INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Head Motion Generation with Synthetic Speech: a Data Driven Approach

Najmeh Sadoughi and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
nxs137130@utdallas.edu, busso@utdallas.edu

Abstract

To have believable head movements for *conversational agents* (CAs), the natural coupling between speech and head movements needs to be preserved, even when the CA uses synthetic speech. To incorporate the relation between speech head movements, studies have learned these couplings from real recordings, where speech is used to derive head movements. However, relying on recorded speech for every sentence that a virtual agent utters constrains the versatility and scalability of the interface, so most practical solutions for CAs use text to speech. While we can generate head motion using rule-based models, the head movements may become repetitive, spanning only a limited range of behaviors. This paper proposes strategies to leverage speech-driven models for head motion generation for cases relying on synthetic speech. The straightforward approach is to drive the speech-based models using synthetic speech, which creates mismatch between the test and train conditions. Instead, we propose to create a parallel corpus of synthetic speech aligned with natural recordings for which we have motion capture recordings. We use this parallel corpus to either retrain or adapt the speech-based models with synthetic speech. Objective and subjective metrics show significant improvements of the proposed approaches over the case with mismatched condition.

Index Terms: Speech driven animation, head movements, text to speech synthesis

1. Introduction

Head motion and speech are two important communicative channels that are tightly coupled to convey the intended message [1, 2]. Head movements change from one person to another, according to their mood, personality, and intended message. The resulting head movement is temporally synchronized with the prosodic and syntactic structure of the speech [3]. This synchrony is not only crucial for natural perception, but also to accomplish other communicative tasks such as increasing the intelligibility of the message [4]. Therefore, it is important to carefully model the relationship between speech and head motion to create believable *conversational agents* (CAs).

To create head movements for CAs, studies have used rule based systems [5, 6] or data driven systems [2, 7]. Rule based methods define several rules for generating head movements based on the content of the message. They choose head movements from a set of hand crafted head movements defined in their system. Data driven approaches usually learn the distribution of the head movements from natural recordings, synthesizing behaviors by sampling from this distribution. Data driven models have the advantage of generating non repetitive behaviors, spanning the range of movements observed during human interactions, which is difficult to achieve with rule based systems [8].

This work was funded by NSF (IIS: 1352950).

Data driven frameworks usually rely on speech prosody to derive head motion for CAs, due to the close relationship between prosodic features and head movements [2, 9]. This speech-driven approach requires natural speech. The range of applications for CAs is growing, so it is not always feasible to have access to natural recorded audio. Instead, these systems use *text-to-speech* (TTS), which provides a flexible and scalable solution. While the use of synthetic speech is not a problem for rule-based systems, as the rules depend on the semantic and syntactic content in the message, it poses a major challenge for speech driven models. This paper proposes strategies to leverage existing speech-driven models for head motion generation for cases relying on synthetic speech.

For speech driven frameworks, the straightforward solution to generate head motion driven by synthetic speech is to use the model trained with natural speech recordings and test it with synthetic speech. However, this approach does not address the differences between synthetic and natural speech, creating a mismatch that may result in adverse effects on the generated head motion. To mitigate this problem, this paper proposes a novel approach for training speech driven models. We generate a parallel corpus where we synthesize the transcription of natural speech, for which we have motion capture recordings. The synthetic speech is then time-aligned so that it is synchronized with the original recordings, and, therefore, with the head motion sequences in the database. Using this parallel corpus, we retrain or adapt the speech-driven models using the aligned synthetic speech. We assess the results using objective and subjective metrics, demonstrating the benefits of the proposed compensation approaches.

2. Related Work

Several studies have reported the high correlation between speech prosody and head movements [1, 4, 10, 11]. At the utterance level, Munhall et al. [4] reported a correlation of $\rho = 0.63$ between head motion and fundamental frequency (F0), and $\rho = 0.324$ between head motion and RMS energy. Busso et al. [11] reported an average of $\rho = 0.7$ canonical correlation between head motion and speech prosodic features (F0, energy and their first and second order derivatives), at the utterance level. Kuratate et al. [10] reported similar results, showing a correlation of $\rho = 0.83$ between head movements and F0. Graf et al. [3] showed that there is a high co-occurrence between the prosodic events and major head movements. The high correlation of speech and head movements leads to enhancement in speech intelligibility [4], as head motion signals syntactic boundaries and stress. To have more believable CAs, the relation between head motion and speech has to be carefully considered.

An interesting approach to preserve the relation between speech and head motion is to use speech-driven models to generate head motions. Busso et al. [2] used *vector quantization* (VQ) to quantize the space of head movements, and used *hid-*

http://dx.doi.org/10.21437/Interspeech.2016-419

den Markov models (HMMs) to model the relationship between speech prosody and head motion during different emotional states. Sargin et al. [12] used *parallel HMMs* (PHMMs) to automatically segment and cluster the joint representation of head movement and speech prosody. Levine et al. [13] used *hidden conditional random fields* (HCRF) to learn the temporal relationship between kinematic features of head movements and speech features. Chiu et al. [14] used *hierarchical restricted Boltzmann machines* (HCRBMs) to learn the next position of head in time using the previous two positions and the speech prosody features. Le et al. [15] proposed to learn a joint model of speech prosody and each of the head motion kinematic features, separately. Assuming that these distributions are independent, they solve a non constrained optimization problem to get the maximum posterior probability at each time frame given the previous two head positions and current speech prosodic features. Mariooryad and Busso [9] designed several *dynamic Bayesian networks* (DBNs) to jointly model head and eyebrow movements with speech prosody features.

When a CA uses synthetic speech, however, almost all the proposed methods correspond to rule-based systems, where modeling the local relationship between speech and head motion is non trivial. To the best of our knowledge, there is only one previous study, where they used speech driven models to generate head movements from synthetic speech. Welbergen et al. [16] used the model proposed by Le et al. [15] and trained the model on 7.4 minutes of the IEMOCAP corpus using natural speech. For synthesis, they derived the models using synthetic speech, and used subjective evaluations to evaluate the generated head movements. However, this approach creates a mismatch between training (natural speech) and testing (synthetic speech) conditions. Our study proposes a novel framework to leverage speech-driven models using synthetic speech, which effectively reduces the training-testing mismatch.

3. Motivation and Resources

3.1. Overview

This study proposes novel approaches to use speech-driven models for head motion generation with synthetic speech. We start with the DBNs proposed by Mariooryad and Busso [9] (Sec. 3.3). The baseline setting (C1) corresponds to the approach proposed by Welbergen et al. [16], where we use the DBNs trained with natural speech which are then tested with synthetic speech. To mitigate this training-testing mismatch in condition C1, we proposed two approaches that require a parallel corpus with synthetic speech locally aligned to natural speech for which we have motion capture data (Sec 4.1). The first approach (C2) consists of training the models from scratch using the time-aligned synthetic signal. We avoid mismatch by training and testing the models with synthetic speech (Sec. 4.2). The second approach (C3) consists of building the models with original, natural speech, which are then adapted to the synthetic speech, reducing the mismatch (Sec. 4.3).

3.2. IEMOCAP Corpus

This study uses the *interactive emotional dyadic motion capture* (IEMOCAP) corpus [17], which is a motion capture database recorded from five dyadic interactions between an actor and an actress. The corpus includes audio, video and motion capture data from 10 actors. The dyadic interaction consists of script based and improvisation scenarios, which are designed to induce a variety of emotions. The motion capture data comprises the facial markers, head, and hand markers. The study in Busso et al. [17] describes the details of the corpus.

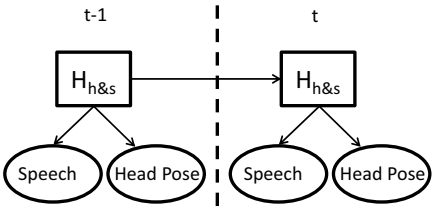

Figure 1: The dynamic Bayesian representation of the model.

This study considers 270.16 minutes of data, where we exclude segments with overlapped speech. From the markers, we consider the three rotation angles describing head poses at 120 fps. From speech, we consider the *fundamental frequency* (F0), intensity, and their first and second order derivatives, following our previous work [2, 9, 18, 19]. We extract F0 and intensity, using Praat over window size of 40ms with 23.3ms overlap (e.g., 60 fps). We increase the sample rate for the acoustic features to 120fps using interpolation. Finally, we normalize the acoustic features using a mean shift normalization per subject, where we normalize the variance by using the global variance derived from the whole data.

3.3. Speech-Driven Animation with DBNs (C1)

This study builds upon the jDBN-3 model proposed by Mariooryad and Busso [9], which is illustrated in Figure 1. This is a generative model, where the continuous node $Speech$ represents speech prosodic features, the continuous node $HeadPose$ represents three head angles, and the discrete node $H_{h\&s}$ represents the joint configuration state between prosodic features and head movements. The continuous variables $Speech$ and $HeadPose$ are modeled with Gaussian distributions, where they have full covariance matrices, but are assumed to be independent of each other. This model is trained by alternating between inferring the probability of hidden states using the forward and backward algorithm, and then updating the parameters of the models to maximize the expected likelihood of the observations. All the variables are available during training. However, during testing the $HeadPose$ variable is considered as missing, and it is estimated by entering the prosodic features as evidence.

Training of this model uses the *expectation maximization* (EM) algorithm, and, therefore, it finds a local optimum. Due to the sensitivity of the EM to initialization, we initialize the states in the model using the *Linde-Buzo-Gray vector quantization* (LBG-VQ) [20]. This approach helps learning the parameters of the model, such that they represent the data better. Therefore, the resulting model tends to generate movements that have higher range of variations.

4. Proposed Framework

4.1. Parallel Corpus with Timely Aligned Synthetic Speech

The goal of the parallel corpus is to retrain or adapt the speech-driven models with synthetic speech. For this purpose, we need synthetic speech that not only has the same lexical content of the original speech in the IEMOCAP database, but also it is time-aligned at the word level. We rely on the same framework used by Lotfian and Busso [21], where the parallel corpus was used as neutral reference models to contrast expressive speech.

The approach starts by creating synthetic speech using the transcriptions of the IEMOCAP sentences. We use *Open Mary* which is an open source *text-to-speech* (TTS) toolkit.We use a male voice and a female voice. Then, we align the synthesized signal matching the word timing of the original signal. We implement this step with Praat, which warps the timing of

the speech signal, while maintaining the fundamental frequency level of the voiced phonemes, using the *pitch synchronous overlap add* (PSOLA) technique [22]. The final step is to replace segments in the synthetic speech that have zeros. These are silence segments, which are problematic when learning our models. We fill these segments with silence recordings collected under similar conditions.

We extract F0, and intensity from the parallel corpus as well. We also perform a mean shift normalization per voice, where the standard deviation is set to match the standard deviation of neutral sentences of the IEMOCAP corpus.

4.2. Retraining the Models with the Parallel Corpus (C2)

Since the synthetic signal conveys the same lexical information as the original speech, and their word content are time- aligned, the synthetic signal is also time-aligned with the motion capture data. This parallel corpus allows us to retrain our models from scratch using synthetic speech. By training and testing the jDBN3 models with synthetic speech, we avoid the mismatch observed in condition C1, which is our baseline (Sec 3.1). This approach is referred to as C2.

A potential problem for the C2 model is that the correlation between the synthetic speech and the original head motion sequences may not be strong enough for our models to learn. The synthetic speech used in this work is emotionally neutral, conveying limited range of variability. Therefore, our DBNs may not fully capture the dependencies between prosodic features and head movements. The second approach proposed in this paper addresses this problem.

4.3. Adapting the Models with the parallel corpus (C3)

Instead of building the jDBN3 from scratch, the second approach adapts the models trained with natural speech. First, we train the jDBN3 models using natural speech, capturing the complex relationship between head motion and speech. Then, we use the parallel corpus to adapt the models, reducing the mismatch between train and test conditions.

We adapt the model originally trained with natural recordings with *maximum a posteriori* (MAP) adaptation using the parallel corpus. The only modality that is different is speech, so the adaptation updates only speech related parameters, which are the mean and covariance matrices of the Gaussian distributions for the *Speech* mode. We use Normal-Wishart prior, since it is the conjugate prior when the likelihood is a Gaussian distribution [23, 24]. Equations 1 and 2 provide the mean and covariance matrices updates, where μ_i is the mean of the i^{th} state, $\bar{x_i}$ is the expectation of the mean for state i using the new observations, n is the weight of new observations for state i, n_p is the weight associated with prior, μ_{pi} is the prior mean of state i, and Σ_{pi} is the prior covariance matrix of state i. Note that we consider the same weights across different states. Since we create the parallel corpus for the entire audio recordings, we consider $\frac{np}{np+n}$ as 0.5. We consider two separate cases, where we adapt only the mean (C3-1), and when we adapt the mean and the covariance matrix (C3-2). We use 80% of the data for training the model, and 20% for finding optimal number of iterations, which is set to four for C3-1, and two for C3-2.

$$\mu_i = \frac{n_p \mu_{pi} + n\bar{x_i}}{n_p + n} \tag{1}$$

$$\Sigma_i = \frac{n_p(\Sigma_{pi} + (\mu_{pi} - \mu_i)(\mu_{pi} - \mu_i)^t) + n\overline{(x_i - \bar{x_i})(x_i - \bar{x_i})^t}}{n_p + n} \tag{2}$$

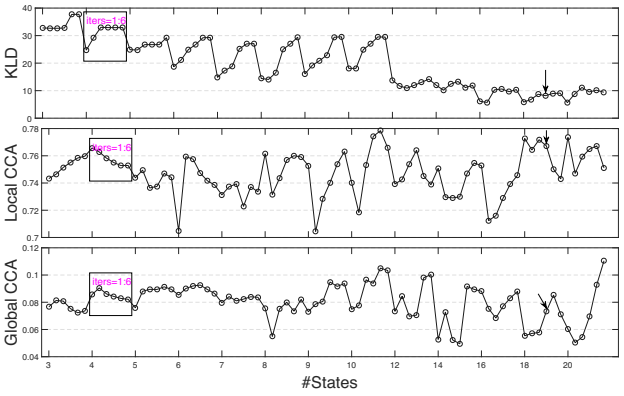

Figure 2: Optimizing the number of states and the number of iterations.

5. Experimental Evaluation

This section describes the implementation of the approach (Sec. 5.1) and the experimental evaluation, which includes objective metrics (Sec. 5.2) and subjective metrics (Sec. 5.3).

5.1. Optimization of the Parameters

An important problem is to set the parameters of the models, which are the number of states of the discrete node $H_{h\&s}$ (Fig. 1), and the number of iterations. We use several metrics to optimize the synthesized head movements generated by the model. First, we estimate *canonical correlation analysis* (CCA) between the original and the synthesized head movements. CCA finds affine transformations for two multivariate sets of temporal data, projecting them into a common space where the correlation of their projections are maximized. We measure CCA at the utterance level (i.e., a set of transformations per sentence), and at the global level (i.e., a single set of transformations across the entire recordings). We estimate the global level CCA by concatenating all the utterances. While estimating CCA over longer sequences reduces the correlation, this global metric indicates whether the model preserves inter-utterance variations. We also use the *Kullback-Leibler divergence* (KLD), between p and q, which are the distributions of head movements for the original and synthesized sequences, respectively. This metrics measures the lost information resulting from using the synthesized movements instead of the original head movements.

We use 80% of the data for training the model, and 20% for finding the optimal parameters. Figure 2 shows the result for KLD, local CCA, and global CCA when we train the models with original speech and head motion data. For each state, the figure shows the performance for six iterations (see highlighted blocks). The analysis reveals that 18 states with 4 iterations yields low KLD value and high CCA values. We select this configuration for the models C1 and C3 (the model that is adapted with the parallel corpus). For C2, we reevaluate this analysis by training the model with synthetic speech, selecting 16 states and 4 iterations.

5.2. Objective Evaluations

We first evaluate the quality of the generated head motion sequences using objective metrics using a five-fold cross-validation framework. As a reference, the turn-based CCA between original head motion and prosodic features is ρ =0.77, which demonstrates the coupling between prosodic features and head motion. Table 1 gives the results, where $CCA_{s\&h}$ is the

Table 1: Objective assessment using CCA and KLD (*: significant difference with the corresponding metric for C1, where $p < 0.05$, †: significant difference with the corresponding metric for C1, where $p < 0.01$ – two tailed t-test).

	turn based		global		
	$CCA_{s\&h}$	CCA_h	$CCA_{s\&h}$	CCA_h	KLD
M1	0.8615	0.7478	0.3286	0.0756	8.4617
C1	0.8103	0.7452	0.2473	0.0787	8.3530
C2	0.7901†	0.6997†	0.3294	0.0412	4.7579
C3-1	0.8399†	0.7514	0.2891	0.0697	8.6299
C3-2	0.8189*	0.7478	0.2475	0.0633	9.3203

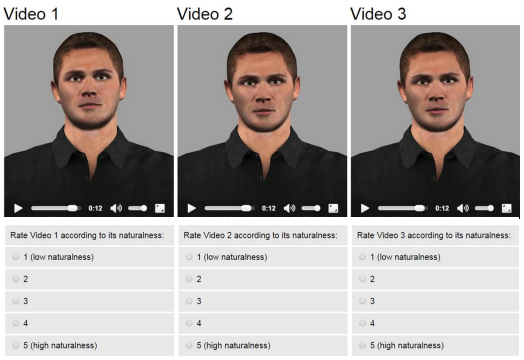

Figure 3: Interface used in subjective evaluation of the synthesized head movements.

CCA between the synthesized head motion sequence and the prosodic features used as input (synthetic or natural), and CCA_h is the CCA between the original and synthesized head motion sequences. We report the CCA at the turn level, and at the global level after concatenating all the sentences. The table also reports the KLD between the synthesized and original head movements.

The first row provides the reference case M1, where the model is trained and tested with natural recordings. The result shows a high local $CCA_{s\&h}$ and CCA_h, showing that the model was successful in capturing audiovisual coupling. The KLD between the original and synthesized head movements is 8.4617, which is explained by the dependency of head motion on other factors such as personality [25], and speech content [26]. For C1, there is a drop in performance compared with M1, due to the mismatch introduced by training the models with natural speech, but testing them with synthetic speech. For C3, there is an improvement for local and global $CCA_{s\&h}$, when adapting the parameters using the synthesized speech (C3-1, C3-2), especially when we only adapt the mean (C3-1). For C2, we observe that the KLD decreases compared to C1.

5.3. Subjective Evaluations

We also evaluate the proposed models using perceptual evaluations. We use Smartbody [27] to visualize the synthesized head movements. Smartbody is an animation rendering tool, which can decode *Biovision Hierarchy* (BVH) files. It can also create lip movements when we provide the phonetic alignment. To limit the number of videos to be evaluated, we consider 20 videos for each of the models achieving the best performance according to the objective evaluation: C1, C2, and C3-1. For each video, we evaluate two consecutive speaking turns of the CA to make the video longer and to incorporate context. When the CA is listening, the states of the models are set to idle, so

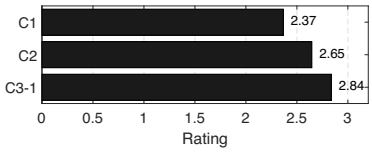

Figure 4: Average naturalness ratings given by the evaluators.

we do not incorporate backchannels. During those segments, we include the original speech of the interlocutor to make the dialog clear. For the perceptual evaluations, we show the evaluators three videos of the same segment, generated with C1, C2, and C3-1.

We used *Amazon mechanical turk* (AMT) for performing the perceptual evaluations. We ask them to rate the videos using a 5 Likert-type scale where 1 means low naturalness and 5 means high naturalness. Figure 3 shows the interface for the evaluation. We only ask each evaluator to complete the evaluation of 10 out of the 20 segments to reduce fatigue (10 segments $\times$ 3 videos = 30 videos). We randomized the order and placement of the videos. We only show the questions after the three videos are thoroughly played, so the evaluators have to watch the videos before answering the questions. Furthermore, we only invite evaluators who have provided reliable rankings in our previous studies on AMT. We recruited 30 evaluators, so each video is independently annotated by 15 raters.

Figure 4 gives the average scores assigned to the C1, C2, and C3-1 models. The perceptual scores indicates that retraining (C2) or adapting (C3-1) the models with the parallel corpus of synthetic speech effectively increases the naturalness perception of the videos, compared to the baseline model (C1). We evaluate the distribution of the rankings assigned to the three groups, concluding that their distributions are not Gaussian. Therefore, we use the Kruskal-Wallis test, which is a non-parametric test to assess whether the differences between the medians of two or more distributions are similar (null hypothesis) or different. The Kruskal-Wallis test shows that these three groups are statistically different ($p < 1.2e^{-6}$). Pairwise comparisons using Tukey's procedure reveals that C1 and C3-1 are statistically different ($p < 7.4e^{-7}$) and C1 and C2 are statistically different ($p < 3.5e^{-3}$). The test does not reveal statistical differences between C2 and C3-1. Reducing the mismatch between test and train conditions by using the proposed parallel corpus that retrains or adapts the models produces more natural head movements.

6. Conclusions

This paper proposed a novel approach to generalize a speech-driven model for head motion generation using synthetic speech. The approach starts by creating a corpus of synthetic speech with time-aligned signals conveying the same lexical content as the natural recordings. This parallel corpus is used to retrain or adapt the model to the synthetic speech (C2, and C3). This approach reduces the mismatch that is present when synthetic speech is directly used to drive the models originally trained with natural speech (the approach that serves as our baseline model, C1). Both objective and subjective evaluations demonstrate the benefits of using the proposed approaches, resulting in more natural head motion sequences.

Our future work includes adding emotional behaviors into our models. We are also considering including other facial gestures (e.g., eyebrow motion) and hand gestures. We are also working on constraining the generated behaviors on the underlying discourse function of the message. This will allow us to generate data-driven behaviors with meanings.

7. References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, October 2005.

[8] M. E. Foster, "Comparing rule-based and data-driven selection of facial displays," in *Workshop on Embodied Language Processing, Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 1–8.

[9] 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.

[10] 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.

[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] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, August 2008.

[13] 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.

[14] C.-C. Chiu and S. Marsella, "How to train your avatar: A data driven approach to gesture generation," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, Eds. Reykjavik, Iceland: Springer Berlin Heidelberg, September 2011, vol. 6895, pp. 127–140.

[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] 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.

[17] 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.

[18] 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.

[19] N. Sadoughi and C. Busso, "Retrieving target gestures toward speech driven animation with meaningful behaviors," in *International conference on Multimodal interaction (ICMI 2015)*, Seattle, WA, USA, November 2015, pp. 115–122.

[20] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan 1980.

[21] R. Lotfian and C. Busso, "Emotion recognition using synthetic speech as neutral reference," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 4759–4763.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, December 1990.

[23] K. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," University of British Columbia, Technical Report, October 2007. [Online]. Available: https://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf

[24] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[25] A. B. Youssef, H. Shimodaira, and D. A. Braude, "Head motion analysis and synthesis over different tasks," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, Eds. Edinburgh, UK: Springer Berlin Heidelberg, October 2013, vol. 8108, pp. 285–294.

[26] J. Lee and S. Marsella, "Nonverbal behavior generator for embodied conversational agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, Eds. Marina del Rey, CA, USA: Springer Berlin Heidelberg, October 2006, vol. 4133, pp. 243–255.

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