Speech-Driven Animation Constrained by Appropriate Discourse Functions

Najmeh Sadoughi[1], Yang Liu[2], Carlos Busso[1]
1. Multimodal Signal Processing (MSP) Laboratory
2. Human Language Technology Research Institute
The University of Texas at Dallas, Richardson TX 75080, USA
Emails: nxs137130@utdallas.edu, yangl@hlt.utdallas.edu, busso@utdallas.edu

ABSTRACT

Conversational agents provide powerful opportunities to interact and engage with the users. The challenge is how to create naturalistic behaviors that replicate the complex gestures observed during human interactions. Previous studies have used rule-based frameworks or data-driven models to generate appropriate gestures, which are properly synchronized with the underlying discourse functions. Among these methods, speech-driven approaches are especially appealing given the rich information conveyed on speech. It captures emotional cues and prosodic patterns that are important to synthesize behaviors (i.e., modeling the variability and complexity of the timings of the behaviors). The main limitation of these models is that they fail to capture the underlying semantic and discourse functions of the message (e.g., nodding). This study proposes a speech-driven framework that explicitly model discourse functions, bridging the gap between speech-driven and rule-based models. The approach is based on *dynamic Bayesian Network* (DBN), where an additional node is introduced to constrain the models by specific discourse functions. We implement the approach by synthesizing head and eyebrow motion. We conduct perceptual evaluations to compare the animations generated using the constrained and unconstrained models.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Conversational Agent*

Keywords

Conversational Agent; Head motion; Eyebrow movement; dynamic Bayesian network; Discourse function

1. INTRODUCTION

Conversational agents (CAs) provide powerful opportunities to better interact and engage with the users. They can communicate verbal and nonverbal information, which can be close to natural human interactions. CAs with naturalistic behaviors can be used in mobile interfaces, devices for the hearing impaired, animated movies, animations, intelligent tutoring systems, and entertainment industry. The challenge is to generate avatars with appropriate naturalistic nonverbal behaviors, responding to the underlying flow of the interaction.

Nonverbal behaviors are expressed through the body posture, hand gestures, head movements, eyebrow movements, facial expression and gaze [15]. People use these behaviors to highlight their emotional state, portray their thoughts, refer to concrete objects, illustrate abstract ideas, and clarify the intended semantic content. In order to create believable CAs, a precise and deep understanding of the relationship between speech and gestures is required. McNeill [20] mentioned that 90% of gestures occur when people are speaking. Moreover, speech and gestures are synchronized at different levels (i.e., phonemes, words, phrases, sentences) [10]. Gestures such as raising the eyebrows, nodding the head, shaking the head and frowning are timely aligned with their underlying verbal message. A CA should capture this complex coupling between gestures and speech.

Previous work on CAs are mainly divided into two categories: rule-based systems [2, 10, 15, 19], and data-driven systems [3, 5, 9, 16–18]. Rule-based systems usually define rules about the behaviors based on the contextual information during the dialog. Data-driven systems rely on recorded data to synthesize behaviors. An appealing approach to drive the animation is speech. Speech conveys rich information that a CA can leverage to synthesize naturalistic behaviors. Prosodic information conveyes the intonation and emphasis which is important to model the variability and complexity of the timings of the behaviors. However, when only speech is used to drive the animation, the behaviors may not capture the contextual information, ignoring the discourse functions of the gestures (nodding for affirmation, shaking head for negation, frowning for disapproval). This study addresses this limitation by proposing a framework for speech-driven animations that explicitly models the underlying discourse functions.

This paper presents our preliminary attempt to bridge the gap between speech-driven approaches and rule-based systems, exploiting the advantages of both methods. The paper proposes to constrain a speech-driven animation system by the underlying discourse function of the message, providing more appropriate behaviors. The models are built upon the *dynamic Bayesian network* (DBN) proposed in Mariooyiad and Busso [18], which captures the relationship between

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 ACM 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.
ICMI'14, November 12–16, 2014, Istanbul, Turkey.
Copyright 2014 ACM 978-1-4503-2885-2/14/11 ...$15.00.
Replace this line with the http://dx.doi.org/10.1145/2663204.2663252.

head movements, eyebrow movements and speech prosody features. We propose to add an extra variable reflecting the discourse function, which constrains the synthesized behaviors. We present perceptual evaluations to compare the constrained and unconstrained speech driven models.

2. RELATED WORK

Some of the frameworks for CA proposed in previous studies depend on carefully designed rules to synthesize appropriate gestures. Cassell et al. [10] designed a system with specific rules to generate facial expressions, head movements, and hand gestures, which were synchronized with speech [10]. In their follow up work, they designed the *behavior expression animation toolkit* (BEAT), which receives text as input to generate appropriate gestures and speech for the animation [11]. Kopp et al. designed a unified framework named *behavior markup language* (BML), for generating animations based on rules [15]. This effort allows other researches working on *embodied conversational agents* (ECAs) to use this language to combine various rules to generate behaviors (gaze, facial expression, head and hand movements). Marsella et al. [19] designed a system to synthesize animations that relies on speech and its transcription. They conducted acoustic and syntactic analysis on the inputs, generating animations based on some predefined behavior mapping. After synthesizing the gestures, they proposed heuristic rules to remove conflicting behaviors. They showed that animations generated using their approach are perceived more appropriate than prosody driven beats.

Studies have shown the relationship between head movements, eyebrow movements and speech [8,13]. Graf et al. [13] showed that head movements and facial expressions are correlated with speech prosody features. They showed that despite variations of amplitude and direction of head movements, the timing of these movements are correlated with prosodic features. They also identified correlation between the rise of eyebrows and prosodic features. These studies have motivated us to drive facial animations using speech as an alternative approach to rule-based methods [5–7, 12].

Brand [3] proposed a mapping approach from HMMs to generate animations from speech. Cao et al. [9] designed a framework based on a generative model to synthesize facial expression with appropriate lip synchronization using speech prosody features. Busso et al. [5,6] synthesized head movements, using HMMs and vector quantization to generate head movements based on speech prosody features. Mariooryad and Busso [18], proposed a framework based on DBNs to model the relationship between head movement, eyebrow movement, and speech. This generative model synthesized head and eyebrow movements using prosody features. Le et al. [16] proposed a live speech-driven head, gaze, and eyelid motion synthesis framework. They designed a *Gaussian mixture model* (GMM) to capture the static kinematic properties of head movement and speech prosody features. Then, they used a gradient descent scheme to find the optimum mapping for head motion synthesis. Levine et al. [17] designed a framework based on *conditional random fields* (CRFs) to use speech prosody features to find the kinematic properties of joint movements.

Rule-based systems define certain predefined behaviors which are not likely to capture the rich and complex relationship and variability between modalities. If we rely solely on rule-based systems, the animation will tend to display repetitive behaviors, with desynchronization between gestures and speech. Data-driven systems have the potential to capture variations of behaviors observed in the corpus, and to produce gestures that are timely synchronized with speech. Several studies have used speech prosodic features to generate facial behaviors. Although prosody is directly affected by emphasis, and emotional state of the speaker, the generated behaviors may not properly respond to the underlying discourse function in the dialog (e.g., nodding for affirmation, shaking head for negation, frowning for disapproval). This study proposes a systematic framework to incorporate the benefits of rule-based systems and speech-driven animation by constraining the models with appropriate discourse functional classes.

3. DATABASE AND ANNOTATION

The study relies on the *interactive emotional dyadic motion capture* (IEMOCAP) corpus [4]. This database comprises dyadic interactions between two actors (a female and a male), during improvisation and scripted scenarios. The scenarios are designed to evoke different emotional reactions from the actors. The corpus consists of five sessions (10 actors), and includes motion capture data, videos, and audio recordings of the interactions. This study only uses the improvisation scenarios for the first session.

The study aims to constrain speech-driven models by the underlying discourse functions. For this purpose, it is important to annotate the data with discourse function classes. The annotation was conducted with ANVIL [14], which is a flexible tool to annotate multiple events in the video. We started with the segmentation boundaries of the speaking turns available in the IEMOCAP corpus. Some segments were split to better localize the behaviors. The discourse function classes selected for the study were inspired by previous studies [19, 22]. Instead of a full semantic analysis of the input transcription, Marsella et al. [19] considered only semantic classes that tend to generate gestures. In our study, we also restrict the set to discourse function classes which we expect to produce nonverbal behaviors. The toolkit GRETA lists an extensive set of mappings between discourse function rules and distinct configurations for facial expressions and head movements [22]. From the list, we select the following discourse function classes for annotation: *affirmation*, *negation*, *question*, *statement* (opinion-statement), *contrast*, *uncertain*, *appreciation*, *request*, *command*, *suggest*, *warn*, and *inform*. However, we did not have enough samples for all of these classes. Therefore, this study only considers four dialog act classes: *affirmation* (90 turns), *negation* (53 turns), *question* (112 turns), and *statement* (158 turns). These classes are prone to generate gestures, and therefore provide the perfect starting point to study the role of discourse functions in the synthesis of behaviors for CA.

4. STATISTICAL ANALYSIS

We conduct a statistical analysis on the data, before constraining the speech-driven models by the four discourse function classes selected in Section 3. The purpose of the analysis is to identify whether the behaviors observed for each discourse function class present characteristic patterns. The study focuses on head and eyebrow motion, so this analysis considers only these gestures. For head motion, we study pitch, yaw, and roll movements and their circular

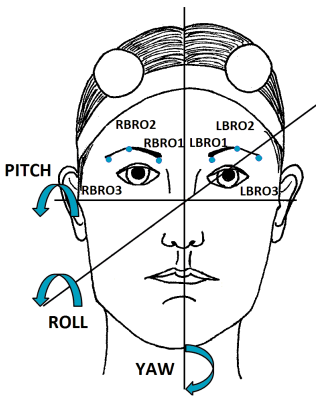


Figure 1: The movements analyzed in this paper.

Table 1: Features that are significantly different for each discourse function class (ANOVA). Figure 1 describes the features for head and eyebrow motion.

Question vs. Non-Question	
Pitch	$F(1,452) = 8.58$, $p = 0.004$
Roll	$F(1,452) = 7.05$, $p = 0.008$
Pitch Velocity	$F(1,452) = 7.05$, $p = 0.008$
Affirmation vs. Non-Affirmation	
LBRO3	$F(1,464) = 7.87$, $p = 0.005$
RBRO3	$F(1,464) = 10.42$, $p = 0.001$
Pitch Velocity	$F(1,464) = 6.74$, $p = 0.0097$
Negation vs. Non-Negation	
Yaw	$F(1,419) = 5.17$, $p = 0.023$
Pitch Velocity	$F(1,419) = 4.99$, $p = 0.026$
Statement vs. Non-Statement	
Pitch Velocity	$F(1,470) = 4.30$, $p = 0.038$

velocity. For eyebrow motion, we study the position and velocity of the markers placed on the eyebrows: LBRO1, LBRO2, LBRO3, RBRO1, RBRO2, and RBRO3 (see Fig. 1).

The analysis separately considers the four discourse function classes. For a given class (e.g., *question*), we group all the turns that do not belong to that class (e.g., *non-questions*). The patterns on the features for these two groups are then compared (*question* versus *non-question*). Table 1 reports the one way *analysis of variance* (ANOVA) results for the features that are found significantly different for each discourse function class. We assert significance when $p < 0.05$. Since the features have different dynamic ranges, we also estimate the standard deviation of the features. This measure characterizes changes in the dynamic range of the features induced by a discourse function class. Then, we estimate a one-way ANOVA over these values and we report the results in Table 2.

The results show that there are significant differences between the mean and standard deviation of *Yaw* for *negation* versus *non-negation*. This result is expected since people tend to shake their head for disagreement. Likewise,

Table 2: Features in which their standard deviations are significantly different for each discourse function class (ANOVA). Figure 1 describes the features for head and eyebrow motion.

Question vs. Non-Question	
Pitch	$F(1,452) = 7.33$, $p = 0.0071$
Roll	$F(1,452) = 5.37$, $p = 0.0210$
Affirmation vs. Non-Affirmation	
Yaw	$F(1,464) = 9.45$, $p = 0.0022$
Roll	$F(1,464) = 19.94$, $p = 10^{-5}$
LBRO1	$F(1,464) = 6.47$, $p = 0.0113$
LBRO3	$F(1,464) = 5.69$, $p = 0.0174$
RBRO1	$F(1,464) = 6.95$, $p = 0.009$
RBRO3	$F(1,464) = 7.84$, $p = 0.005$
Pitch Velocity	$F(1,464) = 21.36$, $p = 4.93 \times 10^{-6}$
Yaw Velocity	$F(1,464) = 12.74$, $p = 3.94 \times 10^{-4}$
Negation vs. Non-Negation	
Yaw	$F(1,419) = 7.42$, $p = 0.0067$
Roll Velocity	$F(1,419) = 8.75$, $p = 0.0033$
Yaw Velocity	$F(1,419) = 53.21$, $p = 1.51 \times 10^{-12}$
RBRO3	$F(1,419) = 6.19$, $p = 0.0132$
Statement vs. Non-Statement	
Roll Velocity	$F(1,470) = 4.46$, $p = 0.0351$

there are significant differences in *Pitch Velocity* for *affirmation* vs. *non-affirmation*. People tend to nod their head for agreement. Also, features describing head movements present statistical differences when people ask questions.

There are few differences in the features for *statement*. Therefore, we expect that models constrained by this discourse function class may not produce perceptive changes on the behaviors. Therefore, we exclude this class for the analysis. Likewise, we exclude the class *negation* since we only have 53 turns which are not enough to train the DBN models (Sec. 5).

5. SPEECH-DRIVEN ANIMATION

This section introduces the acoustic and visual features used to synthesize the behaviors from speech. It also describes the framework to synthesize the CA.

5.1 Mapping between Facial Markers & FAPs

We use the toolkit Xface [1] to generate the animations using the facial markers. Xface is an open source software complying with the MPEG-4 standard. Xface accepts *facial animation parameters* (FAPs) as input. There are 68 FAPs defined in the MPEG-4 standard, where two of them correspond to high-level parameters (visemes and emotions), and the others are low-level parameters. The FAPs are defined in terms of displacement of *facial feature points* (FPs) with respect to their reference position [23]. Most of the facial markers in the IEMOCAP database were selected following the locations of the FPs. Therefore, it is possible to establish a linear transformation from facial markers to FAPs. In particular, we use the approach described in Mariooryad and Busso [18].

5.2 Acoustic Features

The acoustic features to drive the facial animation correspond to prosodic features: fundamental frequency (F0) and the *RMS* energy. We also include their first and second derivatives forming a 6D feature vector. F0 contours have discontinuities during unvoiced regions, which introduce undesired breaks in the input signal. We address this problem by interpolating the F0 contour. Another challenge in modeling acoustic and facial features is the differences in the time resolution. The sampling frequency for the speech signal is 16 KHz, where the acoustic features are extracted from short-time windows every 16.67 ms (60 frames per second). The sampling rate for the motion capture data is 120 Hz. We address this problem by upsampling the acoustic features to 120 Hz. Therefore, the acoustic and facial features are aligned.

5.3 Speech-Driven Models using DBN

The baseline system to model the relationship between facial behaviors and speech correspond to one of the approaches proposed by Mariooryad and Busso [18]. They designed three DBNs, which make different assumptions on the relationship between speech and facial features (eyebrow and head motions). The study demonstrated that the *jDBN3* model, described in Figure 2(a), achieved the best performance using both objective and subjective evaluations. Therefore, we selected this model as our baseline system (i.e., unconstrained models). The node *Speech* represents the acoustic features. The node *Head&Eyebrow* corresponds to a joint variable describing the configuration of head and eyebrow positions. These two nodes are connected to the parent node $H_{h\&e}$, which is a discrete state space variable defining possible speech-gesture configurations (i.e., codebook). During training, all the variables are provided to the system. During synthesis, only the *Speech* variable is available, and the $H_{h\&e}$ node is estimated from the model.

We build upon the *jDBN3* model to incorporate constraints on discourse function classes. We propose to add an extra observation node, *Discourse_ Function*, which is a binary variable indicating the presence or absence of a given discourse function class. Figure 2(b) describes this model, which is referred to as *C-jDBN3*. The addition of this variable provides a novel framework to effectively constrain the model by the underlying discourse function. The *Discourse_Function* node affects the distribution of the hidden states. During inference, introducing evidences about the presence of the discourse function will constrain the model by increasing the posterior probability of appropriate hidden states associated with that discourse class.

Notice that separate models are used to model each of the discourse function classes. The reason for not modeling all of the discourse classes together is the limited data available for the evaluation. The framework is implemented with the Bayes Net Toolbox for MATLAB [21].

5.3.1 Inference

Inference in the *C-jDBN3* is based on Forward-Backward algorithm. The observation variables during learning and inference are different. During learning, we have full observation which includes *Speech*, *Head&Eyebrow*, and *Discourse_Function*. During synthesis, the observation is limited to *Speech* and *Discourse_Function*. The learning inference uses full observation probability (Equation 1), and

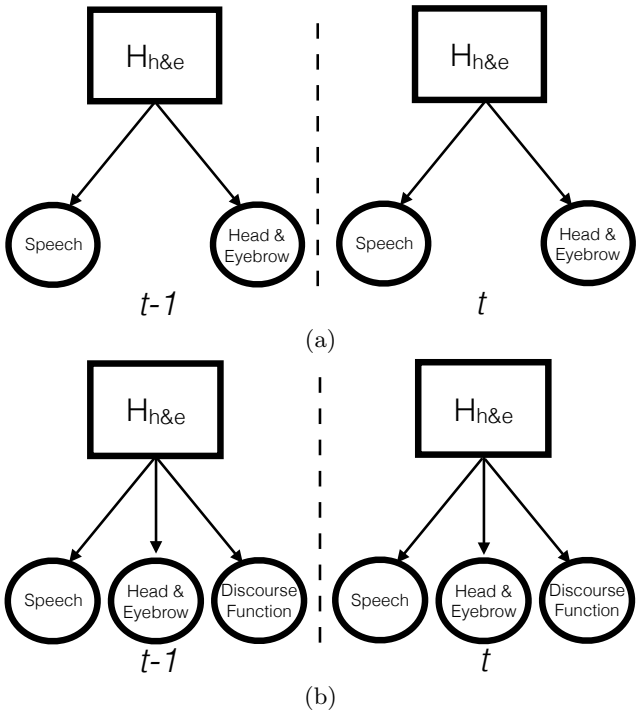


Figure 2: Illustration of (a): *jDBN3*, and (b): *C-jDBN3*. $H_{h\&e}$ represents a single, hidden state space variable, $Speech$ represents the prosodic features extracted from speech, $Head\&Eyebrow$ represents the head and eyebrow features, and $DiscourseFunction$ represents the constraint applied to the model.

the synthesis inference uses partial observation probability (Equation 2).

$$\begin{aligned} O_t^F(i) =& P(Speech_t|H_{h\&e_t}=i) \\ & \cdot P(Head\&Eyebrow_t|H_{h\&e_t}=i) \\ & \cdot P(Discourse\_Function_t|H_{h\&e}=i) \end{aligned} \quad (1)$$

$$\begin{aligned} O_t^P(i) =& P(Speech_t|H_{h\&e_t}=i) \\ & \cdot P(Discourse\_Function_t|H_{h\&e}=i) \end{aligned} \quad (2)$$

5.3.2 Synthesis

After inference with partial observation, the expected values of head and eyebrow features are derived according to Equation 3. These trajectories are used to synthesize the movements using either the unconstrained models (*jDBN3*) and the constrained models (*C-jDBN3*). $\gamma(i)$ is the posterior probability of the i^{th} state generating the observation, and $\mu_{h\&e}(i)$ is the mean of head and eyebrow features for the i^{th} state.

$$E\left[Head\&Eyebrow|Speech, Discourse\_Function\right] = \sum_{i=1}^{n} \mu_{h\&e}(i)\gamma(i) \quad (3)$$

With the *C-jDBN3* model, the *Discourse_Function* node changes the posterior probability of the hidden states given the observation, affecting in a principled manner the ex-

Model	Constraint	States#	Params#	Train Turns#
C-jDBN3	Question	6	527	95
jDBN3	Question	6	521	95
C-jDBN3	Affirmation	6	527	74
jDBN3	Afirmation	6	521	74

Table 3: This table shows the specifications for training the two models, constrained and unconstrained one, for (*affirmation* and *question*).

pected value of head and eyebrow movements. The model generates appropriate movements according to the specific discourse function class.

After finding the expected value for head and eyebrow motion, we implement an interpolation scheme to remove abrupt transitions or jerky movements. The approach selects key points, which are used to interpolate the trajectory. More details about this smoothing approach are given in Mariooryad and Busso [18].

Notice that we synthesize the expected eyebrow and head movements given the observation. Therefore, the behaviors synthesized using these models do not produce movements spanning the same dynamic range as the original behaviors in the database. As a result, some of the movements are not visually perceived in Xface. We address this problem by scaling the behaviors to recover the original standard deviation observed over the training data. The calculated scaling factor is then downscaled by 0.4, since a test data might not include the whole range of movements appeared in the training set. This constant factor is selected empirically. Note that this process only scales the trajectories. It does not interfere with the correlation between the synthesized and original behaviors.

6. EXPERIMENTS AND RESULTS

To assess the performance of the constrained model *C-jDBN3*, we demonstrate that the model can create naturalistic behaviors. We compare the results of the *C-jDBN3* models with the videos synthesized with the *jDBN3* model. For comparison, we also create animations with the actual sequences directly extracted from the motion capture data. We refer to these videos as *original*. The videos for *C-jDBN3* and *jDBN3* are created as follows. First, we implement a five-fold cross-validation scheme, where in each fold four partitions are used to train the models, and one partition is used to create the animations. This scheme maximizes the usage of the data, preserving separate partitions for training and testing the models. Second, we balance the number of samples used for training the *C-jDBN3* models (e.g., same number of samples for *question* and *non-question* classes). Notice that there are more samples not belonging to the discourse function class. Therefore, we randomly select samples to balance the number of samples for both conditions. Third, we use the same set of sentences to train *C-jDBN3* and *jDBN3* models. Table 3 reports the configuration settings for the models.

6.1 Subjective Evaluation of the results

Our aim is to analyze how the synthesized animation using the *C-jDBN3* models are perceived in comparison with the ones generated with the unconstrained *jDBN3* models. We

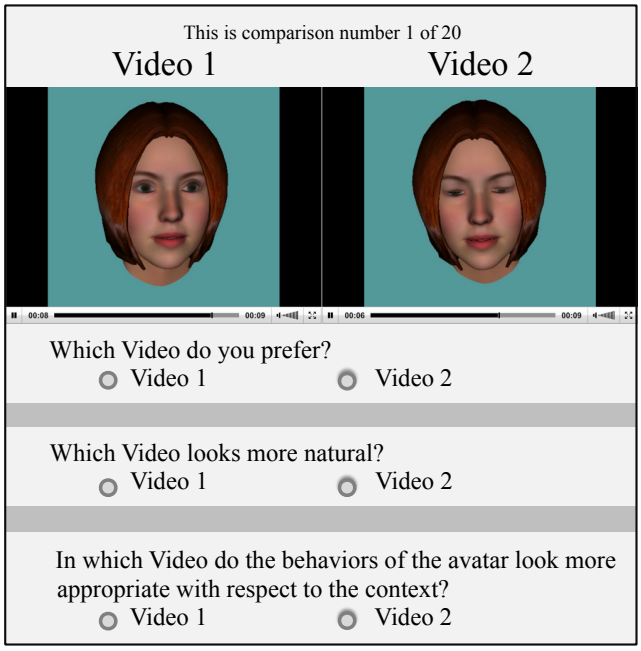
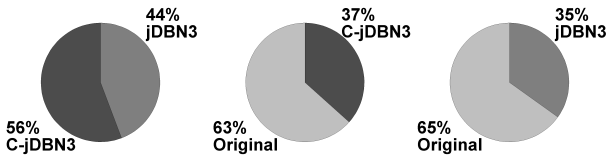


Figure 3: The interface used for the perceptual evaluation.

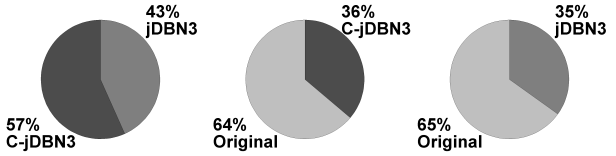
define a perceptual evaluation task in which each evaluator watches two videos with the same audio file, but with different animations. We consider three conditions: *C-jDBN3*, *jDBN3* and *original*. Therefore, the task requires 3 evaluations per sentence (*C-jDBN3* versus *jDBN3*, *C-jDBN3* versus *original*, and *jDBN3* versus *original*). After watching a pair of videos, the evaluator is asked to answer three preference questions: (1) which video do you prefer? (2) which video looks more natural? (3) in which video do the behaviors of the avatar look more appropriate with respect to the context?

For each discourse function class (*question*, *affirmation*), we selected 20 different sentences for the evaluations. Therefore, the evaluations consider 60 comparisons per discourse function class. To reduce fatigue, we split these 60 comparison into three sets with 20 comparisons. Each evaluator completes one of these sets. We rely on crowdsourcing for the perceptual evaluation – *Mechanical Turk* (MT). Figure 3 shows the *graphical user interface* (GUI) to evaluate the people's preference in pairwise comparisons. We recruit 60 different evaluators. Each comparison is evaluated by 10 evaluators. The position of the videos (left or right) are randomized for each evaluator. The agreement between the evaluators is low given the subtle differences between the generated videos (it is a difficult task).

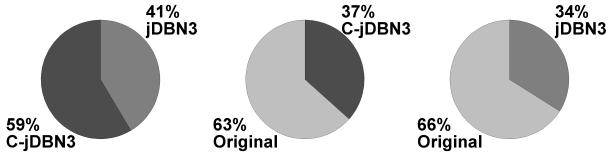
A common problem observed when using crowdsourcing is noisy data provided by cheaters (evaluators interested in the payments who do not care about the task). We address this problem by setting a threshold on the duration of the evaluation of each pair of videos (see Equation 4). If the duration of the task is less than T, we discard the evaluations, since it is unlikely that the evaluators watched both videos before providing their preferences (the evaluation is not reliable). This threshold discarded 6% of the comparison pairs for *Question*, and 10.1% of the comparisons for



(a) Which video do you prefer?

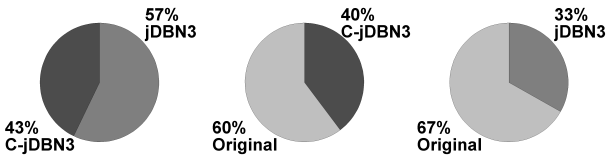


(b)Which video looks more natural?

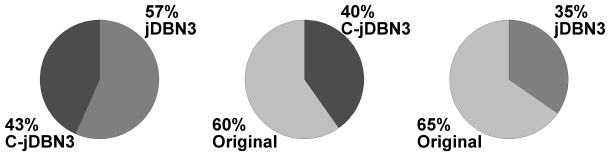


(c) Which video is more appropriate given the context?

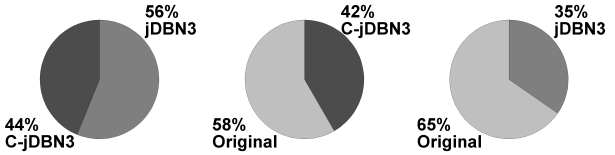
Figure 4: The result of perceptual evaluation for the pairwise comparison tasks where the constraint is *Question*.



(a)Which video do you prefer?



(b)Which video looks more natural?



(c) Which video is more appropriate given the context?

Figure 5: The result of perceptual evaluation for the pairwise comparison tasks where the constraint is *Affirmation*.

Affirmation. For consistency, we recollected the perceptual evaluations for these pairs.

$$T = 2 \times duration_{video} + 3(seconds) \quad (4)$$

To evaluate the results from MT, we conducted an informal evaluation with two evaluators from our laboratory following the same approach. The inter-evaluator agreement in controlled conditions was similar to the ones achieved with MT. Therefore, we conclude that the data from MT is reliable.

Figure 4 and 5 shows the results of the perceptual evaluation. These figures gives the preference for each pairwise comparison. Figure 4 gives the results when *C-jDBN3* is constrained by *question*, and Figure 5 gives the results when *C-jDBN3* is constrained by *Affirmation*. In these two figures, the first, second and third rows correspond to the results of the first, second and third questions, respectively.

For the discourse function class *question*, Figure 4(a) shows that in 56% of the times the evaluators preferred the animations generated with *C-jDBN3* over the ones generated with *jDBN3*. There is a 95.5% probability that this proportion is greater than chances (50%). From the second question (Fig. 4(b)), the evaluators considered 57% of the times the animations generated with the *C-jDBN3* model more natural than the ones generated with the unconstrained *jDBN3* model. There is a 97.6% probability that the proportion for this comparison is greater than chances. Finally, the evaluators considered in 59% of the time that the *C-jDBN3* model was more appropriated than the ones generated with *jDBN3*. There is a 99.45% probability that the proportion for this comparison is greater than chances. Figure 4 also shows that the *C-jDBN3* model is more selected than the unconstrained models when they are compared with the original sequences (indirect comparison).

For the discourse function class *affirmation*, the results are not conclusive. Figure 5(a) shows that the evaluators preferred *C-jDBN3* 43% of the times over the animations generated with *jDBN3*. Similar results are observed for the question about naturalness (Fig. 5(b)) and appropriateness (Fig. 5(c)). These results are not consistent with the indirect evaluations, where we compare the synthesized models with the animations generated with the original data. This indirect comparisons indicate that the constrained models are selected more frequently than the unconstrained models. Our future work will continue to explore the benefits of using constrained models for *affirmation*.

7. CONCLUSIONS AND DISCUSSION

This paper proposed to constrain speech-driven animation models by the underlying discourse functions. We focus on head and eyebrow motion. The statistical analysis demonstrated significant changes in these behaviors across discourse function classes (*affirmation*, *negation*, *question*, and *statement*). This analysis implies that, in addition to speech, other important factors should be included to synthesize natural behaviors. Then, we proposed a DBN, where a variable was added to constrain the behaviors. Perceptual evaluations for the discourse function class *question* showed that the animations with constrained models are perceived more preferable, natural and appropriate than the animations with the unconstrained models. For the class *affirmation*, the results are not conclusive. Although in the direct comparison the evaluators preferred the unconstrained models over the constrained models, indirect comparisons with the original videos reveal a different trend. When evalua-

tors compared the synthesized animations with the original animations, the constrained models were selected more often than the unconstrained models. This result may be due to lack of samples to train the DBN models. The statistical analysis revealed that patterns for eyebrow movements change for *affirmation*. However, the perceptual contribution of eyebrow movements in the synthesized animations is less dominant than the perceptual contribution of head motion [18]. Furthermore, eyebrows have subtle movements which can be captured by increasing the number of parameters. We expect that a better parametrization of eyebrows may help in producing distinctive behaviors that evaluators can perceive.

This study validates the idea of constraining speech-driven models by the discourse function. We need more data to further explore this research direction. We are currently collecting motion capture recordings where the scenarios are carefully designed to elicit specific discourse functions. The corpus will play a key role in extending the proposed constrained speech-driven models. It will offer the opportunity to consider other relevant discourse function classes such as *contrast*, *negation*, *warn*, and *uncertain*. The perceptual evaluation reveals that the raters preferred the animation with the *original* sequences over the ones synthesized with the DBN models (see Fig. 4 and 5). We expect that future improvements on the proposed models can help in reducing this gap.

8. ACKNOWLEDGMENTS

This work was funded by NSF (IIS 1352950).

9. REFERENCES

[1] K. Balci. Xface: MPEG-4 based open source toolkit for 3D facial animation. In *Conference on Advanced Visual Interfaces (AVI 2004)*, pages 399–402, Gallipoli, Italy, May 2004.

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

[3] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, pages 21–28, New York, NY, USA, 1999.

[4] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.

[5] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.

[6] C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, July 2005.

[7] C. Busso, Z. Deng, U. Neumann, and S. Narayanan. Learning expressive human-like head motion sequences from speech. In Z. Deng and U. Neumann, editors, *Data-Driven 3D Facial Animations*, pages 113–131. Springer-Verlag London Ltd, Surrey,United Kingdom, 2007.

[8] C. Busso and 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):2331–2347, November 2007.

[9] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24:1283–1302, October 2005.

[10] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and 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)*, pages 413–420, Orlando, FL,USA, 1994.

[11] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer, 2004.

[12] Z. Deng, C. Busso, S. Narayanan, and U. Neumann. Audio-based head motion synthesis for avatar-based telepresence systems. In *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP 2004)*, pages 24–30, New York, NY, October 2004. ACM Press.

[13] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. Visual prosody: Facial movements accompanying speech. In *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, pages 396–401, Washington, D.C., USA, May 2002.

[14] M. Kipp. ANVIL - a generic annotation tool for multimodal dialogue. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, Aalborg, Denmark, September 2001.

[15] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents*, pages 205–217. Springer, 2006.

[16] B. H. Le, X. Ma, and Z. Deng. Live speech driven head-and-eye motion generators. *IEEE transactions on visualization and computer graphics*, 18(11):1902–1914, 2012.

[17] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics (TOG)*, 29(4):124, 2010.

[18] S. Mariooryad and C. Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2329–2340, October 2012.

[19] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35. ACM, 2013.

[20] D. McNeill. *Hand and Mind: What gestures reveal about thought*. The University of Chicago Press, Chicago, IL, USA, 1992.

[21] K. Murphy. Bayes net toolbox for matlab. https://code.google.com/p/bnt/, October 2007.

[22] I. Poggi, C. Pelachaud, F. Rosis, V. C. B., and Carolis. Greta. a believable embodied conversational agent. *Multimodal Intelligent Information Presentation*, pages 3–25, 2005.

[23] A. Tekalp and J. Ostermann. Face and 2-d mesh animation in MPEG-4. *Signal Processing: Image Communication*, 15(4):387–421, January 2000.