Speech-driven head motion generation from waveforms

Speech Communication 159 (2024) 103056

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech-driven head motion generation from waveforms

JinHong Lu *, Hiroshi Shimodaira

Centre for Speech Technology Research, University of Edinburgh, United Kingdom

ARTICLE INFO

Keywords:
Neural network
Head motion synthesis
Waveform

ABSTRACT

Head motion generation task for speech-driven virtual agent animation is commonly explored with handcrafted audio features, such as MFCCs as input features, plus additional features, such as energy and F0 in the literature. In this paper, we study the direct use of speech waveform to generate head motion. We claim that creating a task-specific feature from waveform to generate head motion leads to better performance than using standard acoustic features to generate head motion overall. At the same time, we completely abandon the handcrafted feature extraction process, leading to more effectiveness. However, the difficulty of creating a task-specific feature from waveform is their staggering quantity of irrelevant information, implicating potential cumbrance for neural network training. Thus, we apply a canonical-correlation-constrained autoencoder (CCCAE), where we are able to compress the high-dimensional waveform into a low-dimensional embedded feature, with the minimal error in reconstruction, and sustain the relevant information with the maximal cannonical correlation to head motion. We extend our previous research by including more speakers in our dataset and also adapt with a recurrent neural network, to show the feasibility of our proposed feature. Through comparisons between different acoustic features, our proposed feature, Wav_{CCCAE}, shows at least a 20% improvement in the correlation from the waveform, and outperforms the popular acoustic feature, MFCC, by at least 5% respectively for all speakers. Through the comparison in the feedforward neural network regression (FNN-regression) system, the Wav_{CCCAE}-based system shows comparable performance in objective evaluation. In long short-term memory (LSTM) experiments, LSTM-models improve the overall performance in normalised mean square error (NMSE) and CCA metrics, and adapt the Wav_{CCCAE}feature better, which makes the proposed LSTM-regression system outperform the MFCC-based system. We also re-design the subjective evaluation, and the subjective results show the animations generated by models where $Wav_{CCCAE}$was chosen to be better than the other models by the participants of MUSHRA test.

1. Introduction

Communication, whether in whatever way, is deemed an essential part in existing civilisation, consisting of verbal and nonverbal forms. Hadar et al. (1983) argue that one important nonverbal form, head motions, directly contribute to speech production. Research into head motions could be going through enormous changes as the synthesis of head motion moves towards fully operational and interactive implementation of its potential. Such niche technology is being tested to apply both head motion and lip-syncing to the creation of a more human-like avatar. However, compared with lip-syncing, we may not reach the point at which head motions could be easily captured and analysed due to a weak link between speech and head motion and a complex collective of speech, emotion, intention, and stance.

Stimulated by methods of input, speech-based (Busso et al., 2005; Gregor et al., 2007) and text-based (Zhang et al., 2007; Jia et al., 2014) approaches are then endowed with research priority. In the process of text mining, a generation algorithm is operating to predict the head movement. On the other hand, a speech-based approach mainly targets collecting acoustic features after audio acquisition, trying to build a model linking head movement and acoustic signature. With such a system running, researchers can obtain the predicted head movement as output while taking the acoustic patterns as input.

The speech-based approach has been investigated with different inputs, e.g. MFCC (Ding et al., 2015), EMA (Ben Youssef et al., 2014), F0 and intensity (Sadoughi and Busso, 2018) etc. The reason for choosing these acoustic features is mainly due to the analysis of the correlation between acoustic features and head motion, which has been researched in previous studies (Hadar et al., 1983; Kuratate et al., 1999; Yehia et al., 2002; Munhall et al., 2004; Busso and Narayanan, 2007; Ishi et al., 2007; Ishi et al., 2014; Sadoughi and Busso, 2017). As all of these acoustic features are derived from speech waveforms, it would be a good consideration to choose waveforms as the input to the system.

* Corresponding author.
E-mail address: jinhong.l@ed.ac.uk (J. Lu).

https://doi.org/10.1016/j.specom.2024.103056
Received 9 February 2023; Received in revised form 4 February 2024; Accepted 25 February 2024
Available online 1 March 2024

This can leverage the advantages of each acoustic feature, maximising the correlation between features and head motion by increasing the amount of information available to predict head motion.

Waveforms intended to probe head motion are rarely noticed on account of their high dimensionality and staggering quantity of irrelevant information, implicating potential cumbrance for neural networks training and high demand for hardware support. A canonical-correlation-constrained autoencoder (CCCAE) (Lu and Shimodaira, 2020) is then proposed by us to address problems related to high dimensionality and intricate information. It is expected to identify low-dimensional patterns from waveforms by training hidden layers to minimise errors in reconstruction as well as to maximise the canonical correlation with head motion. Predictions for head motion are then possible after the processing approach of the extracted low-dimension features happening in another regression neural network. Features obtained in that way proved to be more useful than those acquired from a previous standard autoencoder, with comparisons with other acoustic features in place. In this paper, we continue from the previous research to explore at a deeper level. First, we have done a feature analysis among features to further show the effectiveness of the objective loss of CCA at the personality level. Then we expand the training data to include more speakers for the model feasibility and upgrade the regression model from FNN to LSTM. In the meantime, we also build an external architecture (Haag and Shimodaira, 2016) for comparison. Last, we conduct an extensive subjective evaluation for two aspects, motion appropriateness and model assessment.

The rest of the paper is organised as follows. Section 3 describes the architecture of the proposed approach. Sections 4, 5 and 6 discuss the setting of the experiments and analyse the results. Subjective Evaluation is presented in Section 6.3. Finally, the overall conclusion, limitations and future work are presented in Section 7.

2. Previous and related work

Researchers have been exploring the relationship between speech and gestures over several decades (Birdwhistell, 1952; Bolinger, 1983; Bolinger and Bolinger, 1986; Hadar et al., 1983; Kuratate et al., 1999; Yehia et al., 2002; Munhall et al., 2004; Busso and Narayanan, 2007; Ishi et al., 2007; Ishi et al., 2014; Sadoughi and Busso, 2017). The findings strongly suggest that there is a link between speech and head motion. With this supporting evidence, speech-driven head motion prediction was investigated with different acoustic features (Ben Youssef et al., 2014; Ding et al., 2015; Haag and Shimodaira, 2016; Sadoughi and Busso, 2018). However, our assumption is that regarding those handcrafted features (e.g. MFCCs, Fbank etc.) that are extracted from waveforms, it is unclear to us that whether the handcrafted information is related to the head motion. Thus, to make full use of the information in the original observations, it is apparent to input waveforms directly into the neural network.

2.1. Correlation between speech and head motion

One of the earliest studies concerning the correlation between prosody and gesture was created by Birdwhistell (1952). It suggested that there is an alignment between gestural movements and intonation. Bolinger (1983) and Bolinger and Bolinger (1986) observed that gestures followed pitch contours up and down, in their main direction of movement. Hadar et al. (1983) also showed that speakers' head movements move along with the changes of the prosody, which peaks and falls more noticeably in cases of high intensity.

In more recent years, further analyses have been made. Kuratate et al. (1999) found that the sentence-level correlation between fundamental frequency (F0) and head motion was 0.83, but they also claimed that this analysis was sensitive to the absolute values, rather than the spatiotemporal patterning, of head postures. Yehia et al. (2002) analysed the correlation between head motion and speech over the fundamental frequency (F0) by experimenting with one American English speaker (ES) and one Japanese speaker (JS), monitoring their reading speech utterances. They showed that the correlation among F0 and the 6 DOF (degrees-of-freedom) (3 DOF for rotation and 3 DOF for translation) of head motion was between 0.39 and 0.52 for ES, and between 0.22 and 0.30 for JS, which, on average, is less than 0.50. Munhall et al. (2004) reported that the correlations between head motion (in 6 DOF) and pitch, and amplitude of the talker's voice, were almost always over 0.50, on average about 0.63 in sentence-level, in Japanese read-speech utterances. Busso and Narayanan (2007) presented an Audio-Visual Mapping Framework, which maps the acoustic features onto the facial features space, producing the estimated facial features through an affine minimum mean square error estimator (AMMSE). These estimated facial features were then used to compute the Pearson's correlation with the real facial features. They showed that head motion and mel-frequency cepstral coefficients (MFCCs) had a strong sentence-level correlation of 0.8 after the mapping, for which short-sentenced scripted audio data was used.

Overall, the above studies have demonstrated a high correlation between scripted-speech and head motion. However, our experiments pointed out that the variation of the degree in head motion is much larger in natural human conversation and is impossible to spot such strong correlation between speech and head motion in the same degree. There are other studies to support our hypothesis as well. Ishi et al. (2007) showed an analysis of spontaneous dialogue speech data from one Japanese female speaker and claimed that a strong relationship could not be found between head motion and prosodic features. Sadoughi and Busso (2017) reported that the original head movements and speech (F0 and energy with their delta and delta-delta features) had a global correlation of 0.1931 with the dyadic interactions data.

2.2. Speech-driven head motion system

Researchers have been investigating different acoustic features and their combinations to resolve the weak correlation between speech and head motion for decades. Ben Youssef et al. (2014) built HMM-based acoustic-to-articulatory inverse mapping to predict the articulatory features from speech. The estimated articulatory feature vectors are represented by the trajectories of (x,y)-coordinates of the 6 active EMA coils and are subsequently used to predict head motion through multi-stream HMMs. In their results, they showed that the estimated articulatory feature vectors are more correlated with head motion than acoustic features in local CCA. They also showed that the correlation between the estimated head motion, using articulatory features, and original head motion (or speech), is higher than the estimated head motion using acoustic features. Ding et al. (2015) explored acoustic features (including LPC, MFCC, and filter bank (FBank)) with deep neural networks and showed that the FBank-based systems achieved the highest correlation between the predicted and original head motions. Haag and Shimodaira (2016) built bottleneck features from the combination of MFCC and EMA features, which were treated as an input to deep-BiLSTM and used to generate head motion. Sadoughi and Busso (2018) built a conditional-GAN with BLSTM using F0 intensity (plus first and second derivatives) as an input feature to predict head motion. They claimed that the proposed system outperformed the normal BLSTM architecture models. Greenwood et al. (2017) proposed CVAE-BiLSTM and fed the Fbank features as a condition to the decoder for generating head motion. Ahuja et al. (2022) proposed a new architecture to combine diffusion and discriminator together in gesture generation using text and audio features. They showed that the proposed approach can effectively address the shift in crossmodel grounding and the output distribution from the source to the target speaker with only a few minutes of data. Fares et al. (2022) proposed a transformer-based multimodal with text and speech F0 input to generate facial expression and head motion. They claimed that the results of head motion generation are

lower in difference and higher in correlation to the ground truth with the proposed architecture compared to the LSTM baseline.

All these previous studies indicate that the combinations of the different features outperform a single feature such as MFCC. However, those different features are intrinsically extracted from waveforms. This gives the motivation of the present study, which aims to directly extract useful features from waveforms to predict head movements. We do not consider those latest architectures, because it is not the purpose of the present study to evaluate different speech features with the latest architecture. Instead, we aim to extend the previous research Lu and Shimodaira (2020) and conduct more detailed analyses and evaluations of the proposed feature.

It should be noted that there is a recent trend in the speech community towards utilisation of representation learning techniques, which can automatically learn an intermediate representation of the input signal that better suits the task at hand, hence leading to improved performance. A common approach for implementing such representation learning is to build an upstream model and a downstream model (Chung and Glass, 2020; rahman Mohamed et al., 2022). The upstream model utilised information extracted from the input data itself, and then the downstream model used either the learned representation from the frozen upstream model or fine-tuned the entire pre-trained model in a supervised phase. However, both ways of training have drawbacks in that either the learned representation was not related to the downstream tasks or the fine-tuning process required heavy computational cost. Thus we would like to propose a system which could extract meaningful representation related to the downstream task without fine-tuning. Our recent work (Lu and Shimodaira, 2020; Lu et al., 2021) brought the idea of the Correlational Neural Network (CorrNN) to fulfil our goals.

2.3. Correlational neural network

CCA was first introduced by Hotelling (1936) and Anderson (2009) to find linear projections of two random vectors that are maximally correlated in standard statistics. CCA is useful in learning representations of two data views such that each view's representation is simultaneously the most predictive of, and the most predictable by, the other (Andrew et al., 2013). Andrew et al. (2013) raised the idea of a deep Canonical Correlation Analysis (DCCA), which correlates the two resulting embedded features into a common subspace with a strong linear relationship. Chandar et al. (2015) and Wang et al. (2016) extended the idea of DCCA into a Deep Canonically Correlated Autoencoder (DCCAE), which is to not only maximise the correlation between the two 'bottleneck' features, but also minimise the reconstruction error of the autoencoders. They proved DCCAE to be effective in cross-language tasks and multi-view feature learning, and capable in the creation of highly correlated feature pairs. The differences between the above work and the present work are that both input and output in our work are streaming data and the correlation between speech and head-motion features is much weaker. These differences make the task to be more careful in designing and training the system.

3. Proposed system

Waveforms have been successfully applied in several speech tasks, and one of the more common ways of extracting a representation from a waveform in these end-to-end systems is to use generative models (e.g. autoencoder, GAN) (Phan et al., 2020; Chorowski et al., 2019) and convolutional-based neural networks (Sainath et al., 2015; Ghahremani et al., 2016; Tüske et al., 2018; Hoshen et al., 2015; Loweimi et al., 2020). These approaches have shown a significant improvement by creating and using task-specific features over other approaches using hand-crafted features (e.g. MFCC and Filter bank (Fbank)). Another advantage of these end-to-end systems is that they completely abandon the hand-crafted feature extraction process (Chung and Glass, 2020; rahman Mohamed et al., 2022). However, it is not easy to directly apply them in our head motion generation task because previous literature shows a weak correlation between speech and head motion (Ishi et al., 2007; Sadoughi and Busso, 2017). The standard bottleneck representation/direct training from waveforms may extract irrelevant information and hinder the generation process (Lu and Shimodaira, 2020). To resolve this problem, we proposed building CCCAE (Lu and Shimodaira, 2020; Lu et al., 2021) on top of the correlational neural network, compressing the input into a low-dimensional representation to raise high correlations between the compressed representations and the target data stream.

Fig. 1 shows an overview of our proposed system, which can be viewed as three main modules: (1) a canonical-correlation-constrained autoencoder (CCCAE) for compressing from a high dimension to distribute to a low dimension whilst sustaining correlated information between the waveform input and head motion; (2) a regression neural network for generating the head motion from the compact and correlated embedding; (3) a neural-network-based post filter for constructing smooth head motion from the generated output. In the training procedure, we apply mean square error (MSE) normalised by the variance of the ground truth for these three models.

3.1. Waveform embedding

Chandar et al. (2015) and Wang et al. (2016) started the idea of compressing a set of two data streams into a common subspace with autoencoders. Wang et al. (2016) proposed DCCAE, which consists of two autoencoders for the two data streams respectively. His idea was to not only maximise the correlation between the two 'bottleneck' features, but also to minimise the reconstruction error of the autoencoders. In our study, head motion does not require further dimensional reduction with an autoencoder as it is computed with a time series of rotation vectors of three dimensions. CCCAE (Lu and Shimodaira, 2020), in our present work differs from the original idea of DCCAE (Wang et al., 2016); we only employ a single autoencoder, in which hidden layers are trained in optimising the combination of the correlation between the embedded features and head motion, and the reconstruction errors. Thus, the task of the study now is to project waveforms to a subspace with a single autoencoder, with which the resulted embedded features have a high correlation with head motion. Those more advanced architectures such as VAE/CVAE (Greenwood et al., 2017; Kingma and Welling, 2014; Sohn et al., 2015) are not in our consideration because this type of autoencoders tend to map the input to a latent space that corresponds to the parameters of a variational distribution, whereas our aim is to do feature extraction and information retrieval.

The following objective function is used to train our proposed CCCAE,

$$\mathrm{Obj}_{\mathrm{CCCAE}} = \mathrm{Obj}_{\mathrm{AE}} - \alpha\,\mathrm{CCA}\left(f(\boldsymbol{X}), \boldsymbol{Y}\right) \tag{1}$$

$$\mathrm{Obj}_{\mathrm{AE}} = \sum_{t} \|\mathrm{X}_t - p\left(f\left(\mathrm{X}_t\right)\right)\|^2 \tag{2}$$

where the input waveform vector is viewed as $\boldsymbol{X}_t$ at a time stamp t to the encoder , while $f(\,)$ represents the projection with the encoder, $p(\,)$ represents, therefore, the reconstruction with the decoder, $\boldsymbol{X}$ represents the input sequences of waveform vectors and $\boldsymbol{Y}$ denotes the corresponding head motion vectors, and CCA$(\,)$ is the canonical correlation function. $\alpha \geq 0$ is the weighting factor, where $\alpha = 0$ corresponds to a standard autoencoder with an MSE loss function, $\mathrm{Obj}_{\mathrm{AE}}$. For $\mathrm{Obj}_{\mathrm{CCCAE}}$, the α is set to be 1.

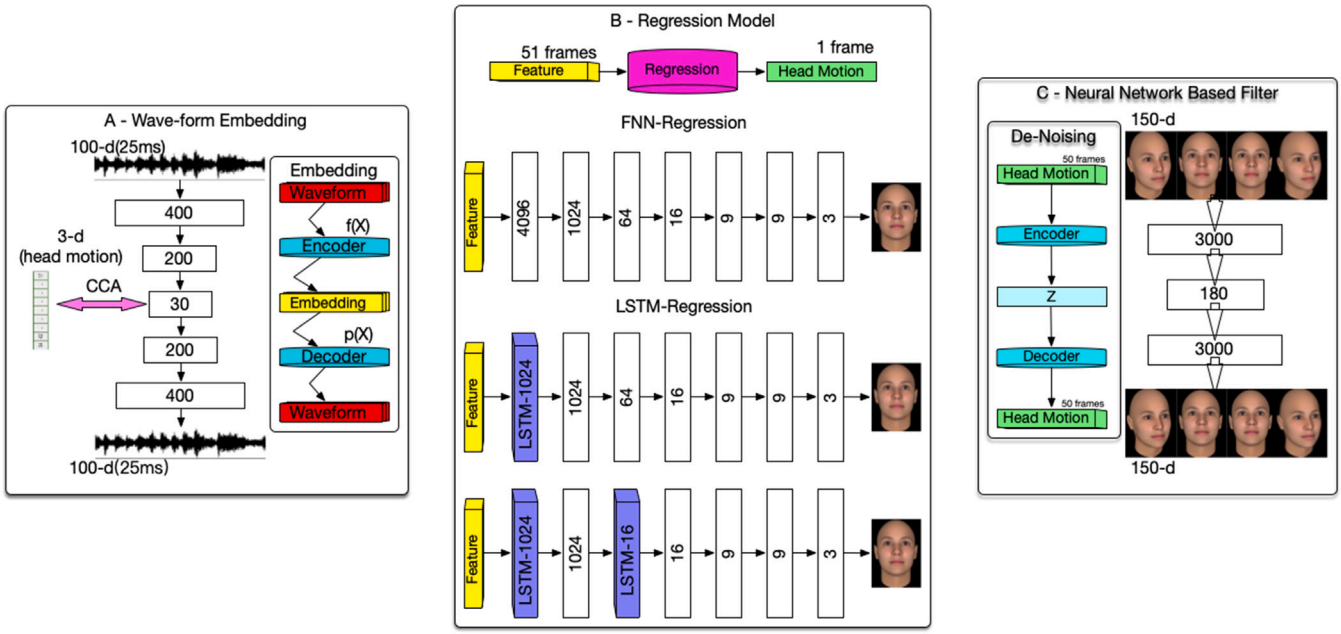


Fig. 1. Overview of the proposed system comprised of three modules: (A) waveform embedding with CCCAE, (B) head motion regression from the features, (C) post filter with an autoencoder.

3.2. Head motion regression

To predict head movements from the proposed feature, we considered two model architectures depending on the use of context. A feed-forward neural network (FNN) was employed as the model that uses short temporal information with a sliding time window, whereas a long short-term memory (LSTM) network was chosen as the one that uses longer temporal information. We did not consider the auto-regressive (AR) model or other advanced models, because it is not the purpose of the present study to seek better models.

3.2.1. Feed-forward neural network (FNN)

The regression model is constructed by 7 feed-forward layers with different numbers of hidden nodes to predict head motion from the waveform embedded features, shown in Fig. 1(B). The architecture and hyperparameters are the same as in the previous work (Lu and Shimodaira, 2020), and we take it as the baseline model.

3.2.2. LSTM

Different from the FNN baseline model, we have built two LSTM-based models: 1-Layer-LSTM (replaced 4096-FNN) and 2-Layer-LSTM (replaced 4096-FNN and 64-FNN).

3.3. Post-filter

Previous research (Haag and Shimodaira, 2016; Ding et al., 2015; Busso and Narayanan, 2007; Sadoughi and Busso, 2018) commonly applied a linear filter or key frames filter to smooth the generated head motion as the generated trajectories were noisy or discontinuous due to the nature of speech. However, there are limitations of such linear/key frame filters. These filters do not know the characteristic of the head motion trajectories, which may result in filtering the key frames instead. Therefore, we built a neural network based de-noising autoencoder based on our previous study (Lu and Shimodaira, 2019), which suggests training the model with the 'clean' data.

4. Experimental setup

4.1. Data

The dataset used in this work is the University of Edinburgh Speaker Personality and Mocap Dataset (Haag and Shimodaira, 2015), which contains 13 native English speaking semi-professional actors' expressive dialogues. These actors were asked to perform with extroverted, introverted and natural speaking styles in non-scripted and spontaneous dialogues. Ishi et al. (2014) proved that intra- and inter-speaker variability can have the effect of varying head motion. Speaker-dependent practice was carried out in our training for each speaker as this is common in speech-driven head motion synthesis (Busso et al., 2005; Ding et al., 2015; Sadoughi and Busso, 2018). We selected a total of six speakers, three males (Speaker A, B, C) and three females (Speaker D, E, F), from the dataset whereas the remaining seven speakers' recordings do not fulfil the amount of training data in our work. A total of 10 data files (around 50 min long) for each speaker were split into 6/2/2 for training, validation and testing respectively without considering the speaking styles of each data file. IEMOCAP, a benchmark dataset in the field, was chosen to be an external data against which to evaluate our proposed method. The female speaker's recordings of Section 1 in IEMOCAP were selected and followed the same method of splitting the data as Sadoughi and Busso (2018).

Speech Features Audio in the database was recorded with a headset microphone at 44.1 kHz with 32-bit depth and a MOTU 8pre mixer. Separate recording channels were used for the two speakers and a synchronisation signal was recorded on a third channel in the mixer. For the purpose of this work, the audio signal was downsampled to 4 kHz prior to the feature extraction. In this work, we followed the same configuration as the previous work (Lu and Shimodaira, 2020), 25 ms windows with 10 ms shifting, to pre-process and extract the acoustic features, 100-dimension of waveform and 39-MFCCs, for each speaker. Moreover, we also extracted 6-dim F0+Energy, 27-dim Fbank and 100-dim waveform to enable feature analysis with the proposed feature and

MFCCs. All of these features were normalised in terms of variance for each dimension.

Head Motion Features Movements of the head as a 3D rigid-body were recorded with the NaturalPoint Optitrack motion capture system at a 100 Hz sampling rate. From the marker coordinates, rotation matrices for the head motion were computed using singular value decompositions (Soderkvist and Wedin, 1994), which were further converted to rotation vectors of three dimensions.

4.2. Model setup

Fig. 1 shows the depth and width of the models, which are decided by the preliminary experiments we conducted. To ensure the robustness of the model training and the relationship between speech and head motion was learnt by the model through the training, we only applied the speaking frame of the target speaker for head motion prediction. During the inference time, we made use of all the frames (including both speaking and listening) to the model for generating head motion trajectories. The model notations described below are used in the rest of the sections, where 'XX' donates a speech feature such as MFCC.

- Wav_{AE}: Embedded features extracted from waveform with the standard autoencoder, Obj_{AE}
- $\text{Wav}_{\text{CCCAE}}$: Embedded features extracted from waveform with the proposed CCCAE, $\text{Obj}_{\text{CCCAE}}$ with $\alpha = 1$
- FNN_{XX}: FNN-regression model trained with XX feature
- LSTM^{1}_{XX}: Regression model with 1-Layer-LSTM trained with XX feature
- LSTM^{2}_{XX}: Regression model with 2-Layers-LSTM trained with XX feature
- BiLSTM_{XX}: Regression model that adapts from external (Haag and Shimodaira, 2016) with few modifications trained with XX feature

FNN_{MFCC}, FNN_{AE}, and $\text{FNN}_{\text{CCCAE}}$ use the FNN-regression network in Fig. 1(B) to generate head motion and the difference is the size of the input layer of each model. Tensorflow version 1.12 was used in building and training with Adam optimisation (learning rate 0.0002) on a GPU machine and a multi-CPU machine. Layer-wide pre-training technique (Takaki and Yamagishi, 2016) was also employed during training.

In the inference time, test data of the same speaker is first fed to the trained CCCAE to extract the embedded feature, and then to generate head motion frame by frame from the trained regression model. The generated head motion is then concatenated 50 time frames to be a distinct head motion and fed to the de-noising model consecutively. Lastly, the overlap-add method is applied to average the filtered head motion for animation.

4.3. Objective measures

To evaluate the similarity between two sequences of vectors, we employed a normalised mean-squared error (NMSE), where MSE is normalised by the variance of ground truth, local canonical correlation analysis (local CCA) (Haag and Shimodaira, 2016), and KL (Kullback–Leibler) divergence. The difference between global CCA and local CCA is that global CCA measures the correlation over the whole sequence, whereas local CCA only calculates the sub sequence's CCA score within a time window and then takes the mean value of all the obtained scores. The reason of selecting local CCA is that there is rarely linear correlation held over long sequences, which calculated by global CCA, as the head motion trajectories are changing over times. We used a time window of 300 frames (or 3 s) with a 50% overlap. We use the following formula to calculate local CCA:

$$T = \{T_1, T_2, T_3, \ldots ..., T_n\}$$
$$\mathrm{r}_{Average} = \frac{1}{|T|}\frac{1}{d}\left(\sum_{t\in T}\sum_{i=1}^{d}\mathrm{corr}\left(A^{[i]}X_{[t:t+l-1]}, B^{[i]}Y_{[t:t+l-1]}\right)\right) \quad (3)$$

where $A^{[i]}$, $B^{[i]}$ are the ith canonical coefficients obtained in the global CCA, d is the dimension of features, n denotes the number of time windows and T_k, $k = 1.....n$, is the start frame of the kth time window such that $T_1 = 0$, $T_2 = 150$, $T_3 = 300$, and so on. l is the window length.

KL divergence is used to measure the similarity between two probability distributions. It is useful in our evaluation because it shows whether there is the capacity for common patterns in the acoustic features and personal dependency of the different speakers. Such personal dependency would result in distinct pattern distribution in the later motion generation. In the evaluation, we applied symmetric KL divergence, which is defined below:

$$\text{Symmetric}_{KL} = \mathrm{D}_{KL}(P \parallel Q) + \mathrm{D}_{KL}(Q \parallel P) \quad (4)$$

where P, Q are the two probability distributions.

Table 1
Comparison of different widths of $\text{Wav}_{\text{CCCAE}}$, where NMSE and local CCA are calculated between $\text{Wav}_{\text{CCCAE}}$ and the original head motion for Speaker A. $_{xk}$ represents the audio sample rate.

Width	NMSE			CCA		
	Train	Valid	Test	Train	Valid	Test
15_{4k}	0.411	0.507	0.480	0.245	0.216	0.219
30_{4k}	0.173	0.239	0.221	0.264	0.234	0.248
60_{4k}	0.233	0.261	0.250	0.220	0.194	0.194
30_{16k}	0.219	0.278	0.247	0.267	0.282	0.281

5. Analysis of the speech feature

In the following section, we first tackled the high dimensional problem of the waveform. Next, we made comparison between hand-crafted audio features and our proposed feature in two aspects: (1) the correlation between audio feature and original head motion; (2) the analysis of the features in speaker dependency.

5.1. Width of the proposed embedding

High dimensionality has been affecting the popularity of the usage of waveform as the input to neural networks, even though the waveform contains the original information of the acoustic features. Here, we seek to resolve this problem by using our proposed model, CCCAE. In the previous section, CCCAE has been described as not only reducing the dimension of the input feature effectively, but also maximising the correlation between the embedding feature and the target. We first explore the possible dimensions of embedding features with sizes of 15, 30 and 60 compared to the dimension of the waveform is 100. This could give us a clear idea about the trade-off between the recovery of waveforms and the correlating information.

Looking at Table 1, the result of the validation set shows that the higher the dimension of the embedding feature is, the better the recovery of the waveform is. It is clear that the size 15 is the worst in term of recovering the waveform as there is too little information. On the other hand, the size 60 is the least correlated to the head motion because there is still too much irrelevant information. Overall, the results show that the size of sample 30 is the best choice to provide the clearest results. The result of the test set is provided as well but is not involved in selecting the architecture. Furthermore, we argue that

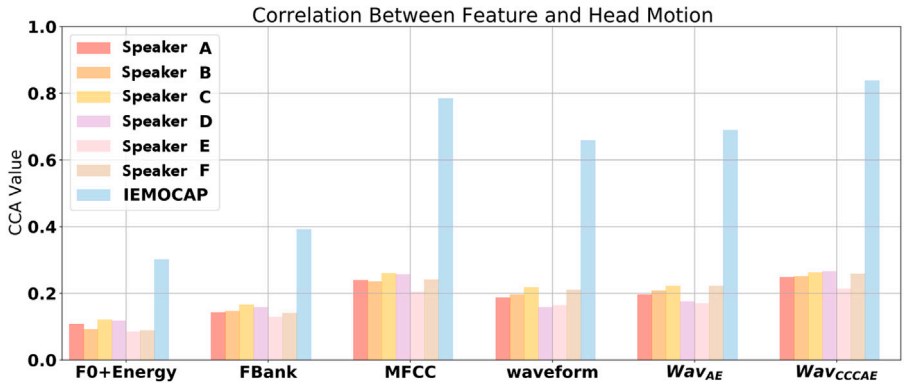


Fig. 2. Local CCA between speech features and original head motion for the test set.

Table 2
Average symmetric KL divergence over speakers, to indicate the similarity of the feature distribution in all dimensions whether there is common pattern in the acoustic feature among speakers.

Measure	Feature					
	F0+Energy	FBank	MFCC	Waveform	Wav_{AE}	Wav_{CCCAE}
Symmetric KL	6.59	11.87	6.00	5.69	6.96	8.98

audio with the sample rate of 4 kHz has a huge information loss, we also investigate the audio with the sample rate of 16 kHz with the same architecture. We notice that there is not much difference in terms of NMSE with the size of 30 for different sample rates. The local CCA for the valid and test data have about 20% and 16% improvement respectively. However, due to the computational cost, we can only use the audio with the sample rate of 4 kHz in the rest of the paper.

5.2. Local CCA between speech feature and original head motion

A basic correlation analysis of local CCA was carried out between speech features and head motions before the regression training and evaluation. Results of local CCA for each speech feature and for each speaker are displayed in Fig. 2. The findings suggest that F0+Energy gives the smallest score, and MFCC achieves the largest in the hand-crafted features, and Wav_{CCCAE} achieves the largest in all the features. Comparing the waveform and the proposed feature, we can see a large improvement (at least 30% is achieved on average) in the results of the test set with Wav_{CCCAE} for each speaker, but only a small improvement is made with Wav_{AE}.

In the meantime, we have used an external dataset, IEMOCAP (Busso and Narayanan, 2007), to evaluate our proposed model. We then calculated a similar correlation result between MFCC and head motion for the findings reported by Busso and Narayanan (2007). Our proposed feature has an improvement of 6% respectively for MFCC and 27% respectively for waveform, respectively.

5.3. Analysis of the features

We visualised the features using T-SNE, where the proposed features were taken from the third layer of DCCCAE (where the layer is specified as 30-d in Fig. 1). Observing at Fig. 3, we noticed the pattern that the sparser the point that each speaker's feature is, the lower the CCA between the feature and the head motion. Moreover, Wav_{AE} and Wav_{CCCAE} show the effects of gathering these points compared with waveforms. This is also shown in Table 2, which shows the average symmetric KL divergence between the speakers. In each feature, the smaller the value is, the larger the overlapped area of the distributions. In terms of the values, we can observe that the value of Wav_{CCCAE} is much larger than Wav_{AE}, this refers that a smaller overlapped area for Wav_{CCCAE}. Thus, we believe that Wav_{CCCAE} has a better effect than Wav_{AE} showing a clear distinct cluster for each speaker. We also notice that FBank feature has the largest value in this KL divergence result, but it has the second lowest correlation in Fig. 2. This implies that little speaker-independent information is carried in terms of head motion. It is because those bank pass filters are designated to capture the information related to the human vocal tract, which is one of the main distinguishing characteristics of individual (Chougule et al., 2014).

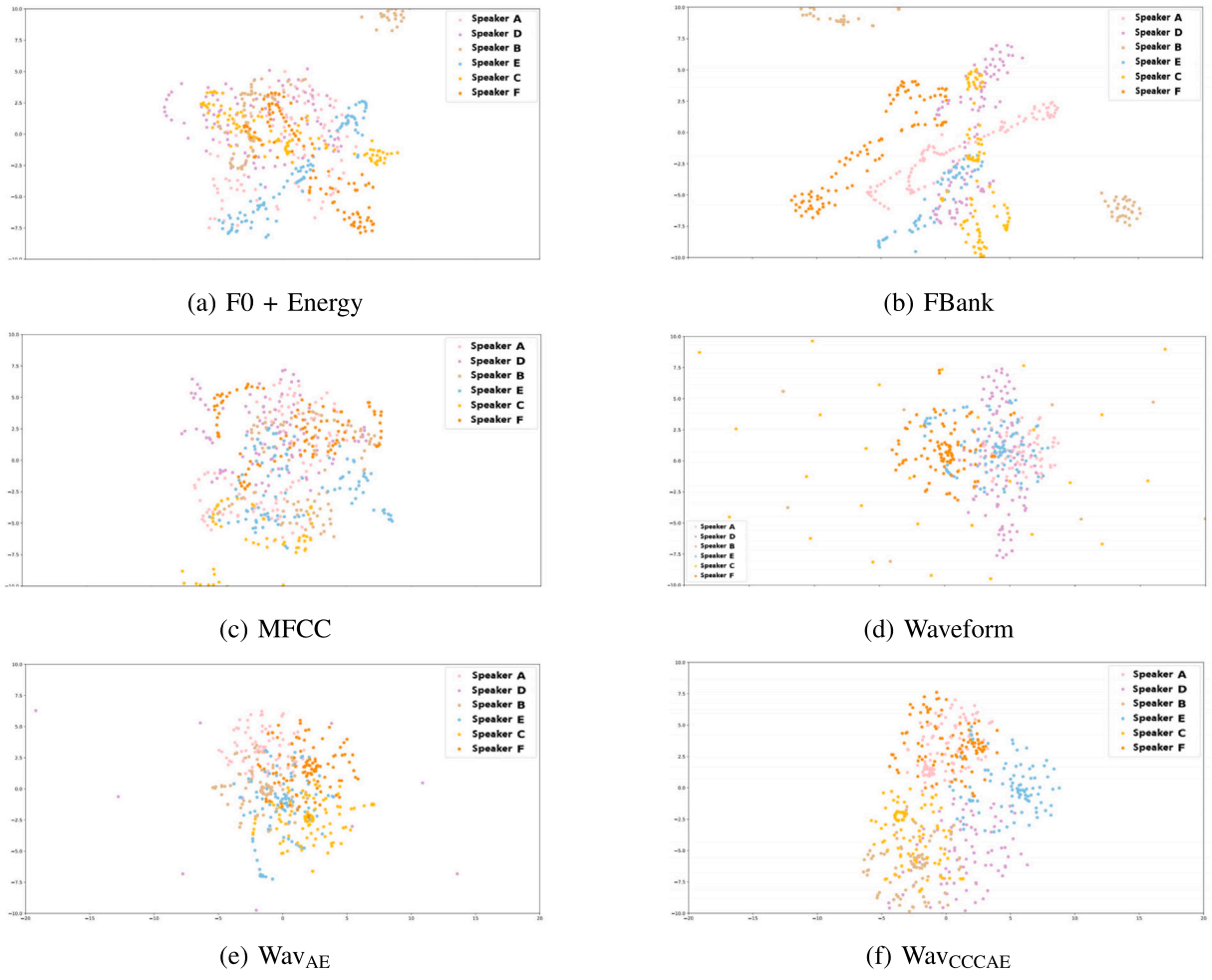
We also assume that each speaker has their own person dependent mannerisms and this affects the head movement in multiple ways, but there are still some patterns of the head movement that remain unchanged in all speakers. With the CCA loss objective, Wav_{CCCAE} shows a well-organised and distinct distribution of each speaker's feature data as there are some feature points where speakers overlapping each other (key properties of the head movement were not changed), and some feature points are spread in different directions (this was person dependent). This distribution has not been shown in the graph of any other feature. The value of Wav_{CCCAE} in Table 2 further demonstrates these surmisings, it is the second highest value and we believe that the overlapped areas show the properties of the head motion amongst all the speakers. As the correlation between this feature and head motion is still unclear, future academic study could develop these areas of research.

6. Evaluation of the head motion synthesis

Beside showing correlation analysis, which was presented in the previous section, we also investigated the effectiveness of the feature by building a neural network to predict head motion using those audio features. In this section, we built a simple FNN to generate head motion, then continued to apply the state-of-the-art architecture, LSTM. Lastly, we carried out a subjective evaluation. We selected MFCC, Wav_{AE}, and Wav_{CCCAE}, where these features were outstanding in the basic analysis, to use in the later evaluation of the regression models in the following section.

6.1. Evaluation of predicted head motion from speech

Fig. 4 reveals how NMSE and local CCA with the ground truth (original head motion) are involved between FNN system trained with different features, which are used to investigate the evaluation of predicted motion. Another coping strategy, which is expected to seek a



(a) F0 + Energy (b) FBank

(c) MFCC (d) Waveform

(e) Wav_{AE} (f) $\text{Wav}_{\text{CCCAE}}$

Fig. 3. T-SNE visualisation of the feature distribution for speaker A–F, to visualise whether there is common pattern in the head motion among speakers.

Table 3
The local CCA between the ground truth and randomised sequences of another speaker, to show the lowest bound of the CCA between two head motion streams.

Measure	Speaker						
	A	B	C	D	E	F	IEMOCAP
Unsynchronised CCA	0.14	0.11	0.11	0.11	0.10	0.10	0.12

chance score, is also developed on the grounds of well-computed local CCA between existing motion and randomised sequences that characterise totally different and unsynchronised speakers. The hypothesised chance score for the speakers is shown in Table 3.

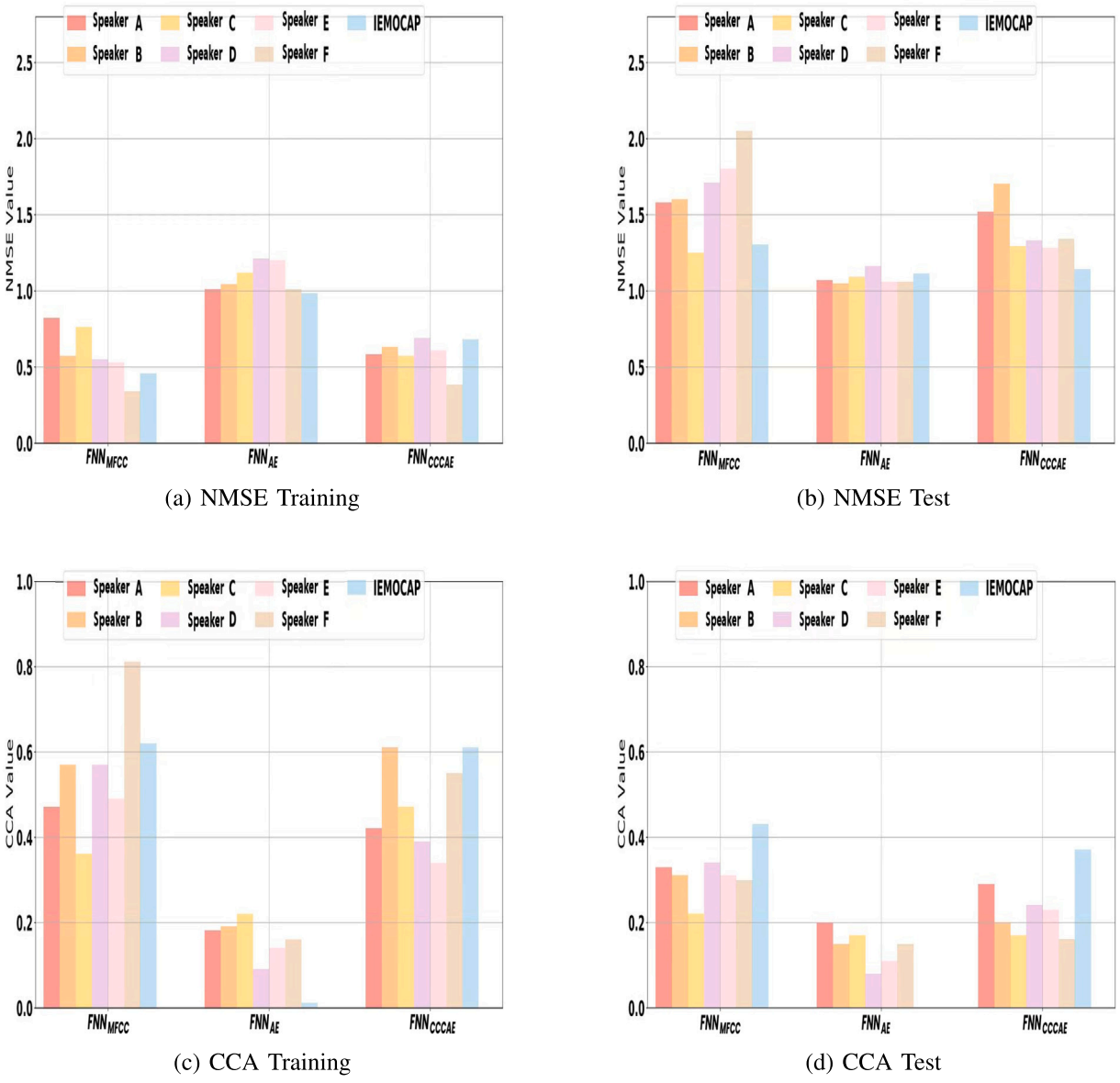
It is notable that, regardless of the lowest NMSE, the result of FNN_{AE} could be biased. Little movement of predicted head motion directly results in NMSE being close to 1.0. This explains why the chance score mechanism is better than FNN_{AE} for all speakers. $\text{FNN}_{\text{CCCAE}}$ has a better performance for most of the speakers except Speaker B and Speaker C in terms of NMSE. However, FNN_{MFCC} achieved the highest local CCA for all speakers. This suggests that $\text{FNN}_{\text{CCCAE}}$ and FNN_{MFCC} have different strength in different metric domains. Overall, the local CCA of FNN_{MFCC} and $\text{FNN}_{\text{CCCAE}}$ in the test dataset is higher than the chance scores.

6.2. Comparison between FNN and LSTM

LSTM is proven to be good in dealing with sequential data from the existing literature. Reflecting upon the above experiments, our proposed regression model in FNN achieves reasonable results. Thus, we would like to investigate how much improvement would be gained by switching FNN to LSTM. In this experiment, we tried two versions of 1-Layer-LSTM (replaced 4096-FNN), named LSTM^1, and 2-Layer-LSTM (replaced 4096-FNN and 64-FNN), named LSTM^2.

From Figs. 5 and 6, lower NMSE/higher CCA indicates a better performance for the models. LSTM^1 and LSTM^2 models have better results than FNN models in MFCC and $\text{Wav}_{\text{CCCAE}}$. Results for the models with Wav_{AE} reflect that there is not much difference for the models switching to LSTM compared to those who do not. A reason for this could be that Wav_{AE} is a low correlated feature so, even with the advantage of LSTM, the model hardly maps the acoustics features to the head motions. Surprisingly, LSTM^2 models perform worse than LSTM^1 in the IEMOCAP as the NMSE is higher and CCA is lower; this may indicate a vanishing gradient for deeper models and higher correlation features compared to Speaker A and D. Moreover, LSTM^1 and LSTM^2 models show better adaption for highly correlated features. $\text{LSTM}^1_{\text{CCCAE}}$ and $\text{LSTM}^2_{\text{CCCAE}}$ outperform FNN_{MFCC} in CCA for Speaker A and Speaker D.

Besides making comparisons among our proposed models, we also built an external algorithm (Haag and Shimodaira, 2016) for feature



(a) NMSE Training (b) NMSE Test

(c) CCA Training (d) CCA Test

Fig. 4. Comparison of different features in terms of performance of head motion prediction for different speakers, where NMSE (Figure a and b) and local CCA (Figure c and d) are calculated between predicted head motion and ground truth.

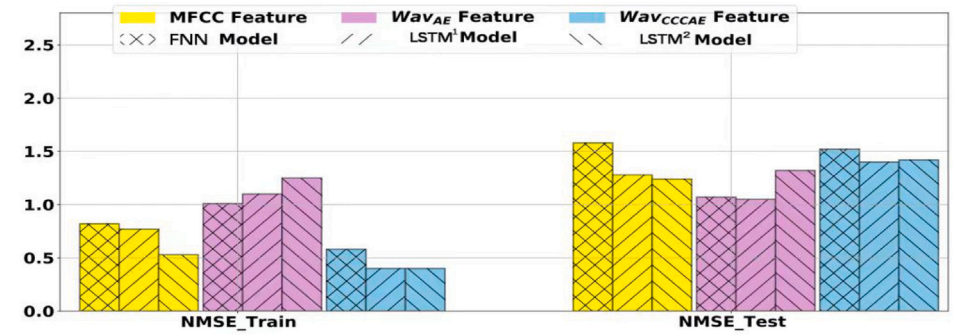
validation. Shown at Table 4, the external algorithm is even better at boosting the proposed feature to reach the highest local CCA for Speaker A without much work on model optimisation. This further proves that Bi-LSTM with forward and backward content information (Huang et al., 2015) could also adapt to the proposed embedded features. Nevertheless, the same performance is achieved with Wav_{AE}.

6.3. Subjective evaluation

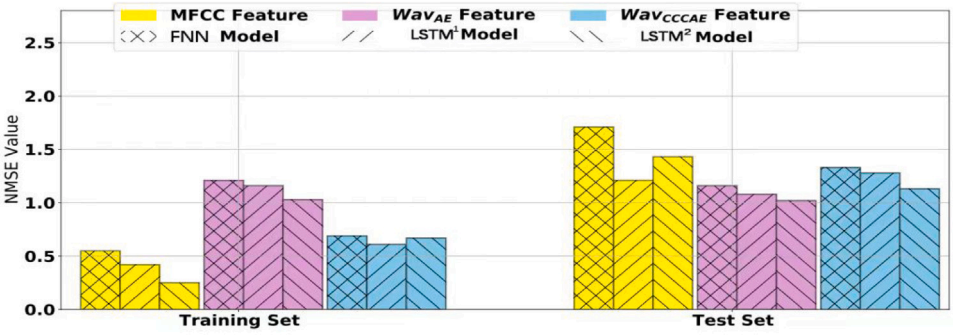
Objective evaluation only shows the numerical differences between the ground truth and the generated head motion, whereas subjective evaluation is able to reflect the opinions of the human observers on whether the generated motion is a close match to human-likeness. Compared to our previous work (Lu and Shimodaira, 2020), which only evaluated the performance with the criteria of naturalness, we validated our models' performance in the following subjective studies. We evaluated our models in two regards:

- Appropriateness — This study mainly focused on the correlation between the speech audio and the animated motion by asking the participants 'How appropriate are the head motions for the speech?'
- Model Assessment — This study asked participants to select 'Which of the following head motions are the most natural', intending to investigate which model architecture generates the most natural head movement using the same input features.

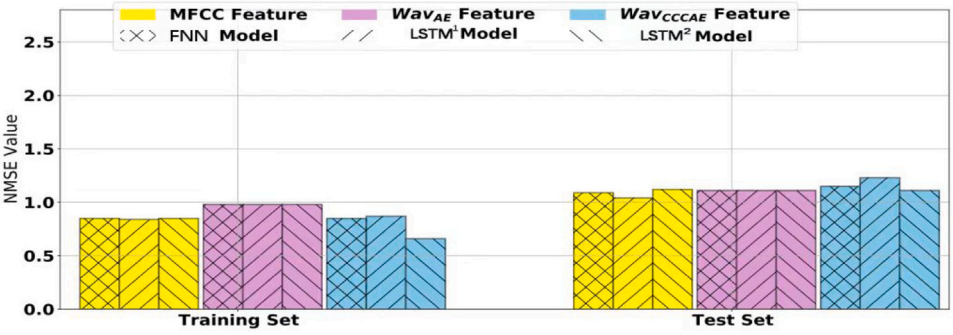
Jonell et al. (2020) indicate that we can trust the online platforms, as there is no difference between the in-lad and the Prolific platform in terms of the perceptual evaluation results. Therefore, we conducted our evaluation over an online platform entirely. A larger group of 50 participants was recruited in this work to ensure the reliability through the crowdsourcing platform Prolific, restricted to a set of English-speaking countries, and native speakers only, in this evaluation



(a) Speaker A



(b) Speaker D



(c) IEMOCAP

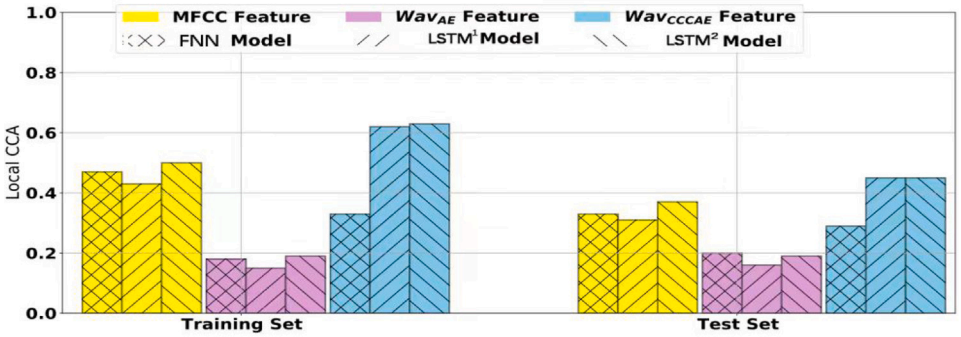
Fig. 5. Comparison of FNN and LSTM systems in terms of performance of head motion prediction, where NMSE is calculated between predicted head motion and ground truth. FNN model, LSTM[1]: 1-Layer-LSTM that replaces 4096-FNN, LSTM[2]: 2-Layers-LSTM that replaces 4096-FNN and 64-FNN.

and they were asked to evaluate both studies. Video samples of the animation are available on the web.[1]

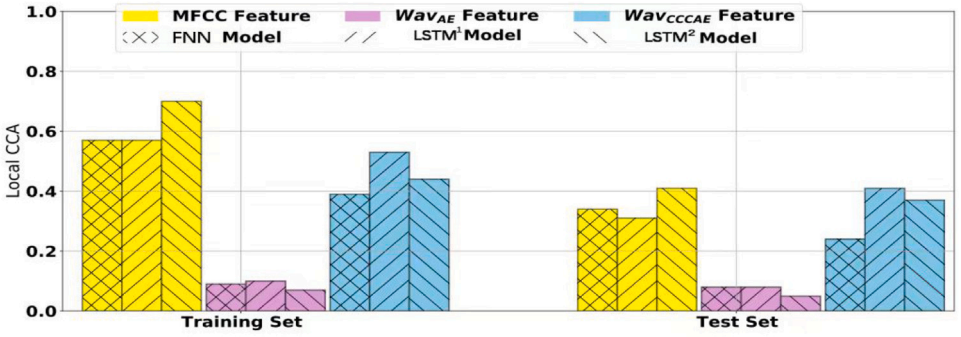
How appropriate are the head motion for the speech?

A perceptual test was carried using a similar method to Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) (International Telecommunication Union, 2015). Compared to the mean opinion score (MOS) test, MUSHRA is able to obtain a better quality to scores with a minimal number of participants. We created the head motion animations with the randomly selected audio samples in the test set using 8 models: Ground Truth (GT), Anchor and both FNN and LSTM models trained with three selected features respectively (FNN_{AE}, FNN_{CCCAE}, FNN_{MFCC}, $LSTM^2_{AE}$, $LSTM^2_{CCCAE}$, $LSTM^2_{MFCC}$). A total of 10 audio samples from each Speaker A and Speaker D is selected (160 animations are generated in total) and each animation lasts 8–12 s long. The anchor in MUSHRA is to calibrate the scale of the scores, where the minor artifacts are not badly penalised. The creation of the anchor is to select a different stream of head motion from another speaker with different utterances, where the resulted anchor animations are natural in term of head motion, but unsynchronised with the audio. Furthermore, a reference animation is provided as well, but it is generated with a different audio utterance to the evaluated one. This reference video was used to reinforce to the participants how to recognise what an appropriate head motion associated with speech audio looks like. The evaluation was performed so that each participant was assigned 10 test questions and the animations of each test question were shuffled so as to be displayed in a random order.(Fig. 7). Each participant then was requested to watch each animation carefully and wholeheartedly, and gave a score, between 0–100, for each animation. Compared to the original MUSHRA, we did not force the participants to rate the anchor to be the worst one or the ground truth to be the best. We requested that participants to score at least one of the animations to the value of 100 to indicate that is the 'ground truth'. Moreover, an attention check

[1] https://homepages.inf.ed.ac.uk/s1569197/phd_project_demo/.



(a) Speaker A



(b) Speaker D

(c) IEMOCAP

Fig. 6. Comparison of FNN and LSTM systems in terms of performance of head motion prediction, where local CCA is calculated between predicted head motion and ground truth. FNN model, LSTM[1]: 1-Layer-LSTM that replaces 4096-FNN, LSTM[2]: 2-Layers-LSTM that replaces 4096-FNN and 64-FNN.

for each speaker was incorporated during the test questions for each participant, which involved displaying a text sentence in the video such as, 'Please rate this video XX'. This 'XX' would be a specific number between 10 to 100, and the participant had to set the corresponding slider to the requested value in order to get through the attention check.

The result is displayed in Fig. 9. From both speakers, we can initially observe that GT scored the highest, and the anchor scored about 38. This indicates that the participants were able to consistently determine the most synchronicity and the non-synchronicity between the head motion and audio. Our proposed models with Wav_{CCCAE} outperformed MFCC models and Wav_{AE} models. Participants had different opinions on the performance between MFCC models and Wav_{AE} models for Speaker A and Speaker D respectively.

The head motion generated from MFCC achieves better in the objective, but lower in the subjective than the head motion generated from Wav_{CCCAE}. A possible reason for this is that while the speaker is in listening, MFCC is a spectral feature and does not effectively represent silence information on the absolute magnitude spectrum after filter extraction and log operation, whereas waveforms are well-presented. This affects models with MFCC predicting active head motion whereas models with Wav_{CCCAE} produce minor head movements whilst listening. An example is shown in Fig. 10, which demonstrates noise in the silence region shown from the Log Energy curve, and the active head motion is generated by FNN_{MFCC}. Another observation from Fig. 10 is that there is a minor head motion in the ground truth, but not in our proposed model for the silence region. This is another reason why the objective result of FNN_{CCCAE} performed worse than FNN_{MFCC}. Participants may have felt that active head motion went against natural human instincts while listening. Even though the ground truth showed animated head motion in the listening region as well, participants still preferred the ground truth over models with Wav_{CCCAE}, which indicates that the head motion generated by MFCC is unnatural. This also suggests that an objective approach is quantifiable, whereas subjective approaches are open to greater interpretation based on personal feeling (Leahu et al., 2008).

Table 4
Comparison of different systems in terms of performance of head motion prediction, where NMSE and local CCA are calculated between predicted head motion and ground truth. Red bold represents the best result for Speaker A and blue bold represents the best result for Speaker D.

System	Speaker	Training		Test	
		NMSE	CCA	NMSE	CCA
FNN_{AE}	A	1.00	0.17	1.06	0.21
	D	1.15	0.09	1.14	0.09
FNN_{CCCAE}	A	**0.55**	0.42	1.39	0.35
	D	0.66	0.39	1.24	0.32
$LSTM^2_{AE}$	A	1.25	0.19	1.32	0.19
	D	1.03	0.07	**1.02**	0.05
$LSTM^2_{CCCAE}$	A	0.58	0.33	1.58	0.29
	D	**0.61**	**0.53**	1.28	**0.41**
$BiLSTM_{AE}$	A	1.02	0.0	**1.05**	0.0
	D	1.01	0.0	1.07	0.0
$BiLSTM_{CCCAE}$	A	0.72	**0.52**	1.73	**0.41**
	D	0.81	0.45	1.64	0.38

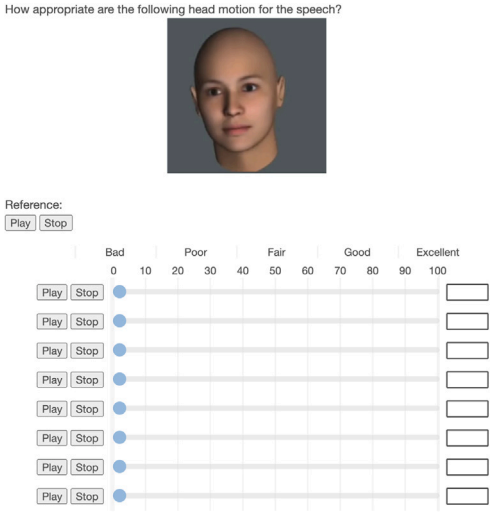


Fig. 7. A screenshot of a MUSHRA question from the evaluation interface. Each animation was generated with the same audio utterances, but different in input features and model architecture.

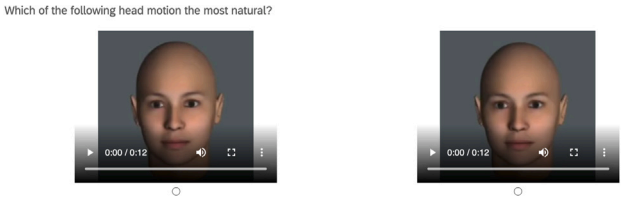


Fig. 8. A screenshot of a A/B test from the evaluation interface. Both animation was generated with the same input feature, but different in the model architecture. Right: LSTM, Left: FNN.

We also applied the significance test (paired t-test) to the mean score distributions across different pairs of the models. We compared in three perspectives, (1) whether the ground truth motion is significantly different from the predicted ones (GT VS $LSTM^2_{CCCAE}$ and GT VS $LSTM^2_{MFCC}$); (2) whether the LSTM model is significantly different to the FNN model ($LSTM^2_{CCCAE}$ VS FNN_{CCCAE} and $LSTM^2_{MFCC}$ VS FNN_{MFCC}); (3) whether Wav_{CCCAE} is better than MFCCs ($LSTM^2_{CCCAE}$ VS FNN_{MFCC} and $LSTM^2_{CCCAE}$ VS $LSTM^2_{MFCC}$).

Observing the results of the significance test in Fig. 9, $LSTM^2_{CCCAE}$ significantly outperforms the models trained with MFCCs, and the difference between GT and $LSTM^2_{CCCAE}$ is not statistically significant in Speaker A. However, $LSTM^2_{CCCAE}$ is only comparable to the models trained with MFCCs and worse than GT in Speaker D. Lastly, the difference between the LSTM models and the FNN models in both speakers is not statistically significant. This implies that their performances are comparable.

Which of the following head motions is the most natural? We conducted this second study using an A/B test to select which video is more natural than the other (Fig. 8). Our intention was to compare the feed-forward neural network and recurrent neural network with the same input features.

In Fig. 11, participants selected that $LSTM^2_{CCCAE}$ was always better than FNN_{CCCAE}, whereas they had different preference for MFCC models in both speakers. From the results we observe that it is clearly shown that LSTM always performed better in Speaker A, as well as generating more preferable head motion with the proposed feature Wav_{CCCAE} according to participants. Moreover, the results of the proposed features Wav_{CCCAE} and Wav_{AE} were consistent in both studies as the LSTM was better than FNN for Wav_{CCCAE} in both speakers, and LSTM was better than FNN in Speaker A but not in Speaker D for Wav_{AE} in the above MUSHRA study as well.

7. Conclusion

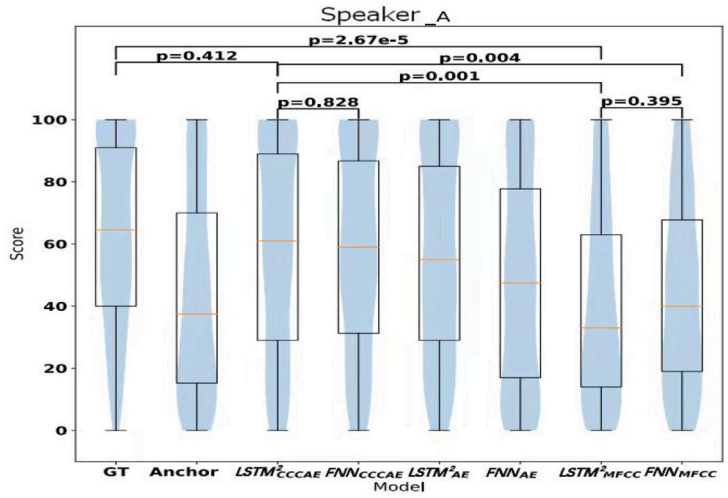
In this paper, we extended our previous research with additional data in training, feature analysis and an advance regression model, and further proved the effectiveness of Wav_{CCCAE}, which extracted data from the waveform with CCCAE. From the objective evaluations, we can conclude that (1) the proposed feature, Wav_{CCCAE}, is more strongly correlated than Wav_{AE} and other popular spectral features such as MFCC and Fbank amongst different speakers; (2) in the test data, the FNN_{CCCAE} achieved better in NMSE, but worse in local CCA than FNN_{MFCC}; (3) the analysis of the features distribution amongst the speakers showed a clear distinct cluster for each speaker in Wav_{CCCAE} only; (4) the LSTM-based regression models were able to boost the overall performance in NMSE and CCA, adapt better with the proposed feature (Wav_{CCCAE}) than MFCC; (5) MUSHRA suggests that the animations generated by models with Wav_{CCCAE} were chosen to be better over the other models by the participants of MUSHRA test. (6) A/B test further that the LSTM-based regression model adapts better with the proposed feature Wav_{CCCAE}. (2), (3) and (4) suggest that with the help of the CCCAE, Wav_{CCCAE} has the potential to be one of the task-specific features for generating head motion, which achieves state-of-the-art results. Overall, the models trained with the proposed feature, Wav_{CCCAE}, show better or comparable performance than the models trained with MFCCs.

CRediT authorship contribution statement

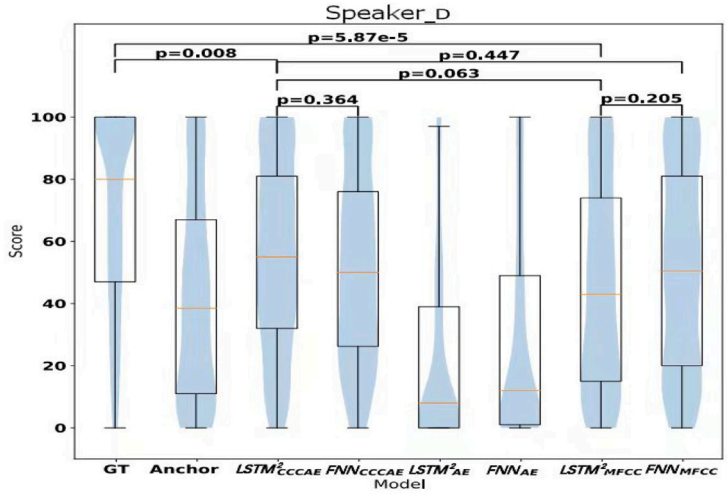
JinHong Lu: Writing – original draft, Writing – review & editing, Methodology, Project administration, Validation, Visualization. **Hiroshi Shimodaira:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.



(a) Speaker A's Appropriateness score



(b) Speaker D's Appropriateness Score

Fig. 9. The Boxplot of the MUSHRA score for both speakers' animation of each model — horizontal line indicates the median with confidence interval. The values between a pair of systems are the *P*-value to indicate the statistical significance.

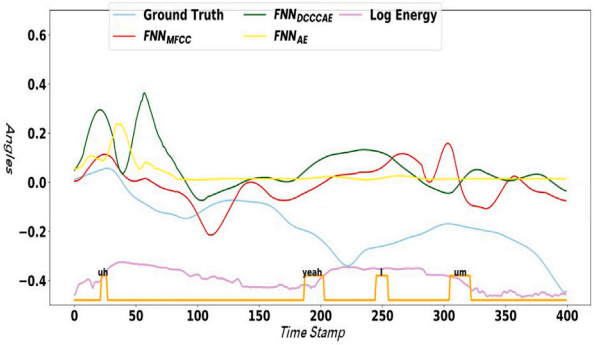


Fig. 10. An example of trajectory-Y generated from different models. The square wave at the bottom indicates whether the speaker is speaking (Up) or listening (Down). The text above the square wave is the corresponding transcript.

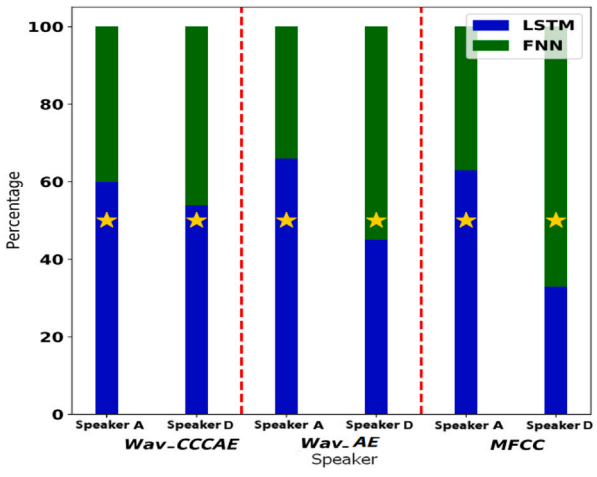


Fig. 11. The barplot of the A/B test for different model architectures. The star position indicates the 50% border line.

Data availability

Data will be made available on request.

References

Ahuja, C., Lee, D.W., Morency, L.-P., 2022. Low-resource adaptation for personalized co-speech gesture generation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 20534–20544.

Anderson, T., 2009. An Introduction to Multivariate Statistical Analysis, third ed. Wiley India Pvt. Limited, [Online]. Available: https://books.google.co.kr/books?id=1iF0CgAAQBAJ.

Andrew, G., Arora, R., Bilmes, J., Livescu, K., 2013. Deep canonical correlation analysis. In: Dasgupta, S., McAllester, D. (Eds.), Proceedings of the 30th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, Vol. 28, PMLR, Atlanta, Georgia, USA, pp. 1247–1255, no. 3.

Ben Youssef, A., Shimodaira, H., Braude, D., 2014. Speech driven talking head from estimated articulatory features. In: Proc. ICASSP. pp. 4573–4577.

Birdwhistell, R., 1952. Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture. Department of State, Foreign Service Institute.

Bolinger, D., 1983. Intonation and gesture. Amer. Speech 58 (2), 156–174.

Bolinger, D., Bolinger, D., 1986. Intonation and its Parts: Melody in Spoken English. Stanford University Press.

Busso, C., Deng, Z., Neumann, U., Narayanan, S., 2005. Natural head motion synthesis driven by acoustic prosodic features: Virtual humans and social agents. Comput. Anim. Virtual Worlds 16, 283–290.

Busso, C., Narayanan, S., 2007. Interrelation between speech and facial gestures in emotional utterances: A single subject study. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2331–2347.

Chandar, S., Khapra, M.M., Larochelle, H., Ravindran, B., 2015. Correlational neural networks. CoRR abs/1504.07225.

Chorowski, J., Weiss, R.J., Bengio, S., van den Oord, A., 2019. Unsupervised speech representation learning using WaveNet autoencoders. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (12), 2041–2053.

Chougule, S.V., Chavan, M.S., Gaikwad, M.S., 2014. Filter bank based cepstral features for speaker recognition. In: 2014 IEEE Global Conference on Wireless Computing & Networking. GCWCN, pp. 102–106.

Chung, Y.-A., Glass, J., 2020. Generative pre-training for speech with autoregressive predictive coding. pp. 3497–3501.

Ding, C., Xie, L., Zhu, P., 2015. Head motion synthesis from speech using deep neural networks. Multimedia Tools Appl. 74 (22), 9871–9888.

Fares, M., Pelachaud, C., Obin, N., 2022. Transformer network for semantically-aware and speech-driven upper-face generation. In: 2022 30th European Signal Processing Conference. EUSIPCO, pp. 593–597.

Ghahremani, P., Manohar, V., Povey, D., Khudanpur, S., 2016. Acoustic modelling from the signal domain using CNNs. In: INTERSPEECH. pp. 3434–3438.

Greenwood, D., Laycock, S., Matthews, I., 2017. Predicting head pose from speech with a conditional variational autoencoder. pp. 3991–3995.

Gregor, H., Hiroshi, S., Yamagishi, J., 2007. Speech driven head motion synthesis based on a trajectory model. In: In Proc. SIGGRAPH.

Haag, K., Shimodaira, H., 2015. The university of edinburgh speaker personality and mocap dataset. In: Facial Analysis and Animation Proceedings. Vienna.

Haag, K., Shimodaira, H., 2016. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In: Intelligent Virtual Agents. pp. 198–207.

Hadar, U., Steiner, T., Grant, E., Rose, F., 1983. Head movement correlates of juncture and stress at sentence level. Lang. Speech 26 (2), 117–129.

Hoshen, Y., Weiss, R.J., Wilson, K.W., 2015. Speech acoustic modeling from raw multichannel waveforms. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4624–4628.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28 (3/4), 321–377, [Online]. Available: http://www.jstor.org/stable/2333955.

Huang, Z., Xu, W., Yu, K., 2015. Bidirectional LSTM-CRF models for sequence tagging.

International Telecommunication Union, 2015. 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.

Ishi, C.T., Haas, J., Wilbers, F.P., Ishiguro, H., Hagita, N., 2007. Analysis of head motions and speech, and head motion control in an android. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 548–553.

Ishi, C.T., Ishiguro, H., Hagita, N., 2014. Analysis of relationship between head motion events and speech in dialogue conversations. Speech Commun. 57, 233–243.

Jia, J., Wu, Z., Zhang, S., Meng, H.M., Cai, L., 2014. Head and facial gestures synthesis using PAD model for an expressive talking avatar. Multimedia Tools Appl. 73 (1), 439–461.

Jonell, P., Kucherenko, T., Torre, I., Beskow, J., 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. IVA '20, Association for Computing Machinery, New York, NY, USA, [Online]. Available: https://doi.org/10.1145/3383652.3423860.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational Bayes. In: CoRR . abs/1312.6114.

Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., Yehia, H., 1999. Audio-visual synthesis of talking faces from speech production correlates. In: Eurospeech'99. Vol. 3, pp. 1279–1282.

Leahu, L., Schwenk, S., Sengers, P., 2008. Subjective objectivity: Negotiating emotional meaning. pp. 425–434.

Loweimi, E., Bell, P., Renals, S., 2020. On the robustness and training dynamics of raw waveform models. In: Proceedings of Interspeech 2020. International Speech Communication Association, pp. 1001–1005, Interspeech 2020, INTERSPEECH 2020 ; Conference date: 25-10-2020 Through 29-10-2020.

Lu, J., Liu, T., Xu, S., Shimodaira, H., 2021. Double-DCCCAE: Estimation of body gestures from speech waveform. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 900–904.

Lu, J., Shimodaira, H., 2019. A neural network based post-filter for speech-driven head motion synthesis. abs/1907.10585.

Lu, J., Shimodaira, H., 2020. Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder. In: Proc. Interspeech 2020. pp. 1301–1305.

rahman Mohamed, A., yi Lee, H., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T.N., Watanabe, S., 2022. Self-supervised speech representation learning: A review. IEEE J. Sel. Top. Sign. Proces. 16, 1179–1210.

Munhall, K., Jones, J., Callan, D., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychol. Sci. 15, 133–137.

Phan, H., McLoughlin, I.V., Pham, L., Chén, O.Y., Koch, P., De Vos, M., Mertins, A., 2020. Improving GANs for speech enhancement. IEEE Signal Process. Lett. 27, 1700–1704.

Sadoughi, N., Busso, C., 2017. Speech-driven animation with meaningful behaviors. CoRR abs/1708.01640.

Sadoughi, N., Busso, C., 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In: ICASSP. pp. 6169–6173.

Sainath, T.N., Weiss, R.J., Senior, A.W., Wilson, K.W., Vinyals, O., 2015. Learning the speech front-end with raw waveform CLDNNs. In: INTERSPEECH.

Soderkvist, I., Wedin, P., 1994. Determining the movements of the skeleton using well-configured markers. J. Biomech. 26, 1473–1477.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, Vol. 28, Curran Associates, Inc., pp. 3483–3491.

Takaki, S., Yamagishi, J., 2016. A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis. In: ICASSP. pp. 5535–5539.

Tüske, Z., Schlüter, R., Ney, H., 2018. Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing. In: ICASSP. pp. 4859–4863.

Wang, W., Arora, R., Livescu, K., Bilmes, J.A., 2016. On deep multi-view representation learning: Objectives and optimization. CoRR abs/1602.01024.

Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. J. Phon. 30 (3), 555–568.

Zhang, S., Wu, Z., Meng, H., Cai, L., 2007. Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar, Vol. 4, pp. IV–837.