Double-DCCCAE: Estimation of Sequential Body Motion Using Wave-Form - AlltheSmooth

JinHong Lu
The University of Edinburgh
Edinburgh, UK
jinhong.l@sms.ed.ac.uk

TianHang Liu
The University of Edinburgh
Edinburgh, UK
T.Liu-34@sms.ed.ac.uk

ShuZhuang Xu
The University of Edinburgh
Edinburgh, UK
S.Xu-27@sms.ed.ac.uk

Hiroshi Shimodaira
The University of Edinburgh
Edinburgh, UK
H.Shimodaira@ed.ac.uk

ABSTRACT

Different types of inputs (e.g., speech, text, motion) to motion estimation have been widely investigated in frame-based system, but these do not reflect the temporal relationship between speech and motion. It is common in the literature to use multiple frames of input to estimate one frame of motion data, this would increase the input's dimension dramatically if one must estimate sequential motion. We also show that the correlation gets weaker between multiple blocks of speech and multiple frames of motion data. To resolve the problems, we extend our previous work and propose a frame-based system that estimates the motion in a sequential manner, double deep canonical correlation constrained autoencoder (Double-DCCCAE), which encodes sequential features (speech/motion) into frame-based embedded features with error and canonical correlation analysis (CCA) loss. The learnt motion embedded feature is estimated from the learnt speech-embedded feature through a simple feed-forward neural network and further decoded back to the sequential raw motion. Our proposed feature pair showed higher correlation than spectral features with motion data, and our model was more preferred than the baseline model (BA) in terms of human-likeness and had similar appropriateness.

KEYWORDS

neural networks, speech, body motion, conversational virtual agent

1 INTRODUCTION

When we are in conversation, a large quantity of motions (such as gesture, body, and head) are spontaneously emitted [1, 2]. These motions are transmitted as non-verbal signals to the listeners, and help the listeners to better understanding what is being expressed [3, 4]. As such, human motion is a key factor for the conversational agents or social robots to interact with us, and act human [5, 6].

To tackle the motion learning challenge for the agents/robots, researchers has explored in many directions.

Speech-Driven: Kucherenko *et al.* [7] implemented a frame-based speech-to-motion mapping with encoder-decoder DNN. The author applied representation learning to learn a motion embedding z with auto-encoder, and then learnt a mapping from the speech features s to the learnt motion representation z with DNN. The synthesised motion was generated by converting the predicted z through the decoder. Ginosar *et al.* [8] showed the results of generating motion sequence in a GAN-RNN system. The proposed generative model learnt to predict the temporal stack of poses from the given audio input, while an adversarial discriminator ensures that the predicted motion was both temporally coherent and in the style of the speaker.

Text-Driven: Yoon *et al.* [9] found that the natural language was useful to predict a frame-by-frame poses with a GRU-Auto-Encoder. The author first converted the speech's text to word embedding as input to the encoder. The encoder captured the speech context, and the results were transmitted to the decoder. The decoder then focused on the specific words instead of whole text with a soft attention mechanism when it generated poses.

Motion-Driven: Ghosh *et al.* [10] proposed a system that generates body motion with a deep LSTM-RNN and a de-noising auto-encoders (DAE). The de-noising auto-encoders were trained to reconstruct the body motion from a 'drop-out' body motion in frame-based systems. The LSTM-RNN was used to predict the body poses from a given pose. The predicted poses were filtered by the DAE and recursively served as input for the next time step.

These previous studies show the potential of using different types of input to predict body motion in frame-based systems, but body motion is a continuous and temporal datatype. Generating motion in a frame-based system does not reflect the temporal relationship between speech and body motion (or seq2seq motion). Possible reasons of diminishing the interests of temporal-based system are 1) Literature shows that generating a single frame of motion requires multiple frames of speech information[7, 11]. Thus, the frame's total number of speech information increases dramatically if the system generates multiple frames of motion at once. The hardware limitation does not allow us to perform such experiments. 2) The correlation between multiple frames of speech information and a frame of body motion is not strong, not to mention that the correlation gets weaker after stacking blocks of multiple speech information to generate multiple frames of body motion. The result of the experiment conducted in this study also shows that the correlation gets weaker. 3) RNN may be a reasonable solution for the sequence-to-sequence data estimation/prediction. However, we cannot ignore the weakness of the RNN for long time-step data in terms of gradient vanishing and exploding [12, 13].

To resolve the speech and motion frame-based problems, we propose Double-DCCCAE, a frame-based system, but the system able to estimate temporal sequence in this paper. Our proposed

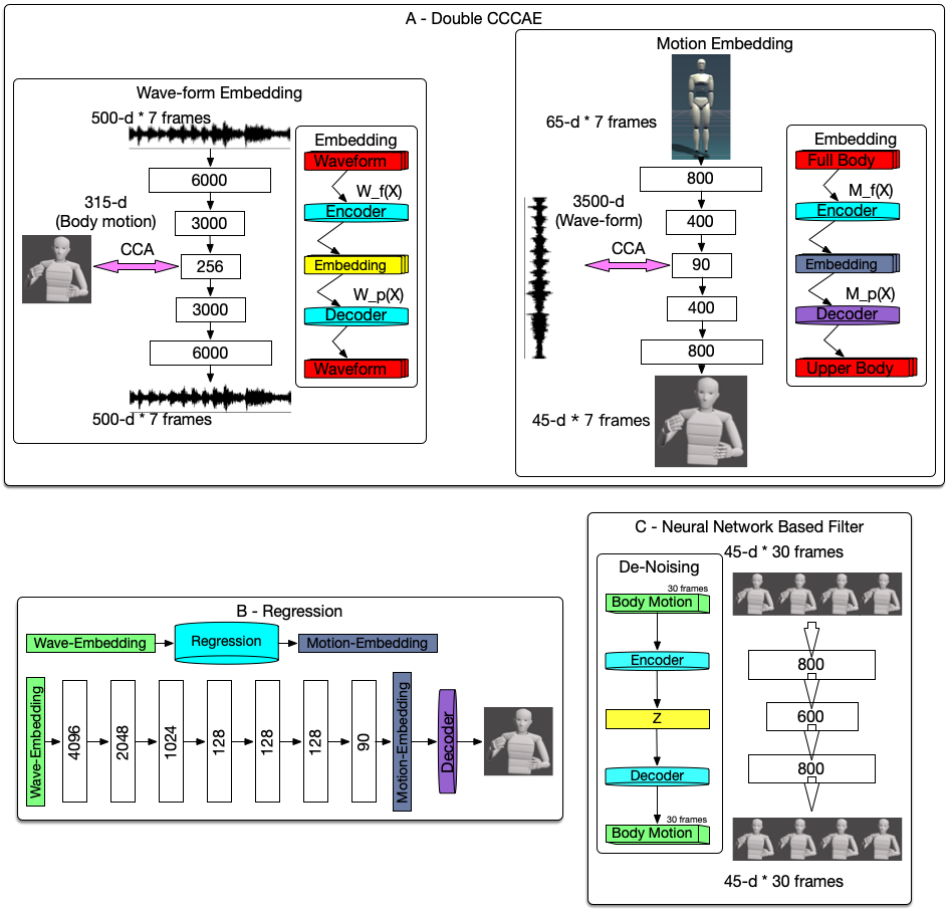

Figure 1: Overview of the proposed system comprised of three modules: (A) embedding with Double-CCCAE, (B) DNN-based sequential motion embedding regression from the wave-form embedded features, (C) post filter with an autoencoder.

system consists of 3 parts, a double deep canonical correlation auto-encoder, a frame-based regression, a post-filter. The auto-encoders are used to compress the information of the sequential data (e.g, speech information or body motion), as well as maintain possible higher correlations with other sequential data. The frame-based regression predicts the sequential motion embedding in a frame-by-frame manner from the wave embedding. The predicted frame-based motion embedding is further decoded by the trained decoder and interpreted as the sequential body motion movements. Lastly, we apply an NN-based filter to smooth the generated movements. We show that the features obtained with the proposed approach are more highly correlated than raw wave-form and MFCC. We submit our model to the GENEA2020 challenge and evaluate it with other participants' models and baseline models in a subjective test.

2 PROPOSED MODEL

Our proposed system can be separated into three modules: 1) double canonical-correlation-constrained autoencoders (CCCAE) for compressing the high-dimensional input (e.g., wave-form, body motion) to the distributed embedding of low dimensions, 2) a regression model for predicting the sequential motion embedding from the wave embedding, and 3) a post-filtering autoencoder for reconstructing smooth body motion. The overall framework of our proposed model is shown in Figure 1.

2.1 Double CCCAE

The framework of an autoencoder for a set of two data streams is proposed by Chandar *et al.* [14] and Wang *et al.* [15]. DCCAE [15] consists of two autoencoders and optimises the combination of the canonical correlation between the learnt "bottleneck" representations and the reconstruction errors of the autoencoders. In our previous work, we compressed high-dimensional wave-forms to low-dimensional and correlated embedding, with head motion using a single autoencoder of correlated neural network (CNN) [11].

We extend our work here to apply two CNN autoencoders for two reasons: 1) To compress sequential body motion data into a frame-based embedding. 2) The dimension of the body motion in this work is much higher than the head motion in our previous work. However, our work here is different from the aforementioned research studies, in which [14, 15] compressed the two stream into one common and correlated space using two autoencoders; on the other hand, we propose to compress the streams into different spaces with different correlated objects. We thus employ two autoencoders in which hidden layers are trained in such a way as to

not only minimise the reconstruction error but also maximise the canonical correlation with body motion. Thus, instead of projecting the two features to a common subspaces, we project the two features to two identical subspaces to ensure the embedded features are well correlated with the objective features.

We do not consider more advanced architectures such as variational auto-encoders (VAE)/conditional-VAE (CVAE)[16–18], because standard AE is more effective in this task as the generative models, VAE/CVAE, are usually harder in training due to KL vanishing[19].

We train each proposed CCCAE with the following objective function:

$$\begin{aligned} \mathrm{Obj}_{\mathrm{CCCAE}} = \frac{1}{N}\sum_{t} \|\boldsymbol{X}_{t-3tot+3} - p(f(\boldsymbol{X}_{t-3tot+3}))\|^2 \\ -\alpha\,\mathrm{CCA}\left(f(\boldsymbol{X}_{t-3tot+3}), \boldsymbol{Y}_{t-3tot+3}\right) \end{aligned} \quad (1)$$

In the above equation, N represents the number of data point, $\boldsymbol{X}_{t-3tot+3}$ represents the input feature vector at a time instance t to the encoder, $f(\,)$ represents the projection with the encoder, $p(\,)$ represents the reconstruction with the decoder, $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the whole sequences of feature vectors and objective feature vectors, respectively, and $\alpha \geq 0$ is the weighting factor, wherein $\alpha = 0$ corresponds to a standard autoencoder with an MSE loss function.

2.2 Regression Model

The idea of predicting motion embedding from speech was proposed by Kucherenko *et al.* [7]. This framework first applies representation learning to learn a motion representation in a frame-based system. Further, it encodes speech to the learnt motion representation and decodes the same through the motion decoder. We extend this idea to a frame-based model in a sequential manner with our highly correlated features estimated by the proposed Double-DCCCAE. We map the wave-form embedding to the motion embedding and decode through the motion decoder, but the decoded raw motion is temporal.

A simple feed-forward deep neural network is applied here for the regression from the wave-form embedded features to the motion embedded feature. We do not consider RNN (e.g., LSTM, GRU) because the present study focuses on decoding a sequential motion movement from a frame-based embedding feature and frame-based mapping between the two embedded features does not require a temporal relationship.

2.3 Neural-Network-Based Filter

The generated trajectories have the movement with minor jerkiness due to the nature of the speech, which can be viewed as the noisy data. It is common to apply post-processing to smooth output [11]. Ding [20] has applied an MLPG algorithm [21] to generate smooth trajectories; Sadoughi [22] smoothed the rotations by converting motion sequences into quaternions and then selecting 15 key points per second, interpolating the intermediate frames [23]; and Hagg applied a 3-order polynomial smoothing filter on the output [24]. However, these smoothing methods have the common problem that there is a trade-off between the smoothness and the accuracy of the filtered body motion. The accuracy here means the similarity between filtered motion and the ground truth because the post-filters may over-smooth the motions and cause the filtered motion to be stationary. We trained a neural-network-based post-filter to overcome these problems in the present study [7, 11, 25].

Unlike from the linear filters based on identifying the impulse transfer function that satisfies the requirements of the filter specification, it was expected that the neural-network-based post-filter would transfer noisy motions into the hidden presentation and reconstruct smooth motions based on the hidden presentation. The uncoordinated movements deliberately acted out by humans are always unavailable or expensive. A feasible method to create noisy data is either by applying dropout to the clean data for making the data discontinuous or adding Gaussian noise to the clean data for making the motions fluctuate [25]. However, CCCAE adopted clean data to synthesise more smooth motions [11]. In this paper, we also explored the effects of different types of data on the filter.

3 EXPERIMENT

3.1 Dataset

We have been provided with the Trinity Speech-Gesture Dataset [26] as the database for GENEA2020 challenge. A male native English speaker was involved in the collection of the dataset. For the audio, the actor produced spontaneous and natural conversational speech without interruptions, that is, without verbal cues from a conversation partner. Moreover, the actor chose the topic he would like to speak on in the conversation with a happy disposition and included a large quantity of gesture motions.

The actor addressed a person situated behind the camera to give him the visual feedback of a conversation partner. Each recording take was approximately 10 minutes long. The author captured 23 takes, totalling 244 minutes of data (provided for training in the challenge).

The author captured the actor's motion with a 53 marker setup and 20 Vicon cameras at 59.95 frames er second (FPS). The audio was recorded at 44 kHz.

Speech Feature: First, we down-sampled the audio rate from 44 kHz to 4 kHz. Raw wave-form vectors were extracted with a window of 125 ms and 67 ms shifting, which resulted in 500 dimensions. Further We extracted the MFCC12_E_D_A feature set from OpenSMILE toolkit. This configuration extracted Mel-frequency Cepstral Coefficients from 100 ms audio frames (sampled at a rate of 50 ms) (Hamming window). It computed 12 MFCC (1-12) from 26 Mel-frequency bands, and applies a cepstral liftering filter with a weight parameter of 22, and the log-energy was appended. 13 delta and 13 acceleration coefficients were appended to the features as well.

Body Motion: The motion data was stored in the BioVision Hierarchy format (BVH). The BVH data describes motion as a time sequence of Euler rotations for each joint in the defined skeleton hierarchy. In the present study, these Euler angles were converted to a total of 64 global joint positions in 3D. Some recordings had a different frame rate than others; therefore, we down-sampled all recordings to a common frame rate of to 20 FPS. Moreover, as the challenge required, we were asked to synthesise the upper body only, which included 45 out of 65 global joint positions. For the purpose of fast convergence in training, we applied standard

normalisation (zero mean and unit variant) to the data at each rotation of the joints.

Experiment Setup: We extracted 25 seconds of the video-audio data in the middle of each provided training file, totalling about 9.5 minutes as the validation data, and the rest of data were used in training. For the testing data, another 10 audio files (with transcripts), totalling about 20 minutes, were provided from the challenge without the motion data.

We conducted preliminary experiments to decide the depth and width of the Double-CCCAE and regression models, which are shown in Figure 1. The post-filter AE will be discussed below.

Training was conducted on a GPU machine and a multi-CPU machine with Pytorch version 1.5 by mini-batch training using Adam optimisation (learning rate 0.0002) [27], the batch size is 4096, and the epoch is 500. Lastly, the motion-decoder was fine-tuned while training with the regression model.

In the evaluation, test data was fed to the trained regression model, and motion embedding was predicted frame by frame and converted to sequential through the motion-decoder. After that, the output of the prediction model was then joined to form distinct head motion with the overlap-add method and concatenation of 30 time frames, which were fed to the post-filtering autoencoder. The final output for animation was generated with the overlap-add method again.

3.2 Objective Evaluation

Given the fact that body motion is loosely associated with speech and is non-deterministic, it was crucial for us to explore appropriate evaluation measures (e.g. MSE, CCA etc.) [7, 9] for the regression models. Thus, we only conducted a subjective evaluation with other participants of the challenge. For the feature analysis, we employed local CCA [11, 24] with a time window of 300 frames or approximately 13 seconds. For post-filter analysis, we applied mean-squared error (MSE) to measure the value differences to ensure the filtered motion was as smooth and natural as the ground truth.

Feature Analysis: As mentioned in the introduction, we conducted a basic correlation analysis between speech features and body motion in local CCA. Looking at Table 1, it noted that the raw wave-form feature gets a weaker correlation with body motion when stacking more frames. The correlation of MFCC feature remains in the range between 0.4 and 0.5. Our proposed 2 embedded features achieved the highest correlation, a clear and large improvement over the raw wave-form and MFCC.

Filter Analysis: We built a filter-based denoising autoencoder. We tried different frame number $\sim \{30, 50, 70\}$ and frame dimension $\sim \{45, 135\}$. In this experiment, 135-dimension refers to 45-dimension of upper body joints with delta and delta-delta features. We also explored the training model on the ground truth (O), and the ground truth data plus with Gaussian noise (N) of standard deviation 0.1. The dropout rate was set as 0.1 in the training for both. We selected the best model based on the lowest average of MSE on the validation set (without noise) and self-subjective evaluation (minimal fluctuation of synthesised animation). Table 2 shows that the model with the 30-frame-number and 45-frame-dimension obtained the best results.

Table 1: Local CCA of stacking multiple frame between speech information and body motion

Feature	Width	CCA		
		Train	Valid	Test
Wave-form	1	0.624	0.631	-
	3	0.483	0.490	-
	5	0.418	0.426	-
	7	0.	0.004	-
MFCC	1	0.417	0.416	-
	3	0.510	0.512	-
	5	0.522	0.528	-
	7	0.491	0.498	-
Proposed	-	0.751	0.839	-

Table 2: MSE loss for each hyparameter set. The abbreviations of data type indicate original data (O) and noisy data (N).

Hyparameters			MSE Loss		
frame num	frame dim	Data	Train	Valid	Test
70	45	O	0.154	0.252	-
		N	0.153	0.250	-
	135	O	0.151	0.254	-
		N	0.154	0.256	-
50	45	O	0.129	0.189	-
		N	0.130	0.190	-
	135	O	0.136	0.203	-
		N	0.135	0.201	-
30	45	O	0.091	0.120	-
		N	0.090	0.119	-
	135	O	0.110	0.144	-
		N	0.109	0.142	-

4 SUBJECTIVE EVALUATION

We submitted our proposed model to the GENEA2020 challenge and they conducted a perceptual test inspired by the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [28] through the crowd-sourcing platform Prolific (formerly Prolific Academic) in two aspects, human-likeness and appropriateness [29].

There were total of 9 models: 5 models from the participants (including us), 2 baseline models [7, 9], 1 ground truth model and 1 anchor model. The following abbreviations are used to represent each model in the evaluation:

- N : Ground truth.
- M : Anchor (mismatched) natural motion capture from the actor, corresponding to a different speech segment than that played together with the video. This ensures the production of very high-quality motion (same as N), but whose behaviour is completely unrelated to the speech.
- BA : The baseline system [7], which only takes speech audio into account when generating system output
- BT: The baseline system [9], which only takes text transcript information (including word timing information) into account when generating system output
- S... : Participants' submissions (ours is SB).

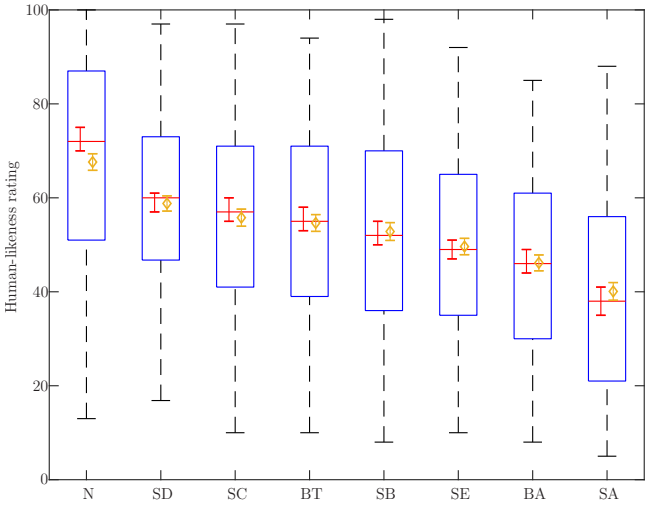

Figure 2: Boxplot visualising the ratings distribution in the human-likeness study. Red bars are the median ratings (each with a 0.01 confidence interval), and yellow diamonds are the mean ratings (also with a 0.01 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each system

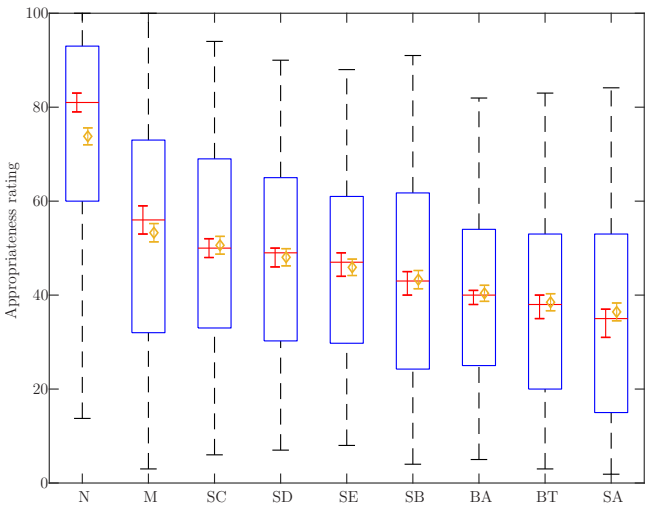

Figure 3: Boxplot visualising the ratings distribution in the appropriateness study. Red bars are the median ratings (each with a 0.01 confidence interval), yellow diamonds are mean ratings (also with a 0.01 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each system.

The evaluation was processed such that every participant was assigned about 10 different speech segments and the corresponding generated motion videos of each segment from different systems. Further, each participant was asked to watch each video and give a score on a 0- to 100-point rating scale that was divided into successive 20-point intervals, which were labelled (from best to worst) 'Excellent', 'Good', 'Fair', 'Poor', and 'Bad'. A total of 125 participants in each study were recruited and asked to follow the instructions to rate each video.

The results of the human-likeness and appropriateness evaluations are shown in Figure 2 and Figure 3, respectively. In terms of sample median, our model (SB) was rated third in human-likeness and fourth in appropriateness among the participants' submissions. Moreover, the sample median score of our model was above BA but below than BT in terms of human-likeness, and above both baselines in appropriateness. These results suggest that our proposed embedded features effectively improved the model generalisation compared to BA, which had similar model structure and ideas as us. Another interesting point noted here is that our model had a larger value range than other models (except N and M). This may indicate that participants had two extreme viewing points in our model, or different output clips were rated more extreme than for other systems.

Figure 4 visualises the (partial) ordering of conditions induced by the significance tests in each study. In the human-likeness, the result shows our model (SB) was not statistically significantly different to BT, but better than BA. In appropriateness, there was not much difference between BA and our model, but better than BT.

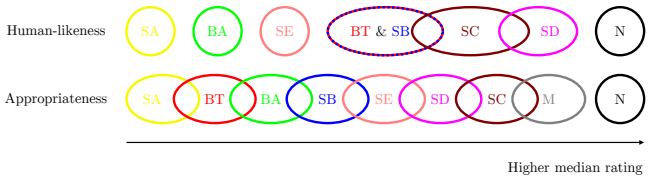

Figure 4: Significance of differences between conditions in the two studies. Each conditions is an ellipse; if two ellipses overlap (or, in one case, coincide), that means that the corresponding conditions were not statistically significantly different at the 0.01 level after Holm-Bonferroni correction. There is no scale on the axis here since the plot only is designed to visualise the partial ordering induced by the significance tests (i.e., ordinal information only).

5 CONCLUSION

In this paper, we extended our previous work to propose a new architecture. The proposed model not only creates highly correlated feature pair but also estimates sequential raw motion data in a frame-based manner. From the objective evaluation, we concluded that Double-DCCCAE enables the creation of a more correlated feature pair, diminishing the side-effect of stacking multiple blocks of speech information and motion data. We showed extensive experiments to select an appropriate neural-network-based filter. Our proposed filter demonstrated a good smoothing effect to the predicted motion in the visualisation. In the subjective evaluation, our model was more preferred than the baseline model (BA) in terms of human-likeness and had similar appropriateness, suggesting that the high correlated feature pair and the sequential estimation helped in improving the model generalisation.

REFERENCES

[1] U. Hadar, T. Steiner, E. Grant, and F. Rose, "Head movement correlates of juncture and stress at sentence level," *Language and Speech*, vol. 26, no. 2, pp. 117–129, 1983.

[2] D. Mcneill, "Hand and mind: What gestures reveal about thought," *Bibliovault OAI Repository, the University of Chicago Press*, vol. 27, 06 1994.

[3] M. Knapp, J. Hall, and T. Horgan, *Nonverbal Communication in Human Interaction*. Cengage Learning, 2013. [Online]. Available: https://books.google.co.uk/books?id=-g7hkSR_mLoC

[4] D. Matsumoto, M. Frank, and H. Hwang, *Nonverbal Communication: Science and Applications: Science and Applications*, ser. EBSCO ebook academic collection. SAGE Publications, 2013. [Online]. Available: https://books.google.co.uk/books?id=PeOeu3qFFTIC

[5] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 708–713.

[6] M. Salem, F. Eyssel, K. J. Rohlfing, S. Kopp, and F. Joublin, "To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, pp. 313–323, 2013.

[7] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *International Conference on Intelligent Virtual Agents (IVA '19)*. ACM, 2019.

[8] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019.

[9] Y. Yoon, W. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4303–4309.

[10] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, "Learning Human Motion Models for Long-term Predictions," *CoRR*, vol. abs/1704.02827, 2017. [Online]. Available: http://arxiv.org/abs/1704.02827

[11] J. Lu and H. Shimodaira, "Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder," *ArXiv*, vol. abs/2002.01869, 2020.

[12] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[13] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, p. III–1310–III–1318.

[14] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational Neural Networks," *CoRR*, vol. abs/1504.07225, 2015. [Online]. Available: http://arxiv.org/abs/1504.07225

[15] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On Deep Multi-View Representation Learning: Objectives and Optimization," *CoRR*, vol. abs/1602.01024, 2016. [Online]. Available: http://arxiv.org/abs/1602.01024

[16] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder," in *Interspeech*, 08 2017, pp. 3991–3995.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *CoRR*, vol. abs/1312.6114, 2014.

[18] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3483–3491. [Online]. Available: http://papers.nips.cc/paper/5775-learning-structured-output-representation-using-deep-conditional-generative-models.pdf

[19] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating kl vanishing," 2019.

[20] 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.

[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1315–1318.

[22] N. Sadoughi and C. Busso, "Novel realizations of speech-driven head movements with generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6169–6173.

[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, 2007.

[24] 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.

[25] J. Lu and H. Shimodaira, "A neural network based post-filter for speech-driven head motion synthesis," *arXiv preprint arXiv:1907.10585*, 2019.

[26] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, ser. IVA '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 93–98. [Online]. Available: https://doi.org/10.1145/3267851.3267898

[27] 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

[28] "Method for the subjective assessment of intermediate quality level of coding systems," recommendation ITU-R BS.1534 https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-1-200301-S!!PDF-E.pdf.

[29] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data," in *Proceedings of the International Workshop on Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents*, ser. GENEA '20, 2020. [Online]. Available: https://genea-workshop.github.io/2020/