Predicting Head Pose in Dyadic Conversation

David Greenwood, Stephen Laycock, and Iain Matthews

University of East Anglia, Norwich, United Kingdom
david.greenwood@uea.ac.uk
s.laycock@uea.ac.uk
iain.matthews@uea.ac.uk

Abstract. Natural movement plays a significant role in realistic speech animation. Numerous studies have demonstrated the contribution visual cues make to the degree we, as human observers, find an animation acceptable. Rigid head motion is one visual mode that universally co-occurs with speech, and so it is a reasonable strategy to seek features from the speech mode to predict the head pose. Several previous authors have shown that prediction is possible, but experiments are typically confined to rigidly produced dialogue.
Expressive, emotive and prosodic speech exhibit motion patterns that are far more difficult to predict with considerable variation in expected head pose. People involved in dyadic conversation adapt speech and head motion in response to the others' speech and head motion. Using Deep Bi-Directional Long Short Term Memory (BLSTM) neural networks, we demonstrate that it is possible to predict not just the head motion of the speaker, but also the head motion of the listener from the speech signal.

Keywords: speech animation, head motion synthesis, visual prosody, dyadic conversation, generative models, BLSTM, CVAE

1 Introduction

Speech animation involves transforming and deforming a character model, temporally synchronised to an audible utterance to give the appearance that the model is speaking. Given the close relationship between speech and gesture, the problem is challenging, as human viewers are very sensitive to natural human movement [24]. Practical applications of speech animation, for example computer games and animated films, often rely on motion capture devices or hand keyed animation. Demand for realistic animation within these domains is high and both of these approaches are expensive and time consuming, providing considerable incentive for automation of the process. Embodied Conversational Agents (ECAs) for education, training, entertainment or Human-Computer Interaction (HCI) require realistic motion in both speaking and listening modes. More recently, increasing interest in Virtual Reality (VR) and Augmented Reality (AR) applications provide further stimulus for the development of predictive animation systems.

Human discourse essentially flows in two modes: the explicit mode of audible speech, and a more supportive visual mode where non-verbal gestures complement and enhance the audible mode. Research suggests that speech and visual gesture stem from the same internal process and share the same semantic meaning [22, 7]. Speaker head motion is a rather interesting aspect of visual speech. Head motion has been shown to contribute to speech comprehension [25], yet unlike the articulators, it is under independent control. As the audio channel contains the most complete information stream in an utterance, it is a reasonable strategy to seek a mapping from within this modality that might enable plausible predictions of head pose. Indeed, there is significant measurable correlation between speech and speaker head motion [5] that has motivated a number of authors to choose this approach.

2 Previous Work

Initial speech to head motion prediction strategies took the approach of clustering motion patterns and assigning class labels [11, 5]. Hidden Markov Models (HMMs) were trained for each cluster, modelling the relation between the speech features and head motion. These approaches rely on a suitable labelling of motion units, either manually or automatically; a challenging problem in itself.

In recent years, the Graphics Processor Unit (GPU) has enabled efficient training of Deep Neural Networks (DNNs), and within many aspects of speech and language processing, DNNs are now state of the art [19, 10, 9]. DNNs were proposed as a modelling strategy for head motion prediction by Ding *et al.* [12]. Using a deep Feed-Forward Neural Network (FFN) regression model to predict Euler angles of nod, yaw and roll, they were able to report advantages over the previous HMM based approaches and were able to avoid the problem of clustering motion. Deep FFNs are a powerful modelling tool, capable of learning complex non-linear mappings, however, they are limited in their ability to model long term temporal data.

The Long Short Term Memory (LSTM) network [18], has been used to great effect in many disciplines arguably related to the speech to head pose problem. Graves [16], demonstrated the ability of LSTM networks to model long term structure by predicting discrete text values, and by predicting the real values of hand-writing trajectories. Another example by Sutskever *et al.* [28] reports state of the art performance for the language translation task. Ding *et al.* [13] introduced Bi-Directional Long Short Term Memory (BLSTM) networks to the head motion task, noting improvements over their own earlier work [12]. More recently Haag [17] uses BLSTMs and Bottleneck features [14] and noted a subtle improvement.

Yngve [31] introduced the term "backchannels" to describe listener interaction providing acoustic and visual signals that inform turn taking. Later, Allwood *et al.* [1] suggested this linguistic feedback conveys perception, comprehension and empathy. Ward & Tsukahara [29] gave evidence that audible speaker prosody offers cues for backchannel response from the listener.

There have been a number of listener models described in the literature. Cassel *et al.* [6] report a comprehensive rule-based model that triggers backchannels from multi-modal input. Watanabe *et al.* [30] describe a rule-based speech driven interactive agent. Nishimura *et al.* [26] presented a decision tree model driven by prosodic audio features. Morency *et al.* [23] demonstrated a data driven sequential probabilistic model using HMMs and Conditional Random Fields (CRFs). Bevacqua *et al.* [3] introduced a model with personality traits.

Generative models [20, 27] trainable with back propagation [2], have taken an important step in learning. These models can perform probabilistic inference and make diverse predictions. For example, Bowman *et al.* [4] employed a Variational Autoencoder (VAE) for natural language generation. Given the diverse expectation of head pose during conversation, either as speaker or listener, generative probabilistic models represent an encouraging prospect for head pose prediction.

3 Corpus

Head motion prediction studies typically use data that is not widely available. At the present time there are few significant multi-modal corpora freely available, that are suitable for modelling any rigid gesture with speech. For our own research we developed a corpus as described in this section.

3.1 Data Collection

We invited two actors, one female (speaker A), one male (speaker B) to recite from a scripted set of short conversational scenarios. The actors were encouraged to speak emotively and emphatically in order to provide natural, expressive and prosodic speech. In all, 3600 utterances were captured, giving a total of around six hours of speech.

We used six synchronised cameras, with three cameras aimed at each actor. Video frequency was 59.94 Frames per Second (FPS) and audio was recorded at 48 kHz then down sampled to 16 kHz. Each actor had 62 landmarks distributed about the face, which along with 58 natural feature landmarks such as eyes and lip edges, were tracked with Active Appearance Models (AAMs) [21]. With the cameras arranged such that left and right stereo pairs were formed on each actor, we were able to derive 3D models. The 3D models were stabilised by selecting the least deformed points and, using Procrustes analysis [15], rigid motion was separated from deformation. The rotations are about the X,Y and Z axes of a right handed coordinate system, with Y pointing up.

3.2 Motion Statistics

After data collection, we pre-processed our rigid motion modalities, to leave a global mean of zero and a global unit standard deviation. We took basic statistical measures (standard deviation, maximum and minimum values, and mean)

for each individual utterance for head rotation and delta 1 and 2 (first and second derivatives) of head rotation. We were able to identify significant outliers as failed reconstructions which were subsequently removed from the corpus. We show in Figure 1 the delta 1 for X, Y, Z Euler angles, for each actor, during speaking and listening, for the entire corpus. In Table 1, we show the median of the absolute minima and maxima for each rotation mode, to give an overview of the dynamic properties of our corpus.

Fig. 1. The standard deviation, maximum, minimum and mean delta 1 for head rotation angles, from our entire corpus. We can observe characteristic differences in each actor, for speaking and listening.

Table 1. For the entire corpus, we summarise the head motion deltas with the median of the absolute minima and maxima for each rotation mode.

	Speaker A			Speaker B			Listener A			Listener B		
	x	y	z	x	y	z	x	y	z	x	y	z
Delta 1	0.15	0.33	0.12	0.18	0.37	0.12	0.06	0.23	0.07	0.07	0.12	0.06
Delta 2	0.06	0.17	0.06	0.10	0.25	0.05	0.03	0.11	0.03	0.05	0.08	0.04

3.3 Audio Feature Extraction

We used a sliding frame over the time domain audio signal of $2/59.94$ s with an overlap of $1/59.94$ s, matching the sampling rate of our motion data. Following

convention, each frame was multiplied by a Hamming window. Although we have experimented with many audio features, for this report we use the log of the filter bank values as described by Deng *et al.* in [10]. Under this scenario we have a feature vector temporally aligned with the 3 Euler angles: nod (x), yaw (y) and roll (z). We normalise all features to have unit variance and zero mean.

4 Model Topology

Our modelling strategies feature LSTM networks, and although there are many variations to consider, we use the implementation in the popular Keras framework [8]. We describe each of our modelling strategies in the following subsections, along with our observations for their respective advantages and disadvantages.

4.1 Bi-Directional Long Short Term Memory (BLSTM)

Our application of the BLSTM differs from Ding *et al.* [13]. Instead of predicting one motion coefficient at each time step, we predict a short span: $1 \leq k \leq 29$. This allows observation of frame-wise variation in prediction and permits options on recombining each frame. For this report we simply take the mean at each predicted time step. Notably, we do not apply any post process to the prediction such as smooth filtering. We observed distinct motion events in our data $> 500\ ms$ and to ensure capturing these events the receptive field was $29 \leq n \leq 129$ time steps, $n/59.94\ s$. This network works well for a single actor, and less well for multiple actors where we observe greater variation at each predicted time step. We can see in our statistics plots(Figure 1) that each actor has individual motion characteristics, we speculate that a significantly larger corpus might allow this model to separate this variation.

4.2 Conditional Variational Autoencoder (CVAE)

A VAE comprises an encoder and a decoder. The encoder, $Q_\theta(z|x)$, seeks to represent input data x in a latent space z with weights and biases θ, where the encoder outputs the parameters of a Gaussian probability density. The decoder, $P_\phi(x|z)$, with weights and biases ϕ, transforms the parameters to the distribution of the original data. Our Conditional Variational Autoencoder (CVAE) model adds a conditioning element to the VAE, such that the decoder is $P_\phi(x, c|z)$, and we use a deep BLSTM for both the encoder and decoder. Recall, we regard head pose as having a diverse, one to many relationship to any utterance. The generative model here permits sampling from a normal distribution during prediction, giving the option of multiple predictions for any given utterance. Further, this model performs well with multiple actors.

5 Model Training

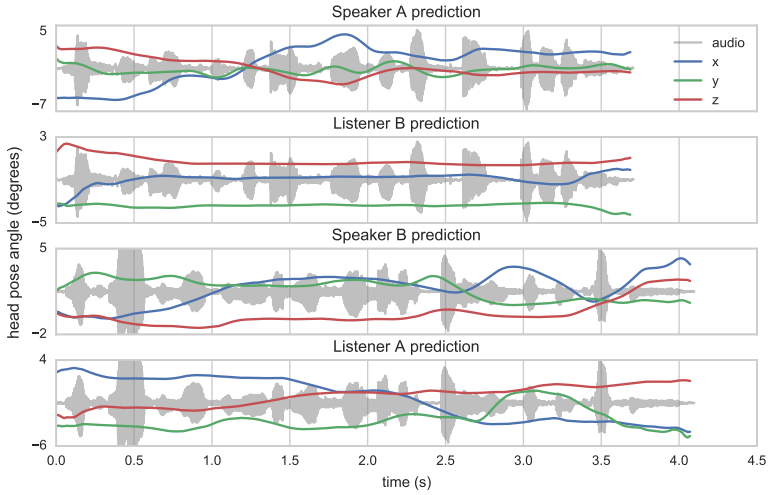
We trained the networks on our data, split 85% for training, 10% for validation and 5% for testing. Our objective function is Mean Squared Error (MSE), except for the CVAE model which has a custom objective function: the sum of the reconstruction loss and the Kullback-Leibler divergence [20]. Our optimising function is *RMSprop*, and we set an initial learning rate of 10^{-3}. Training continues until no further improvement on the validation set, with a patience of 5 epochs. Model weights are saved at each epoch. We reload the best weights, decrement the learning rate by a factor of 10 until 10^{-5}, finally stopping at the best validation error. We then select the model with the lowest overall validation error. The total number of examples presented to the network at training time depends somewhat on the value of span k and time steps n, and is in a range approximately 7×10^4 to 3×10^5. For this report, we trained models on each single speaker, each individual listener, speaker A and B combined and listener A and B combined.

6 Evaluation

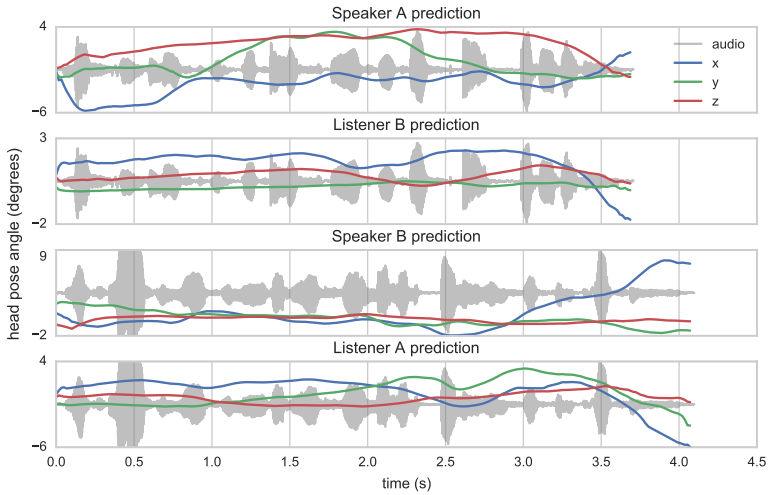
Subjective testing has been commonly used to evaluate speech animation quality. However, such tests are often small scale and can lack statistical significance. Furthermore, for the period of time such systems have been developed, now some decades, the subjective tests invariably confirm the proposed system. This suggests such testing strategies might be unreliable. Empirical measurements utilised so far can also have problems. Previous authors have used point wise measures such as MSE or Canonical Correlation Analysis (CCA) against a true example to assess results. Head motion during speech does not have a deterministic outcome. If a speaker were head shaking to express disagreement, a phase shift would affect MSE, but not necessarily the plausibility of the motion. Conversely, CCA on the X, Y, Z rotations would show strong correlation for head shake against head nod at the same phase and frequency, even though the sentiment is opposite. In the event we had a reliable empirical measure, comparison with existing systems remains difficult, due to the lack of standard multi-modal corpora. Consequentially, we assess our predictions by comparing the dynamic statistics to those of our entire data set, that we show in Figure 1 and Table 1.

7 Results

For each of our models we make predictions using examples from our data that have been randomly and fairly selected. None of the test examples have been involved in the training of any model nor have any been used to select the best model. A further constraint on the test examples is that for each speaker, the corresponding listener is not involved in training or selection. Reconstruction simply involves presenting a test utterance and forward propagating through each network. Each resulting motion coefficient has 1 to k values, from which we take the mean.



(a) BLSTM predictions.



(b) CVAE predictions.

Fig. 2. Example predictions from our models, discussed in Section 7. The BLSTM model is trained on each individual speaker and listener, whereas for the CVAE, we use a single model to predict both speakers, and a single model to predict both listeners.

7.1 BLSTM

We show some example predictions from our BLSTM network in Figure 2a. For speaker A, the utterance: "This is the most ridiculous spiritual quest I've ever been on." and for speaker B the utterance: "It's laughable to me that you

assume I have any interest in touching you." The head pose angle is predicted from the same utterance for both speaker and listener. We show a summary of the motion deltas in Table 2. We observe that our results fall within a small factor of the global summary in Table 1. Generally motion is a little smoother than our recorded motion, which we attribute partially to noise in the data collection, and to variation at each predicted time step. We note that head motion corresponds to events in the audio, both for speaker and listener. For these predictions the models were trained for single speaker and single listener, a total of four individual models.

7.2 CVAE

For our generative model, we use the same utterances as in 7.1. Here we train the speaker model on both actors, and the listener model on both actors. We find our trajectory statistics are closest to our corpus for these models (Table 2) and observe the prediction responds very well to the audio, matching key prosodic events of an expressive utterance. We make predictions from this model by sampling from the unit Gaussian space and conditioning with our example audio features. A parameter for this model, not present in the earlier models, is the size of the latent space. For this report we show a model with z in 3 dimensions, which we found to have no disadvantage to larger space.

Table 2. For the predictions from our models discussed in Section 7.1 and 7.2, we summarise the deltas with the median of the absolute minima and maxima for each rotation mode.

	Speaker A			Speaker B			Listener A			Listener B		
	x	y	z	x	y	z	x	y	z	x	y	z
BLSTM Delta 1	0.14	0.13	0.15	0.07	0.07	0.05	0.05	0.10	0.09	0.06	0.03	0.05
BLSTM Delta 2	0.05	0.07	0.06	0.02	0.03	0.02	0.03	0.04	0.04	0.03	0.02	0.02
CVAE Delta 1	0.08	0.15	0.15	0.10	0.12	0.10	0.12	0.11	0.11	0.08	0.06	0.06
CVAE Delta 2	0.03	0.05	0.07	0.05	0.04	0.06	0.05	0.06	0.07	0.03	0.03	0.03

8 Conclusion

The question of what represents appropriate or plausible head motion during speech is unclear. Subjectively, we have observed certain key events support viewer acceptance, but we have not yet been able to identify exactly why this is the case. We do know however, that it is important to have correct motion [25], and also that we can identify when it's not correct [24]. We have taken a decision to offer an alternative assessment for model predictions by showing statistics for the entire utterance. Developing a universal measurement of correct head motion, or indeed more broadly gesture, is an open and difficult problem, and we are actively pursuing this goal.

Our most interesting results come from the CVAE model, that addresses the diverse expectation of speech to head motion. We can predict a number of plausible motion trajectories by choosing new values for z, but with the same audio features. Quicktime movie files are provided in the supplementary material showing examples from our models.

In this paper we have presented our work on predicting head pose in dyadic conversation. We described our corpora, and presented modelling strategies that offer diverse but plausible outcomes for audio input. The LSTM has been a powerful tool in speech and language modelling, and as the encoder-decoder in our CVAE has shown great utility. We feel that generative models offer great promise to this field and we continue working in this area.

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of semantics 9(1), 1–26 (1992)
2. Bengio, Y., Laufer, E., Alain, G., Yosinski, J.: Deep generative stochastic networks trainable by backprop. In: Proceedings of The 31st International Conference on Machine Learning. pp. 226–234 (2014)
3. Bevacqua, E., De Sevin, E., Hyniewska, S.J., Pelachaud, C.: A listener model: introducing personality traits. Journal on Multimodal User Interfaces 6(1-2), 27–38 (2012)
4. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. CoNLL 2016 p. 10 (2016)
5. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)
6. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. pp. 520–527. ACM (1999)
7. Cassell, J., McNeill, D., McCullough, K.E.: Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. Pragmatics & cognition 7(1), 1–34 (1999)
8. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
9. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: An overview. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8599–8603. IEEE (2013)
10. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al.: Recent advances in deep learning for speech research at Microsoft. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 8604–8608. IEEE (2013)
11. Deng, Z., Narayanan, S., Busso, C., Neumann, U.: Audio-based head motion synthesis for avatar-based telepresence systems. In: Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence. pp. 24–30. ACM (2004)
12. Ding, C., Xie, L., Zhu, P.: Head motion synthesis from speech using deep neural networks. Multimedia Tools and Applications pp. 1–18 (2014)

13. Ding, C., Zhu, P., Xie, L.: Blstm neural networks for speech driven head motion synthesis. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
14. Gehring, J., Miao, Y., Metze, F., Waibel, A.: Extracting deep bottleneck features using stacked auto-encoders. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 3377–3381. IEEE (2013)
15. Gower, J.C.: Generalized procrustes analysis. Psychometrika 40(1), 33–51 (1975)
16. Graves, A.: Generating sequences with recurrent neural networks. CoRR abs/1308.0850 (2013), http://arxiv.org/abs/1308.0850
17. Haag, K., Shimodaira, H.: Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In: International Conference on Intelligent Virtual Agents. pp. 198–207. Springer (2016)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
19. Huang, J.T., Li, J., Gong, Y.: An analysis of convolutional neural networks for speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4989–4993. IEEE (2015)
20. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems. pp. 3581–3589 (2014)
21. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
22. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
23. Morency, L.P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Intelligent Virtual Agents. pp. 176–190. Springer (2008)
24. Mori, M.: The uncanny valley. Energy 7(4), 33–35 (1970)
25. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychological science : A journal of the American Psychological Society / APS 15(2), 133–137 (2004)
26. Nishimura, R., Kitaoka, N., Nakagawa, S.: A spoken dialog system for chat-like conversations considering response timing. In: International Conference on Text, Speech and Dialogue. pp. 599–606. Springer (2007)
27. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of The 31st International Conference on Machine Learning. pp. 1278–1286 (2014)
28. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
29. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in english and japanese. Journal of pragmatics 32(8), 1177–1207 (2000)
30. Watanabe, T., Okubo, M., Nakashige, M., Danbara, R.: Interactor: Speech-driven embodied interactive actor. International Journal of Human-Computer Interaction 17(1), 43–60 (2004)
31. Yngve, V.H.: On getting a word in edgewise. In: Chicago Linguistics Society, 6th Meeting. pp. 567–578 (1970)