Joint Learning of Facial Expression and Head Pose from Speech

David Greenwood, Iain Matthews, Stephen Laycock

School of Computing Science
University of East Anglia
United Kingdom

david.greenwood@uea.ac.uk, iain.matthews@uea.ac.uk, s.laycock@uea.ac.uk

Abstract

Natural movement plays a significant role in realistic speech animation, and numerous studies have demonstrated the contribution visual cues make to the degree human observers find an animation acceptable. Natural, expressive, emotive, and prosodic speech exhibits motion patterns that are difficult to predict with considerable variation in visual modalities. Recently, there have been some impressive demonstrations of face animation derived in some way from the speech signal. Each of these methods have taken unique approaches, but none have included rigid head pose in their predicted output.

We observe a high degree of correspondence with facial activity and rigid head pose during speech, and exploit this observation to jointly learn full face animation and head pose rotation and translation combined. From our own corpus, we train Deep Bi-Directional LSTMs (BLSTM) capable of learning long-term structure in language to model the relationship that speech has with the complex activity of the face. We define a model architecture to encourage learning of rigid head motion via the latent space of the speaker's facial activity. The result is a model that can predict lip sync and other facial motion along with rigid head motion directly from audible speech.

Index Terms: Speech Animation, Deep Learning, LSTM, BLSTM, RNN, Audiovisual Speech, Shape Modelling, Lip Sync, Uncanny Valley, Visual Prosody

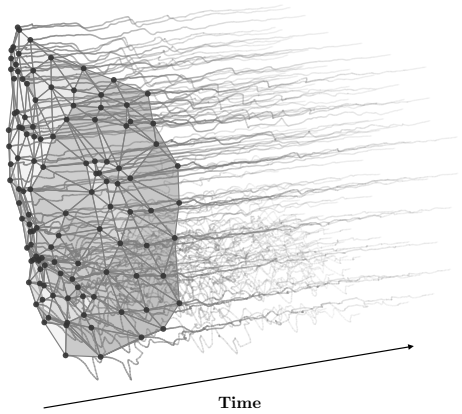

Figure 1: *Facial activity is complex during speech. Here we illustrate our shape model deforming over time to give an impression of the correspondence different regions of the face have while speaking.*

1. Introduction

We can describe speech animation as deforming and transforming a character model, temporally aligned to an audible utterance, to give the impression the character is speaking. The task is very challenging, as mismatches between visual speech and audio can change what a viewer believes they heard [1], and speaker head pose can affect comprehension [2]. In addition to these effects, we, as human viewers, can experience feelings of reduced empathy or even revulsion if the animation is not quite right [3]. Production level speech animation, such as is found in mainstream movies and video games, often use performance capture or teams of skilled animators. Both of these approaches are time consuming, expensive and lack scalability. This provides considerable motivation to develop techniques to automate the process of high fidelity speech animation.

In this work we model the complete facial activity during speech, along with the rigid pose of the speaker's head. We extend our earlier work on speaker head pose [4] by modelling six Degrees of Freedom (DoF): the rotations of nod, yaw and roll and the translations on those axes. Head pose has properties that make it difficult to model directly from speech. There is high measurable correlation between speech audio and head pose, yet a speaker repeating an utterance several times may move his head in significantly different manner on each repetition. In our own corpus we observe a closer correspondence between facial activity and head pose during speech by modelling head pose directly from facial features rather than from audio. We hypothesise the modal gap is smaller between facial activity and head pose as the anatomical, physical and kinetic constraints are closer.

To exploit this observation, we first train a model to predict the facial animation from audio features, then in a second stage, encourage the model to learn head pose from the latent representation of the facial activity by using a separate objective for each mode.

2. Related Work

2.1. Facial Animation

The automatic production of realistic speech animation is a long held goal of many areas of graphics, speech and language research, and work extends deeply into the literature. Lewis and Parke [5] describe a lip syncing model based on Linear Predictor Coefficients (LPCs) to predict *visemes* [6], the visual counterpart of phonemes. In this early work they acknowledge the importance of other aspects of speech animation, notably head pose, for fully expressive automated character animation. They also remark on the strong perceptual effects that we now refer to as the 'Uncanny Valley' [3].

Linguistic based methods to produce plausible facial ani-

mation have been developed over several decades [7, 8], either 3D mesh [9] or 2D video [10] based. Their common requirement is some form of alignment of the phoneme content either as transcript or by prior processing with external tools [11]. The complex relationship between co-articulated phonemes and visemes is defined as a many-to-many mapping in the work of Taylor *et al.* [12] superseding the static shape-to-shape model of Fisher [6].

Data-driven, or machine learning based models that rely only on the input of audio have a similarly lengthy history. Voice Puppetry [13] is a notable example that uses Hidden Markov Models (HMMs) for trajectory sampling. Most recently Suwajanakorn *et al.* [14] use a regression model of Long Short Term Memory (LSTM)[15] networks to produce highly plausible 2D lip animation. Karras *et al.* [16] employ a deep neural network combining fully connected layers and Convolutional Neural Networks (CNNs) to model facial animation with emotional content.

The recent work by Taylor *et al.* [11], Suwajanakorn *et al.* [14] and Karras *et al.* [16], arguably represent the state of the art for data driven facial animation, and these three works appeared in the literature at the same time. Interestingly, all three had hand animated head pose applied to reduce perceptual dissonance.

2.2. Head Pose

Head pose during speech is another aspect of visual speech with a rich history in the literature. Early studies approached the problem by categorically labelling clustered head motion patterns [17, 18, 19]. HMMs were trained for each cluster, modelling the relation between the speech features and cluster label. Hofer [20, 21] observes the limitations of the frame wise approach of his predecessors, and proposes a trajectory based model. More recently Ben-Youssef [22] proposed an improved clustering for motion. All of these approaches rely on a suitable labelling of motion units, either manually or automatically, which is a challenging problem in itself.

Ding *et al.* [23] introduce a deep Feed-Forward Neural Network (FFN) regression model to predict Euler angles of nod, yaw and roll. They report advantages over the previous HMM based approaches and were able to avoid the problem of clustering the motion. Deep Bi-Directional Long Short Term Memory (BLSTM) models appear in Ding *et al.* [24], where they report improvements over their own earlier work. More recently Haag [25] uses BLSTMs and Bottleneck features [26]. In our own earlier work [4], we use a BLSTM based Conditional Variational Autoencoder (CVAE) to model the many-to-many mapping of speech to head pose prediction, both for speaker, and for the head pose of the listener in dyadic conversation [27].

3. Corpus

We believe clean, unbiased data are an important part of data-driven face learning, so we collect and process data to develop a corpus as described in this section.

3.1. Data Collection

We hired two actors, one female (Subject A), one male (Subject B) to recite from a scripted set of short conversational vignettes. 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.

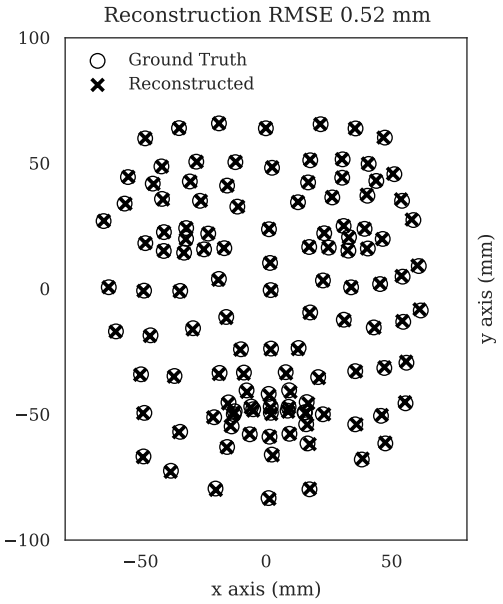

Figure 2: *We can represent greater than 98% of the facial variance with 8 Principal Components. We show the Root Mean Square (RMS) reconstruction error for those 8 components and the original shape. We use this measure as a quantitative evaluation of our predictions.*

We used six cameras to record with synchronised frame timing, with three cameras aimed at each actor. Recording frequency was 59.94 Frames per Second (FPS) and resolution 1280×720 pixels (720p). Audio was recorded simultaneously at 48 kHz and later down sampled to 16 kHz mono.

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) [28, 29]. With the cameras arranged such that left and right stereo pairs were formed on each actor, we were able to derive 3D shape models. The shape models were stabilised by selecting the least deformed points and, using Procrustes analysis [30], rigid motion was separated from deformation. The translations and rotations are about the x, y and z axes of a right handed coordinate system, with y pointing up. For this report, we use the data collected for Subject A only.

3.2. 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 Filter Bank (LogfBank) of 40 filters as described by Deng *et al.* in [31]. The resulting feature vector is now centrally and temporally aligned with the recorded motion of the deforming facial landmarks and the six DoF of rigid head pose. Finally, we normalise our input data to have unit variance and zero mean.

3.3. Shape Parametrisation

The activity of the face is highly complex during speech. Figure 1 illustrates the deformation of the shape model as Subject A is

speaking. Although complex, one can clearly see considerable correspondence in different regions of the face. This observation motivates us to borrow from our AAMs, and use Principal Component Analysis (PCA) to reduce the dimensionality of the shape model. We find we can retain more than 98% of the variation in the entire corpus for our Subject's shape model in 8 Principal Components. Figure 2 shows the shape model decomposed to 8 components then reconstructed, compared with the original mean shape model. The Root Mean Square Error (RMSE) is acceptably small at ≈ 0.5 mm, and the plot shows that there are no significant outliers. We will use the reconstruction error measured in this way to evaluate the accuracy of the predicted animation.

4. Model Description

Carrying on from our previous work [4], we use Deep BLSTMs to predict the facial deformation and the six DoF of rigid head pose combined. Clearly, much of the activity of the orofacial region has significant correspondence with speech production. Other regions of the face, along with head pose, have also been shown to have a relationship with speech [32]. Our initial experiment was to consider how well we could predict face animation with a deep BLSTM, with audio features as input and our 8 PCA values as output. We observed good modelling, particularly of the more significant components. We further experimented with predicting head pose from facial expression, and observed improved performance over direct prediction from audio features. We hypothesise that the facial activity during speech closes the modal gap to head pose, i.e., the motion of the face is controlled by anatomy and limited by kinetic constraints, and so is the rigid motion of the head. When we try to model head pose directly from audio features we can not force the model to learn via that space.

Simply concatenating the rigid pose and shape values and training a Deep BLSTM did not provide the results we had seen with independently trained models. So we describe a forked model, with separate objectives for the six DoF head pose values and the PCA expression values. This allows independent control of each of these modalities, both in the topology, and in the training of the model. We found our best results were achieved by developing a model that only predicted the PCA values, then forking the model late in the latent space to a new stack of layers, with output to head pose values. Figure 3 illustrates the topology of the network. Our experience with these networks so far has been that the number of trainable parameters is limited by the quantity of our data. We find a properly converged model has a layer of 250 hidden units at input, with four subsequent layers, tapering in hidden units to 50, to the PCA objective. The head pose branch can be as little as two layers of 30 hidden units, much smaller than a model for predicting head pose alone. Recall, we use BLSTM, so the number of hidden units is doubled, as the count is for each direction.

4.1. Training

We trained the networks on our data, split 90% for training, 5% for validation and 5% for testing. Our test examples have been excluded from the corpus from the outset and have never been used for training or model selection. We divide our data into short sections of $t = 129$ samples, starting each new section at $t_0, t_1, t_2, \ldots$, giving a total count of examples $\approx 8 \times 10^4$. Examples shorter than 129 samples are discarded, not padded.

Both our objective functions are Mean Squared Error

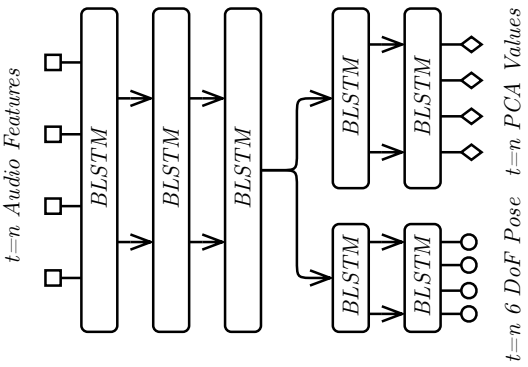

Figure 3: *The topology of the deep BLSTM model. We pre-train the route to the PCA expression values, then train the whole model with separate objectives for the PCA values and the 6 DoF of the rigid head pose.*

(MSE), and our recent experiments use *adam* [33] for optimisation, with the parameters: $lr = 0.001$, $beta_1 = 0.9$, $beta_2 = 0.999$, $epsilon = 1 \times 10^{-8}$, $decay = 0.0$. Comparing identical models, *adam* converges more quickly than *rmsProp* [34] for our task. Training continues until no further improvement on the validation set, with a patience of 5 epochs. We first train a model with the sole objective of the PCA expression values, then load those weights to the lower layers of our forked network (Figure 3). We recommence training of the entire network now with two objectives. Interestingly, the loss for PCA expression values continues to descend from this point. While we monitor both losses, our early subjective tests indicate viewers discriminate on the overall animation quality more by face accuracy than head pose, so we train until no further validation improvement on the PCA fork, with a patience of 5 epochs. We use the Keras framework [35], with Tensorflow [36] back end. In this report, the models are trained on one Subject, A, from our corpus.

5. Results

Table 1 shows the results of predictions for held out utterance examples. We quantitatively evaluate our predictions in the following way: We use CCA to measure correlation for each predicted example by projecting to one base and calculating Pearson's r for the projection to the base. CCA > 0.5 represents significant correlation, and CCA $= 1.0$ is maximum correlation. We show CCA for the 8 predicted PCA component values, CCA for the head rotation values, and CCA for the head translation values. We report RMSE for the reconstructed PCA shape model for the whole utterance measuring the error in millimetres (mm). We report RMSE for head rotation in degrees, and head translation in mm. We find CCA the more valuable measure for head pose as it indicates comparable modulation by the audio waveform, whereas a uniform offset in the trajectory can increase RMSE without adversely effecting the quality of the prediction. For the facial activity we desire *both* high correlation and low RMSE.

For qualitative assessment, we show plots of the trajectories of the first three Principal Components (Figure 4) and the rotation angles of nod (x), yaw (y) and roll (z) (Figure 5). On the

Table 1: *For a quantitative evaluation of our predictions we show six scenes held out from our corpus. We show the reconstruction RMSE in mm for our shape model for the entire utterance, along with Canonical Correlation Analysis (CCA) for the true and predicted PCA components. We show the same measure for the six DoF of head pose, though the head pose rotation error unit is degrees.*

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE PCA Reconstruction	1.46	1.37	2.84	1.68	1.24	1.88
RMSE Pose Rotation	2.68	3.73	3.93	4.27	3.27	1.81
RMSE Pose Translation	2.85	1.81	3.84	2.84	3.64	2.57
CCA PCA Components	0.97	0.96	0.96	0.98	0.98	0.96
CCA Pose Rotation	0.90	0.74	0.79	0.85	0.89	0.94
CCA Pose Translation	0.79	0.69	0.72	0.54	0.78	0.96

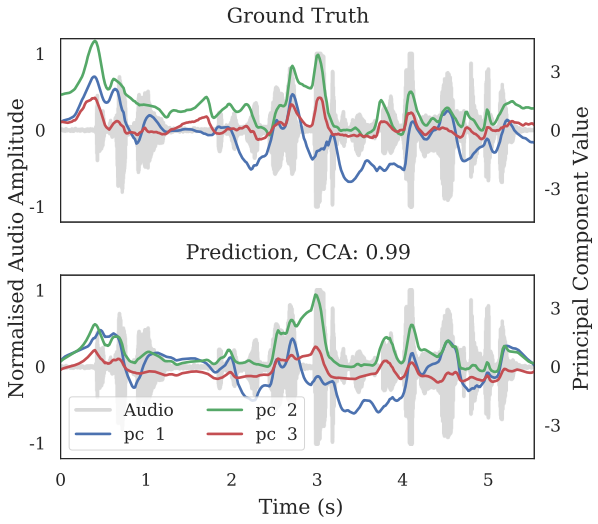

Figure 4: *The ground truth and prediction of the first three Principal Components. The first three components are largely associated with the orofacial area. We show CCA for the components plotted in PCA space. We can see clearly how the components are modulated by the audio. Qualitatively, one can observe how closely the component values in the prediction follow the ground truth.*

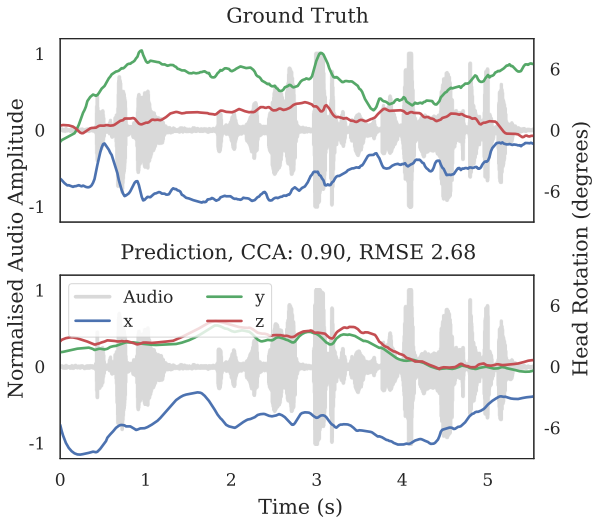

Figure 5: *The ground truth and prediction of the rigid head pose angles. We observe how head pose angle is also modulated by the audio, but has somewhat more diverse expectation. Quantitatively, for the three rotation axes we show CCA and RMSE (degrees).*

plot in Figure 4 we report CCA for just the three plotted components. These components largely relate to the orofacial area and indicate lip sync performance. On Figure 5 we report measure in the same way as Table 1. For further qualitative evaluation we render animations of the data, examples of which can be found in the supplementary material for this paper.

6. Discussion

Future work involves seeking a generalisation of our method so we can not only train on multiple speakers from within our corpus, but also predict speakers from outside our corpus. Our technique works equally well for each speaker individually, so we are optimistic regarding that goal. A general system for character animation would need to drive any reasonable character, which may or may not have human-like features, *and* coexist within an industry standard production pipeline. Rig retargeting [11] is a technique for sampling a deforming mesh to learn animator friendly blend-shape weights, thus merging the pipeline forward of our parametrised shape model. A limitation of our corpus, rather than our method, is a small quantity of silence. Short pauses during speech are modelled well currently, but we do not have sufficient data to model pensive duration, retrospection and so on that occur in silence; yet these natural activities do involve animation. As our model makes low latency predictions directly from LogfBank audio speech features that can be processed quickly, we expect to extend the work to *real time* prediction of complete facial animation.

The concept of forcing a model to learn via an intermediate modality presents new ideas for tackling other visual modes that are under independent control in addition to head pose.

7. Conclusions

We have described our corpus, which we have used to train a data driven deep BLSTM model for predicting a complete character head animation solely from audio speech features input. Our low latency model predictions include accurate lip sync, animation of all the facial features, and rigid head pose rotations and translations.

8. References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.

[2] 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 : A journal of the American Psychological Society / APS*, vol. 15, no. 2, pp. 133–137, 2004.

[3] M. Mori, "The uncanny valley," *Energy*, vol. 7, no. 4, pp. 33–35, 1970.

[4] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder," *Proc. Interspeech 2017*, pp. 3991–3995, 2017.

[5] J. P. Lewis and F. I. Parke, "Automated lip-synch and speech synthesis for character animation," *SIGCHI Bull.*, vol. 17, no. SI, pp. 143–147, May 1986. [Online]. Available: http://doi.acm.org/10.1145/30851.30874

[6] C. G. Fisher, "Confusions among visually perceived consonants." *Journal of speech and hearing research*, vol. 11 4, pp. 796–804, 1968.

[7] J. Lewis, "Automated lip-sync: Background and techniques," *Computer Animation and Virtual Worlds*, vol. 2, no. 4, pp. 118–122, 1991.

[8] W. Mattheyses and W. Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication*, vol. 66, pp. 182–217, 2015.

[9] L. Wang, W. Han, F. K. Soong, and Q. Huo, "Text driven 3d photo-realistic talking head," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1997, pp. 353–360.

[11] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 93, 2017.

[12] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*. Eurographics Association, 2012, pp. 275–284.

[13] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 21–28.

[14] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 95:1–95:13, Jul. 2017.

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

[16] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 94:1–94:12, Jul. 2017.

[17] Z. Deng, S. Narayanan, C. Busso, and U. Neumann, "Audio-based head motion synthesis for avatar-based telepresence systems," in *Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*. ACM, 2004, pp. 24–30.

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

[19] 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, 2007.

[20] G. Hofer and H. Shimodaira, "Automatic head motion prediction from speech data," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[21] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," in *ACM SIGGRAPH 2007 posters*. ACM, 2007, p. 86.

[22] A. Ben Youssef, H. Shimodaira, and D. A. Braude, "Articulatory features for speech-driven head motion synthesis," *Proceedings of Interspeech, Lyon, France*, 2013.

[23] C. Ding, L. Xie, and P. Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, pp. 1–18, 2014.

[24] C. Ding, P. Zhu, and L. Xie, "Blstm neural networks for speech driven head motion synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] 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*. Springer, 2016, pp. 198–207.

[26] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.

[27] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose in dyadic conversation," in *International Conference on Intelligent Virtual Agents*. Springer, 2017, pp. 160–169.

[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[29] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[30] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[31] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8604–8608.

[32] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: facial movements accompanying speech," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, May 2002, pp. 396–401.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[34] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, 2012.

[35] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[36] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/