predicted behavior to the behavior in the test dataset at each discrete window. The accuracy of prediction was on average 54.57% for gaze targets and 62.77% for gesture type. While prediction accuracy is significantly better than chance (25% for gaze and 20% for gestures), it might be further improved by considering more features. Moreover, the highest and lowest accuracies in the cross validation for gaze were 68.23% and 46.48%, respectively, and for gestures were 69.76% and 44.25%, respectively. These results suggest a large variation in how people employed behaviors during narration. Collecting more data to train the model might improve predictive accuracy.

3.6. Generating Robot Behavior

At the *domain* level, the robot's gestures were designed based on our observations of human narrators' gestures. While gesture performance at a given gesture point varied slightly among narrators, narrators displayed semantically similar elements. For example, to describe "beating (paper) with a stick," participants displayed one-handed or two-handed up and down movements, but at different speeds and with different hand angles with respect to the ground. For each unique gesture point, we created one robot gesture that captured the common elements we observed from the human narrators. Robot gestures were created through puppeteering, which involved manually moving the robot's arms while recording key frames of the gestural trajectories. Gesture libraries were created for the four common gesture types to include all the gestures that might be used in the narration task.

To generate robot behaviors, simple mechanisms were used to link components at the feature and domain levels. Lexical affiliates in the robot's speech were manually annotated and tagged with gesture types. The mechanism for linking inferred gesture types to actual robot gestures functioned as follows. For each inferred gesture, we first checked whether the robot's speech included lexical affiliates that temporally overlapped with the gesture within a window of 2,000 ms—1,000 ms before and 1,000 ms after the beginning of the gesture. No lexical affiliates or a lexical affiliate for a gesture type that did not match with the inferred gesture type within this window prompted the robot to randomly select and perform a gesture from the gesture library of the inferred gesture type. If the gesture type of the found lexical affiliate was consistent with the inferred gesture type, the lexical affiliate was linked to an actual gesture using an association mechanism.

Speech and inferred gaze and gesture motions were synchronized using the Robot Behavior Toolkit [9], a Robot Operating System (ROS) module for controlling multimodal robot behaviors. Figure 5 illustrates a sample sequence of speech, gaze, and gesture behaviors that were controlled by the proposed learning-based approach.

4. EVALUATION

In addition to the model evaluation described above, we evaluated how robot behaviors produced by the learning-based approach shaped people's perceptions of the robot in a human-robot interaction study. The evaluation sought to test our central hypothesis that robot behaviors produced by learned parameters would improve interaction outcomes, specifically improved perceptions of the robot in terms of immediacy, naturalness, effectiveness, likability, and credibility and improved ability of the participants to retell the robot's story, over baseline behaviors such as behaviors produced by randomly generated parameters or no behaviors, while resulting in outcomes that are comparable to those elicited by conventional modeling approaches.

4.1. Study Design, Task, and Procedure

The evaluation study used the same narration task as the one used in the modeling study. We manipulated the method used to control the robot's gaze and gesture behaviors while keeping the robot's speech the same across conditions. We designed a between-participants study, in which each participant was randomly assigned to one of the following conditions:

1. In the *learning-based* condition, the learned DBN described in the previous section directed the robot's gaze and gestures.
2. In the *unimodal* condition, the robot only used the speech channel to verbally present the narration topic without gaze or gesture behaviors. This condition served as the experimental control.
3. In the *random* condition, the same network structure as described in the previous section was used to generate behaviors. However, the model instead used randomly generated parameters, introducing temporal and spatial randomness in the produced behavior.
4. The *conventional* condition involved directing the robot's gaze and gestures based on designer-specified parameters for aligning behaviors. In particular, the parameters specified how different types of gestures were aligned with speech features (i.e., lexical affiliates) according to the literature on human

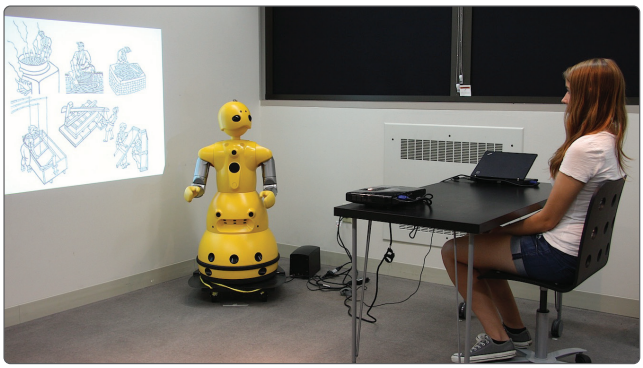


Figure 6: An experimenter demonstrating the setup of the evaluation study.

communication [21]. The parameters also specified gaze-gesture relationships by extracting distributions representing where a speaker looked while performing different types of gestures. Every time its gesture state changed, the robot determined its new gaze target based on these distributions. The behaviors generated using this approach predicted significant improvements in learning and perceptions of the robot [10].

After obtaining informed consent, the experimenter directed participants to a controlled laboratory environment, where each participant listened to the robot narrate the process of making paper (Figure 6). The robot's narration lasted approximately six minutes. Following the narration, participants completed a distractor task that lasted approximately five minutes and then took a quiz on the presented information. Participants then retold the narration and filled out a post-experiment questionnaire that evaluated their perceptions of the robot. Participants received $5 for their participation. The entire experiment took approximately 30 minutes per participant.

4.2. Participants

A total of 29 participants (16 males, 13 females), whose ages ranged 18–38 ($M = 22.62, SD = 4.35$), were recruited from the University of Wisconsin–Madison campus. There were six, eight, seven, and eight participants in the learning-based, unimodal, random, and conventional conditions, respectively. Participants reported relatively low familiarity with robots ($M = 2.48, SD = 1.53$) and with the process of making paper ($M = 1.69, SD = 1.11$) in seven-point rating scales.

4.3. Measurement and Analysis

We used a post-experiment questionnaire to evaluate the participants' perceptions of the robot's behavior. Four measurement scales were developed using seven-point questionnaire items. *Immediacy*, defined as psychological distance between individuals [22], assessed how close participants felt to the robot and how engaging they thought the robot was (3 items, Cronbach's $\alpha = .79$). *Naturalness* gauged how natural the robot's motions were (5 items, Cronbach's $\alpha = .84$). *Effectiveness* measured how participants perceived the robot's effectiveness as a presenter (4 items, Cronbach's $\alpha = .87$). *Likability* evaluated how likable the robot was (8 items, Cronbach's $\alpha = .88$). An additional item measured the robot's *credibility* using a question about whether or not the robot provided sufficient information for the participant to answer the quiz questions. Moreover, we asked participants to choose from a list of 20 adjectives to describe the robot's overall behavior. The list of 20 adjectives consisted of 10 positive and 10 negative adjectives. Two manipulation-check items were included to ensure that manipulations in the trained model to create the unimodal and random conditions were successful.

In addition to questionnaire evaluation, we evaluated participants' performance in retelling the information that the robot presented, calculating measures of *familiarity* with the narration topic, their use of *body language*, and an *overall evaluation* of presenter effectiveness. Three raters who were blind to the experimental conditions rated video recordings of the participants' retelling performance on these measures using seven-point scales. Inter-rater reliability analysis using intra-class correlation coefficient (ICC) [36] as a measure revealed high correlations among the three raters on measures of familiarity of content (ICC(3,3)=.884), use of body language (ICC(3,3)=.897), and overall evaluation (ICC(3,3)=.771).

While we did not develop any hypothesis regarding the effect of the robot's behaviors on participants' learning of the information presented by the robot, we included an exploratory measure that involved a quiz consisting of 18 questions on this information. This measure explored whether or not the manipulations in the robot's behaviors affected participants' recall of the narrated information.

One-way fixed-effects analysis of variance (ANOVA) tests, using the manipulation in the robot's behaviors as the fixed factor, were conducted to analyze the manipulation checks, questionnaire measures, quiz data, and retelling evaluation. Planned many-to-one multiple comparisons used the Dunnett's method, considering the learning-based approach as the comparison baseline, to assess how the unimodal, random, and conventional conditions compared against the learning-based condition. Additionally, we performed Dunnett's tests, considering the unimodal and random conditions as baselines, using data on manipulation checks to verify whether the learned model was successfully manipulated to create the unimodal and random conditions. To determine whether behaviors produced by the learning-based and conventional approaches are comparable, we followed guidelines suggested by Walker and Nowacki [39] and Julnes and Mohr [12], applying a conservative equivalence margin of 0.50 (i.e., $p > .50$) to the comparisons between these two conditions.

4.4. Results

4.4.1. Manipulation checks

To check whether our manipulations to create the unimodal behavior condition was successful, we asked participants whether or not they perceived the robot to be motionless. The analysis of variance found a significant effect of our manipulation on this measure, $F(3,25) = 22.88, p < .001$. Comparisons using the Dunnett's test showed that participants in the unimodal condition perceived the robot to be more motionless than those in the learning-based, $p < .001$, random, $p < .001$, and conventional, $p < .001$, conditions. We also asked participants whether the robot's motions appeared random to verify that we successfully created the random condition. The analysis found a significant effect of our manipulation on this measure, $F(3,25) = 6.12, p = .003$. Comparisons showed that participants in the random condition perceived the robot's motions to be more random than those in the learning-based, $p = .005$, unimodal, $p = .017$, and conventional, $p = .002$, conditions did.

4.4.2. Perceptions of the robot

We found that our manipulation had a significant effect on the perceived immediacy, $F(3,25) = 6.91, p = .002$, naturalness, $F(3,25) = 5.77, p = .004$, effectiveness, $F(3,25) = 4.59, p = .011$, and likability, $F(3,25) = 6.44, p = .002$, of the robot, but not on participants' perceptions of the robot's credibility, $F(3,25) = 2.07, p = .130$. Comparisons further revealed that the learning-based and conventional approaches showed equivalence in measures of perceived immediacy, $p = .923$, naturalness, $p = .917$, effectiveness, $p = .999$, likability, $p = .722$, and credibility, $p = .645$.

Comparisons also showed that participants in the learning-based condition rated the robot to have higher immediacy, $p = .002$, and to be more natural, $p = .003$, more effective, $p = .027$, and more likable, $p = .023$, than those in the unimodal condition. However,

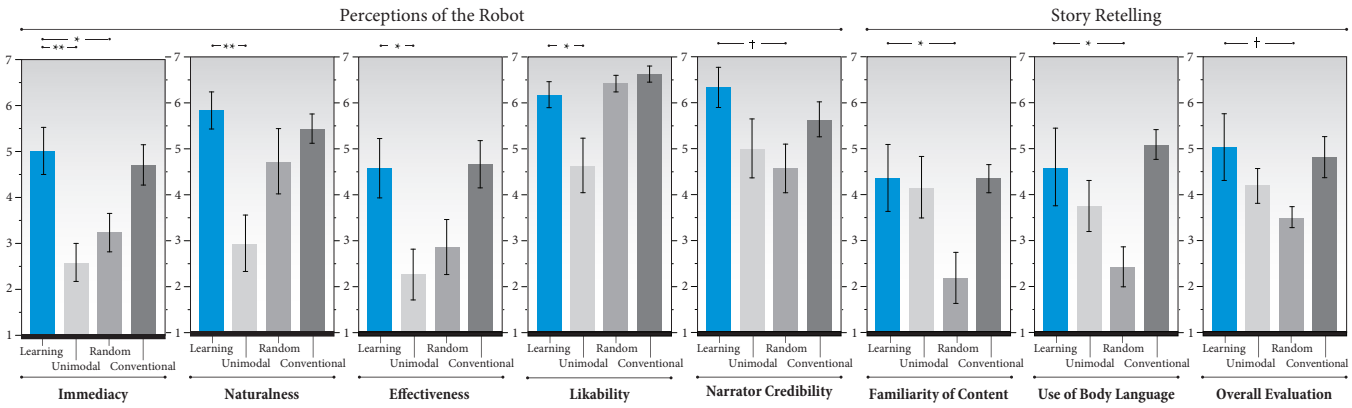


Figure 7: Results on perceptions of the robot and retelling performance. Only significant results are marked. (NS), (†), (*), (**), and (***) denote $p > .10$, $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively.

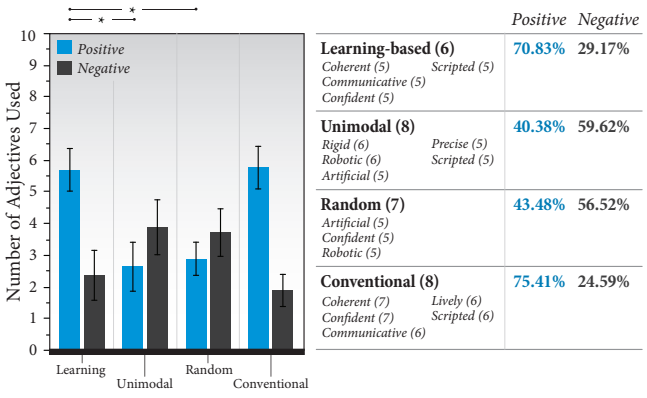
no significant differences were found between the learning-based and unimodal conditions in perceived credibility, $p = .191$. While the comparisons showed that participants in the learning-based condition perceived the robot to have higher immediacy, $p = .032$, and marginally more credibility than in the random condition, $p = .072$, no differences were found between the learning-based and random conditions in measures of naturalness, $p = .378$, effectiveness, $p = .131$, and likability, $p = .931$. Figure 7 summarizes these results.

Our analysis also found a significant effect of our manipulation on the number of positive adjectives that participants used to describe the robot's behaviors, $F(3,25) = 6.44, p = .002$. Comparisons showed that participants in the learning-based condition used more positive adjectives than those in the unimodal, $p = .013$, and random, $p = .027$, conditions did. The learning-based and conventional conditions demonstrated equivalence in the use of positive adjectives, $p = .999$ (Figure 8) . However, our manipulation did not have a significant effect on the number of negative adjectives used, $F(3,25) = 1.86, p = .162$. Comparisons showed that participants in the unimodal, $p = .354$, random, $p = .460$, and conventional, $p = .947$, conditions used a similar number of negative adjectives in describing the robot's behavior to those in the learning-based condition did. Figure 8 shows the top three adjectives used to describe the robot's behaviors.

4.4.3. *Retelling performance*

We found a significant effect of our manipulation on participants' perceived familiarity with the narration topic, $F(3,25) = 3.37, p = .034$, and effective use of body language, $F(3,25) = 4.67, p = .010$, but not on the overall evaluation of the participant as an effective presenter, $F(3,25) = 2.16, p = .118$. Comparisons revealed equivalence between the learning-based and conventional approaches with respect to familiarity of the topic, $p = 1.000$, effective use of body language, $p = .848$, and overall evaluation of presenter effectiveness, $p = .977$, as shown in Figure 7.

Comparisons also showed that participants in the learning-based condition were rated to be more familiar with the topic, $p = .042$, and to more effectively use body language, $p = .033$, than those in the random condition. They were also rated marginally higher in overall evaluation than those in the random condition, $p = .084$. However, comparisons did not find significant differences in participants' familiarity with the topic, $p = .986$, effective use of body language, $p = .569$, and overall performance as an effective presenter, $p = .447$, between the learning-based and unimodal conditions (Figure 7).



	Positive	Negative
Learning-based (6) *Coherent (5)* *Scripted (5)* *Communicative (5)* *Confident (5)*	70.83%	29.17%
Unimodal (8) *Rigid (6)* *Precise (5)* *Robotic (6)* *Scripted (5)* *Artificial (5)*	40.38%	59.62%
Random (7) *Artificial (5)* *Confident (5)* *Robotic (5)*	43.48%	56.52%
Conventional (8) *Coherent (7)* *Lively (6)* *Confident (7)* *Scripted (6)* *Communicative (6)*	75.41%	24.59%

Figure 8: Results on participants' use of positive and negative adjectives in describing the robot's behavior. Top three choices of adjectives and percentages of used positive and negative adjectives are listed. Values in parentheses indicate how many participants used the adjective to describe the robot. Only significant results are marked. (*) denote $p < .050$.

4.4.4. *Cognitive assessment*

The manipulation in the robot's behaviors did not have a significant effect on participants' information recall in the quiz, $F(3,25) = 0.98, p = .418$. No significant differences were found between the learning-based condition and the unimodal, $p = .621$, random, $p = .866$, and conventional, $p = .986$, conditions in comparisons.

4.4.5. *Summary and Discussion*

Overall, our results showed that participants in the learning-based condition rated the immediacy, naturalness, effectiveness, and likability of the robot higher than those in the unimodal condition did. Participants in the learning-based condition were also rated higher in their retelling of the narration topic compared with those in the random condition. Additionally, the learning-based and conventional conditions showed equivalence in participants' perceptions of the robot and retelling performance.

While behaviors in the unimodal condition negatively affected how participants perceived the robot in terms of immediacy, naturalness, effectiveness, and likability, it did not affect participants' perceived credibility of the robot, retelling performance, or information recall. The following excerpts from post-experiment interviews provide some insight into these results:

> *"The robot didn't particularly move, which I didn't think was realistic, and also it didn't make eye contact with me."*

> *"When it switched steps, it would be more engaging if it would [like] point to the new step..."*

> *"I was more focused on what the robot was saying and the screen..."*

While participants in the unimodal condition noticed that the robot's behaviors were not natural, they still paid attention to the verbal information and the projected illustration to understand the presented information, demonstrating comparable results in perceived credibility, retelling of the topic, and learning.

Participant comments also provided insight into why the random manipulation did not affect their perceptions of the robot but affected their retelling performance, as illustrated in the excerpts below:

> *"The motion was kind of distracting ... sometimes maybe it used hand gestures in a way that I wouldn't necessary to use that ... [but] I feel like people do that sometimes too..."*

> *"It was very distracting... what are you [the robot] trying to do, I don't know what you [the robot] are trying to do. It was just odd."*

Most participants in the random condition found the robot's behaviors to be distracting, which suggests that participants may have found it difficult to focus on the information presented by the robot, potentially leading to poor retelling performance. The discrepancy between quiz results and topic familiarity in retelling may reflect the effects distraction may have had on their learning.

However, while some of the participants in the random condition perceived that the robot's behaviors deviated from social norms, others found these behaviors to be acceptable. These results are consistent with previous work on robots' use of gestures to support speech, which found no differences in participants' ratings of a robot displaying gestures that were semantically incongruent with its speech and one that displayed congruent behaviors in measures of how lively, active, engaged, communicative, and fun-loving they perceived the robot to be [34].

The participants perceived the robot in the learning-based condition as showing higher immediacy—displaying more engagement and psychological closeness—than they did in the unimodal and random conditions. This result is consistent with research in education, which shows that instructors with high immediacy use more gestures and greater eye contact with students [22].

Finally, our results suggested that the behaviors driven by the learning-based and conventional approaches demonstrated similar effectiveness in every measure. Participant comments also highlight similar limitations both approaches have in generating natural, humanlike behaviors, as suggested by the following comments on the robot's gaze behavior, the first by a participant in the learning-based condition and the other two from participants in the conventional condition:

> *"It seemed not to sometimes make eye contact when I expected it to. There were a few periods where it was talking and talking straight at the screen and not making eye contact."*

> *"What it [the robot] could do more is to look as if it is an actual human were presenting. It could look more at the screen [be]cause that's how I feel a lot people do."*

> *"There were a few times I noticed it pointed to the screen without looking at it."*

These comments also highlight the variability in people's expectations of the behaviors of an effective presenter and suggest that gaze behaviors controlled by the learning-based and conventional approaches will require further improvement to meet the expectations of a broader population of users.

5. GENERAL DISCUSSION

5.1. Design Implications

To generate robot behaviors that enable natural and effective human-robot interaction, researchers and designers have constructed models of multimodal behavior from human data through an inspection-based process that identifies behavioral patterns and extracts alignment parameters for behaviors in different modalities. While this process has been successfully used to design robot behaviors that achieve predicted outcomes (e.g., [9, 10, 29, 40]), the inspection-based approach to identifying and extracting behavioral parameters does not scale well to more challenging modeling tasks, such as modeling relationships among a large number of multimodal behaviors from large datasets. The learning-based approach presented in this paper facilitates this process by automatically learning behavioral patterns and alignment parameters from annotated human data and promising greater scalability to large datasets and a large number of behavioral modalities. Additionally, the learning-based approach offers formalisms in the form of probabilistic representations of the design space for robot behaviors, which might facilitate the validation of the learned models and extensions of these models using methods such as active learning. While our results show that the learning-based and conventional approaches reach comparable effectiveness in generating robot behaviors, we expect the learning-based approach to better capture the dynamics of multimodal behavior when more data and behavioral modalities are considered.

5.2. Limitations and Future work

A key limitation of our work is the assumption that the Markov process that underlies the production of multimodal behavior is discrete, following an empirically determined window size of 500 ms. However, human behavior is inherently more continuous. While the use of this discrete window yielded acceptable behaviors, a consideration of human behavior as a continuous process in modeling might yield better outcomes. We also assumed that the duration of state transitions were identical across different behavioral modalities, although different behaviors might have different state duration. For example, gaze behavior might change more frequently than gestures. Methods that explicitly model duration, such as the use of a hidden semi-Markov model (HSMM), might be more appropriate for modeling these temporal characteristics.

In designing robot behaviors, we considered the speaker's behaviors independent of the recipient's behaviors. While this approach might be adequate for a narration scenario, extending our work to more interactive scenarios such as conversations will require jointly modeling recipient and speaker behaviors using modeling approaches such as the coupled hidden Markov model (CHMM). We also focused on gaze and gesture as nonverbal channels, while other nonverbal behaviors, such as head nods and facial expressions, play a key role in communication.

In our proposed approach, the specific network structure of the model, which we constructed based on literature on human communication, has a significant effect on the success of the model in achieving natural, effective humanlike behavior, and alternative network structures or structures that are learned directly from data might achieve different or better results. Finally, this work used simple, proof-of-concept mechanisms to link components in the feature and domain levels (Figure 2). Future work could extend the learning-based approach to jointly learn at both feature and domain levels and the associations between the two levels, which might facilitate the generation of richer and more accurate robot behaviors.

6. CONCLUSION

Human behavior offers a rich space for designing natural and effective behaviors for robots. Researchers and designers have explored ways of capturing the patterns in which humans behave as models and heuristics to control robot behaviors. However, these approaches usually involve a laborious process of inspecting large volumes of data to identify temporal patterns in behaviors and alignments across behaviors in different modalities. To facilitate this process, we proposed a *learning-based* modeling approach that uses probabilistic graphical models (PGMs) to automatically learn distribution and alignment parameters from data on human behavior. In this paper, we demonstrated a specific instantiation of this approach that used a dynamic Bayesian network (DBN) to model speech, gaze, and gesture behaviors in a narration task and to estimate distribution and alignment parameters for these behaviors to enable a humanlike robot to perform the same narration task. We then evaluated the effectiveness of our approach in achieving natural, humanlike robot behaviors by comparing it against a number of baselines. The results showed behaviors generated by a learned model to be more effective than no behavior and random behavior baselines in a number of measures and equally effective as those controlled by a set of heuristics. Our findings suggest that the proposed learning-based approach offers design outcomes that are comparable to those obtained by heuristics-based approaches while promising significant reduction in the effort involved in constructing models of multimodal behavior and greater scalability to more complex modeling tasks toward achieving richer and more natural human-robot interactions.

7. ACKNOWLEDGMENTS

This research was supported by National Science Foundation awards 1017952 and 1149970 and an equipment loan from Mitsubishi Heavy Industries, Ltd. We would like to thank Jilana Boston, Brandi Hefty, Ross Luo, Catherine Steffel, and Lindsay Jacobs for their help in our work.

8. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, 1976.

[2] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1):3–55, 1999.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society*, pages 1–38, 1977.

[4] J. F. Ferreira, M. Castelo-Branco, and J. Dias. A hierarchical bayesian framework for multimodal active perception. *Adaptive Behavior*, 20(3):172–190, 2012.

[5] C. Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press New York, 1981.

[6] Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82(1):B1–B14, 2001.

[7] C. C. Heath. Gesture's discrete tasks: Multiple relevancies in visual conduct in the contextualization of language. In P. Auer and A. Di Luzio, editors, *The contextualization of language*, pages 102–127. John Benjamins, 1992.

[8] A. Henderson, F. Goldman-Eisler, and A. Skarbek. Sequential temporal patterns in spontaneous speech. *Language and Speech*, 9:207–216, 1966.

[9] C.-M. Huang and B. Mutlu. Robot behavior toolkit: Generating effective social behaviors for robots. In *Proc. HRI 2012*, 2012.

[10] C.-M. Huang and B. Mutlu. Modeling and evaluating narrative gestures for humanlike robots. In *Proc. RSS 2013*, 2013.

[11] J. Jaffe, L. Cassotta, and S. Feldstein. Markovian model of time patterns of speech. *Science*, 144:884–886, 1964.

[12] G. Julnes and L. B. Mohr. Analysis of no-difference findings in evaluation research. *Evaluation Review*, 13(6):628–655, 1989.

[13] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita. An affective guide robot in a shopping mall. In *Proc. HRI 2009*, pages 173–180, 2009.

[14] A. Kendon. Looking in conversation and the regulation of turns at talk: A comment on the papers of g. beattie and d. r. rutter et al. *British Journal of Social and Clinical Psychology*, 17:23–24, 1978.

[15] A. Kendon. *Gesticulation and Speech: Two Aspects of the Process of Utterance*. Mouton De Gruyter, 1980.

[16] A. Kendon. Do gestures communicate? a review. *Research on Language and Social Interaction*, pages 175–200, 1994.

[17] D. Kollar and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[18] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[19] J. Lee and S. Marsella. Modeling speaker behavior: a comparison of two approaches. In *Proc. IVA 2012*, pages 161–174, 2012.

[20] M. Lee, S. Kiesler, and J. Forlizzi. Receptionist or information kiosk: how do people talk with a robot? In *Proc. HRI 2010*, 2010.

[21] D. McNeill. *Hand and Mind*. The University of Chicago Press, 1992.

[22] A. Mehrabian. *Silent messages*. Wadsworth, 1971.

[23] A. S. Meyer, A. M. Sleiderink, and W. J. M. Levelt. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66(2):B25–B33, 1998.

[24] Y. Mohammad and T. Nishida. Learning interaction protocols using augmented baysian networks applied to guided navigation. In *Proc. IROS 2010*, pages 4119–4126, 2010.

[25] Y. Mohammad, T. Nishida, and S. Okada. Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction. In *Proc. IROS 2009*, pages 2537–2544, 2009.

[26] L.-P. Morency. Modeling human communication dynamics. *IEEE Signal Processing Magazine*, 27(5):112–116, 2010.

[27] K. Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.

[28] K. P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[29] B. Mutlu, J. Forlizzi, and J. Hodgins. A storytelling robot: Modeling and evaluation of humanlike gaze behavior. In *Proc. HUMANOIDS 2006*, pages 518–523, 2006.

[30] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM TiiS*, 1(2):12:1–12:33, 2012.

[31] V. Ng-Thow-Hing, P. Luo, and S. Okita. Synchronized gesture and speech production for humanoid robots. In *Proc. IROS 2010*, pages 4617–4624, 2010.

[32] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. Providing route directions: design of robot's utterance, gesture, and timing. In *Proc. HRI 2009*, pages 53–60, 2009.

[33] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances. In *Proc. ICMI 2007*, pages 255–262, 2007.

[34] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, pages 201–217, 2012.

[35] E. Schegloff. On some gestures' relation to speech. In J. M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in conversational analysis*, pages 266–296. Cambridge Univ. Press, 1984.

[36] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[37] J. Sidnell. Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction*, 39(4):377–409, 2006.

[38] J. Streeck. Gesture as communication i: Its coordination with gaze and speech. *Communication Monographs*, 60(4):275–299, 1993.

[39] E. Walker and A. S. Nowacki. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26:192–196, 2011.

[40] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka. Precision timing in human-robot interaction: coordination of head movement and utterance. In *Proc. CHI 2008*, pages 131–140, 2008.