Speech Communication 95 (2017) 87–99

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Meaningful head movements driven by emotional synthetic speech

Najmeh Sadoughi*, Yang Liu, Carlos Busso

The University of Texas at Dallas, Richardson, TX 75080, United States

ARTICLE INFO

Article history:
Received 13 January 2017
Revised 1 June 2017
Accepted 28 July 2017
Available online 29 July 2017

Keywords:
Conversational agents
Head motion with meaning
Speech-driven animation

ABSTRACT

Speech-driven head movement methods are motivated by the strong coupling that exists between head movements and speech, providing an appealing solution to create behaviors that are timely synchronized with speech. This paper offers solutions for two of the problems associated with these methods. First, speech-driven methods require all the potential utterances of the *conversational agent* (CA) to be recorded, which limits their applications. Using existing *text to speech* (TTS) systems scales the applications of these methods by providing the flexibility of using text instead of pre-recorded speech. However, simply training speech-driven models with natural speech, and testing them with synthetic speech creates a mismatch affecting the performance of the system. This paper proposes a novel strategy to solve this mismatch. The proposed approach starts by creating a parallel corpus either with neutral or emotional synthetic speech timely aligned with the original speech for which we have the motion capture recordings. This parallel corpus is used to retrain the models from scratch, or adapt the models originally built with natural speech. Both subjective and objective evaluations show the effectiveness of this solution in reducing the mismatch. Second, creating head movement with speech-driven methods can disregard the meaning of the message, even when the movements are perfectly synchronized with speech. The trajectory of head movements in conversations also has a role in conveying meaning (e.g. head nods for acknowledgment). In fact, our analysis reveals that head movements under different discourse functions have distinguishable patterns. Building on the best models driven by synthetic speech, we propose to extract dialog acts directly from the text and use this information to directly constrain our models. Compared to the unconstrained model, the model generates head motion sequences that not only are closer to the statistical patterns of the original head movements, but also are perceived as more natural and appropriate.

1. Introduction

Head movements during conversation play an important role to convey verbal and non-verbal information. For instance, people use prototypical head movements, such as nods for backchannel (Cassell et al., 1994), affirmation, and emphasis (Marsella et al., 2013). The rhythmic beat associated with head movements increases speech intelligibility, as head movements help to parse sentences (Munhall et al., 2004). But the role of head movements is not limited to conveying the lexical message. Head movements also convey the emotional state of the speaker (Pelachaud et al., 1996; Busso and Narayanan, 2007; Busso et al., 2007b). Due to the multifaceted role of head movements in conversations, it is important to synthesize head movements for *conversational agents* (CAs) capturing the relation between head motion and verbal and non-verbal information.

Studies have shown that including head movements for CAs increases the level of perceived naturalness (Busso et al., 2007a; Mariooryad and Busso, 2012b), and the level of warmth and competence (van Welbergen et al., 2015). To generate head movements for CAs, studies usually rely on specifying rules based on the content of the message (Cassell et al., 1999; DeCarlo et al., 2004; Pelachaud et al., 1996). The key limitation of rule-based systems is the repetitiveness of the movements, since the behaviors are constrained by a given set of rules. Also, the timing between head motion and verbal and non-verbal events are not easily captured with rules. As an alternative approach, the strong coupling between speech and head movements (Munhall et al., 2004; Busso et al., 2005; Graf et al., 2002) has motivated data driven methods. Similar to head movements, speech is modulated by the emotional state of the speaker, conveying intonational, rhythmic, and syntactic information. Speech-driven strategies usually include models to capture the strong correlation with head motion, exploring this relationship during synthesis (Busso et al., 2007a; Mariooryad and Busso, 2012b; Sadoughi et al., 2014; Deng et al., 2004; Levine et al., 2010; Sargin et al., 2007; Chiu and Marsella, 2011; Le et al., 2012;

* Corresponding author.
E-mail address: nxs137130@utdallas.edu (N. Sadoughi).

http://dx.doi.org/10.1016/j.specom.2017.07.004
0167-6393/© 2017 Elsevier B.V. All rights reserved.

Ding et al., 2015a; Haag and Shimodaira, 2016). Head movements generated with speech-driven models are not bounded by rules, and can capture the variability shown in their training data. Furthermore, synchronization between speech and head movements is learned from the data which is another key advantage over rule-based systems. However, there are still important open challenges when we rely on speech-driven models. This paper proposes novel solutions to address two of these challenges: lack of generalization due to the need for speech, and lack of meaning in the behaviors.

The first challenge addressed in this study is the need for speech to synthesize head motion. With the increased use of CAs in different domains, it is difficult to pre-record all the utterances that a CA may use. *Text-to-speech* (TTS) systems offer a flexible solution to scale the conversations for CAs. Using synthetic speech is straightforward for rule-based systems, since the rules are derived from the text. However, it is an issue for speech-driven systems, since they rely on acoustic features. Speech-driven frameworks are trained with natural speech that is tightly connected with the head movements. The straightforward approach to address this problem is to extract acoustic features from synthetic speech during testing, as proposed by van Welbergen et al. (2015). The models are trained on the original speech and tested with synthetic speech to generate head movements. However, this solution creates a mismatch problem between the train and test distributions of the speech features, and can have adverse distortions on the generated head movements. Furthermore, when emotionally neutral TTS is used, the synthetic speech signal will lack the characteristic emotional fluctuations observed during daily interactions, generating monotonic head motion sequences. Building on our previous work (Sadoughi and Busso, 2016), this paper proposes elegant solutions to use synthetic speech to generate head motion sequences. For each sentence, we start by creating a synthetic speech conveying the same lexical content, which is timely aligned with the original recordings for which we have motion capture data. We refer to the aligned synthetic speech with the corresponding motion capture data as *parallel* corpus. We use this parallel corpus to either retrain or adapt the models trained with the original recordings. We evaluate the approach with neutral and emotional TTS, where we match the emotional content of the original speech to create an emotional parallel corpus. We demonstrate the benefit of the proposed solutions with objective and subjective evaluations, which show statistically significant improvements in natural perception on the synthesized head movements with the proposed framework compared with the baseline method that has the mismatch between the train and test conditions.

The second challenge addressed in this study is the lack of meaning in the head motion sequences since the models are agnostic about contextual information. Although the correlation between the prosodic information and head movements is high, the trajectory of the head movements may convey lexical and discourse information. For instance, people tend to shake their heads to show disagreement with their interlocutors. Relying only on prosodic features can result in head movements that are synchronized with speech, but are uncorrelated with the meaning of the message. Even worse, it may contradict the meaning (e.g., head shake for affirmation). We propose to constrain our model based on the *dialog acts* (DAs) in the message. We start by automatically extracting several dialog acts from the text, grouping them into broad categories, which we refer to as discourse functions. We analyze differences between head movement patterns under four discourse functions: *affirmation, negation, question*, and *backchannel*. The results from the analysis show statistical differences between the pattens associated with each category. These results emphasize the need to constrain our speech-driven models on these discourse functions. We use an architecture based on *dynamic Bayesian networks* (DBNs) introduced in Sadoughi and Busso (2017) to constrain the models based on the underlying discourse functions. While these DBNs were originally created for speech-driven models, this study demonstrates that the constrained models can also be effectively used when the models are driven by synthetic speech, relying on the proposed training framework. Using synthetic speech, the results with objective and subjective evaluations show that the proposed models are effective in capturing the statistical patterns associated with each discourse function, resulting in head motion with meaning.

The paper is organized as follows. Section 2 discusses previous studies relevant to this study. Section 3 motivates the study, introducing an overview of the proposed framework. Section 4 describes the resources of this study including the corpus, the text-to-speech system, the dialog act recognizer, the baseline DBN model, and the objective metrics used to evaluate the generated head movements. Section 5 presents the proposed approaches to mitigate the train and test mismatch when synthetic speech is used instead of natural speech, reporting the results of our objective and subjective evaluations. Section 6 presents the proposed approach to constrain the models on the underlying discourse function, and reports objective and subjective evaluations of the constrained model driven by synthetic speech, demonstrating the improvements over unconstrained models. Section 7 concludes the paper summarizing the main contributions of the study and discussing open research directions.

2. Related work

2.1. Relationship between head motion and speech

Studies have shown strong coupling between prosodic features and head movements. Kuratate et al. (1999) found a correlation of $\rho = 0.88$ between the fundamental frequency (F0) and head movements while analyzing the recordings from an English speaking subject. Similar results were reported by Munhall et al. (2004), where they found a correlation of $\rho = 0.63$ between the fundamental frequency and head movement, and a correlation of $\rho = 0.324$ between RMS energy and head movements. Busso and Narayanan (2007) found a local coupling of 0.70 in terms of *canonical correlation analysis* (CCA) between head motion and prosodic features including fundamental frequency, intensity, and their first and second order derivatives. Replicating the same analysis, Mariooryad and Busso (2012a) found a CCA of 0.79 between prosodic features and head movements in the improvised recordings of the IEMOCAP corpus (Busso et al., 2008).

2.2. Speech-driven frameworks

There are several studies proposing speech-driven models for head movement generation. Some of these studies have focused on blending original head motion segments, where the segments are chosen based on the input speech. For instance, Chuang and Bregler (2005) proposed to generate head movements by storing segmented F0 contours and their corresponding head motion contours. For synthesis, they limited the search space by finding a reduced set with similar sentence level features extracted from the fundamental frequency. A second search identified the most similar F0 contours in the reduced set. The corresponding motion capture recordings of the selected F0 contours are combined and blended to form the final head motion trajectory for the query. Deng et al. (2004) created a feature vector from speech. Their approach finds the *K nearest neighbors* (KNN) to the input speech from the stored audio segments. Then, they combine the head movement sequences of the selected KNNs by using an objective function as a weighted sum of several criteria, including the distance between the input audio features and a nearest neighbor's stored audio features, the smoothness of the generated head movements, and the distance of the predicted head movements from the true head movements over some known key frames. Their approach is solved with dynamic programming searching through the space of recorded audio features and head movements to minimize the overall cost function.

Many speech-driven studies have relied on probabilistic models such as *hidden Markov models* (HMMs), DBNs, and *Gaussian mixture models* (GMMs). For instance, Busso et al. (2007a) quantized the space of head movement using vector quantization. They learned the coupling between prosodic features and head movements using emotion dependent HMMs. Mariooryad and Busso (2012b) designed several DBNs to learn the relationship between prosodic features, head and eyebrow movements. They showed that jointly modeling head and eyebrow motions results in more natural movements than the ones obtained by separately generating these behaviors. Levine et al. (2009) designed a speech driven framework to generate body movements from speech prosodic features. They clustered the motion data for head, arms, and lower body into subunits, which are considered as hidden states of HMM. The speech is segmented into syllables, and prosodic features are extracted for each segment. The model predicts the most probable motion given the speech, and the previous hidden state. In a subsequent study,Levine et al. (2010) proposed an approach to select a gesture and model its kinematics. The kinematics is inferred from speech. They proposed a *hidden conditional random field* (HCRF) to capture the coupling between prosodic features and 14 joint movements, including head rotations. Given the input speech, the HCRF is used to infer a sequence of hidden clusters over the joint movements. The inferred hidden layers are searched with reinforcement learning to find an optimum policy for selecting the motion segments. Le et al. (2012) proposed to use GMM to model the relationship between the fundamental frequency and intensity and kinematic features of head movement. Assuming independence between the head pose, its velocity and acceleration, they maximized the posterior probability of the kinematic features using gradient descent. Ding et al. (2013) used *fully parameterized hidden Markov models* (FPHMMs) to generate head and eyebrow movements from prosodic speech. FPHMMs are built upon *contextual HMMs* (CHMMs), changing not only the mean and covariance of the hidden state, based on the current context, but also their transition probabilities. For the contextual variables, they relied on the average of the speech features over a sliding window. Objective evaluation of their results showed better performance for joint modeling of the head and eyebrow movements rather than their separate modeling. Hofer and Shimodaira (2007) proposed an approach based on HMM for predicting head movements from speech. The task was predicting classes of head movements (i.e. head shake, head node, shift and none) by training class-dependent HMMs, similar to speech recognition tasks. The class dependent HMMs are trained separately for speech and head pose streams while the transition matrices are tied between the models. During testing, the model trained with speech is used to derive the most probable sequence of motion class labels.

Recent studies have also used deep learning for head movements generation. Ding et al. (2014) showed with objective metrics that *deep neural networks* (DNN) driven with speech features generates better head motion sequences than the ones generated with HMMs. DNN allowed pre-training with unlabeled data, which improved their performance. Lan et al. (2016) proposed a deep learning framework based on *bidirectional long short term memory* (BLSTM) to map acoustic features into facial and head movements. Haag and Shimodaira (2016) extracted bottleneck features using DNN. The input of the DNN was a segment of audio features and its output was the head rotations with their first and second order derivatives for the middle frame. The activation of an intermediate layer (i.e., bottleneck feature) was concatenated with audio features. This feature vector was the input of a BLSTM which predicted the head movements. The evaluation showed that the head motion sequences of the BLSTM were better than the ones generated with the DNN model.

2.3. Head motion frameworks driven by synthetic speech

Most studies on speech-driven methods for facial animation have focused on using natural speech as input. Hence, the sentences uttered by the CAs have to be pre-recorded, which limits the applications of these methods. To overcome this problem, van Welbergen et al. (2015) proposed to use synthetic speech to create head movements. They used the probabilistic model proposed by Le et al. (2012) for modeling the relationship between speech and head movements. They trained the models with 7.4 min of the IEMOCAP corpus. During synthesis, they extracted speech features from synthetic speech to drive the models. The key challenge in this approach is the mismatch between training and testing conditions, due to the clear differences in the acoustic space between natural and synthetic speech. In our preliminary study (Sadoughi and Busso, 2016), we proposed a solution to handle the mismatch created by training with natural speech and testing with synthetic speech. We used the DBN model proposed by Mariooryad and Busso (2012b) to capture the coupling between prosodic features and head movements. To cope with the mismatch, we proposed to build a parallel corpus of synthetic speech aligned with our original recordings (see Fig. 1). As a result, the synthetic speech is also aligned with the original head motion sequences derived from the motion capture data, which is key in this framework. We used the parallel corpus to either retrain the model from scratch, or adapt the model originally trained with natural recordings. The evaluation results showed that both methods improved the results compared to the baseline model which has the train and test mismatch. Our study builds upon this framework, as described in Section 5.

2.4. Meaningful behaviors by speech-driven models

Speech-driven models relying only on prosodic features may create head motions that are perfectly synchronized with speech but contradict the meaning of the message, generating behaviors without meaning. Adding meaning to behaviors is the key advantage of rule-based systems, where the rules for head motion are dictated by contextual information (Cassell et al., 1994; 2003; Poggi et al., 2005). An appealing solution is to combine rule-based and data-driven methods, overcoming these limitations, bridging the gap between these methods. There are few studies that have attempted this approach. Marsella et al. (2013) mapped the input speech and its transcription into behaviors using a multi-step framework. Their framework starts by performing a syntactic analysis on the text to infer the potential communicative functions. They defined a dictionary of functions which maps communicative functions into behaviors (e.g. big nod for *affirmation*, shake for *negation*, tilt right for *contrast*, small nod for *emphasis*). They also performed an analysis over the audio to estimate the emotional content of the utterance. The result of the analysis affects the selection of the behaviors, providing emotion-related rules (e.g. side to side head movements instead of front to back head movements for sadness).

In a preliminary study (Sadoughi et al., 2014), we studied the role of discourse information on behaviors. We manually annotated four discourse functions including *affirmation, negation, question*, and *statement* for all the recordings in the first session of the IEMOCAP database (one male and one female speaker). The analysis of these discourse functions showed that there are differences

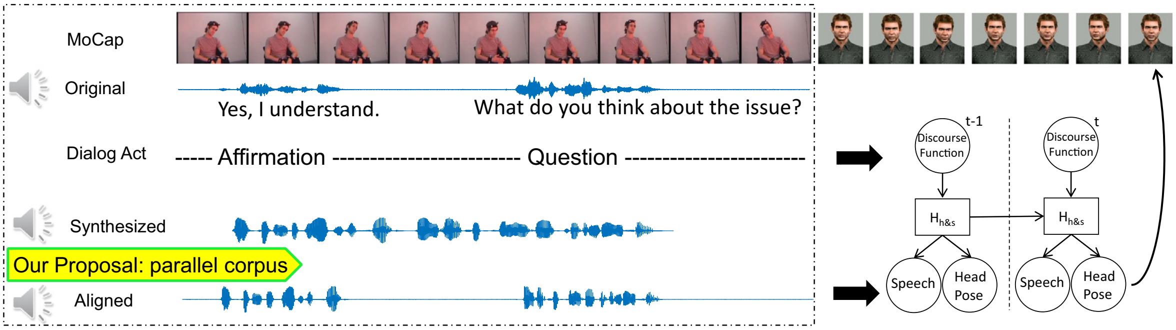


Fig. 1. Overview of the proposed method.

between the head and eyebrow movements under these discourse functions. These results suggested that constraining the models on the discourse function of the sentence was an effective approach to generate gestures with meaning, where the actual rules are directly learned from the data. We have explored these ideas with models driven with natural speech (Sadoughi et al., 2014; Sadoughi and Busso, 2017). Building on the DBNs proposed by Mariooryad and Busso (2012b) (see Section 4.6), we added an extra child variable to model two discourse functions: *question* and *affirmation* (Sadoughi et al., 2014). For *question*, the constrained models were perceived more natural than the baseline models without constraints. However, the results were inconclusive for *affirmation*. Since the study relied on a small portion of the IEMOCAP database, Sadoughi and Busso (2017) extended the analysis on the MSP-AVATAR corpus (Sadoughi et al., 2015). We considered four discourse functions: *affirmation, negation, question*, and *suggestion*. The paper analyzed the differences for head and arm movements across the discourse functions, proposing an alternative DBN to effectively constrain the models on the underlying discourse function. The results showed improvements in terms of subjective and objective metrics when we constrained the models on the discourse functions or behaviors.

3. Motivation and proposed approach

Previous studies have demonstrated the benefits of using speech-driven models to synthesize head motion sequences. This paper proposes novel solutions for head motion models driven by synthetic speech. The proposed frameworks improve the generalization and contextualization of these models by (1) handling the mismatch condition that exists for speech driven head movements when we use synthetic speech using emotional synthetic speech, and (2) effectively constraining the synthetic speech driven models on the underlying discourse function using the dialog acts directly extracted from text. These contributions represent transformative solutions for speech-driven head motion models. Although the study focuses on these two important problems, it is important to notice that head motion also depends on other factors such as age, gender, personality, and culture. These factors are not considered in this study and are left as future extensions of this work.

Fig. 1 gives an overview of the proposed approach with these two main contributions. This section presents the motivation of the proposed solution, describing the general ideas. We discuss the details in Sections 5 and 6.

3.1. Use of speech synthesis

Using synthetic speech to drive head motion models trained with natural speech creates a mismatch problem. In Sadoughi and Busso (2016), we presented preliminary evaluations to address this mismatch problem. We proposed the use of a parallel corpus with emotionally neutral synthetic speech that is timely aligned with the original recordings, which is used to either retrain the models from scratch or adapt the trained model with original recordings to synthetic speech. Although the parallel corpus with neutral synthetic speech may have similar intonation and beats as the original recordings, it lacks the range of acoustic changes associated with emotion. If the synthetic speech is neutral, the correlation between the parallel corpus and the motion capture recordings might not be strong enough for the model to learn. This paper builds upon this work, extending the analysis to incorporate emotional speech. We explore the use of emotional speech to build the parallel corpus that not only is timely aligned with the original speech, but also is consistent with its emotional content (as described by arousal, valence and dominance scores). This approach reduces even more the mismatch between training and testing conditions, improving the results as described in Section 5.

3.2. Adding discourse information

An open question is whether constraining the models on the discourse functions is also useful for models driven by synthetic speech. One clear advantage in driving the models with synthetic speech is that we can automatically label dialog acts from text, removing the need for manual annotations, which is time and labor demanding. The approach also provides an elegant solution to automatically learn rules for each discourse function from the data, producing behaviors that preserve the statistical patterns of head motion observed during daily interactions. This study relies on a classifier, which automatically assigns the underlying dialog act labels to the sentences. This approach allows us to extend the analysis over the entire IEMOCAP corpus. We analyze head movements under different discourse functions by grouping similar dialog acts. The objective and subjective analysis of the synthesized head movements (Section 6) shows that when constrained on the discourse functions our models capture similar patterns as shown in the original data, resulting in more natural and appropriate head movements.

3.3. Overview of the proposed system

Fig. 1 gives the overview of the proposed system. We start by creating a parallel corpus which not only is aligned with the original speech recordings, but also carries similar emotional content as described by the arousal, valence and dominance scores. We use the emotional parallel corpus to retrain or adapt the models. The target text is analyzed by our dialog act classifier, assigning the corresponding discourse function classes. This information is used by our constrained models generating natural and meaningful behaviors that are driven by synthetic speech.

4. Resources

4.1. The IEMOCAP corpus

This study uses the *interactive emotional motion capture* (IEMOCAP) database (Busso et al., 2008), which is a multimodal corpus designed for studying emotions. This corpus contains video, audio, and motion capture recordings from 10 actors during dyadic interactions. The corpus was recorded with improvised and script based scenarios. The corpus was segmented into speaking turns, where each turn was emotionally annotated in terms of the attributes arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) by two evaluators, and categorical emotions (happiness, sadness, anger, neutral state, excited, frustrated, disgusted, fearful, surprised, other), by three annotators. We assign the average scores across annotators for arousal, valence and dominance provided to each speaking turn. The corpus also includes transcriptions, with time information for turns, words and phonemes derived with forced alignment. We used 270.16 min of the IEMOCAP corpus consisting of turns with non-overlapping segments.

4.2. Creating the parallel corpus with MaryTTS

The proposed approach relies on creating an emotional parallel corpus with synthetic speech. We use MaryTTS (2017), an open source *text-to-speech* (TTS) toolkit. MaryTTS provides word timings, which we use to align the synthetic speech to the original speech signal. It also has an option to create emotional speech, which is controlled by specifying the values for continuous emotional attributes (e.g., valence, arousal, and dominance). This feature makes this TTS a perfect toolkit for our study.

Lotfian and Busso (2015) proposed to use aligned synthesized sentences as a neutral reference to contrast expressive speech cues in emotion recognition tasks. We adopted the same approach to align the synthetic speech with the original recordings, where the speech signal for each word is warped to match the word duration in the original sentence. For this purpose, we used the *pitch synchronous overlap add* (PSOLA) technique (Moulines and Charpentier, 1990) implemented in Praat, which warps the speech signal while maintaining its fundamental frequency. We replace segments without speech with prerecorded silence segments under similar conditions, avoiding zero-value regions which create problems to our probabilistic models. Fig. 1 illustrates this method, displaying in order the original speech, the synthesized signal, and the time-aligned synthesized signal.

To demonstrate the importance of using an emotional TTS, we create the parallel corpus under two conditions: *emotional parallel corpus*, where we match the arousal, valence and dominance scores assigned to the sentences in the IEMOCAP corpus, and, *neutral parallel corpus*, where the sentences are synthesized without specifying the emotional attributes. Since the parallel corpora are aligned with the original sentences, they are also tightly aligned with the head motion sequences in the original corpus.

4.3. Dialog act recognition

The proposed models described in Fig. 1 respond to the underlying dialog act, which are automatically detected from the text. We build a dialog act classifier for this purpose. We used the *Switchboard dialog act* corpus of telephone conversations (Jurafsky et al., 1997) for training a dialog act recognizer. This corpus contains dialog act annotations from a set of 42 labels for 1155 five-minute conversations, containing 205,000 utterances and 251,187 distinct words. The algorithm that we use for dialog act recognition uses bag of words features, and calculates the entropy of the words (uni-grams, and bi-grams) across all the dialog acts, and ranks them accordingly. The top 1000 words from uni-gram, and the top 1000 words from bi-gram categories are chosen as the features. We train a *conditional random field* (CRF) classifier with these features. CRFs are popular models to discriminate sequences of varying sizes. We used the pocket CRF toolkit implemented by Qian (2010). This framework reaches an accuracy of 72.6% on the test set of the switchboard, which is 7% below the state-of-the-art performance reported on this corpus for this task (Ribeiro et al., 2015). This performance is good enough for our application.

The sentence boundaries are detected by finding the end-of-sentence punctuations in the transcriptions of the IEMOCAP corpus including period, question mark, and exclamation points. We use the timings from the word alignment to segment the sentences. We use the CRF classifier to assign a dialog act label to each of the segmented sentences. We consider four groups of discourse functions including *question, affirmation, negation* and *backchannel*. Each of these discourse functions comprises one or multiple dialog act labels:

- *affirmation* includes “agree”, “yes answers” and “affirmative non-yes answers”
- *negation* includes “no answers”
- *question* includes “question”, “or-clause”, “wh-question”, “rhetorical question”, and “declarative wh-question”
- *backchannel* includes “acknowledge/backchannel”, and “acknowledge answer”

The backchannel class corresponds to short segments where the speaker verbally communicates active listening using words such as “yeah” and “right”. Jurafsky et al. (1997) provides examples for each of these dialog acts. After inspecting the results, we identified sentences with the word *no* which were wrongly labeled as “agree”. We moved those sentences into the *negation* group. Overall, the total number of sentences are 231 for *affirmation*, 128 for *negation*, 1091 for *question* and 135 for *backchannel*.

4.4. Audiovisual features

Previous studies on head motion synthesis have shown that prosodic features including *fundamental frequency* (F0) and *intensity* are highly correlated with head motion and can be used to generate head movements (Busso and Narayanan, 2007; Busso et al., 2007a). Using Praat, we extracted the F0 and intensity over window size of 40 ms, with 23.3 ms overlap (60 fps). To synchronize the motion capture recording (120 fps), with the audio features, we up-sample the audio features. Also, to avoid introducing discontinuities of F0 in our models, we linearly interpolated the unvoiced regions. We z-normalized the prosodic features extracted from the natural speech and emotional parallel corpus by using speaker-dependent mean and global variance. We follow a similar approach for the features extracted from the neutral parallel corpus, where the global variance is scaled to match the variance of neutral segments in the IEMOCAP corpus. We represent head motion using three Euler angles, extracted from the motion capture recordings using *singular value decomposition* (SVD). The details are explained in Busso et al. (2008). We do not consider head translation in this study.

4.5. Objective metrics to assess performance

As an objective metric, we use local and global *canonical correlation analysis* (CCA), and *Kullback–Leibler divergence* (KLD) to evaluate the generated head movements. CCA finds linear transformations (A, and B) for two multidimensional variables X and Y which transform the variables into a new orthogonal space, where the correlations between the transformed variables (i.e. A^TX, and B^TY)

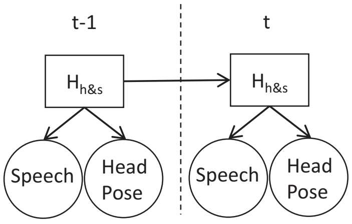


Fig. 2. The diagram of the DBN proposed by Mariooryad and Busso (2012b). We build the proposed models starting from this structure.

are maximized. In the new space, the dimensions are sorted according to correlation value between the two variables, where the first dimension has the largest correlation value. We use the correlation between the two variables along the first dimension in the new space as a criterion of performance. Local CCA measures the CCA over each turn, averaging the results across all the speaking turns. Global CCA finds the CCA after concatenating the sequences for all the speaking turns, estimating a single CCA value. The values for global CCA are usually lower than the ones for local CCA, since it uses a single transformation across the entire data (i.e., matrices *A*, and *B*). We measure CCA between the original and synthesized head movements (CCA_h), and between the input speech used to drive the models (natural or synthetic speech) and the synthesized head movements (CCA_{hs}). High values of CCA indicate that the generated sequences follow the temporal patterns of the original head motion sequences (CCA_h) and they are timely coupled with the input speech (CCA_{hs}). As a reference, the local CCA for the original head movements and the original prosodic features is $CCA = 0.7664$, which is very high.

Our second metric is KLD, which gives an asymmetric distance between two distributions *p*, and *q*. $KLD(p||q)$ is the amount of information lost when *q* is used to represent *p*. We evaluate the KLD between the synthesized head movements (*q*) and the original head movements (*p*). Ideally, this value should be as small as possible, indicating that the generated head sequences have similar distributions as the original head motion sequences.

4.6. Speech-driven models with dynamic Bayesian network

The proposed models correspond to DBN, building on the structure proposed by Mariooryad and Busso (2012b). Before introducing the proposed structure, we summarize the base model described in Fig. 2. The *Speech* node represents the prosodic features, and the *Head Pose* node represents the head pose (three Euler angles) at each time frame. The $H_{h\&s}$ node is a hidden state representing the possible joint configurations observed in the data between prosodic features and head poses (e.g., head-speech codebook). When the *Speech* node is initialized, the $H_{h\&s}$ node changes affecting the state of the *Head Pose* node. Since the two visible nodes are continuous variables, their conditional probabilities are modeled with Gaussian distributions. Furthermore, we assume that the transitions follow a Markov property of order one (i.e. each state is conditioned based on only one past state). This model does not require the features within each modality to be uncorrelated, since it uses a full covariance matrix for prosodic features and head pose. However, it assumes that *Speech* and *Head Pose* nodes are independent given $H_{h\&s}$.

The model is composed of initial probabilities of the states, the transition probabilities between the states, and the conditional observation probabilities. We optimize the parameters of the DBN with *expectation maximization* (EM). Assuming initial values for the parameters and training samples, we can infer the posterior probabilities of the states given each data (E-step), and use those values to update the parameters of the model to maximize the log-likelihood of the model (M-step). We run the forward-backward algorithm (Murphy, 2002) to infer the posterior probabilities during the E-step. During the M-step, we update the parameters using the closed form solutions derived by maximizing the log-likelihoods (Murphy, 2002).

For inference during training, we have access to both observation nodes (*Speech, Head Pose – full observation*). During synthesis, we only have observations for the *Speech* node (*partial observation*). We get the expected value for the *Head Pose* node at each time frame, given the values for the *Speech* node for the entire sentence,

$$\begin{aligned} &\gamma_{i,t} = P(H_{h\&s} = i | Speech_{1:T}) \\ &HeadPose = \Sigma_{i=1}^{n} \mu_i \times \gamma_{i,t} \end{aligned} \tag{1}$$

where, $H_{h\&s}$ represents the hidden state, $Speech_{1:\,T}$ represents the prosodic features for a sentence with *T* frames, $\gamma_{i,\,t}$ is the posterior probability of the model given the input speech, and μ_i is the mean of head pose for the i^{th} state.

As aforementioned, the DBN's parameters are optimized using the EM algorithm, which is sensitive to initialization. Therefore, we initialize the states using *Linde–Buzo–Gray vector quantization* (LBG-VQ) (Linde et al., 1980). This initialization expedites learning, and reaches a better likelihood for the model (Sadoughi and Busso, 2017).

After getting the expected outputs from the models, we smooth the head movement sequences, following the same approach proposed by Busso et al. (2005). The synthesized head motion sequence is downsampled to 15 fps, defining key points. Then, the head Euler angles of these key points are transformed into quaternions, interpolating the rest of the frames using spherical cubic interpolation.

4.7. Training DBNs with natural speech

We train the DBNs in Fig. 2 using the original speech in the IEMOCAP corpus. We refer to this model as DBN-O. We use this model as reference, and also as the initial model for the adaptation scheme described in Section 5.2. We perform an eight-fold cross validation, using 6 folds for training, 1 fold for testing, and 1 fold for validation.

During the cross-validation, we use the validation set to conduct a grid search over the number of states for $H_{h\&s}$. We consider three to 22 states, setting the number of iterations for the EM algorithm during training of the DBNs to four, since our previous experiments showed that four iterations is usually enough (Sadoughi and Busso, 2016). This process is separately optimized for each fold. Fig. 3 displays the local and global CCA_h and the KLD, as a function of the number of states for one of the validation sets. We choose the number of states such that the values for global and local CCA_h are high and the value for KLD is low. For the iteration shown in Fig. 3, we chose 16 states, since KLD is saturated, and the values for local and global CCA_h are high. We perform the same grid search across the remaining seven cross-validations. The first row of Table 1 shows the average (18.25) and standard deviations (1.669) of the number of states for the DBN-O model (i.e., train and test with original speech). It also provides the performance measured with global and local *CCAs* and KLD. The results show that the synthesized head movements are highly correlated with the original head movements and prosodic features. The KLD for this model is given as a reference, highlighting that head motion patterns are not fully determined by prosodic features.

5. Head movements driven by synthetic speech

This section describes the methods used to assess the use of emotional and neutral parallel corpora to handle the mismatch of

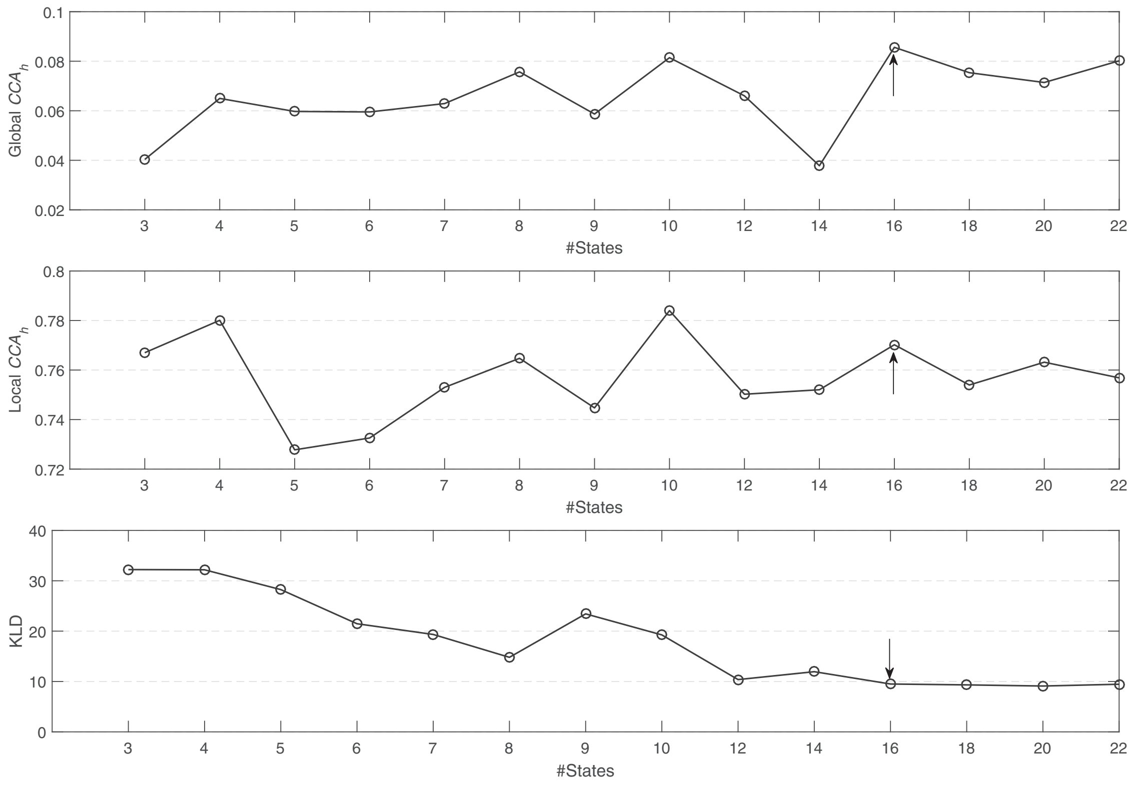


Fig. 3. Selection of the number of states for $H_{h\&s}$ using validation set. The figures show three objective metrics for one fold of the cross validation. The arrows indicate the selected value, which for this fold is 18.

Table 1
Objective evaluation of the DBNs over the test set. An asterisk denotes that the local CCA values for the retrain and adapt conditions are significantly higher ($p < 0.05$) than the local CCA for the corresponding mismatch condition (i.e., E or N).

Data	Condition	#States	Local		Global		KLD
		MEAN (STD)	CCA_h	CCA_{hs}	CCA_h	CCA_{hs}	
Original	DBN-O	18.25 (1.7)	0.7538	0.8615	0.1078	0.3945	8.6699
Neutral Parallel	DBN-mismatch-N	18.25 (1.7)	0.7402	0.7721	0.0357	0.1815	8.5156
	DBN-retrain-N	12.87 (3.0)	0.7183	0.8126*	0.0831	0.3768	6.9763
	DBN-adapt-N	18.25 (1.7)	0.7363	0.8094*	0.0393	0.2496	8.4847
Emotional Parallel	DBN-mismatch-E	18.25 (1.7)	0.7596	0.7932	0.0470	0.2109	8.7373
	DBN-retrain-E	18.00 (1.1)	0.7561	0.8453*	0.1227	0.4912	6.6541
	DBN-adapt-E	18.25 (1.7)	0.7617*	0.8312*	0.0456	0.2884	8.7362

training the head motion models with natural speech, and testing the models with synthetic speech. The baseline methods correspond to mismatched conditions, where the models trained with original speech are tested with neutral synthetic speech (DBN-mismatch-N) or emotional synthetic speech (DBN-mismatch-E). To handle this mismatch, we proposed two approaches; retraining the model from scratch with the neutral parallel corpus (DBN-retrain-N) or the emotional parallel corpus (DBN-retrain-E), and adapting the model trained with the original speech using the neutral parallel corpus (DBN-adapt-N) or the emotional parallel corpus (DBN-adapt-E).

5.1. Retraining DBNs with parallel corpora (DBN-retrain)

The parallel corpora are timely aligned with the original head motion sequences. Therefore, we can train the DBNs in Fig. 2 from scratch using the parallel corpora. This approach removes the mismatch between train and test conditions in our baseline methods (DBN-mismatch-N and DBN-mismatch-E), since it uses synthetic speech to create the models. To evaluate the contribution of using emotional synthetic speech, we separately evaluate this approach building the DBNs with the neutral parallel corpus (DBN-retrain-N) and emotional parallel corpus (DBN-retrain-E).

5.2. Adapting the DBN with parallel corpora (DBN-adapt)

Instead of training the models from scratch, we also propose to adapt the model trained with the original recordings (DBN-O - Section 4.7), using the audio from the parallel corpora. The key idea is to initialize the model with speech-driven models that capture the relationship between speech and head motion. After that, we reduce the mismatch by adapting these models to capture the acoustic properties of synthetic speech. We view adaptation as a more conservative approach than training from scratch since the starting model captures the actual relationship between speech and head motion. We implement this approach using the neutral parallel corpus (DBN-adapt-N), and the emotional parallel corpus (DBN-adapt-E).

We use *maximum a posteriori* (MAP) adaptation to adapt the models originally trained with natural speech. The models are adapted using synthetic speech instead of the original speech. The adaptation is only applied to the speech parameters of the model, since the only difference is the acoustic features. Notice that the conditional probability is modeled with Gaussian distributions (Section 4.6). Therefore, the speech parameters for state i are the mean vector (μ_i) and covariance matrix (Σ_i). The joint conjugate prior density is Normal–Whishart (Gauvain and Lee, 1994).

Our previous study showed that mean adaptation alone was better than mean and covariance adaptation (Sadoughi and Busso, 2016), so this study only uses mean adaptation. This is a standard expectation-maximization formulation which is derived by maximizing the posterior probability, resulting in Eq. (2) (Murphy, 2007). This equation gives the closed form solution for updating the mean of the states, given the new data, where μ_{pi} is the prior mean of the i^{th} state, and $\bar{x}_i$ is the estimated mean of the i^{th} state given the new data, n_p is the weight of the prior estimation and n is the weight associated with the new estimation. Note that when updating the parameters we use the same weights across the states proportional to the amount of adaptation data (n_p). Since we synthesize the speech for all the recordings, the size of the original and synthesize speech is the same, resulting in $\frac{n_p}{n_p+n} = 0.5$. Therefore, μ_i is just the average of μ_{pi} and $\bar{x}_i$.

$$\mu_i = \frac{n_p\mu_{pi} + n\bar{x}_i}{n_p + n} \tag{2}$$

The number of iterations for MAP adaptation is separately determined using the validation set, for each fold in the cross-validation. The average number of iterations used for MAP is 2.635 ($STD = 0.744$) for DBN-adapt-N, and 2.125 ($STD = 0.641$) for DBN-adapt-E.

5.3. Experimental evaluation using objective metrics

Table 1 shows the results in terms of the objective metrics described in Section 4.5. The first row corresponds to the DBN-O model, where we train and test the model with the original recordings. This case does not have a mismatch, serving as a reference. For local CCAs, we evaluate statistical differences with the t-test, asserting significance with p-value $<$ 0.05. For global CCAs and KLD, we only have one value per case, preventing us from conducting statistical tests.

The DBN-mismatch-N and DBN-mismatch-E models introduce a mismatch by using synthetic speech. To estimate the values of CCA_h and CCA_{hs} for these cases, the duration of the sequences have to be synchronized with the original sequences. Therefore, we use the synthetic sentences from the neutral parallel corpus (DBN-mismatch-N) and the emotional parallel corpus (DBN-mismatch-E). When we generate the head motion sequences with synthetic speech without any compensation (DBN-mismatch-N and DBN-mismatch-E), the values for global and local CCA drop (except local CCA_h for DBN-mismatch-E). The results show a statistically significant drop in local CCA_{hs} for the DBN-mismatch-N models compared to DBN-O ($p = 3.8 \times e^{-115}$). By using emotional parallel corpus, the correlations increase, but they are still lower than the values obtained by using the original recordings (for local CCA_{hs}: $p = 2.9 \times e^{-76}$).

When we compare the model trained from scratch with the neutral parallel corpus (DBN-retrain-N) with its mismatched condition (DBN-mismatch-N), we observe that the local CCA_{hs} increases ($p = 1.1 \times e^{-17}$), the global CCA_h and CCA_{hs} increase, and the KLD decreases. All these results suggest that DBN-retrain-N is a better model than DBN-mismatch-N. We unexpectedly observe that local CCA_h decreases to 0.7183 ($p = 1.7 \times e^{-13}$). We hypothesize that head movement is affected by emotion. The neutral parallel corpus ignores emotional fluctuations on the acoustic features reducing the correlation between the original and synthesized head motions. In fact, when we use the emotional parallel corpus (DBN-retrain-E), we observe that the local CCA_h increases to 0.7561, overcoming this problem. The model DBN-retrain-E provides the best performance across all the metrics, except local CCA_h, for which the differences are not statistically significant. These results reveal the benefits of using emotional TTS. The values for all the CCA metrics are comparable with the ones obtained by the models generated with the original speech (DBN-O).

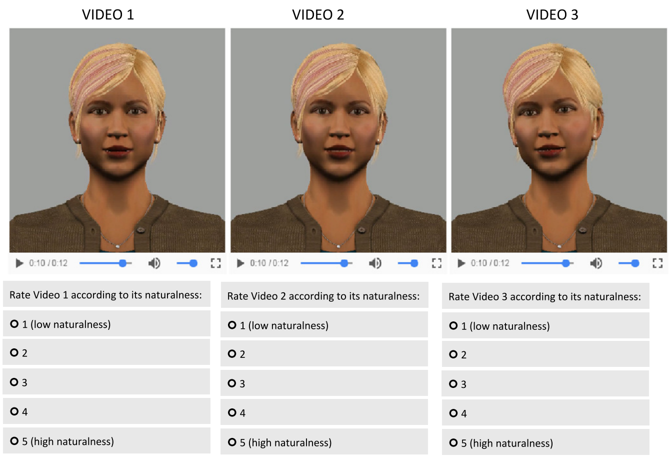
When we adapt the models using the parallel corpora (DBN-adapt-N and DBN-adapt-E), the results are in general better than the mismatched conditions (DBN-mismatch-N and DBN-mismatch-E), but worse than the models trained from scratch (DBN-retrain-N and DBN-retrain-E). Overall, we also observe that using emotional parallel corpus (DBN-adapt-E) gives better results than using the neutral parallel corpus (DBN-adapt-N).

5.4. Experimental evaluation using subjective evaluation

We also report results comparing the models using subjective scores obtained with crowdsourcing evaluations. Subjective evaluations provide convincing evidences to assess the quality of the head motion sequences generated with the models. The audio plays an important role in assessing the naturalness of the videos. Therefore, it is important to compare cases where the audio is constant across models. For this reason, we separately evaluate the models with neutral parallel corpus and emotional parallel corpus. In the first evaluation, we compare the models DBN-mismatch-N, DBN-retrain-N and DBN-adapt-N. In the second evaluation, we compare the models DBN-mismatch-E, DBN-retrain-E and DBN-adapt-E. For the mismatched models, we use neutral (DBN-mismatch-N) or emotional (DBN-mismatch-E) synthetic speech without the alignment step.

For the evaluation, we used Smartbody (Thiebaux et al., 2008), which is an animation toolkit that can render BVH files, and uses the phonetic time alignment to generate synchronized lip motion. We use the time alignments of the phonemes provided by MaryTTS. The evaluation uses the test sentences from the IEMOCAP database. For the evaluation, we replace one of the speaker with a CA. We concatenated two consecutive speaking turns per speaker always starting with the CA in listening mode to give the annotators enough context for each video that we generated. The CA is in idle mode while listening to the interlocutor. In each of the two evaluations, we generated 20 videos for each condition. These videos were perceptually assessed with crowdsourcing evaluations, using *Amazon mechanical turk* (AMT). To increase the quality in the evaluation, we only allowed annotators who had performed well in our previous studies relying on crowdsourcing evaluations (Burmania et al., 2015; Sadoughi and Busso, 2017). Fig. 4 shows the interface of our evaluation. We give three videos at a time to each evaluator showing the same interaction under the three conditions (e.g., DBN-mismatch-N, DBN-retrain-N and DBN-adapt-N). The audio and other facial behaviors are the same, and the only difference is the head motion sequences. To avoid lazy evaluators completing the task without even watching the videos, we only show the questionnaires after they have thoroughly watched the three videos. We also randomize the presentation order of the three conditions for each task.

The evaluators are asked to annotate the level of naturalness from the videos, using a likert scale from 1 (low naturalness) to 5 (high naturalness). Each evaluator is given 30 videos in total (10 × 3 conditions) to avoid fatigue. In total we have 60 evaluators, where 30 evaluators annotated the videos using the neutral parallel corpus and 30 evaluators annotated the videos using the emotional parallel corpus. Each video is evaluated by 15 raters. We estimate the inter-evaluator agreement from the annotations. The Cronbach's alpha is $\alpha = 0.60$ for the videos with neutral syn-



VIDEO 1

Rate Video 1 according to its naturalness:

○ 1 (low naturalness)

○ 2

○ 3

○ 4

○ 5 (high naturalness)

VIDEO 2

Rate Video 2 according to its naturalness:

○ 1 (low naturalness)

○ 2

○ 3

○ 4

○ 5 (high naturalness)

VIDEO 3

Rate Video 3 according to its naturalness:

○ 1 (low naturalness)

○ 2

○ 3

○ 4

○ 5 (high naturalness)

Fig. 4. The user interface for the perceptual evaluation using crowdsourcing.

thetic speech, and $\alpha = 0.76$ for the videos with emotional synthetic speech.

Fig. 5 gives the average ranking by the annotators for the two evaluations. The Kolmogorov–Smirnov test on the distributions of the ratings for the groups revealed that the ratings are not Gaussian. Therefore, we used the Kruskal–Wallis test to assess statistical differences between the conditions. When using neutral parallel corpus (Fig. 5a), the test reveals that the three conditions come from different distributions ($\tilde{\chi}^2(2, 897) = 56.1$, $p = 6.5 \times e^{-13}$). Pairwise comparisons of the results show that values for DBN-mismatch-N are significantly lower than the values for DBN-retrain-N ($p = 1.7 \times e^{-5}$) and DBN-adapt-N ($p = 1.4 \times e^{-3}$). We did not find significant differences between DBN-retrain-N and DBN-adapt-N. When using emotional parallel corpus (Fig. 5b), we observe similar results. The Kruskal–Wallis test shows significant differences between the conditions ($\tilde{\chi}^2(2, 897) = 22.5$, $p = 1.3 \times e^{-5}$). Pairwise comparisons show that the values for DBN-mismatch-E are significantly lower than the values for DBN-retrain-E ($p = 9.7 \times e^{-10}$) and DBN-adapt-E ($p = 8.7 \times e^{-9}$). We did not find statistical differences between the values for DBN-retrain-E and DBN-adapt-E.

The result from the subjective evaluations shows that retraining the models from scratch or adapting them to the synthetic speech improves the generated head movements in terms of naturalness. Moreover, comparisons between the neutral (Fig. 5a) and emotional (Fig. 5b) cases show that the videos generated with emotional synthetic speech are perceived with higher level of naturalness. We take this result with caution since the differences may be due to speech quality, better coupling between head motions and speech, or a combination of both effects.

6. Discourse functions & head movements

The second contribution of this study is adding contextual information to the models to create behaviors with meaning. We analyze the head movement patterns associated with four discourse function classes considered in this study (Section 4.3). We build on the DBN-retrain-E model which is the best model described in Section 5, presenting a systematic framework to constrain the head motion sequences by the underlying discourse function. The proposed framework aims to capture the consistent patterns on the head trajectories that exist when the ten actors in the corpus communicated a message associated with a specific discourse function (e.g. head shaking during negations). The movements are clearly speaker and cultural dependent, so results from a similar database collected from other individuals may provide different data-learned patterns. However, this study focuses on the formulation of the problem, showing that it is possible to constrain the text driven models to convey behaviors that are characteristic of a given discourse function.

6.1. Role of discourse function on head motion

We extract the mean and standard deviation of the head pose trajectories under different discourse functions to explore possible inter-class differences. First, we group all the sentences in the IEMOCAP corpus according to the discourse functions. The column "Original" in Figs. 6 gives the mean values for pitch (−up, + down), yaw (−left, + right) and roll (−clockwise, + counterclockwise) of the original head motion sequences as a function of the discourse functions. The figure illustrates that for *question* the head pose is oriented in a look-up direction. This result is in line with the rule used by Cassell et al. (1994), where the CA looks up for *question*.

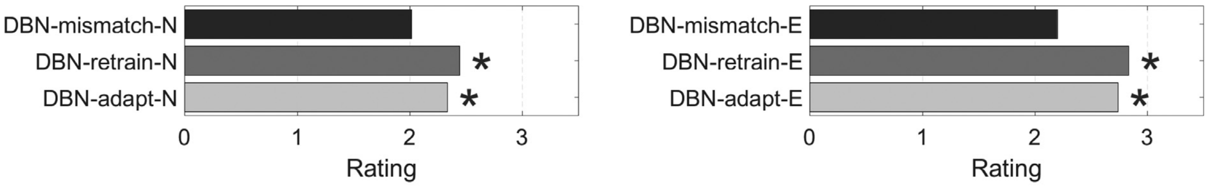
We also study the standard deviation of head movements as a function of discourse functions to identify potential differences in the dynamic of the trajectories. The average durations of the sentences are 0.51 s ($STD = 0.36$) for *affirmation*, 0.46 s ($STD = 0.35$) for *negation*, 1.12 s ($STD = 0.91$) for *question*, and 0.36 s ($STD = 0.16$) for *backchannel*. These duration differences can influence the standard deviations estimated over each segment. Furthermore, our models give the same weight to each frame, so it is important to keep segments with consistent duration in the analysis when we compare the head motion patterns generated by our model. Therefore, we randomly select segments of 200 ms (24 frames) from each sentence, discarding the ones with shorter lengths. We estimate the standard deviations over these segments, grouping the results as a function of the discourse functions (203 segments for *affirmation*, 101 segments for *negation*, 1040 segments for *question* and 115 segments for *backchannel*). The column "Original" in Figs. 7 gives the average of the standard deviations for pitch, yaw and roll for each discourse function.

The Kruskal-Wallis test shows that the mean value of the standard deviations for pitch, yaw, and roll are different, for all the discourse functions (*affirmation*: $\tilde{\chi}^2(2, 606) = 24.2$, $p = 5.7 \times e^{-6}$, *negation*: $\tilde{\chi}^2(2300) = 18.5$, $p = 9.8 \times e^{-5}$, *question*: $\tilde{\chi}^2(2, 3117) = 105.5$, $p = 1.2 \times e^{-23}$, *backchannel*: $\tilde{\chi}^2(2342) = 15.0$, $p = 5.5 \times e^{-4}$). Fig. 7 indicates with color-coded asterisks the results of pairwise comparisons between pitch, yaw and roll angles. An asterisk on top of the bar indicates that the mean value for this bar is significantly higher than the mean value of the bar associated with the asterisk's color (we assert significance at p-value < 0.05). For example, consider the results for the original head movements for *affirmation*. The mean standard deviation for pitch and yaw are statistically higher than the mean standard deviation for roll, but the difference between pitch and yaw angles is not statistically significant. For *negation*, yaw movements are more prominent than pitch and roll movements, as we expected more head shakes. For *question*, we see more prominent movements for pitch and yaw angles than for roll. For *backchannel*, the pitch movements are more prominent than the movements for yaw or roll, as we expect more head nods. This observation agrees with the rule of nodding during backchannel used by Cassell et al. (1994).

The results from this analysis show statistical differences in head motion patterns as a function of discourse functions. We aim to capture these differences in our model, by constraining the model with discourse functions.

6.2. Constrained DBN model (CDBN)

To capture the characteristic head movements for each dialog act, we used the model proposed by Sadoughi and Busso (2017),



(a) Videos with neutral synthetic speech

(b) Videos with emotional synthetic speech

Fig. 5. Average of the perceived naturalness obtained with crowdsourcing evaluations. The color-coded asterisks denotes statistically higher values for the bar compared with the bar with the same color of the asterisk ($p < 0.05$).

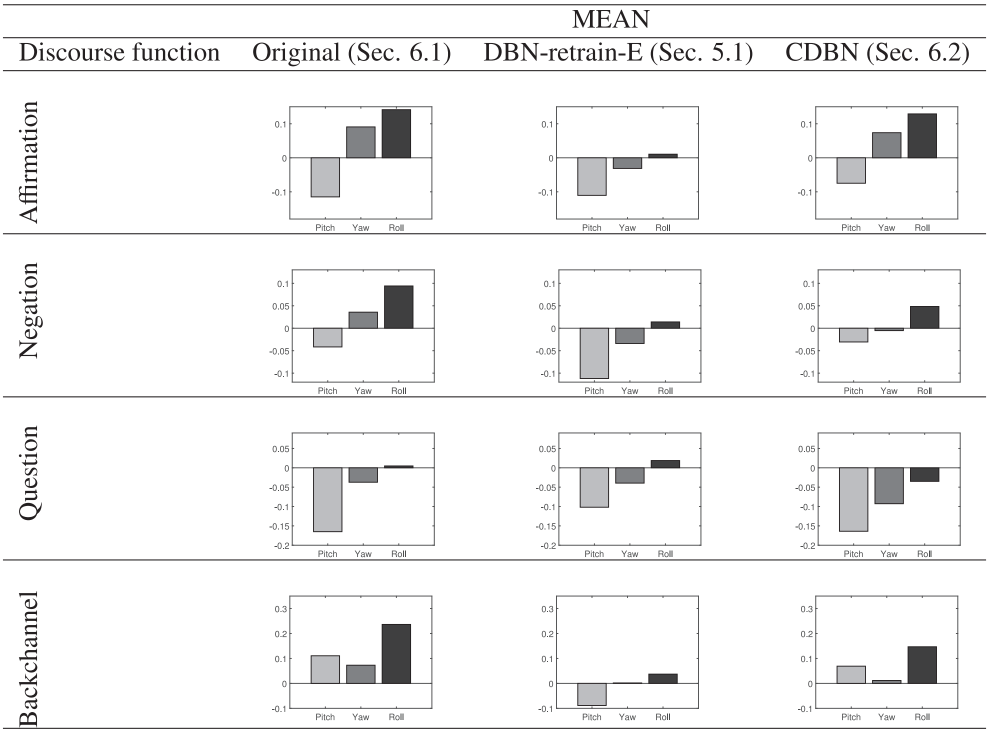


Fig. 6. The average of head angles per discourse function for the original head motion (Section 6.1), and the synthesized sequences using DBN-retrain-E (Section 5.1) and CDBN (Section 6.2).

which we refer to as *constrained DBN* (CDBN). These models were originally created for speech-driven applications. This paper extends the approach to models driven by synthetic speech.

Fig. 8 shows the model, where the hidden state $H_{h\&s}$ is constrained by a discourse function node. As a result of this node, the transition probabilities between the states are also dependent on the current discourse function (See Eq. (3)). The CDBN encodes the differences between discourse functions in this conditional probability (i.e. a_{ijk}), capturing the characteristic patterns associated with each discourse class.

$$a_{ijk} = P[H_t = i | H_{t-1} = j, DiscourseFunction_t = k] \quad (3)$$

During training, we enforce differences between discourse functions by imposing sparsity in the transition matrices associated with each discourse constraint (i.e., we only allow transitions between some of the states for each discourse function). Assume that we have K constraints, where the kth constraint has N_k states. We independently initialize N_k states for constraint k with $k \in \{1, \ldots, K\}$, creating mutually exclusive states for each constraint. Some of these states are very similar, therefore, we merge them to remove their redundancy. Those states become shared between constraints. We use the KLD between the Gaussian distributions attributed to each state as the criterion to merge states. We empirically set a threshold equal to 1, and merge states for which their KLD is less than this threshold. The proposed strategy sparsely represents the configurations of the hidden states associated with each discourse function, where some of the states are shared between two or more discourse functions and some are exclusively associated with one discourse function. Making a sparse state space representation gives the model the capability of learning the characteristics of each category, while coping with unbalanced or limited data. The discourse function node is a discrete variable with five options: *affirmation, negation, question, backchannel*, and *other*. The model also scales when more discourse functions are needed.

6.3. Evaluation and results

We perform a similar eight-fold cross-validation, using six folds for training, one fold for validation and one fold for testing. The validation set is used to determine the number of states per discourse category ($H_{h\&s}$), for each fold. We perform a grid search evaluating from three to 22 states per discourse category, following the same strategy discussed in Section 4.7. We select the number of states that gives high global and local CCA_{hs}, and low KLD. On average, the number of states chosen per discourse function is 13.55 (STD = 4.83), and the total number of states is 65.25 (STD = 7.81).

Since this model uses more states than the DBN in Fig. 2, it may have more transitions requiring further smoothing. Therefore, we

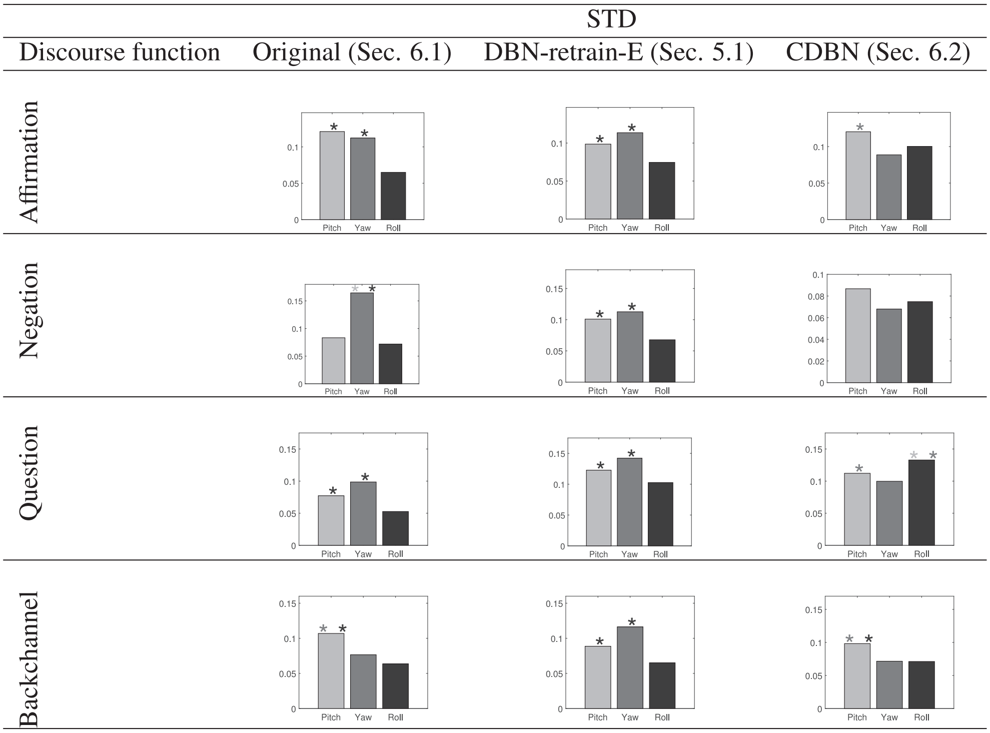


Fig. 7. The average of the standard deviation for each discourse function for the original head movements (Section 6.1), and the synthesized sequences with DBN-retrain-E (Section 5.1) and CDBN (Section 6.2).

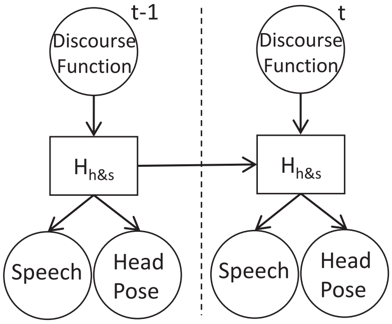


Fig. 8. The *constrained DBN* (CDBN). It introduces explicit dependencies on the discourse functions, creating meaningful behaviors.

reduce the number of key points to 10 fps (Section 4.6 provides the details of the smoothing approach).

The objective metrics for the CDBN over the test set are: local $CCA_h = 0.7836$, local $CCA_{hs} = 0.8480$, global $CCA_h = 0.1304$, global $CCA_{hs} = 0.5452$, and KLD $= 6.7641$. Compared to the unconstrained model, i.e. DBN-retrain-E (Table 1), the results show improvements for local CCA_h ($p = 3.57 \times e^{-41}$), global CCA_h, global CCA_{hs} and KLD. The results are similar for local CCA_{hs}, where we do not observe significant differences.

We evaluate whether the CDBNs are successful in capturing the characteristic patterns associated with the discourse functions. We analyze the mean of the head movements generated by the models DBN-retrain-E and CDBN (Fig. 6). The results show that the average values of the generated behaviors are closer to the original values when we use the CDBN. The results are particularly clear for *affirmation, question* and *backchannel* showing the benefits of constraining the model (the average absolute distance to the mean values of the original data is 0.0855 for DBN-retrain-E, and 0.038 for CDBN).

We also evaluate the standard deviation of the movements generated by the models. Fig. 7 shows the average standard deviation estimated over 200 ms segments. For the DBN-retrain-E models, all the conditions show statistical differences using the Kruskal–Wallis test. For the CDBN, we only observe statistical differences for *affirmation, question* and *backchannel*. The pairwise comparisons of the results are indicated with a color-coded asterisk on top of the bars. The patterns generated by the DBN-retrain-E are indistinguishable across all the discourse functions, highlighting the need for constraining the models on the discourse function. However, the head movements synthesized by CDBN show distinct patterns for movement along pitch, yaw, and roll angles across the discourse functions. For *affirmation* and *backchannel*, the patterns are closer to the original recordings. For other discourse functions, there is room for improvements. Notice that the models generate the expected trajectory given the discourse function and speech patterns. Therefore, we expected that the CDBN captures the average trajectory patterns, which we achieved as displayed on Fig. 6. An open question is to identify strategies that will also preserve the second order statistics in the trajectories.

6.4. Subjective evaluations

We also evaluate the results of the constrained models using subjective evaluations. We wrote five dyadic scenarios for each discourse function, resulting in 20 scenarios. Each scenario consists of a dialog of five turns (to have enough context) between two speakers. The utterances of one speaker belong to one of the target discourse function (e.g., “ A: I am going. B: Where are you going? A: I am going to see Dave. B: Why are you going to see him?” for question). We synthesize the dialog using MaryTTS, and we use the synthetic speech to drive the CDBN and DBN-retrain-E models. The only difference between the animations was the head motion. We used AMT for the perceptual evaluations with a similar protocol as the one used in Section 5.4. This time, we asked two questions for each video: “how natural are the movements of the avatar?”, and “how appropriate are the movements of the avatar with respect to the context of the dialog?” The questionnaires include 5 point

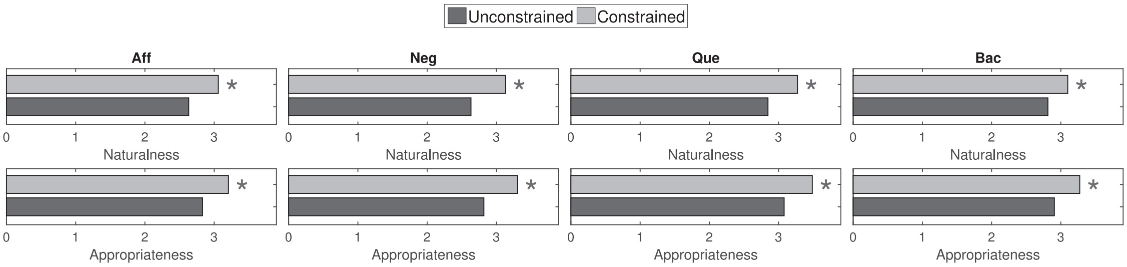


Fig. 9. The average of the ratings given by the raters for naturalness and and appropriateness of the movements. The asterisks denote statistically higher values for the bar ($p < 0.05$).

likert-like scales, where 1 corresponds to low naturalness and appropriateness and 5 corresponds to high naturalness and appropriateness. We recruited 30 evaluators who watched side-by-side the 20 videos generated by CDBN and DBN-retrain-E (40 videos total, each with 30 evaluations). They answered the questions after watching the entire videos. The Cronbach's alpha between the evaluators is $\alpha = 0.77$.

We separately compare the naturalness and appropriateness ratings assigned to the videos. Fig. 9 gives the average ratings for the constrained and unconstrained model for each discourse category. The Kolmogorov–Smirnov test on the ratings demonstrates that the ratings are not Gaussian. Hence, we use the Kruskal–Wallis test for each discourse category to compare the distribution of the ratings. Across discourse functions, the test shows that the distributions of ratings for the constrained model are different from the unconstrained models for both naturalness and appropriateness. The differences are all statistically significant ($p < 0.05$). These results are consistent with the findings observed in the objective evaluations. People perceived the movements generated with the constrained model as more natural and appropriate than the unconstrained model.

7. Conclusions

With the increased use of CAs, it is important to automatize the process of creating natural and appropriate head movements from text. This paper offered solutions to overcome two challenges associated with designing models to capture the relationship between text and head movements, in a speech-driven framework. These problems are lack of generalization due to the need for speech, and lack of meaning in the behaviors. The first problem limits the application of speech-driven models. We proposed to use synthetic speech during test. To avoid the mismatch of training the models with natural speech and testing the models with synthetic speech, we proposed to create a parallel corpus of synthetic speech that is timely aligned with the original speech, and, therefore, with the head motion sequence from the recordings. We used this corpus to retrain the models from scratch, or to adapt the existing models trained with the original speech recordings. We implemented these ideas with neutral and emotional synthetic speech. The emotional fluctuations on the synthetic speech increases the coupling between speech and head movements creating more natural behaviors that are timely synchronized with the input speech. Objective and subjective metrics indicate that the best performance is obtained by training the models from scratch using emotional parallel corpus (DBN-retrain-E).

The second problem addressed in this study is the lack of meaning in the behaviors created by speech-driven models, which may be timely coupled with the speech but convey contradicting meaning. We addressed this problem by constraining the models based on discourse functions, bridging the gap between rule-based and speech-driven systems. The method uses a classifier, which assigns dialog act labels to the sentences based on the transcriptions. We grouped the dialog acts labels into four discourse function classes (affirmation, negation, question and backchannel). The analysis of the head motion sequences in terms of these discourse functions shows statistical differences that have to be considered in the generation of head motion sequences. We accomplished this goal by adding an extra variable that explicitly introduced dependencies on the underlying discourse function. Comparing the synthesized head movement sequences with and without the discourse function constraints shows that the proposed models create behaviors with statistical patterns that are closer to the original head motion sequences. Perceptual evaluations showed that the constrained models generate head motion sequences that are perceived more natural and appropriate than the ones generated with unconstrained models.

This study opens several research directions. Heylen et al. (2006) surveyed studies analyzing the role of head motion during human conversation, listing 25 different roles, including dictating turn-management, enhancing communicative attention, marking contrast between sentences, and communicating the degree of understanding. While this study explored the relationship between head motion and prosodic features as a function of discourse functions, there are many other important head motion functions that we do not consider. Likewise, head motion depends on factors such as personality, age, culture, and gender. Fortunately, the formulation proposed in this study is flexible and can be easily adapted to incorporate these variables. The key challenge in adding these variables is to collect appropriate data for the task. Another extension of this work is to generate head motion sequences when the CA is listening to the user. This is an interesting problem to increase the rapport between the CA and the user (Gratch et al., 2006).

One of the limitations associated with DBNs is state level representation. Even if we increase the number of states, the generated head motion sequences are discontinuous and require post-smoothing techniques. A solution is to use more powerful models, such as the recently popular BLSTM, which have shown to be successful in modeling continuous data, such as head movements (Lan et al., 2016; Ding et al., 2015b; Haag and Shimodaira, 2016), lip movements (Fan et al., 2015), and facial movements (Li et al., 2016). Another limitation of the proposed models is that they are not online, i.e. they require the entire utterance to synthesize the head movement. Using a forward pass during inference, instead of the forward-backward pass, can achieve such a goal. This paper uses MaryTTS for generating speech from text, since it provides all the features required by this study (emotional speech, phonetic alignment), and it is open source. However, using a higher quality TTS can increase the acoustic fluctuations achieving more natural head motion sequences.

Acknowledgments

This work was funded by US National Science Foundation grants IIS-1352950 and IIS-1718944. The authors thank Yandi Xia for his contribution to the dialog act recognizer.

References

Burmania, A., Parthasarathy, S., Busso, C., 2016. Increasing the reliability of crowdsourcing evaluations using online quality assessment. IEEE Trans. Affect. Comput. 7 (4), 374–388. doi:10.1109/TAFFC.2015.2493525.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP: interactive emotional dyadic motion capture database. J. Lang. Resour. Eval. 42 (4), 335–359. doi:10.1007/s10579-008-9076-6.

Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S., 2007a. Rigid head motion in expressive speech animation: analysis and synthesis. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1075–1086. doi:10.1109/TASL.2006.885910.

Busso, C., Deng, Z., Neumann, U., Narayanan, S.S., 2005. Natural head motion synthesis driven by acoustic prosodic features. Comput. Animat. Virtual Worlds 16 (3–4), 283–290. doi:10.1002/cav.80.

Busso, C., Deng, Z., Neumann, U., Narayanan, S.S., 2007b. Learning expressive human-like head motion sequences from speech. In: Deng, Z., Neumann, U. (Eds.), Data-Driven 3D Facial Animations. Springer-Verlag London Ltd, Surrey, United Kingdom, pp. 113–131. doi:10.1007/978-1-84628-907-1_6.

Busso, C., Narayanan, S., 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2331–2347. doi:10.1109/TASL.2007.905145.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., Yan, H., 1999. Embodiment in conversational interfaces: Rea. In: International Conference on Human Factors in Computing Systems (CHI-99). Pittsburgh, PA, USA, pp. 520–527.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Bechet, T., Douville, B., Prevost, S., Stone, M., 1994. 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, pp. 413–420.

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

Chiu, C.-C., Marsella, S., 2011. How to train your avatar: a data driven approach to gesture generation. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K. (Eds.), Intelligent Virtual Agents. In: Lecture Notes in Computer Science, Vol. 6895. Springer Berlin Heidelberg, Reykjavik, Iceland, pp. 127–140. doi:10.1007/978-3-642-23974-8_14.

Chuang, E., Bregler, C., 2005. Mood swings: expressive speech animation. ACM Trans. Graph. 24 (2), 331–347. doi:10.1145/1061347.1061355.

DeCarlo, D., Stone, M., Revilla, C., Venditti, J.J., 2004. Specifying and animating facial signals for discourse in embodied conversational agents. Comput. Animat. Virtual Worlds 15 (1), 27–38. doi:10.1002/cav.5.

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

Ding, C., Xie, L., Zhu, P., 2015a. Head motion synthesis from speech using deep neural networks. Multimed. Tools Appl. 74 (22), 9871–9888. doi:10.1007/s11042-014-2156-2.

Ding, C., Zhu, P., Xie, L., 2015b. BLSTM neural networks for speech driven head motion synthesis. In: Interspeech 2015. Dresden, Germany, pp. 3345–3349.

Ding, C., Zhu, P., Xie, L., Jiang, D., Fu, Z.-H., 2014. Speech-driven head motion synthesis using neural networks. In: Interspeech Singapore, pp. 2303–2307.

Ding, Y., Pelachaud, C., Artieres, T., 2013. Modeling multimodal behaviors from speech prosody. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (Eds.), International Conference on Intelligent Virtual Agents (IVA 2013). Lecture Notes in Computer Science, Vol. 8108. Springer Berlin Heidelberg, Edinburgh, UK, pp. 198–207. doi:10.1007/978-3-642-40415-3_19.

Fan, B., Wang, L., Soong, F. K., Xie, L., 2015. Photo-real talking head with deep bidirectional LSTM. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015). Brisbane, Australia, pp. 4884–4888. doi:10.1109/ICASSP.2015.7178899.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Trans. Speech Aud. Process. 2 (2), 291–298. doi:10.1109/89.279278.

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

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.P., 2006. Virtual rapport. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (Eds.), International Conference on Intelligent Virtual Agents (IVA 2006). Lecture Notes in Computer Science, Vol. 4133. Springer-Verlag Berlin Heidelberg, Marina Del Rey, CA, USA, pp. 14–27.

Haag, K., Shimodaira, H., 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (Eds.), International Conference on Intelligent Virtual Agents (IVA 2016). Lecture Notes in Computer Science, Vol. 10011. Springer Berlin Heidelberg, Los Angeles, CA, USA, pp. 198–207. doi:10.1007/978-3-319-47665-0_18.

Heylen, D., Reidsma, D., Ordelman, R., 2006. Annotating state of mind in meeting data. In: First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006)). Genoa, Italy, pp. 84–87.

Hofer, G., Shimodaira, H., 2007. Automatic Head Motion Prediction from Speech Data. In: Eighth Annual Conference of the International Speech Communication Association. Antwerp, Belgium, pp. 758–761.

Jurafsky D., Shriberg L., Biasca D., 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. https://web.stanford.edu/~jurafsky/ws97/manual.august1.html.

Kuratate, T., Munhall, K. G., Rubin, P. E., Vatikiotis-Bateson, E., Yehia, H., 1999. Audio-visual synthesis of talking faces from speech production correlates. In: Sixth European Conference on Speech Communication and Technology, Eurospeech 1999. Budapest, Hungary, pp. 1279–1282.

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

Le, B.H., Ma, X., Deng, Z., 2012. Live speech driven head-and-eye motion generators. IEEE Trans. Vis. Comput. Graph. 18 (11), 1902–1914. doi:10.1109/TVCG.2012.74.

Levine, S., Krähenbühl, P., Thrun, S., Koltun, V., 2010. Gesture controllers. ACM Trans. Graph. 29 (4), 124:1–124:11. doi:10.1145/1778765.1778861.

Levine, S., Theobalt, C., Koltun, V., 2009. Real-time prosody-driven synthesis of body language. ACM Trans. Graph. 28 (5), 172:1–172:10. doi:10.1145/1618452.1618518.

Li, X., Wu, Z., Meng, H., Jia, J., Lou, X., Cai, L., 2016. Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data. In: Interspeech 2016. San Francisco, CA, USA, pp. 1477–1481. doi:10.21437/Interspeech.2016-364.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. 28 (1), 84–95.

Lotfian, R., Busso, C., 2015. Emotion recognition using synthetic speech as neutral reference. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015). Brisbane, Australia, pp. 4759–4763. doi:10.1109/ICASSP.2015.7178874.

Mariooryad, S., Busso, C., 2012a. Factorizing speaker, lexical and emotional variabilities observed in facial expressions. In: IEEE International Conference on Image Processing (ICIP 2012). Orlando, FL, USA, pp. 2605–2608. doi:10.1109/ICIP.2012.6467432.

Mariooryad, S., Busso, C., 2012b. Generating human-like behaviors using joint, speech-driven models for conversational agents. IEEE Trans. Aud. Speech Lang. Process. 20 (8), 2329–2340. doi:10.1109/TASL.2012.2201476.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., Shapiro, A., 2013. Virtual character performance from speech. In: ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013). Anaheim, CA, USA, pp. 25–35. 10.1145/2485895.2485900.

Mary TTS, 2017. Mary TTS, Version: 5.1.2, http://mary.dfki.de/, Retrieved August 3rd, 2017.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9 (5–6), 453–467.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. 15 (2), 133–137.

Murphy, K., 2002. Dynamic Bayesian Networks: Representation, Inference and Learning Ph.D. thesis. University of California Berkely.

Murphy, K.P., 2007. Conjugate Bayesian analysis of the Gaussian distribution. Technical Report. University of British Columbia. https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.

Pelachaud, C., Badler, N., Steedman, M., 1996. Generating facial expressions for speech. Cogn. Sci. 20 (1), 1–46.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., de Carolis, B., 2005. Greta. a believable embodied conversational agent. In: Stock, O., Zancanaro, M. (Eds.), Multimodal Intelligent Information Presentation. Springer Netherlands, Dordrecht, The Netherlands, pp. 3–25. doi:10.1007/1-4020-3051-7_1.

Qian, X., 2010. Pocket CRF. https://sourceforge.net/projects/pocket-crf-1/. Max margin Markov networks package for sequence labeling tasks.

Ribeiro, E., Ribeiro, R., de Matos, D. M., 2015. The influence of context on dialogue act recognition. arXiv: https://arxiv.org/abs/1506.00839.

Sadoughi, N., Busso, C., 2016. Head motion generation with synthetic speech: a data driven approach. In: Interspeech 2016. San Francisco, CA, USA, pp. 52–56. doi:10.21437/Interspeech.2016-419.

Sadoughi, N., Busso, C., 2017. Speech-driven animation with meaningful behaviors. IEEE Transactions on Audio, Speech and Language Processing Under review.

Sadoughi, N., Liu, Y., Busso, C., 2014. Speech-driven animation constrained by appropriate discourse functions. In: International Conference on Multimodal Interaction (ICMI 2014). Istanbul, Turkey, pp. 148–155. doi:10.1145/2663204.2663252.

Sadoughi, N., Liu, Y., Busso, C., 2015. 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. pp. 1–6. doi:10.1109/FG.2015.7284885.

Sargin, M. E., Erzin, E., Yemez, Y., Tekalp, A. M., Erdem, A. T., Erdem, C., Ozkan, M., 2007. Prosody-driven head-gesture animation. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007). Vol. 2. Honolulu, HI, USA, pp. 672–680.

Thiebaux, M., Marsella, S., Marshall, A. N., Kallmann, M., 2008. Smartbody: behavior realization for embodied conversational agents. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems. Vol. 1. Estoril, Portugal, pp. 151–158.

van Welbergen, H., Ding, Y., Sattler, K., Pelachaud, C., Kopp, S., 2015. Real-time visual prosody for interactive virtual agents. In: Brinkman, W.-P., Broekens, J., Heylen, D. (Eds.), Intelligent Virtual Agents. In: Lecture Notes in Computer Science, Vol. 9238. Springer Berlin Heidelberg, Delft, The Netherlands, pp. 139–151. doi:10.1007/978-3-319-21996-7_16.