Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 00 (2019) 1–19
Speech Communi-cation

Speech-Driven Animation with Meaningful Behaviors

Najmeh Sadoughi, Carlos Busso

The University of Texas at Dallas

Abstract

Conversational agents (CAs) play an important role in *human computer interaction* (HCI). Creating believable movements for CAs is challenging, since the movements have to be meaningful and natural, reflecting the coupling between gestures and speech. Studies in the past have mainly relied on rule-based or data-driven approaches. Rule-based methods focus on creating meaningful behaviors conveying the underlying message, but the gestures cannot be easily synchronized with speech. Data-driven approaches, especially speech-driven models, can capture the relationship between speech and gestures. However, they create behaviors disregarding the meaning of the message. This study proposes to bridge the gap between these two approaches overcoming their limitations. The approach builds a *dynamic Bayesian network* (DBN), where a discrete variable is added to constrain the behaviors on the underlying constraint. The study implements and evaluates the approach with two constraints: discourse functions and prototypical behaviors. By constraining on the discourse functions (e.g., questions), the model learns the characteristic behaviors associated with a given discourse class learning the rules from the data. By constraining on prototypical behaviors (e.g., head nods), the approach can be embedded in a rule-based system as a behavior realizer creating trajectories that are timely synchronized with speech. The study proposes a DBN structure and a training approach that (1) models the cause-effect relationship between the constraint and the gestures, and (2) captures the differences in the behaviors across constraints by enforcing sparse transitions between shared and exclusive states per constraint. Objective and subjective evaluations demonstrate the benefits of the proposed approach over an unconstrained baseline model.

Keywords:
Speech-Driven Animation; Dynamic Bayesian Network; Generation of Meaningful Behaviors.

1. Introduction

Body language is an essential part of face-to-face conversations. People consciously or unconsciously use head motion, hand gestures, and facial expressions while speaking. We use these modalities for multiple purposes including to emphasize ideas, parse sentences into smaller syntactic units, complement verbal information, and express our emotions. Therefore, incorporating naturalistic behaviors that fulfill these communication goals is important in the design of a *conversational agent* (CA) [1]. CAs are playing a relevant role in several fields including business enterprises, healthcare, entertainment, and education. Their use has also increased with new website and mobile applications, providing a great platform for virtual reality, visual aid for hearing impaired individuals, and virtual agents for online shopping [2].

Creating behaviors that are perceived natural while conveying the underlying meaning in the message is challenging. Most studies in this field have relied on either rule-based or data-driven systems [3]. Rule-based systems create

contextual rules to trigger behaviors, emphasizing the semantic and syntactic information [1, 4, 5]. However, the variation of the gestures generated using rules is bounded by the predefined dictionary of handmade gestures [6]. Furthermore, scheduling the movement with speech is challenging [7, 8]. Data-driven approaches learn the behaviors directly from data. There are several studies that have used speech to create behaviors [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Speech prosody is highly correlated with facial expression and head movements [20], so it is possible to generate behaviors that are timely aligned with speech (rhythm, emphasis). However, speech-driven methods do not consider the meaning of the sentence. While the gesture may be perfectly aligned with speech, its meaning may contradict the message (e.g., nodding while saying "no").

This study leverages the advantages of rule-based and speech-driven systems, bridging the gap between these methods by overcoming their drawbacks. We address this problem by constraining our speech-driven model by contextual information to generate behaviors with meaning. This approach relies on *dynamic Bayesian models* (DBNs) capturing the temporal relationship between speech and gestures (in this study, hand and head motion). We introduce the constraints as an extra discrete variable that conditions the state configuration between speech and gestures, modeling the specific behavioral characteristics associated with the given constraint. We demonstrate the potential of the proposed approach with two evaluations, where the constraints are either discourse functions (negation, affirmation, questions and suggestion) or predefined prototypical hand and head gestures. With the discourse functions, we aim to synthesize head movement trajectories that are commonly associated with a given discourse function (e.g., head roll for questions, head shakes for negations). Instead of creating hand-crafted rules, the proposed model learns the statistical patterns from the data. For prototypical gestures, we aim to learn statistical models that generate pre-defined hand and head behaviors, and their joint representations with prosody features. This model plays the role of a behavior realizer in the SAIBA framework [4], and has the potential to be integrated into a rule-based system. We consider three prototypical hand gestures (*To-Fro*, *So-What*, and *Regress*) and head gesture (*Head Nod* and *Head Shake*). The constraints are introduced as input to the system, changing the discrete variable that conditions the generated gestures. During synthesis, the models will create novel realizations of these gestures that are timely synchronized with speech. The proposed models are effective, producing realizations that are perceived more natural than the unconstrained models, bridging the gap between rule-based and speech-driven methods.

The paper is organized as follows. Section 2 describes previous studies that are relevant to our work, emphasizing the contributions of our paper. Section 3 provides an overview of the framework. Section 4 presents the resources used in this study, including the database and features. It also describes the approach used to obtain the annotations for the discourse functions and prototypical behaviors. Section 5 describes the baseline framework. Sections 6 presents the proposed approach, which is evaluated in section 7. Section 8 finalizes the paper, with a summary of the work, and future research directions.

2. Related Work

Several studies have proposed schemes to generate gestures, which can be categorized into rule-based and data-driven methods.

2.1. Rule-Based Systems

Cassell et al. [1] presented one of the early studies on rule-based framework to synthesize behaviors. They defined several rules to generate appropriate behaviors dictated by the meaning of the message. In a later study, Cassell et al. [8] introduced the *behavior expression animation toolkit* (BEAT), which uses text to create animations with appropriate and synchronized gestures. BEAT tags semantic labels in the text, which are mapped into appropriate behaviors by heuristics rules suggested after observing human-human nonverbal displays. The synchronization is decided based on the timing of the words in the text. Poggi et al. [21] introduced GRETA, which is an *embodied conversational agent* (ECA), comprising several modules such as emotional mind, dialog manager, plan enrichment and body generator. The body enrichment module labels text with appropriate behaviors assigning synchronization points, which are realized by the body generator module. GRETA includes a number of predefined gestures which can be exploited to generate animations with specific communicative goals. Kopp and Wachsmuth [22] proposed to find a prominent word or phrase to synchronize speech and gestures. The prominent words convey the communicative goal, creating anchors for the peak of the gesture. Marsella et al. [23] proposed a framework to generate animation

from speech. Their system uses an *automatic speech recognition* (ASR) module to get the transcriptions, which are semantically analyzed to extract communicative goals in the message. They defined a list of behaviors associated with the communicative goals, mapping the text to behaviors. Their system also analyzes emotional cues in speech extracting arousal level, which dictates the selection of the behaviors generated for each communicative goal.

2.2. Data-Driven Systems

An alternative approach to generate behaviors is using data-driven methods that exploit the relationship between body movements and acoustic features (e.g., prosody). Vigot et al. [24] demonstrated that there is a statistically significant correlation between prosodic features and raw body movements. Graf et al. [25] showed that there is correlation between prosodic events and behaviors such as eyebrow and head movements. Busso et al. [20] reported that the correlation between prosodic features and head movements across different emotions are on average more than $\rho = 0.69$, using *canonical correlation analysis* (CCA). Speech and gestures also co-occur. The study from McNeill [26] showed that more than 90% of the gestures occur while speaking, showing the tied connections between these modalities. These results have motivated synthesizing behaviors using speech-driven models.

Busso et al. [11] proposed emotion dependent *hidden Markov models* (HMM) to synthesize head movements with prosodic features. Mariooryad et al. [13] investigated several *dynamic Bayesian networks* (DBNs) to jointly model head and eyebrow movements driven by speech, capturing the dependencies not only between speech and facial behaviors, but also between head and eyebrow motions. Some studies have argued that speech is correlated with the kinematics of the behaviors. Le et al. [16] proposed to jointly model prosodic features and kinematic features of head motion using *Gaussian Mixture Models* (GMM). Levine et al. [14] presented a system to synthesize body movement using *hidden conditional random fields* (HCRFs), modeling the relationship between prosody and kinematic features of the joint rotations. The task was to predict kinematic parameters from speech. They use reinforcement learning to select behaviors in the database that match the inferred kinematics parameters. Bozkurt et al. [27] designed a system for generating upper body beat gestures based on prosodic features. They clustered prosodic features into intonational phrases, and movements into gestural phrases. These units were jointly modeled using a *hidden semi Markov model* (HSMM), which allowed asynchrony between the gestures and prosodic phrases by modeling the state duration of the hidden state. Chiu et al. [15] proposed to use *hierarchical factored conditional restricted Boltzmann machines* (HFCRBMs) which learns how to generate the joint poses for the next frame based on the previous frame conditioned on the prosodic features.

2.3. Hybrid Approaches

Rule-based and data-driven methods have advantages and disadvantages. Rule-based methods do not capture the range of behaviors observed during human interaction, are limited by the stored behaviors, and often result in repetitive behaviors. The synchronization between behaviors and speech is challenging, since they do not learn the synchronization from natural recordings. However, they can consider the meaning of the message to derive appropriate behaviors. Speech-driven methods can capture broader variations of behaviors, learning appropriate synchronization between speech and gestures. However, they may not create appropriate behaviors that match the intended communicative goal. Using pure speech-driven methods may be enough to predict beat gestures but not iconic or metaphoric gestures which are closely related to the message. Bridging the gap between rule-based and data-driven frameworks has the potential to create behaviors that are meaningful, timely synchronized, and representative of the range of variations observed during human interaction.

Studies have attempted to combine both approaches creating hybrid frameworks. Stone et al. [28] designed a hybrid system to generate meaningful behaviors given the text. They jointly segment audio and motion capture recordings into units expressing pre-defined communicative intents. Given an input text, they parse the input into their predefined categories, using dynamic programming to find speech and motion capture segments that have the same communicative goal. The generation with this framework is limited to the stored speech segments. Sadoughi et al. [29] proposed to constrain a speech-driven model based on the discourse functions of the sentence to generate more meaningful head and eyebrow motion. The study was limited to only two discourse functions: *affirmation* and *question*, where the subjective evaluation of the result showed improvements for the constrained model versus the unconstrained model when the constraint was *question*.

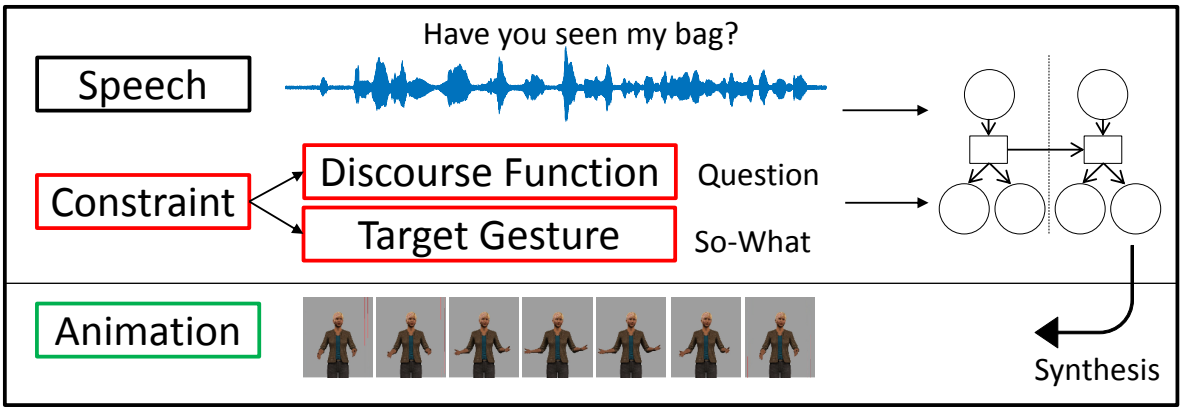


Figure 1. Overview of the proposed system to generate behaviors. In addition to speech, the models take constraints as input (either discourse functions of target gesture), generating meaningful data-driven behaviors.

2.4. Contribution and Relation to our Prior Work

This paper proposes a framework to create meaningful behaviors driven by speech. This framework creates animations based not only on prosodic features, but also on constraints which are either discourse functions (i.e., semantic structure of speech), or prototypical behaviors (e.g., head nods). This study builds upon our previous work, which we summarize in this section.

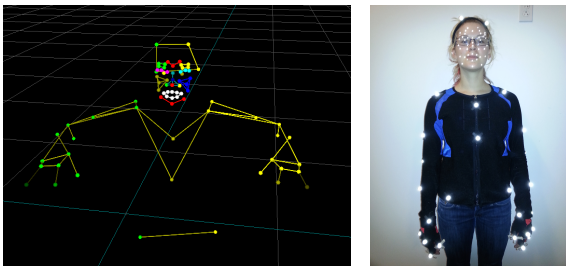
Sadoughi et al. [29] proposed a model to generate speech-driven head and eyebrow movements constrained on discourse functions. The preliminary study tested the constrained model on one session of the IEMOCAP corpus, constraining the models on two dialog acts: *question*, and *affirmation*. The subjective evaluation of the models showed that the behaviors from the constrained models are more preferable, natural and appropriate compared to the unconstrained model. In Sadoughi and Busso [30], we explored the idea of constraining the models using prototypical behaviors such as head nods. The models were trained with gestures directly retrieved from the corpus by providing few examples.

The models presented in our preliminary studies have several limitations. First, the variability of the generated behaviors is limited since the model optimization is susceptible to a poor initialization to reduce the mean square error, often resulting in average trajectories. Second, the structure of the proposed models requires balanced datasets per constraint, which is an unnecessary restriction. This study presents an improved speech-driven model that overcomes these problems by changing the structure of the model and the training strategy, which systematically reduces the confusion between the constraints during training. The contributions of this paper are (1) designing a constrained speech-driven model to generate more meaningful behaviors (Sec. 6.1) and (2) a novel training approach to effectively learn distinct patterns associated with different constraints (Sec. 6.2).

3. Overview

This study aims to improve nonverbal displays of CAs using speech-driven models that are constrained by either the underlying discourse function in the message or prototypical behaviors specified by rule-based systems. In an attempt to create meaningful gestures, Marsella et al. [23] defined several functions based on the content of the speech. These discourse related functions create a mapping between content and behaviors. Likewise, Poggi et al. [21] designed a toolkit with several embedded functions to generate behaviors. The inputs to these mappings are communicative goal of the utterance, which we call discourse functions. These discourse functions are associated with relevant gestures that contribute in understanding the underlying message of the speech.

Figure 1 gives the overview of our system, which takes as input speech and the underlying discourse function or intended gesture, producing meaningful behaviors that are timely synchronized with speech, convey the right message, and display the range of behaviors observed during human interactions. Our framework can take the role of behavior realizer proposed under the SAIBA framework [4], bridging the gap between rule-based and data-driven approaches.



(a) Upper-body markers (b) Actress with markers

Figure 2. The MSP-AVATAR database [31]. One of the subjects wore markers that were tracked with a VICON system.

We aim to answer the following questions in the context of CAs, "do discourse functions affect behaviors?" If so, "is there a principled framework to capture the characteristics behaviors associated with discourse functions?" Knowing the target gesture, "can we effectively create the gesture which is synchronized with and modulated by speech?" We address these questions by exploring four discourse functions: *negation*, *affirmation*, *question* and *suggestion*. We propose a principled speech-driven approach to capture the characteristic behaviors for each discourse function. We also propose to constrain the models with prototypical behaviors. While the framework is general, we evaluate three hand gesture and two head gestures.

4. Resources: Database, Features and Annotations

This section provides a brief description of the corpus and features used in this study, focusing on the annotation process and the method used to retrieve target gestures.

4.1. The MSP-AVATAR Database

We collected the MSP-AVATAR corpus [31] to study the role of discourse functions and gestures. This corpus was collected to provide data to synthesize more meaningful and naturalistic behaviors.

The MSP-AVATAR corpus contains recordings of dyadic interactions based on improvisations of daily scenarios. It encompasses the recordings from six actors interacting in four dyadic interactions. The scenarios are carefully designed such that they include the use of eight discourse functions: *contrast*, *confirmation/negation*, *question*, *uncertainty*, *suggestion*, *giving orders*, *warning*, and *informing*. There are also scenarios prompting the actors to use iconic gestures (e.g., large, small) and deictic gestures for *pronouns* (e.g., "me", "you"). The discourse functions in this corpus are carefully chosen based on previous studies [23, 21], which are likely to elicit characteristic behaviors.

The corpus consists of audio, video and motion capture recordings, collected at the Motion Capture laboratory of the University of Texas at Dallas. In each dyadic session, one of the actors wore 43 facial markers, and a suit in which we attached 28 markers (Fig. 2). Therefore, we have motion capture data for four different subjects. The facial markers include most of the *feature points* (FPs) in the MPEG-4 standard [32]. For the upper body, we follow the position of the markers in the *Vicon skeleton template* (VST). For each of the actors, we used a Lavalier microphone (SHURE MX150) connected to a portable digital recorder (TASCAM DR-100MKII). The microphone recorded the speech at a resolution of 16 bit and a sampling rate of 44.1 KHz. We used two Sony Handycams (HDR-XR100) which recorded at 1,920 $\times$ 1,080 resolution. The cameras were positioned to record the frontal view of each actor, without interfering with the Vicon system. In total, we have 74 sessions with a duration of two hours and fifty eight minutes.

4.2. Motion & Audio Features

The data-driven models take speech features as input, generating the most likely behaviors. This study considers head and hand gestures. We use the upper body joint rotations derived after solving the skeleton of the motion capture recordings in Blade. We consider the pitch, yaw, and roll rotations for the head (i.e., 3 *degree of freedom* (DOF)), arms (3 DOF $\times$ 2) and forearms (2 DOF $\times$ 2). We normalize the motion capture data per subject to reduce the scale mismatches between subjects. We use the z-normalization, where we subtract the mean of a variable, dividing the

result by its standard deviation. This normalization allows the model to learn more generic dependencies between speech and motion capture features. The sampling rate for the motion capture data is 120fps.

The acoustic features correspond to prosodic features, following our previous work [11, 13, 17, 29, 33]. We extract the fundamental frequency and energy using Praat [34], estimating their first and second order derivatives resulting in a 6D feature vector. These features are extracted using 40ms windows every 16.67ms with 23.3ms overlap (i.e., 60fps). We interpolate the unvoiced segments in the fundamental frequency to avoid discontinuities. The feature vector is up-sampled to match the sampling rate of the motion capture data (i.e., 120fps). This approach provides smooth speech features. We address speaker-dependent differences in the acoustic domain by separately normalizing the speech features for each subject (i.e., z-normalization).

4.3. Annotation of Discourse Functions

We manually annotated the 74 sessions, identifying sentences associated with discourse functions. Some of the discourse classes are harder to reliably annotate, so we only consider four classes: asking questions (*question*), showing agreement (*affirmation*), showing disagreement (*negation*), and making suggestions (*suggestion*). The evaluation was conducted with *Amazon mechanical turk* (AMT), using the OCTAB interface designed by Park et al. [35]. This toolkit is suitable for segmental annotations of the videos, where annotators can mark the beginning and end of segments in the videos where they noticed the requested discourse function. To improve the quality of the annotations, our approach identifies good evaluators using a screening phase. We ask the evaluators to annotate the discourse function *questions*, which is one of the easiest tasks. We also annotated this discourse function in our laboratory. We manually compared the annotations provided by each evaluator with our annotations, selecting the ones who provided reasonable answers. Then, we invited the selected evaluators to complete the rest of the assignments for the other three discourse functions. We recruited three annotators per assignment.

We use the method proposed by Zhou et al. [36] to aggregate the annotations coming from different annotators. This method solves a crowdsourcing model which estimates the hidden variables relevant to the difficulty of the tasks and the hidden variables relevant to the reliability of the annotators by using the minimax conditional entropy principle. The approach not only provides the hard labels after fusion, but also gives a confidence level in the assigned label (i.e., the soft assignment). To use this method, we consider our task as a binary classification task where each video frame either belongs to the target discourse function or not (30fps). We derive a soft assignment for each of the four discourse functions using the three evaluations per frame. We only consider the frames where the soft assignments are more than 0.9 for one of the discourse function, increasing the reliability in the labeled segments. Notice that the annotated frames are not mutually exclusive among the discourse functions. If we separately consider the co-occurrences between two or more discourse functions (e.g. *suggest* and *question*) as extra constraints, we would need enough realizations of these combinations. Unfortunately, the total durations of the co-occurrences of labels between two or more discourse functions vary between 0.5s to 310s, which is not enough. For simplicity, we enforce mutually exclusive segments by removing the overlaps, keeping as many segments as possible. This approach results in total durations of 734.4s for *affirmation*, 1,118.7s for *negation*, 1,149.1s for *question*, 1,582.5s for *suggestion*, and 6,111.7s for *other*.

4.4. Prototypical Behaviors

This study demonstrates that it is possible to create prototypical behaviors using data-driven models. While the framework is general, we only consider three prototypical behaviors for hand and two prototypical behaviors for head movements. Figure 3 illustrates the target hand gestures. The behaviors are defined as follows:
Head Nod: One or more pitch rotations of head.
Head Shake: One or more yaw rotations of head.
So-What: Movement of both hands in an arc in an outward manner (Fig. 3(a)).
To-Fro: Movement of both hands form side to side (Fig. 3(b)).
Regress: Movements of hands in circles towards the body (Fig. 3(c)).

The data-driven models require enough examples of these gestures to effectively train the models. Annotating the entire corpus with these prototypical gestures is a major task, as the annotations should include multiple simultaneous behaviors. Instead, we use the supervised framework introduced by Sadoughi and Busso [30] to automatically retrieve these instances from the dataset. The key idea is to annotate few examples of the target behavior, and retrieve the rest of

(a) So-What

(b) To-Fro

(c) Regress

Figure 3. Illustration of the three prototypical hand gestures considered in this study. These gestures were defined by Kipp [37].

Table 1. Prototypical gesture considered in the behavior retrieval framework. Number of examples in the train, and test & develop sets.

Region	Behavior	$\#Samples_{Train}$	$\#Samples_{Test\&Dev}$
HEAD	Nod	56	308
	Shake	39	237
HAND	To-Fro	47	77
	So-What	28	72
	Regress	24	114

the segments until we have enough data to train the models. The approach is a supervised approach that simultaneously solves the segmentation and detection of the target gestures. The first step is downsampling the motion capture sequences using clusters. This is a nonuniform downsampling approach that discards segments without variations while keeping changes in the trajectories. Then, we use a multi-scale sliding window framework that considers windows of different sizes, accounting for variation in the duration of the gestures. The next step is to determine whether the selected segments include the target gesture. The approach consists of two steps. In the first step, we screen the segments using one-class *support vector machine* (SVM), which reduces the potential segments, removing everything that departs from trajectories of the training examples. The second step uses the *dynamic time alignment kernel* (DTAK) algorithm to evaluate the candidate segments in more detail. For DTAK, we use the implementation provided by Zhou et al. [38]. The approach gives the flexibility to add new prototypical examples if needed (e.g., head roll, or pointing hand). We only need to identify few examples that are used to retrieve enough training example to build our models. This framework is a cost-effective, practical and scalable approach.

In Sadoughi and Busso [30], we set the detection threshold by maximizing the f-score. However, for this study it is more important that the selected segments are indeed from the target gestures (i.e, recall rate is less important). Therefore, in this study we set the detection thresholds per subject by maximizing the precision on the developing set.

We manually annotated three sessions per subject to evaluate the behavior retrieval framework ($3 \times 4 = 12$). Table 1 gives the number of examples annotated per target behavior in these 12 sessions (column $\#Samples_{Test\&Dev}$). These 12 sessions are partitioned into development (two session per speaker) and test (one session per speaker) sets using three-fold cross-validation. The development set is exclusively used to set the detection thresholds. Table 2 shows the accuracy of the behavior retrieval framework. The precision rates for head gestures are higher than 96%. For hand gestures the precision rates are higher for *so-what* (80%) and *regress* (90.5%). The precision rate is lower for *to-fro*, since this behavior is more complex.

Since our algorithm independently solves the detection of gestures, it is possible to have overlaps between two or more target gestures. We observe that the durations of these overlaps are 552.4s for head gestures, and between 30.2s

Table 2. The precision rates of the retrieved gestures on the test set.

Region	Behavior	$Samples_{Test}$	
		Precision [%]	Retrieved gestures [#]
ad	Shake	97.10	69

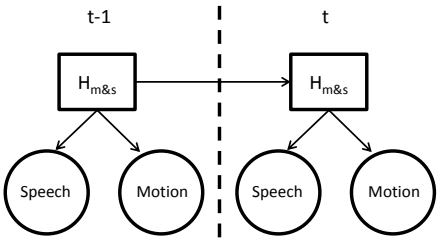


Figure 4. Baseline DBN [13], which jointly models speech features and body movements (head or hand gestures in this paper).

to 91.3s for hand gestures (i.e., 34.1s between *so-what* and *to-fro*, 65.0s between *so-what* and *regress*, 91.3s between *to-fro* and *regress*, and 30.2s between the three hand gestures). Similar to the annotation of discourse functions, we separately remove the overlap segments, resulting in mutually exclusive segments for hand and head gestures. For head gestures, we identify 1029.9s for *shake*, 2056.3s for *nod*. The remaining frames are labeled as *other* (7484.1s). For hand gestures, we identify 201.6s for *so-what*, 448.6s for *to-fro*, and 567.3s for *regress*. The remaining frames are labeled as *other* (9352.7s).

5. Baseline Model

We consider speech-driven methods built with *dynamic Bayesian network* (DBN). This section introduces the original DBN proposed by Marioory ad and Busso [13], which is the building block of the proposed models. This DBN framework also serves as a baseline for our models.

Figure 4 illustrates the baseline model, which was referred to as jDBN3 in Marioory ad and Busso [13]. This structure was the best model to jointly capture not only the relation between speech and facial features, but also the relation between facial features. In the diagram, the circle nodes represent the observation variables and the rectangle nodes represent the hidden variables. In our model, the node *Speech* represents the prosodic features and the node *Motion* represents either hand or head motions. The nodes *Speech* and *Motion* are continuous variables and are modeled with Gaussian distributions. The hidden discrete state variable $H_{m\&s}$ represents the state configuration between speech features and the gesture. It serves as a discrete codebook constraining the speech and gesture space. The transition matrix between the hidden variables is ergodic, where the transition probabilities follow the Markov property of order one. The time unit of the DBN is the time frame in the data (120fps).

In this model, the nodes *Speech* and *Motion* are conditionally independent given $H_{m\&s}$. When the speech features (f_{speech}) are entered in the models, Bayesian inference updates the marginal probabilities of the state configuration node $H_{m\&s}$, affecting the node *Motion*. The evidences entered in node *Speech* are used to predict the gestural feature, which is denoted by f_{Motion} This model preserves the full dependencies of the features within a modality by having full covariance matrices. This section describes the inference and synthesis method.

5.1. Inference

There are differences in the inference process for learning and synthesizing the gestures. During learning, we have access to the observations for the nodes *Motion* and *Speech*, so we use the full observation probability ($P_t^f(i)$) in Equation 1. During synthesis, we only have observations for the node *Speech*, and the task is to predict the variable *Motion*. Therefore, we use partial observation probability ($P_t^p(i)$) in Equation 2.

$$\begin{aligned} P_t^f(i) =& P(f_{Speech_t}|H_{m\&s_t} = i) \\ & \cdot P(f_{Motion_t}|H_{m\&s_t} = i) \end{aligned} \tag{1}$$

$$P_t^p(i) = P(f_{Speech_t}|H_{m\&s_t} = i) \tag{2}$$

5.2. Synthesis

For synthesis of the *Motion* variable, we use the Viterbi algorithm. We find the probability of the i^{th} state, called $\gamma(i)$ given the partial observation sequence and the model for each point in time (Eq. 3, where q_t is the state at time t, O is the partial observation and λ represents the parameters of the model). Equation 4 calculates the expected value for the variable *Motion* given speech features f_{Speech}, where μ_h^i is the mean of the variable *Motion* for the i^{th} state.

$$\gamma_t(i) = P(q_t = i|O, \lambda) \tag{3}$$

$$E\left[Motion_t|f_{Speech}\right] = \sum_{i=1}^{n} \mu_h^i \gamma_t(i) \tag{4}$$

The parameters of the models are learned with conventional *expectation maximization* (EM). Since EM finds local optimum, the initialization is very important. In Mariooryad and Busso [13], we randomly initialized the models. Since the generated behaviors correspond to the expected values given the speech features (Eq. 4), the states may converge to the average position of the behaviors, reducing the range of behaviors generated by the model. To address this problem, we increase the representation of the initial states by using the *Linde-Buzo-Gray vector quantization* (LBG-VQ) technique [39]. This approach is often used while training HMMs, and leads to sparser states which increase the range of behaviors generated by the models.

We smooth the trajectories generated by this network following the approach proposed by Busso et al. [11]. The method selects equidistant key-points. The value of the joint rotations in these key-points are transformed into their quaternion representation, where they are interpolated. The interpolation connects the key-points providing smooth transitions. We implement this method using 12 key-points per second for the hand motion, and 15 key-points per second for the head motion.

6. Proposed Constrained Models

This section describes the proposed model built upon the improved version of the DBN proposed by Mariooryad et al. [13] described in Section 5. The key goal is to introduce constraints to generate meaningful behaviors. The proposed models synthesize behaviors that are not only timely coupled with speech, but also meaningful. The constraints are either discourse functions or predefined prototypical gestures. The proposed framework takes speech and the desired constraint as inputs, producing the synthesized trajectories. The constraints are not hidden states; they are observational states, which are provided to the model as inputs. The constraint is modeled as an integer assigned to each frame, which indicates one of the possible constraints. We set the variable to zero when we do not want to specify any constraint. For example, for head motion using prototypical behaviors, the constraint variable is set to *zero* for no constraint, *one* for head shakes and *two* for head nods. If a sentence is tagged with one of the constraints, all the frames for that sentence are set to the corresponding value. The discourse function constraints bridge the gap between rule-based and data-driven system. The prototypical gesture constraints can serve as the behavior realizer in rule-based systems, capturing the intrinsic variability of each gesture, while preserving their temporal coupling with speech.

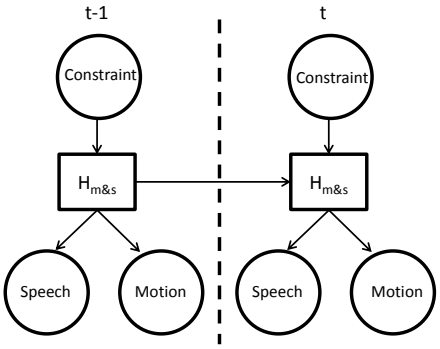


Figure 5. Proposed framework that adds a constraint to generate meaningful data-driven behaviors.

6.1. Adding Constraints to the DBN Model

Figure 5 illustrates the constrained model proposed in this study, which we refer to as *Constrained DBN* (CDBN). The key addition with respect to the baseline model is the node *Constraint* which is introduced as a parent of the hidden state variable $H_{m\&s}$. With this additional node, the state variable is directly conditioned on the given constraint, affecting the relationship between gesture and speech. Effectively, this model has transition matrices, prior probabilities, and state prior probabilities for each constraint, learning the intrinsic characteristics of the gestures conditioned on the given constraint.

This structure is different from the model proposed in our previous work, where the constraint was introduced as a child [29]. By adding the constraint node as a parent we obtain the following advantages: (1) we separately model the prior probabilities of the constraints and their affect on the hidden states, (2) we handle constraint categories with unbalanced training data, and (3) we model a more reasonable cause-effect relationship between the variables.

The constraint added to the baseline model is a discrete observation node, representing the presence of a given constraint for each frame. We add the label *other* when the constraint is not specified as an input. Equation 5 explicitly highlights that the transition probabilities a_{kij} from the previous state to the current state depend on both the previous state and the current constraint:

$$a_{kij} = P\left[q_t = j | q_{t-1} = i, Constraint_t = k\right], \tag{5}$$

where q_t is the state at time t, $Constraint_t$ is the constraint at time t, and a_{kij} is the transition probability between the i^{th} and j^{th} state when the constraint is k. During synthesis, partial observations for this model include speech features (f_{Speech}) and the *Constraint*. Equation 6 defines the expected value for the variable *Motion* using partial inference. In this equation, $c_{1:t}$ represents the constraint sequence for the whole turn, meaning that $\gamma_{c_{1:t}}(i)$ depends not only on $f_{Speech_{1:t}}$, but also on $Constraint_{1:t}$.

$$E\left[Motion | f_{Speech}, Constraint\right] = \sum_{i=1}^{n} \mu_h^i \gamma_{c_{1:t}}(i) \tag{6}$$

6.2. Training Sparse Transition Matrices

The characteristic patterns associated with each constraint are captured by the constraint-dependent transition matrices (a_{kij}). If these transition probabilities are similar, the behaviors generated after imposing the constraints will also be similar, and the model will fail to generate the characteristic patterns of each constraint. As a result, we want to increase the differences in the transition probability assigned to each constraint. For this purpose, we propose a state tying approach to make the conditional transition matrices sparse, which is customized to our problem. Tying the states is an approach used in training HMMs to learn the states topology within each domain [40], to share the states between different domains [41] or to reduce the number of states [42]. Our state tying approach aims to impose sparsity on the transition matrices so the models can better capture the differences across constraints (either discourse

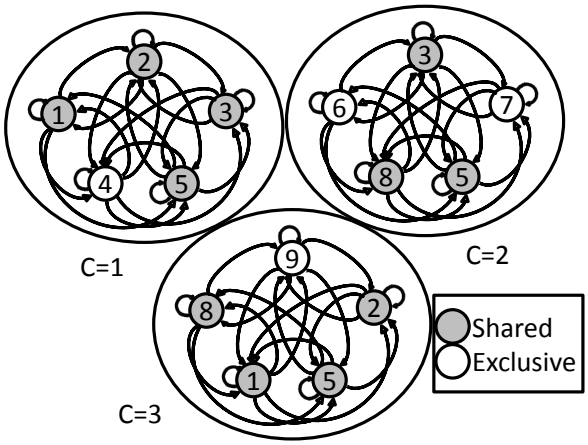


Figure 6. The state graph of the model with sparse transition matrix for a model with 3 constraints and 5 states per constraint (C: Constraint).

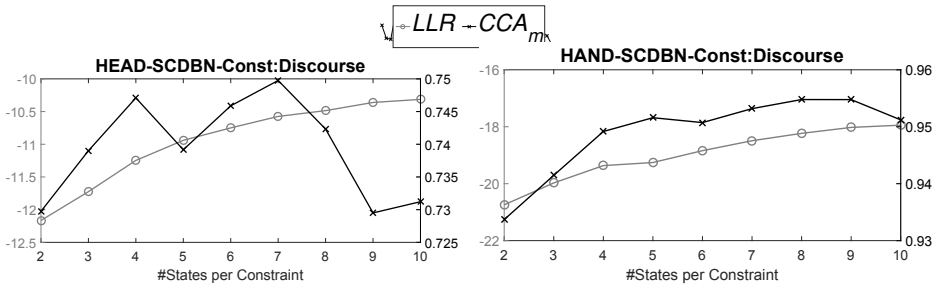
function or prototypical behaviors), while reducing the redundant states. The proposed approach will create states that are exclusive to each constraint, and states that are shared across constraints.

The proposed approach creates N states per constraint in node $H_{m\&s}$, which are separately trained using the data associated with the given constraint (e.g., data annotated with either discourse functions or prototypical gestures). These N states capture the characteristic patterns for each discourse function. If we have K constraints, this step will generate $N \times K$ states. Using all these states is not practical since it unnecessarily increases the number of states in $H_{m\&s}$, and, therefore, the number of parameters. Furthermore, many of these states are redundant. Instead, we merge similar states, creating shared states and constraint specific states. To merge states, previous studies have used *Kullback-Leibler* (KL) divergence [40, 41], *weighted KL* (wKL) divergence [43] or *delta-likelihood* (DL) [42]. In our implementation, we merge similar states using the KL divergence. Notice that wKL and KL are similar for our task since our models use one mixture per state. Since each state is a multivariate Gaussian distribution, we use Equation 7 to find similar states:

$$KL(P_j, Q_i) = \frac{1}{2}\left[\text{tr}(\Sigma_{q_i}^{-1}\Sigma_{p_j}) - \log(\Sigma_{q_i}^{-1}\Sigma_{p_j}) - d + (\mu_{q_i} - \mu_{p_j})^T\Sigma_{q_i}^{-1}(\mu_{q_i} - \mu_{p_j})\right] \tag{7}$$

where P_j and Q_j are the multivariate conditional Gaussian distribution for states p_i, and q_j, with covariance matrices Σ_{p_j} and Σ_{q_i}, and mean vectors μ_{p_j} and μ_{q_i}, and d is the dimension of the Gaussian. We merge states as follows. First, we select all the states associated with a constraint. For each of them, we find the closest state from the states associated with other constraints, as determined by the KL divergence metric. If the difference is less than a threshold (empirically set to 1), we merge the states, becoming a shared state across constraints. We sequentially repeat this process for the states of each of the constraints. Finally, we create a new state which is shared between all the constraints to allow transition between the constraints. The resulting conditional transition matrices for each constraint is sparse, allowing only transitions between the N states plus the aforementioned additional state shared across constraints ($(N + 1) \times (N + 1)$). This is the initialization phase for the model, and the parameters are refined afterward using EM. Figure 6 gives an illustration of the states for a model with 3 constraints and 5 states per constraint. States 1, 2, 3, 5 and 8 are shared across more than one constraint, and states 4, 6, 7, and 9 are exclusive.

To illustrate the importance of these sparse transition matrices, we compare the transition matrices when (1) the N states are shared across constraints, and (2) the N states per constraint are defined using the proposed approach. We estimate these transition matrices for *head shakes*, *head nods*, and *other*, using $N = 8$. The average of the L_∞ distances (Eq. 8) between the transition matrices conditioned on *Head Nods* (A), and *Head Shakes* (B) are 0.018 for option 1 (shared states) and 0.96 for option 2 (sparse matrices). This result shows that the training approach is more successful at capturing the differences between different constraints. Therefore, we rely on this approach for training the CDBN models.



(a) CDBN, constrained on discourse for head (b) CDBN, constrained on discourse for hand

Figure 7. Changes of LLR and CCA as we increase the number of states for the model constrained on discourse functions for hand and head movements.

$$d_\infty(A, B) = \max_{1 \le i \le N} \max_{1 \le j \le N} |a_{ij} - b_{ij}| \tag{8}$$

7. Experiments & Results

This section reports the experiments and the results of the evaluation of the proposed models. The baseline model is the improved framework proposed by Mariooryad and Busso [13] (Sec. 5). We separately train the models for head and hand, since the gesture constraints do not necessarily coincide. We train separate models for two sets of constraints: discourse functions, and prototypical behaviors. This approach results in four separate models. The number of constraints (K) depends on whether we use discourse functions or prototypical behaviors. When the constraints are discourse functions, we have K=5 (*affirmation*, *negation*, *question*, *suggestion*, *other*) for both the head and hand models. When the constraints are prototypical behaviors, we have K=3 for head (*nod*, *shake*, *other*) and K=4 for hand (*to-fro*, *so-what*, *regress*, *other*).

We compare the models using objective and subjective metrics. The subjective and objective evaluation relies on a modified version of a ten-fold cross-validation approach that maximizes the use of the corpus, and reduces the bias in the data partition assigned to the test set. We refer to the actual motion capture recordings of each subject as the *original* trajectories. We also refer to the speech prosodic features as the *speech* features.

7.1. Optimization of Number of States

Before training the models, we need to determine the number of states. Optimizing this parameter is computationally expensive since the CDBN models need to be trained multiple times as we vary N, repeating the approach for each fold in the cross-validation process. Therefore, we simplify the evaluation by setting one of the ten partitions as a validation set. We use the other nine partitions to train the models. This process is conducted once, using the optimal parameters for the rest of the evaluation.

We use two objective metrics. The first metric is the average *canonical correlation analysis* (CCA) between the original trajectory of the behaviors and the synthesized movements (CCA_m, where m is either hand or head motion). CCA projects two multidimensional data into a common space where their correlations are maximized. The value range between 0 and 1, where 1 implies perfect correlation and 0 no correlation between the variables. The estimation of CCA is invariant to the direction of the movement (e.g. head shake and head nod), so this metric is a valuable measurement to evaluate the couplings between the generated and true gestures (CCA is robust to affine transformations of the two variables, which suits our problem). We estimate CCA_m per turn, reporting the average results. The second metric is the average *log likelihood rate* (LLR) of the model ($\frac{\log P(O|\lambda)}{T}$), where T is the number of frames, and $P(O|\lambda)$ is the observation probability given the model parameter λ. We determine the number of state such that the CCA_m and the LLR of the model are both high.

We separately estimate the number of parameters for each model (baseline models for head and hand, CDBN models for head and hand with discourse and gesture constraints). Figure 7 shows an example of the changes observed

Table 3. Objective evaluations on the models trained for head movements. #*States* is the average number of states across different cross-validation iterations, #*States/Const* is the number of states per constraint (e.g., *N*), and #*Params* is the average number of learnable parameters for the model across the cross-validation iterations.

Region		Head Movements		
Model		Baseline	CDBN-Dis	CDBN-Ges
#States		7	28.1	22.2
#States/Const		7	7	8
#Params		300	1000.7	786.0
Global CCA_m	Pseudo-random	0.0611	0.0691	0.0736
	Synthesized	0.0880	0.1149	0.1003
Global CCA_{ms}	Pseudo-random	0.0642	0.0942	0.0641
	Synthesized	0.0833	0.1377	0.0804
KL divergence		7.593	2.718	2.957

for LLR and CCA_m in the validation set for the CDBN model for head and hand motions constrained on discourse functions. This figure shows that the CCA_m and LLR values start to saturate when the number of states is N=7 for head and N=8 for hand. We obtain similar figures for other cases, not reported in the paper, setting the optimal value for N. The row #States/Const in Tables 3 and 4 provides the number of states per constraint (e.g., N) for all the conditions. We use these parameters for the rest of objective and subjective evaluations.

7.2. Objective Evaluation

As we mentioned, we use ten partitions for the experimental evaluation. We avoid using the partition used for validation in the test set. Therefore, we only consider nine folds where the test set in each cross validation is one of the remaining nine partitions. After selecting the test set, we form the training set with the other nine partitions, adding the partition used for validation.

We separately evaluate the models constrained on discourse functions (CDBN-Dis) and prototypical gestures (CDBN-Ges) by comparing the generated trajectories with the baseline model (Sec. 5). We evaluate the generated movements in terms of the global CCA between the original and generated motion sequences (global CCA_m), and the global CCA between the generated motion sequences and the speech sequence (global CCA_{ms}). Global CCA between two signals is estimated after separately concatenating the synthesized sequences for the entire corpus, obtaining one CCA value between them. These results are referred to as *synthesized* in Tables 3 and 4. As a reference, we also estimate the global CCA for pseudo-random sequences, where the synthesized segments are concatenated in random order. Then, they are compared with either the original sequences (global CCA_m), or speech sequences (global CCA_{ms}). As a result, there is not temporal alignment between the corresponding synthesized sequences and the original data for pseudo-random sequences. These results are referred to as *pseudo-random* in Tables 3 and 4. We also use the KL divergence which is estimated over the entire data. The KL divergence between distributions p and q measures the amount of information lost when distribution q is used to represent distribution p. We evaluate the KL divergence between the synthesized movements (q) and the original movements (p). Ideally, this value should be as small as possible, indicating that the generated movement sequences have similar distributions as the original motion sequences.

As a reference, Global CCA_{ms} between the original head movements and speech is 0.1931, and between the original head movements and speech is 0.3275. These values can be considered as upper bounds for the global CCA_{ms} for our experiments. Tables 3 and 4 give the results for the synthesized hand and head movements, respectively. The first observation, as a sanity check, is that the CCA values for the models are higher than the CCA values for pseudo-random baselines. Since we use a single CCA transformation across the entire corpus, it is expected that the values for global CCA are between 0.05 and 0.40. The results show that the constrained model on discourse function, achieves higher global CCAs than the baseline model. As demonstrated by the subjective evaluation, the movements are more natural and appropriate when we constrain the models with discourse functions. The constrained models on prototypical gestures achieve higher CCA values compared to the baseline with the exception of CCA_{ms} for head movement. The results for the KL divergence show improvements on all the constrained models compared with the

Table 4. Objective evaluations on the models trained for hand movements. *#States* is the average number of states across different cross-validation iterations, *#States/Const* is the number of states per constraint (e.g., *N*), and *#Params* is the average number of learnable parameters for the model across the cross-validation iterations.

Region		Hand Movements		
Model		Baseline	CDBN-Dis	CDBN-Ges
#States		12	40	37
#States/Const		12	8	9
#Params		1247	3356.0	3130.0
Global CCA_m	Pseudo-random	0.1081	0.1493	0.1187
	Synthesized	0.2000	0.2474	0.2133
Global CCA_{ms}	Pseudo-random	0.1755	0.2518	0.1841
	Synthesized	0.2569	0.3206	0.2917
KL divergence		1.346	1.176	0.827

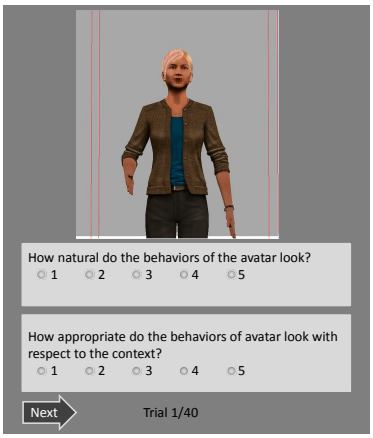


Figure 8. The interface for rating the animations based on appropriateness and naturalness using five-point Likert-like scales. The questions are displayed after the video is played.

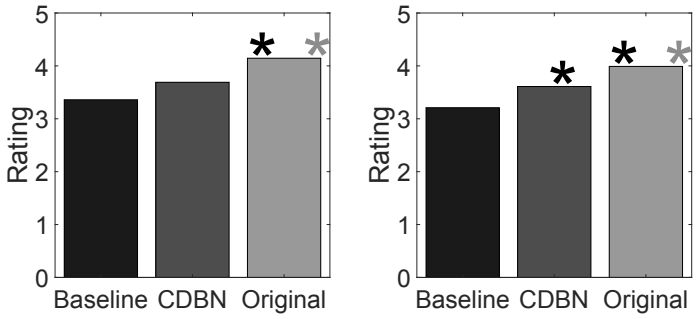
baseline models. The distributions of the generated behaviors are closer to the distributions of the original trajectories, compared to the baseline model.

7.3. Subjective Evaluations

This section reports the subjective evaluations of the behaviors generated with the proposed models. We use the Smartbody toolkit [44] for rendering the movements, where the only variable that we control is the hand gesture and head motion. Everything else is kept consistent across conditions (e.g., facial expressions). We render the animations by matching the selected characters with the gender of the speaker. We separately evaluate the models constrained on either discourse functions or prototypical gestures.

The first part of the subjective evaluation is when we constrain the models on the discourse functions (Sec. 4.3). We evaluate the perceived appropriateness and naturalness of the movements generated by the baseline model and CDBN model constrained on the discourse functions. For each constraint, we randomly selected 10 segments labeled with the corresponding discourse function. To provide enough context, we include the speaking turn preceding the selected turn. The animation is idle when the CA is listening to the other speaker (the MSP-AVATAR corpus consists of dyadic scenarios).

We use *Amazon mechanical turk* (AMT) for the perceptual evaluations, using the interface shown in Figure 8. We display the questions after the video is played to assure that the evaluators do not answer the questions before the video is played. We randomize the order of the videos for each evaluator. We only allow workers from the United States with overall acceptance rate of more than 80%.



(a) Appropriateness-All Constraint (b) Naturalness-All Constraint

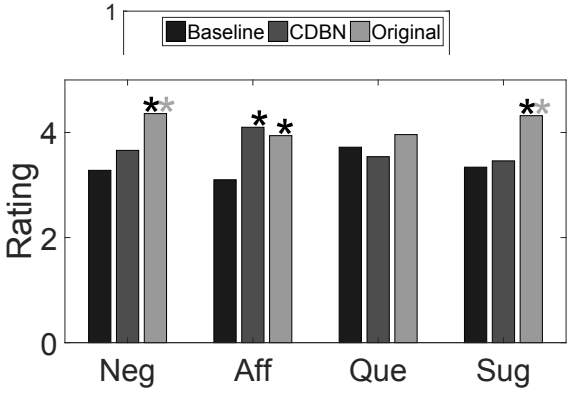
Figure 9. Results of the perceptual evaluations when the constraints are discourse functions. The bars represent the means per condition. Asterisks represent statistically higher values with respect to the bar indicated by the color of the asterisk (p-value<0.05).

We evaluate 120 segments (4 discourse functions ×3 conditions ×10 videos) animated using (1) the original motion capture data (*Original*), (2) the baseline model (*Baseline*), and (3) the CDBN models constrained on discourse functions (*CDBN*). Each video was annotated by five different evaluators. We asked the subjects to rate the animations in terms of naturalness and appropriateness of the movements using a five-point Likert-like scale (Fig. 8). In total, we have 15 evaluators, where nine are females and six are males (average age is 31.1).

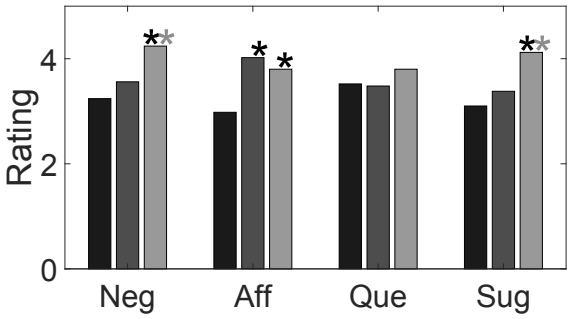
Figures 9(a) and 9(b) give the average of the ratings given by the evaluators for appropriateness and naturalness, respectively. The Cronbach's alpha between the annotators is α =0.48. The Kruskal-Wallis test shows that the videos synthesized by the three conditions are different ($p < 1e^{-10}$). The pairwise comparisons of the results are denoted in the figures with a color coded asterisks. The color of the asterisk indicates that the given condition is statistically higher than the condition associated with the bar with the given color (we assert performance at p-value<0.05). The pairwise comparison of the results shows that the original motion capture recordings are perceived as more natural and appropriate than the animations synthesized by both models ($p < 0.001$). However, the CDBN models are perceived with higher level of appropriateness and naturalness than the baseline models. The difference is statistically significant for naturalness ($p < 0.01$).

We also analyze the performance of the constrained model per discourse functions. Figure 10 gives the results. With the exception of *questions*, the CDBN model improves the perception of naturalness and appropriateness over the baseline models. The difference is statistically different for *affirmation* ($p < 0.001$), where the values for the constrained model are higher than the videos rendered with the unconstrained model. By observing the recordings, we note that subjects tend to use head nods during affirmation. The models can easily capture this gesture creating the intended perceptual results when we synthesize novel realization of this discourse functions. For other discourse functions, our subjects tend to use a broader number of gestures. For instance, for question the subjects tend to move their head up or to the side and open their hands; for suggestion they tend to move one of their hands and nod. Our models capture these behaviors, but obtaining the intended perceptual results during subjective evaluations is harder since these behaviors are subtle. Overall, the consistency in the results reveal that the proposed models and training strategy can effectively capture the range of behaviors characteristic of the given constraint.

The second part of the subjective evaluation is when we constrain the models on the prototypical behaviors (Sec. 4.4). We synthesize 60 segments per gesture (i.e., where the constraint is the target gesture). We randomly choose these segments from the fully annotated sessions. We find the accuracy per gesture by watching the animations generated for these segments, where a success is considered when the generated behavior matches the target gesture. Table 5 gives the accuracy for different head and hand gestures. For example, 85% for *so-what* means that 51 out of the 60 segments included the intended hand behavior. The generated head gestures for *nod* and *shake* match the target gesture more than 80% of times. This high accuracy demonstrates the benefits of the proposed constrained models. For hand movements, the gesture *so-what* has the highest accuracy (85%). The accuracy for *to-fro* and *regress* is not as high. We hypothesize that the accuracy for *regress* may be due to the high variability observed in the training samples. For *to-fro*, the result may be related to the lower precision of the samples retrieved for this



(a) Appropriateness-Per Constraint



(b) Naturalness-Per Constraint

Figure 10. Results of the perceptual evaluations per discourse function. The bars represent the means per condition. Asterisks represent statistically higher values with respect to the bar indicated by the color of the asterisk (p-value<0.05).

prototypical behavior (Table 2). The results from these evaluations are encouraging, suggesting that there is room for improvements. We have included as a supplemental material a video with examples of the animations created with the constrained models.

8. Conclusions

This paper explored the idea of introducing constraints in speech-driven models to generate behaviors with meaning that are timely coupled with speech. We evaluated a unified model with two types of constraints for hand and head movements: discourse functions and prototypical gestures. We incorporated discourse functions into the speech-driven framework to capture the characteristics behaviors associated with each of the four classes considered in the study (negations, affirmations, questions and suggestions). Our model constrained on discourse functions aimed to capture characteristic patterns for the discourse functions (e.g., generating meaningful gestures when people are asking questions). Likewise, we constrained the models with predefined prototypical gestures for head (shake, nod) and hand (so-what, to-fro, regress) gestures. This model can be used by a rule-based system as a behavior realizer. The proposed approach not only creates the appropriate behavior, but also captures the temporal coupling between speech and the synthesized movement, which is not easily achieved with only rule-based systems.

The proposed framework is built upon the DBN models proposed by Mariooryad and Busso [13], providing two important contributions to effectively constrain the models on the underlying discourse function or prototypical gesture. First, we introduce a variable that constrains the state configuration between speech and gestures, capturing the cause-effect relation of the gesture production. The training approach uses a better initialization of the states by using vector quantization, which effectively increases the range of the movements generated by the model. Second, we introduced shared and exclusive states for each of the constraints, creating sparse transition probability matrices.

Table 5. Accuracy of the synthesized gestures using the CDBN framework when we constrain the models on prototypical gestures.

Region	Behavior	Accuracy [%]
HEAD	Nod	81.7
	Shake	80.0
HAND	To-Fro	61.7
	So-What	85.0
	Regress	55.0

Some of the states are shared between constraints, while others are exclusively associated with a constraint. This approach effectively captures the differences in the behaviors across constraints.

The results from the objective and subjective evaluations demonstrated the benefits of the proposed approach. The results of the perceptual evaluation showed significant improvement for the constrained model over the unconstrained baseline model for *affirmation*. The results for prototypical gestures also revealed the potential of the proposed work. The head gestures synthesized by the constrained model generated the target gesture with 80% accuracy. The hand gestures generated by the constrained model showed 85% accuracy for *so-what*. For *to-fro*, and *regress* the accuracies are lower.

The study opens interesting opportunities to increase the role of data-driven models in CAs. For example, the proposed approach can be combined with rules driven from natural recordings [45] to create meaningful and naturalistic gestures. One limitation of the approach is that it requires speech. We are exploring training schemes to extend the models by driving the behaviors using synthetic speech [33, 46]. If we can solve the challenges in using synthetic speech instead of natural speech, we can increase the range of CA applications for data-driven models. With the transcription, we can also infer discourse functions using automatic algorithms. Dialog acts are semantic tags which can be retrieved from the text using supervised classifiers. These tags can then be translated into discourse functions, resulting in an autonomous meaningful behavior generator. Finally, we can address the lower performance for prototypical hand gestures by adding more data, capturing intrinsic variability between people, and by using more powerful frameworks. Advances in deep learning, in particular, offer appealing alternatives for this task [17, 18, 19, 47, 48, 49].

Acknowledgment

The authors would like to thank Sunghyun Park, Philippa Shoemark, and Louis-Philippe Morency for sharing the OCTAB interface. This work was funded by National Science Foundation grants IIS:1718944.

References

[1] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, M. Stone, Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents, in: Computer Graphics (Proc. of ACM SIGGRAPH'94), Orlando, FL,USA, 413–420, 1994.

[2] V. Chattaraman, W.-S. Kwon, J. E. Gilbert, Y. Li, Virtual shopping agents: Persona effects for older users, Journal of Research in Interactive Marketing 8 (2) (2014) 144–162, doi:\bibinfo{doi}{10.1108/JRIM-08-2013-0054}.

[3] N. Sadoughi, C. Busso, Head Motion Generation, in: B. Müller, S. Wolf, G.-P. Brueggemann, Z. Deng, A. McIntosh, F. Miller, W. Scott Selbie (Eds.), Handbook of Human Motion, Springer International Publishing, ISBN 978-3-319-30808-1, 1–25, doi:\bibinfo{doi}{10.1007/978-3-319-30808-1_4-1}, 2017.

[4] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson, Towards a common framework for multimodal generation: The behavior markup language, in: International Conference on Intelligent Virtual Agents (IVA 2006), Marina Del Rey, CA, USA, 205–217, doi:\bibinfo{doi}{10.1007/11821830_17}, 2006.

[5] E. Bevacqua, M. Mancini, R. Niewiadomski, C. Pelachaud, An expressive ECA showing complex emotions, in: Proceedings of the Artificial Intelligence and Simulation of Behaviour (AISB 2007) Annual Convention, Newcastle, UK, 208–216, 2007.

[6] M. E. Foster, Comparing Rule-based and Data-driven Selection of Facial Displays, in: Workshop on Embodied Language Processing, Association for Computational Linguistics, Prague, Czech Republic, 1–8, 2007.

[7] H. Welbergen, A. Nijholt, D. Reidsma, J. Zwiers, Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information, in: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment (INTETAIN 2005), vol. 3814 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Madonna di Campiglio, Italy, ISBN 978-3-540-30509-5, 203–212, doi:\bibinfo{doi}{10.1007/11590323_2}, 2005.

[8] J. Cassell, H. Vilhjálmsson, T. Bickmore, BEAT: the behavior expression animation toolkit, in: H. Prendinger, M. Ishizuka (Eds.), Life-Like Characters: Tools, Affective Functions, and Applications, Cognitive Technologies, Springer Berlin Heidelberg, New York, NY, USA, 163–185, doi:\bibinfo{doi}{10.1007/978-3-662-08373-4_8}, 2003.

[9] M. Brand, Voice puppetry, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999), New York, NY, USA, ISBN 0-201-48560-5, 21–28, doi:\bibinfo{doi}{http://doi.acm.org/10.1145/311535.311537}, 1999.

[10] Y. Cao, W. Tien, P. Faloutsos, F. Pighin, Expressive speech-driven facial animation, ACM Transactions on Graphics 24 (4) (2005) 1283–1302, doi:\bibinfo{doi}{10.1145/1095878.1095881}.

[11] C. Busso, Z. Deng, M. Grimm, U. Neumann, S. Narayanan, Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis, IEEE Transactions on Audio, Speech and Language Processing 15 (3) (2007) 1075–1086, doi:\bibinfo{doi}{10.1109/TASL.2006.885910}.

[12] C. Busso, Z. Deng, U. Neumann, S. Narayanan, Natural Head Motion Synthesis Driven by Acoustic Prosodic Features, Computer Animation and Virtual Worlds 16 (3-4) (2005) 283–290, doi:\bibinfo{doi}{10.1002/cav.80}.

[13] S. Mariooryad, C. Busso, Generating Human-like Behaviors using Joint, Speech-driven Models for Conversational Agents, IEEE Transactions on Audio, Speech and Language Processing 20 (8) (2012) 2329–2340, doi:\bibinfo{doi}{10.1109/TASL.2012.2201476}.

[14] S. Levine, P. Krähenbühl, S. Thrun, V. Koltun, Gesture controllers, ACM Transactions on Graphics 29 (4) (2010) 124:1–124:11, doi:\bibinfo{doi}{10.1145/1778765.1778861}.

[15] C.-C. Chiu, S. Marsella, How to train your avatar: A data driven approach to gesture generation, in: H. H. Vilhjálmsson, S. Kopp, S. Marsella, K. Thórisson (Eds.), Intelligent Virtual Agents, vol. 6895 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Reykjavik, Iceland, ISBN 978-3-642-23973-1, 127–140, doi:\bibinfo{doi}{10.1007/978-3-642-23974-8_14}, 2011.

[16] B. H. Le, X. Ma, Z. Deng, Live speech driven head-and-eye motion generators, IEEE Transactions on Visualization and Computer Graphics 18 (11) (2012) 1902–1914, doi:\bibinfo{doi}{10.1109/TVCG.2012.74}.

[17] N. Sadoughi, C. Busso, Novel Realizations of Speech-driven Head Movements with Generative Adversarial Networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, 6169–6173, 2018.

[18] N. Sadoughi, C. Busso, Expressive Speech-Driven Lip Movements with Multitask Learning, in: IEEE Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, China, 409–415, doi:\bibinfo{doi}{10.1109/FG.2018.00066}, 2018.

[19] N. Sadoughi, C. Busso, Speech-Driven Expressive Talking Lips with Conditional Sequential Generative Adversarial Networks, ArXiv e-prints (2018) 1–13.

[20] C. Busso, S. Narayanan, Interrelation between speech and facial gestures in emotional utterances: a single subject study, IEEE Transactions on Audio, Speech and Language Processing 15 (8) (2007) 2331–2347, doi:\bibinfo{doi}{10.1109/TASL.2007.905145}.

[21] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, B. de Carolis, Greta. A Believable Embodied Conversational Agent, in: O. Stock, M. Zancanaro (Eds.), Multimodal Intelligent Information Presentation, Text, Speech and Language Technology, Springer Netherlands, Dordrecht, The Netherlands, 3–25, doi:\bibinfo{doi}{10.1007/1-4020-3051-7_1}, 2005.

[22] S. Kopp, I. Wachsmuth, Synthesizing multimodal utterances for conversational agents, Computer animation & virtual worlds 15 (1) (2004) 39–52, doi:\bibinfo{doi}{10.1002/cav.6}.

[23] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, A. Shapiro, Virtual character performance from speech, in: ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013), Anaheim, CA, USA, 25–35, doi:\bibinfo{doi}{10.1145/2485895.2485900}, 2013.

[24] R. Voigt, R. J. Podesva, D. Jurafsky, Speaker Movement Correlates with Prosodic Indicators of Engagement, in: Speech Prosody (SP 2014), Dublin, Republic of Ireland, 70–74, 2014.

[25] H. P. Graf, E. Cosatto, V. Strom, F. J. Huang, Visual Prosody: Facial Movements Accompanying Speech, in: Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition, Washington, D.C., USA, 396–401, 2002.

[26] D. McNeill, Hand and Mind: What gestures reveal about thought, The University of Chicago Press, Chicago, IL, USA, ISBN 0-226-56132-1, 1992.

[27] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, E. Erzin, Multimodal analysis of speech prosody and upper body gestures using hidden semi-Markov models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Vancouver, BC, Canada, 3652–3656, doi:\bibinfo{doi}{10.1109/ICASSP.2013.6638339}, 2013.

[28] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, C. Bregler, Speaking with hands: Creating animated conversational characters from recordings of human performance, ACM Transactions on Graphics (TOG) 23 (3) (2004) 506–513.

[29] N. Sadoughi, Y. Liu, C. Busso, Speech-Driven Animation Constrained by Appropriate Discourse Functions, in: International conference on multimodal interaction (ICMI 2014), Istanbul, Turkey, 148–155, doi:\bibinfo{doi}{10.1145/2663204.2663252}, 2014.

[30] N. Sadoughi, C. Busso, Retrieving target gestures toward speech driven animation with meaningful behaviors, in: International conference on Multimodal interaction (ICMI 2015), Seattle, WA, USA, 115–122, doi:\bibinfo{doi}{10.1145/2818346.2820750}, 2015.

[31] N. Sadoughi, Y. Liu, C. Busso, MSP-AVATAR Corpus: Motion Capture Recordings to Study the Role of Discourse Functions in the Design of Intelligent Virtual Agents, in: 1st International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS 2015), Ljubljana, Slovenia, 1–6, doi:\bibinfo{doi}{10.1109/FG.2015.7284885}, 2015.

[32] I. Pandzic, R. Forchheimer, MPEG-4 Facial Animation - The standard, implementations and applications, John Wiley & Sons, ISBN 0-470-84465-5, 2002.

[33] N. Sadoughi, C. Busso, Head Motion Generation With Synthetic Speech: a Data Driven Approach, in: Interspeech 2016, San Francisco, CA, USA, 52–56, doi:\bibinfo{doi}{10.21437/Interspeech.2016-419}, 2016.

[34] P. Boersma, D. Weenink, Praat, a system for doing phonetics by computer, Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, http://www.praat.org, 1996.

[35] S. Park, G. Mohammadi, R. Artstein, L. P. Morency, Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface, in: ACM Multimedia 2012 workshop on Crowdsourcing for multimedia (CrowdMM), Nara, Japan, 29–34, doi:\bibinfo{doi}{10.1145/2390803.2390816}, 2012.

[36] D. Zhou, Q. Liu, J. Platt, C. Meek, Aggregating ordinal labels from crowds by minimax conditional entropy, in: International Conference on Machine Learning (ICML 2014), Beijing, China, 262–270, 2014.

[37] M. Kipp, Gesture Generation by Imitation: From Human Behavior to Computer Character Animation, Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Germany, 2003.

[38] F. Zhou, F. De la Torre, J. K. Hodgins, Aligned cluster analysis for temporal segmentation of human motion, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008), Amsterdam, The Netherlands, doi:\bibinfo{doi}{10.1109/AFGR.2008.4813468}, 2008.

[39] Y. Linde, A. Buzo, R. Gray, An Algorithm for Vector Quantizer Design, IEEE Transactions on Communications 28 (1) (1980) 84–95.

[40] K.-C. Au, K.-W. Cheung, Learning hidden Markov model topology based on KL divergence for information extraction, Lecture notes in computer science (2004) 590–594doi:\bibinfo{doi}{https://doi.org/10.1007/978-3-540-24775-3_70}.

[41] Y. Qian, H. Liang, F. K. Soong, A cross-language state sharing and mapping approach to bilingual (Mandarin–English) TTS, IEEE Transactions on Audio, Speech, and Language Processing 17 (6) (2009) 1231–1239, doi:\bibinfo{doi}{10.1109/TASL.2009.2015708}.

[42] H.-Y. Cho, S. Kim, A New Distance Measure for a Variable-Sized Acoustic Model Based on MDL Technique, ETRI journal 32 (5) (2010) 795–800, doi:\bibinfo{doi}{10.4218/etrij.10.1510.0062}.

[43] A. Ogawa, S. Takahashi, Weighted distance measures for efficient reduction of Gaussian mixture components in HMM-based acoustic model, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, Las Vegas, NV, USA, 4173–4176, doi:\bibinfo{doi}{10.1109/ICASSP.2008.4518574}, 2008.

[44] M. Thiebaux, S. Marsella, A. N. Marshall, M. Kallmann, Smartbody: Behavior realization for embodied conversational agents, in: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1, vol. 1, Estoril, Portugal, 151–158, 2008.

[45] C.-C. Chiu, L. Morency, S. Marsella, Predicting co-verbal gestures: a deep and temporal modeling approach, in: W. Brinkman, J. Broekens, D. Heylen (Eds.), International Conference on Intelligent Virtual Agents (IVA 2015), vol. 9238 of *Lecture Notes in Computer Science*, Springer, Cham, Delft, The Netherlands, ISBN 978-3-319-21995-0, 152–166, doi:\bibinfo{doi}{10.1007/978-3-319-21996-7_17}, 2015.

[46] N. Sadoughi, Y. Liu, C. Busso, Meaningful Head Movements Driven by Emotional Synthetic Speech, Speech Communication 95 (2017) 87–99, doi:\bibinfo{doi}{10.1016/j.specom.2017.07.004}.

[47] K. Haag, H. Shimodaira, Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis, in: D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, A. Leuski (Eds.), International Conference on Intelligent Virtual Agents (IVA 2016), vol. 10011 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Los Angeles, CA, USA, ISBN 978-3-319-47664-3, 198–207, doi:\bibinfo{doi}{10.1007/978-3-319-47665-0_18}, 2016.

[48] X. Lan, X. Li, Y. Ning, Z. Wu, H. Meng, J. Jia, L. Cai, Low level descriptors based DBLSTM bottleneck feature for speech driven talking avatar, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, 5550–5554, doi:\bibinfo{doi}{10.1109/ICASSP.2016.7472739}, 2016.

[49] N. Sadoughi, C. Busso, Joint Learning of Speech-Driven Facial Motion with Bidirectional Long-Short Term Memory, in: J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, S. Kopp (Eds.), International Conference on Intelligent Virtual Agents (IVA 2017), vol. 10498 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Stockholm, Sweden, ISBN 978-3-319-67400-1, 389–402, doi:\bibinfo{doi}{10.1007/978-3-319-67401-8_49}, 2017.