NOVEL REALIZATIONS OF SPEECH-DRIVEN HEAD MOVEMENTS WITH GENERATIVE ADVERSARIAL NETWORKS

Najmeh Sadoughi and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

nxs137130@utdallas.edu, busso@utdallas.edu

ABSTRACT

Head movement is an integral part of face-to-face communications. It is important to investigate methodologies to generate naturalistic movements for *conversational agents* (CAs). The predominant method for head movement generation is using rules based on the meaning of the message. However, the variations of head movements by these methods are bounded by the predefined dictionary of gestures. Speech-driven methods offer an alternative approach, learning the relationship between speech and head movements from real recordings. However, previous studies do not generate novel realizations for a repeated speech signal. Conditional *generative adversarial network* (GAN) provides a framework to generate multiple realizations of head movements for each speech segment by sampling from a conditioned distribution. We build a conditional GAN with *bidirectional long-short term memory* (BLSTM), which is suitable for capturing the long-short term dependencies of time-continuous signals. This model learns the distribution of head movements conditioned on speech prosodic features. We compare this model with a *dynamic Bayesian network* (DBN) and BLSTM models optimized to reduce *mean squared error* (MSE) or to increase concordance correlation. The objective evaluations and subjective evaluations of the results showed better performance for the conditional GAN model compared with these baseline systems.

Index Terms— Head movements synthesis; speech-driven animation; conditional generative adversarial networks.

1. INTRODUCTION

Head movement plays various roles during face-to-face conversations [1, 2]. People use head movements to manage turn taking, signal contrast, emphasize their mood, show hesitation, and communicate backchannels. Previous studies have demonstrated that the inclusion of head movements in *conversational agents* (CAs) increases speech intelligibility [3], enhances the level of warmth and competence of the CA [4], and improves the perceived naturalness of the CA [5]. Therefore, it is important to study frameworks that can help in generating more convincing movements for CAs.

The predominant approach to synthesize head movements is using rule-based frameworks [6,7]. These methods rely on a predefined dictionary of gestures. They define mappings between the communicative goals of the utterance and gestures, which are usually derived from the outcomes of psychological studies. The problem with these methods is that the variations in these methods are limited to the predefined dictionary of the gestures in the system. Hence, the movements may look repetitive after some time. Furthermore, the temporal synchronization between speech and gestures also needs to be specified in the system.

This work was funded by NSF award IIS-1718944.

More than 90% of human gestures occur while speaking [8]. Speech and head movements are combined in a non-trivial manner, externalizing the speakers thought, emotions, and intents. Human recordings of speech and head movements have revealed a strong coupling between these two modalities, showing co-occurrence of head movements with speech prosodic patterns [3, 9, 10]. Due to the high level of synchrony between speech prosodic features and head movements, speech-driven approaches [2, 5, 11–18] are proposed as an alternative to rule-based systems. These methods aim to capture the variations of head movements in real recordings, learning the synchrony between speech and head movements from data.

Most of the previous speech-driven frameworks generate head movements with limited variations or small range of movements. Although previous studies have proposed strategies to improve the range of movements synthesized by the model by incorporating other factors such as prototypical gestures [18], or discourse functions [19], still the common architectures which are supposed to capture the beat-like gestures [8] lack the capability to model the range of movements seen in real recordings.

We propose a model based on *generative adversarial networks* (GANs) [20]. GANs learn the distribution of the data, increasing the range of beat-like head movements derived by the model. We propose to use a conditional GAN, which is conditioned on speech prosodic features. This model can generate multiple novel realizations of head movements for an input speech signal by sampling from the conditional distribution of the data. The objective and subjective evaluations demonstrate the benefit of the proposed model.

2. RELATED WORK

There are several recent studies on speech-driven head movements with *deep neural networks* (DNNs). For instance, Ding et al. [13] explored DNNs implemented with only fully-connected layers, only *bidirectional long-short term memory* (BLSTM) units, and a hybrid approach that is built with fully connected layers and BLSTMs. These models mapped filter bank features of speech to head movements, optimized to minimize the *sum of squared errors* (SSE) between the predictions and original head movements. Their results demonstrated better performance when using only BLSTM model compared with only fully connected DNN. They achieved their best performance by the hybrid approach. Haag et al. [14] proposed to extract bottleneck features extracted with a feed-forward network. They used the bottleneck features along with input speech features in a separate model composed of BLSTM and fully-connected layers, resulting in some improvements. Greenwood et al. [21] built separate deep models for the listening and speaking turns, mapping the filter bank features extracted from speech to head poses. They investigated comparisons between a BLSTM model and a *conditional variational auto-encoder* (CVAE) by illustrating the statistical prop-

erties of the moments for the generated head movements.

Note that most of the previous studies only generate one sequence of head movements per speech signal. However, the mapping between speech prosodic features and beat-like head movements is a one to many problem. We propose to build speech-driven head movement models with conditional GANs, conditioning the model on speech prosodic features. This model learns the conditional distributions of the data. During testing, it generates novel realizations by sampling from this conditional distribution, resulting in many head movements for each speech sentence. The model captures the dynamics of the head movements from the changes in the prosodic features. The noise provided to the model captures the global variation of head pose across time conditioned on the input speech features. During training, we also provide mismatched samples of audio and head movements to the discriminator as extra fake samples. This approach helps the discriminator to learn the underlying coupling between speech features and head movements, which in turn helps the generator synthesize head movements which are coupled with the speech features to better fool the discriminator. Note that we share the model between speaking and listening segments, capturing the transitions between them. We compare the results from our model with the DBN model proposed by Mariooryad and Busso [5], a BLSTM model optimized to reduce the MSE inspired by [13], and a BLSTM model optimized to increase the concordance correlation.

3. RESOURCES AND BACKGROUND

3.1. The IEMOCAP Corpus

For our experiments, we use the *interactive dyadic emotional motion capture* (IEMOCAP) corpus [22], consisting of audio, video, and motion capture data from dyadic interactions between 10 actors. The actors performed script-based and improvisation scenarios, where the scenarios were designed to elicit different emotions. We arbitrarily chose the recordings of the first female subject for our experiments, comprising 14 sessions (1h6m) of dyadic interactions between her and another actor. We divide these sessions into 8 sessions for training, 3 for validation, and 3 for testing.

From the motion capture recordings, we obtain the head rotations for pitch, yaw and roll. From audio, we rely on prosodic features of speech, similar to previous studies on speech-driven head motion synthesis [5, 12, 15, 23]. We extract the fundamental frequency and intensity from speech using Praat over each 40ms window, shifting the window by 16.7ms each time (60 fps). We upsample the prosodic features to match the sampling rate of the motion capture data (120 fps). We provide the first and second derivatives of these features to the models. Furthermore, since we use full sessions of the recordings consisting of listening and speaking segments, we also consider a binary input variable for each frame, where one represents speaking turns, and zero represents listening turns.

3.2. Rendering Head Movements

We rely on Xface [24] for rendering the animations. Xface is an MPEG-4 compatible rendering tool. Most facial markers in the IEMOCAP corpus follow the MPEG-4 standard. We follow the same approach used by Mariooryad and Busso [5] to map the markers to *facial action parameters* (FAPs). When generating the animated clips, all the facial markers are rendered with the original motion capture recordings, replacing only the head movements generated by each model (conditional GAN or baselines).

4. METHODOLOGY

4.1. Bidirectional Long-short Term Memory (BLSTMs)

Speech and head movements are time continuous signals. To capture the temporal and cross modality dependencies, we use *recurrent neural networks* (RNNs). RNNs consider weights between consecutive frames, tied at all frames, which makes their learning tractable. However, vanilla RNNs suffer from exploding or vanishing gradients during training [25]. Hence, alternative versions of RNNs with gating mechanisms such as LSTMs [25] have been proposed to handle these issues. Each LSTM node is associated with a cell variable to keep track of the history. Our model entails three gates: input, forget and output. The input gate uses the previous hidden state (in time) and the input values to determine how much of the input is stored in the cell. The forget gate uses the previous hidden state and input values to determine how much of the current content in the cell has to be forgotten. The output gate uses the previous hidden state, the current cell content, and the input value to determine the output value of the node.

We rely on bidirectional LSTMs, which use previous and future information. In practice this model can be used in real time with a short delay. However, we assume the whole speech signal is given for our experiments. LSTMs can help the model to learn the long and short term dependencies between speech and head movements.

4.2. Generative Adversarial Networks (GANs)

GANs were proposed to learn the distribution of the data using adversarial training [20]. GAN is composed of a generator, and a discriminator. The generator generates data by sampling from a noise distribution (z). The samples from the original data are called real samples (labeled as 1), and the samples generated from the generator are called fake samples (labeled as 0). The generator and the discriminator play a minimax game, where the role of the discriminator is to distinguish between the real and fake samples, and the role of the generator is to fool the discriminator (i.e., generate samples close to the original). To achieve this goal during training, the weights of the generator are frozen and the discriminator weights change to reduce the binary cross entropy loss function to distinguish between the fake and real samples. Then, the discriminator weights are frozen and the generator changes its weights to maximize the cross entropy on the fake samples. This is achieved by changing the labels assigned to the fake samples to one.

4.3. Conditional GAN

Our proposed model is a conditional GAN (Fig. 1). Both the discriminator and the generator are conditioned on the speech features. The generator captures the distribution of head movements conditioned on prosodic features. During testing, we can sample from this conditional distribution by selecting different noise values, generating multiple novel realizations of head movements for one speech signal.

As aforementioned in Section 4.2, the discriminator is trained to distinguish the samples generated by the generator (fake) from the original samples (real). The generator needs to capture the temporal and cross-modality dependencies to fool the discriminator. Therefore, the discriminator has to distinguish two types of fake samples: head pose sequences which do not look realistic, and head pose sequences which do not match the speech features. To account for the first fake samples, we generate samples by sampling from the noise

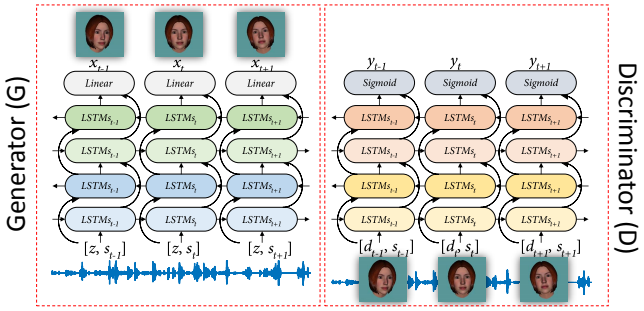

Fig. 1. Conditional GAN used for head movement synthesis, where s_t represents speech features at time t, z represents noise sample, x_t represents the output of the generator, d_t represents head pose at time t and y_t represents prediction by the discriminator at time t.

distribution as is commonly done on GAN. To account for the second fake samples during training the discriminator, we add another type of fake samples inspired by the matching-aware discriminator in the conditional GAN proposed by Reed et al. [26] for text-to-image synthesis. These new sets of fake samples are mismatched segments of audio and visual sequences selected from the original data. Although without using this error the discriminator can learn the couplings between audio and head movements, adding this aspect in the loss function is very helpful to expedite the learning.

Note that in this model the generator is composed of two BLSTM layers with 128 nodes, and a linear output layer tied across the time frames. The discriminator has two layers of BLSTMs with 128 nodes, and a sigmoid output layer tied across time frames. The generator uses the same noise sample across the entire sequence to avoid introducing discontinuities in the generated sequence. The dynamics of the sequence is learned from the time varying speech features provided at each frame. Note that the generator has to learn not only realistic static head poses, but also it has to generate sequences with realistic dynamics. Hence, the discriminator which is supposed to guide the generator towards the data distribution has to learn these aspects. If we only consider the labels at the final frame of the sequence, it may be harder to correct the static errors of the rest of the frames. Furthermore, although LSTMs are supposed to capture long term dependencies, this feature is less effective as the length of the sequence gets very large. Due to all these reasons, we used fake/real labels for all the frames, facilitating the learning of static and dynamic long and short term errors. Our experiments also demonstrate that this approach expedites the learning by the model.

5. EXPERIMENTS

5.1. Baselines

Dynamic Bayesian Network (DBN): We use the DBN, proposed by Marioryad and Busso [5] as one of our baseline models. This model uses two sets of input variable: speech features (*Speech*) and head poses (*Head*). The model uses discrete hidden states to capture the coupling between these two variables. The transitions between the states follow the Markov property of order one. Each of these variables are continuous. Hence, they are modeled with Gaussian distributions with full covariance matrices. During testing, we have partial evidence (only *Speech*), which propagates through the network providing the expected values of *Head* at each frame. They optimized this model on the IEMOCAP corpus, selecting six hidden states for the head model. Their model is optimized only on the speaking turns. However, we use this model for speaking and listening segments. Hence, we add a new node to the model, representing the talking or listening segments, modeled by a binary distribution (i.e., S/L). Figure 2 illustrates this model. We train the model using

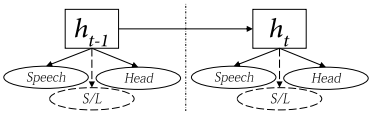

Fig. 2. The DBN model proposed by Marioryad and Busso [5], with an extra input variable S/L.

the *expectation maximization* (EM) algorithm to maximize the log-likelihood of the model on the entire sequence, using the popular forward-backward algorithm. We train the model with 10 iterations.

BLSTM-MSE: Our second baseline model is inspired by the study conducted by Ding et al. [13]. This model is composed of two layers of BLSTMs with 128 nodes for each path, and a linear output layer, in which the weights are tied across the frames. The model is optimized to reduce the MSE between the predicted and original data.

BLSTM-CC: Our third baseline model (BLSTM-CC) has the same structure as BLSTM-MSE. However, the objective function is 1-ρ_c [27]. ρ_c is the *concordance correlation* (CC) between the predictions and the original head poses.

5.2. Experimental Evaluation

We implemented BLSTM models with Keras. We initialize the weights of all the nodes, using the method of Glorot et al. [28]. To avoid overfitting, we use a dropout of 0.20 over all the layers. For optimization, we rely on *adaptive moment estimation* (ADAM), which utilizes the history of the gradient in terms of its first and second moments, making the learning rate more robust during different epochs. We select a learning rate of 0.0001, after experimenting with different values ($[0.1 - 0.0001]$), and observing the changes of loss on the validation set. As our batch size, we use one sample at a time, where the length of the sequence is 1,024 (8.5s). We experimented with shorter sequences ($\sim \{100, 200, 400\}$). However, we found out there are longer term dependencies between head pose frames, which are not captured by the model if the sequence length is short. We noticed 1,024 frames (8.5s) gives reasonable performance on the validation set.

We consider an m dimensional zero mean Gaussian noise with an identity covariance matrix as the noise for the generator. We explored different noise dimensions $\sim \{1, 10, 100\}$, where 10 gave the best result on the validation set. We noticed that pre-training the generator with concordance correlation objective improves the speed of training. Therefore, we pre-train the generator for 10 epochs, and then the discriminator for 5 epochs. Next, we alternate between updating the generator and the discriminator over each batch, training the GAN for 10 epochs. For the baseline BLSTM models, we trained the models for 10 epochs.

We smooth the predicted trajectories following the method proposed by Busso et al. [23]. This method consists of converting the rotations into quaternions, and then selecting 15 key points per second, interpolating the intermediate frames.

5.3. Objective Evaluations

The synthesized results by GAN are usually objectively evaluated by fitting a distribution to the generated samples, and finding the likelihood of the test samples in the distributions [20]. We use the unseen audio samples of the test set, and generate their corresponding outputs by sampling from noise. For this evaluation, we treat each frame of the generated sequences as one sample, resulting in

Table 1. Objective evaluation of the generated head movements in term of log-likelihood (the values in the parentheses are the standard deviations of the metrics, by sampling multiple times from noise).

Type	Model	Log-likelihood	
		MEAN	STD
Baseline	DBN	-121.406	120.976
	BLSTM-MSE	-106.107	113.766
	BLSTM-CC	-38.415	65.410
Proposed	Conditional GAN	**-30.559 (3.2752)**	**48.674 (3.8405)**

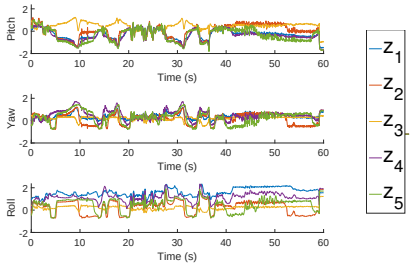

Fig. 3. The figures demonstrate 5 different realizations of three head angles synthesized by sampling from the conditional GAN.

103.7K samples. These samples are then used to fit a Parzen kernel density, where the bandwidth of the kernel is optimized by 3-fold cross validation on the samples. Table 1 gives the average and standard deviation of these values. Since the conditional GAN uses noise as input, we estimate the results for GAN five times by sampling from the noise distribution, reporting the average of the results in terms of the log-likelihood mean and standard deviations. Note that all the pairwise comparisons are statistically different (z-test: p$-value < 0.001$). These results demonstrate the best performance is achieved with our conditional GAN. The results are better for BLSTM compared to DBN. The results clearly demonstrate the benefit of using a concordance correlation-based objective function compared with MSE.

Note that one of the benefits of GAN is generating novel realizations for the same input speech by sampling from noise (z). Figure 3 demonstrates the three head angles synthesized by sampling five times from the conditional GAN, relying on the same input speech but different noise values. Visual inspection shows that the five sequences are reasonable.

5.4. Subjective Evaluations

For subjective evaluation, we randomly select five continuous segments from the test samples with at least 15 seconds of talking duration. The average length of the dialogs is 39.2s. We render the videos with the head movements synthesized by the three baseline models and the conditional GAN. Comparing two videos is usually less susceptible to personal biases than separately rating a stimulus. Therefore, we devise a comparison task consisting of the head movements by conditional GAN with one of the baseline models. For each task, the placement of the videos are randomized. The evaluators are provided with two videos labeled as video 1 and video 2 and the question "Which video has more natural head movements?". To allow the annotators to convey their soft perceptions, we provide multiple choices: 1. "Definitely video 1", 2. "Video 1", 3. "Moderately video 1", 4. "Slightly video 1", 5. "Both look similar", 6. "Slightly video 2", 7. "Moderately video 2", 8. "Video 2", 9. "Definitely video 2".

We use *Amazon mechanical turk* (AMT) to recruit annotators for our evaluations. We only allow the task to be done by annotators who performed well in our previous tasks [2, 29, 30]. This helps avoiding

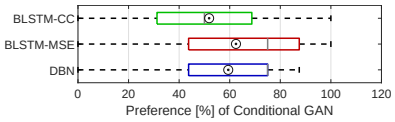

Fig. 4. The figure gives the comparison results between the conditional GAN and the three baseline models.

spammers. Furthermore, the evaluators are shown one pair of videos at a time, where the order is randomized for each person. The question is shown after the evaluator has played the videos, reducing the chance of answering the question before watching the videos.

Each comparison task is composed of evaluating five videos synthesized by the conditional GAN model with the five videos synthesized by one of the baseline models. In total, we have three separate tasks, where we recruit 12 people from AMT for the evaluation (four per task). We replaced three subjects (one per condition) whose average pairwise Cronbach's alpha with the rest of the annotators are less than zero. The Cronbach's alpha values between the annotators are 0.459 (DBN), 0.640 (BLSTM-MSE) and 0.718 (BLSTM-CC). Figure 4 gives the comparison results. On average, people preferred the head movements generated by the conditional GAN model more than 50% over the baselines. We compute the proportion of preferences for the conditional GAN model compared with each of the baseline models. We get the proportions using Equation 1, where a_i is the preference for the i^{th} sample, from n total samples. The proportion of preferences for the conditional GAN compared to the baselines are 0.682 (DBN), 0.737 (BLSTM-MSE) and 0.542 (BSLTM-CC). Proportion test shows significantly higher proportion than 50% for conditional GAN compared with BLSTM-MSE (p$-value < 0.05$). The p-values of this test on the comparisons for DBN and BLSTM-CC are 0.0977 and 0.3860, respectively.

$$p = \frac{Count\,(a_i \geq 50\%)}{[Count\,(a_i \geq 50\%) + Count\,(a_i \leq 50\%)]}, i \in \{1, \cdots, n\} \quad (1)$$

6. CONCLUSIONS & FUTURE WORK

This paper proposed a novel strategy to utilize conditional GAN for head movement synthesis. Beat gestures of head have intrinsic random properties, and conditional GANs provide an appropriate framework to capture them by fitting a conditional distributions to the observed training samples. We propose to condition a GAN model on prosodic features, which are varying from one frame to another, capturing the dynamics of head movements. The input noise for the GAN model captures different variations of head motions under the same prosodic states. After training the model, we generate novel realizations of head movements by sampling from the conditional distribution learned by the model (i.e., prosodic features plus different noise values). The average log-likelihood of the test samples in the fitted distributions of the generated samples by the conditional GAN is higher compared with the baseline models, showing that the model better fits the distribution of the data. The direct comparisons between the conditional GAN model and the three baselines showed higher average preferences for the conditional GAN model. The preference of the proposed model is significantly higher than BLSTM-MSE.

The conditional GAN proposed in this study is shared between speaking and listening segments. Adding more audio-visual features from the interlocutor during the speaking segments may provide more predictive features of the head pose of the listener. Although we propose this model for head movement synthesis, this model can be applied to learn facial movements during conversations.

7. REFERENCES

[1] D. Heylen, "Challenges ahead head movements and other social acts in conversation," in *Artificial Intelligence and Simulation of Behaviour (AISB 2005), Social Presence Cues for Virtual Humanoids Symposium*, Hertfordshire, United Kingdom, April 2005, p. 8.

[2] N. Sadoughi and C. Busso, "Head motion generation," in *Handbook of Human Motion*, B. Müller, S. Wolf, G.-P. Brueggemann, Z. Deng, A. McIntosh, F. Miller, and W. Scott Selbie, Eds. Springer International Publishing, January 2017, pp. 1–25.

[3] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, February 2004.

[4] H. Welbergen, Y. Ding, K. Sattler, C. Pelachaud, and S. Kopp, "Real-time visual prosody for interactive virtual agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, W.-P. Brinkman, J. Broekens, and D. Heylen, Eds. Delft, The Netherlands: Springer Berlin Heidelberg, August 2015, vol. 9238, pp. 139–151.

[5] 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*, vol. 20, no. 8, pp. 2329–2340, October 2012.

[6] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013)*, Anaheim, CA, USA, July 2013, pp. 25–35.

[7] 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)*, Orlando, FL,USA, 1994, pp. 413–420.

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

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

[10] 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*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[11] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, July 2005.

[12] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 124:1–124:11, July 2010.

[13] C. Ding, P. Zhu, and L. Xie, "BLSTM neural networks for speech driven head motion synthesis," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 3345–3349.

[14] K. Haag and H. Shimodaira, "Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis," in *International Conference on Intelligent Virtual Agents (IVA 2016)*, ser. Lecture Notes in Computer Science, D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski, Eds. Los Angeles, CA, USA: Springer Berlin Heidelberg, September 2016, vol. 10011, pp. 198–207.

[15] B. H. Le, X. Ma, and Z. Deng, "Live speech driven head-and-eye motion generators," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 11, pp. 1902–1914, November 2012.

[16] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and E. Erzin, "Multimodal analysis of speech prosody and upper body gestures using hidden semi-Markov models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 3652–3656.

[17] N. Sadoughi, Y. Liu, and C. Busso, "Speech-driven animation constrained by appropriate discourse functions," in *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 148–155.

[18] N. Sadoughi and C. Busso, "Speech-driven animation with meaningful behaviors," *ArXiv e-prints (arXiv:1708.01640)*, vol. abs/1708.01640, pp. 1–13, August 2017.

[19] N. Sadoughi, Y. Liu, and C. Busso, "Meaningful head movements driven by emotional synthetic speech," *Speech Communication*, vol. 95, pp. 87–99, December 2017.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS 2014)*, vol. 27, Montreal, Canada, December 2014, pp. 2672–2680.

[21] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose in dyadic conversation," in *International Conference on Intelligent Virtual Agents (IVA 2017)*, ser. Lecture Notes in Computer Science, J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, Eds. Stockholm, Sweden: Springer Berlin Heidelberg, August 2017, vol. 10498, pp. 160–169.

[22] 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*, vol. 42, no. 4, pp. 335–359, December 2008.

[23] 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*, vol. 15, no. 3, pp. 1075–1086, March 2007.

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

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[26] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning (ICML)*, San Juan, Puerto Rico, May 2016, pp. 1–10.

[27] N. Sadoughi and C. Busso, "Joint learning of speech-driven facial motion with bidirectional long-short term memory," in *International Conference on Intelligent Virtual Agents (IVA 2017)*, ser. Lecture Notes in Computer Science, J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, Eds. Stockholm, Sweden: Springer Berlin Heidelberg, August 2017, vol. 10498, pp. 389–402.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, Sardinia, Italy, May 2010, pp. 249–256.

[29] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5190–5194.

[30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.