ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414660

DOUBLE-DCCCAE: ESTIMATION OF BODY GESTURES FROM SPEECH WAVEFORM

JinHong Lu, TianHang Liu, ShuZhuang Xu, Hiroshi Shimodaira

Centre for Speech Technology Research
The University of Edinburgh
10 Crichton Street Edinburgh EH8 9AB United Kingdom

ABSTRACT

This paper presents an approach for body-motion estimation from audio-speech waveform, where context information in both input and output streams is taken in to account without using recurrent models. Previous works commonly use multiple frames of input to estimate one frame of motion data, where the temporal information of the generated motion is little considered. To resolve the problems, we extend our previous work and propose a system, double deep canonical-correlation-constrained autoencoder (D-DCCCAE), which encodes each of speech and motion segments into fixed-length embedded features that are well correlated with the segments of the other modality. The learnt motion embedded feature is estimated from the learnt speech-embedded feature through a simple neural network and further decoded back to the sequential motion. The proposed pair of embedded features 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 comparable in terms of similar appropriateness.

Index Terms— 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 humans [5, 6].

To tackle the motion-learning challenge for the agents/robots, researchers has explored in different approaches.
Kucherenko *et al.* [7] implemented a speech-to-motion mapping with encoder-decoder DNN. Yoon *et al.* [8] found that the natural language was useful to predict a frame-by-frame poses with a GRU-Auto-Encoder. Ghosh *et al.* [9] proposed a system that generates body motion recursively with a deep LSTM-RNN and a de-noising auto-encoders from a given pose.

These previous studies show the potential of using different types of input to predict body motion pose by pose, 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 the complexity of the model increases as if we would like to generate multiple frames of motion movement [7, 10]. 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 becomes 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 becomes weaker.

To resolve the problems, we propose double deep canonical-correlation-constrained autoencoder (D-DCCCAE), a frame-based system, yet being able to estimate temporal sequence in this paper. Our proposed system consists of three parts, a double deep canonical correlation constrained 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 to the motion data. We submit our model to the GENEA2020 challenge [11] and evaluate it with other participants' models and baseline models in a subjective test.

1.1. Relation To Prior Work

Chandar *et al.* [12] and Wang *et al.* [13] have proposed the idea of Correlated Neural Network (CorrNN), where models the two data streams to be highly correlated in one common subspace through two autoencoders. In our previous work [10], we proposed to use one autoencoder for the head motion estimation; whereas in this present work, we employ two autoencoders for the body motion estimation from speech where context information in both input and output streams is taken in to account. Kucherenko *et al.* [7] has proposed the framework of predicting motion embedding from speech stream and decoding back to the raw motion through a pre-trained decoder, where you can expect high human pose variation of the generated motion. Our work here is different in that we consider context information in the decoder module and the motion is generated in a sequential manner.

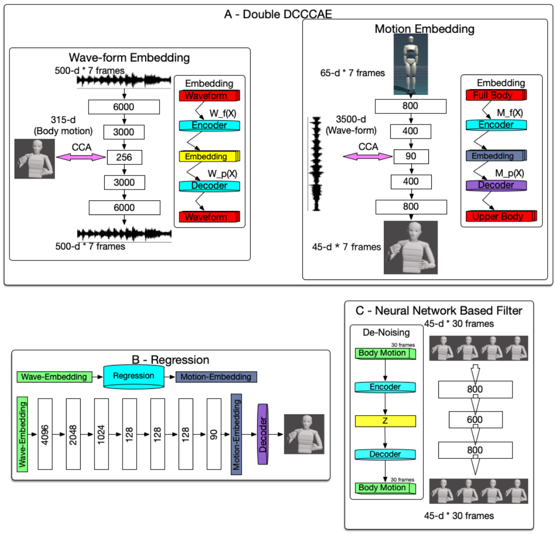

Fig. 1: Overview of the proposed system comprised of three modules: (A) embedding with Double-DCCCAE, (B) DNN-based sequential motion embedding regression from the wave-form embedded features, (C) post filter with an autoencoder.

2. PROPOSED MODEL

Our proposed system can be separated into three modules: 1) double deep canonical-correlation-constrained autoencoders (D-DCCCAE) 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 head motion. The overall framework of our proposed model is shown in Figure 1.

2.1. Double DCCCAE

In our previous work, we compressed high-dimensional wave-forms to low-dimensional and correlated embedding, with head motion using a single autoencoder of CorrNN [10]. However, our work here is different from the aforementioned research studies, in which [12, 13] 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 extend our work here to apply two CorrNN autoencoders for the reason that the dimension of the body motion in this work is much higher than the head motion in our previous work. We compress the information into fixed-length embeddings. 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 separate subspaces to ensure the embedded features are well correlated with the objective features.

We train each proposed DCCCAE with the following objective function:

$$\begin{aligned}\mathrm{Obj}_{\mathrm{DCCCAE}} = \sum_t \|\boldsymbol{X}_{t\pm3} - p(f(\boldsymbol{X}_{t\pm3}))\|^2 \\ -\mathrm{CCA}\left(f(\boldsymbol{X}_{t\pm3}), \boldsymbol{Y}_{t\pm3}\right) \qquad (1)\end{aligned}$$

In the above equation, $\boldsymbol{X}_{t\pm3}$ 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.

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 D-DCCCAE. We map a frame of wave-form embedding to a frame of motion embedding and decode through the motion decoder. The decoded motion is in sequential of multiple frames.

A simple feed-forward deep neural network is applied here for the regression from the wave-form embedded vector to the motion embedded vector. 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 vector and the framed-based mapping between the two embedded features does not have a temporal relationship.

2.3. Neural-Network-Based Filter

The generated trajectories are jerky 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 [10, 7]. We followed the procedure in our previous work [10] and trained a neural-network-based post-filter to overcome these problems in the present study.

3. EXPERIMENT

3.1. Dataset

We have been provided with the Trinity Speech-Gesture Dataset [14] 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. 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 per 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. We extracted 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. Thus, we could not do any objective evaluation for the test data.

We conducted preliminary experiments to decide the depth and width of the Double-DCCCAE, regression models and the post-filter AE, which are shown in Figure 1. 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) [15], the batch size is 4096, and the epoch is 500. Lastly, the motion-decoder was fine-tuned while training with the regression model.

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

Feature	Width	CCA	
		Train	Valid
Wave-form	1	0.624	0.631
	3	0.483	0.490
	5	0.418	0.426
	7	0.	0.004
MFCC	1	0.481	0.481
	3	0.588	0.591
	5	0.602	0.609
	7	0.566	0.574
DCCCAE	1	0.835	0.887
	7	0.687	0.750
D-DCCCAE	7	0.792	0.861

Table 2: Comparison of different systems in terms of performance of body motion prediction, where MSE and local CCA are calculated between predicted body motion and ground truth. '7to7' refers to using 7 frames of the features to estimate 7 frame of the body motion. M_X refers to the regression model is trained with feature X.

Model	Width	Train		Valid	
		MSE	CCA	MSE	CCA
M_{MFCC}	7to7	0.984	0.545	1.202	0.332
M_{DCCCAE}	7to7	0.974	0.563	1.203	0.330
$M_{D-DCCCAE}$	7to7	0.989	0.510	1.203	0.334

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 body 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

To evaluate the performance of the models, we employed mean square error (MSE) and local CCA to compare the estimated and ground truth motion vectors [10, 16].

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

Table 3: Summary statistics of user-study ratings for all conditions in the two studies, with 0.01-level confidence intervals. The human-likeness of M was not evaluated explicitly, but is expected to be very close to N since it uses the same motion clips. F:Input feature, A: Audio feature, T: Text Feature. Our proposed system is SB.

		Human-likeness		Appropriateness	
ID	F	Median	Mean	Median	Mean
N	-	$72 \in [70, 75]$	67.6 ± 1.8	$81 \in [79, 83]$	73.8 ± 1.8
M	-	-	-	$56 \in [53, 59]$	53.3 ± 2.0
BA	A	$46 \in [44, 49]$	46.2 ± 1.7	$40 \in [38, 41]$	40.4 ± 1.8
BT	T	$55 \in [53, 58]$	54.6 ± 1.8	$38 \in [35, 40]$	38.5 ± 1.9
SA	A+T	$38 \in [35, 41]$	40.1 ± 1.9	$35 \in [31, 37]$	36.4 ± 1.9
SB ours	A	$52 \in [50, 55]$	52.8 ± 1.9	$43 \in [40, 45]$	43.3 ± 2.0
SC	A	$57 \in [55, 60]$	55.8 ± 1.9	$50 \in [48, 52]$	50.6 ± 1.9
SD	A+T	$60 \in [57, 61]$	58.8 ± 1.7	$49 \in [46, 50]$	48.1 ± 1.9
SE	A+T	$49 \in [47, 51]$	49.6 ± 1.8	$47 \in [44, 49]$	45.9 ± 1.8

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 waveform and MFCC. Compared to our previous method [10], D-DCCCAE shows improvement over DCCCAE for 7 frames, but is comparable for 1 frame.

Model Comparison: After the analysis above, we conducted the evaluation of the body motion estimation for MFCC, DCCCAE and D-DCCCAE features. Table 2 shows the result of different systems with different number of the input and output frames, where were examined in MSE and local CCA. $M_{D-DCCCAE}$ has the highest MSE and lowest CCA in the train set, but achieves the highest CCA in the valid set and MSE is similar among the three models for the valid set. This indicates that our proposed model has the potential to perform better if the generalisation of the model is better.

4. SUBJECTIVE EVALUATION

We submitted our proposed model to the GENEA2020 challenge [11] and they conducted a perceptual test inspired by the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [17] through the crowd-sourcing platform Prolific (formerly Prolific Academic) in two aspects, human-likeness and appropriateness [18].

There were total of 9 models: 5 models from the participants (including us), 2 baseline models [7, 8], 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 [8], which only takes text transcript information (including word timing information) into account when generating system output
- S... : Participants' submissions (ours is SB).

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 Table 3. We are the one of the two teams, who used audio feature only, in the submissions. 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 suggests that our proposed embedded features effectively improved the model generalisation compared to BA, which had similar model structure and ideas as us.

5. CONCLUSION

In this paper, we extended our previous work to propose a new architecture. The proposed model not only creates a highly correlated feature pair, but also generates sequential raw motion data in a frame-based manner. From the objective evaluation, we concluded that D-DCCCAE enables the creation of a more correlated feature pair, diminishing the side-effect of stacking multiple blocks of speech information and motion data. D-DCCCAE achieves the highest CCA in the valid set in the model comparison. In the subjective evaluation, our model achieved higher score than BA in both aspects (human-likeness and appropriateness) according to the participants' preference, suggesting that the high correlated feature pair and the sequential estimation helped in improving the model generalisation. In future, we could consider exploring higher stacking to unearth the potential of D-DCCCAE.

6. REFERENCES

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

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

[3] M.L. Knapp, J.A. Hall, and T.G. Horgan, *Nonverbal Communication in Human Interaction*, Cengage Learning, 2013.

[4] D. Matsumoto, M.G. Frank, and H.S. Hwang, *Nonverbal Communication: Science and Applications: Science and Applications*, EBSCO ebook academic collection. SAGE Publications, 2013.

[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, Katharina J. Rohlfing, Stefan 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] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström, “Analyzing input and output representations for speech-driven gesture generation,” in *International Conference on Intelligent Virtual Agents (IVA ’19)*. 2019, ACM.

[8] 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.

[9] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges, “Learning Human Motion Models for Long-term Predictions,” *CoRR*, vol. abs/1704.02827, 2017.

[10] JinHong Lu and Hiroshi Shimodaira, “Prediction of Head Motion from Speech Waveforms with a Canonical-Correlation-Constrained Autoencoder,” pp. 1301–1305, 2020, Proc. Interspeech 2020.

[11] “Genea challenge 2020,” Feb 01, 2021, https://genea-workshop.github.io/2020/.

[12] Sarath Chandar, Mitesh M. Khapra, Hugo Larochelle, and Balaraman Ravindran, “Correlational Neural Networks,” *CoRR*, vol. abs/1504.07225, 2015.

[13] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes, “On Deep Multi-View Representation Learning: Objectives and Optimization,” *CoRR*, vol. abs/1602.01024, 2016.

[14] Ylva Ferstl and Rachel McDonnell, “Investigating the use of recurrent motion modelling for speech gesture generation,” in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, New York, NY, USA, 2018, IVA ’18, p. 93–98, Association for Computing Machinery.

[15] Diederik P. Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[16] Kathrin Haag and Hiroshi 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.

[17] “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.

[18] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje 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*, 2020, GENEA ’20.