To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations

Chaitanya Ahuja
cahuja@andrew.cmu.edu
Carnegie Mellon University

Shugao Ma
shugao@fb.com
Facebook Reality Labs, Pittsburgh

Louis-Philippe Morency
morency@cs.cmu.edu
Carnegie Mellon University

Yaser Sheikh
yasers@fb.com
Facebook Reality Labs, Pittsburgh

ABSTRACT

Non verbal behaviours such as gestures, facial expressions, body posture, and para-linguistic cues have been shown to complement or clarify verbal messages. Hence to improve telepresence, in form of an avatar, it is important to model these behaviours, especially in dyadic interactions. Creating such personalized avatars not only requires to model intrapersonal dynamics between a avatar's speech and their body pose, but it also needs to model interpersonal dynamics with the interlocutor present in the conversation. In this paper, we introduce a neural architecture named Dyadic Residual-Attention Model (DRAM), which integrates intrapersonal (monadic) and interpersonal (dyadic) dynamics using selective attention to generate sequences of body pose conditioned on audio and body pose of the interlocutor and audio of the human operating the avatar. We evaluate our proposed model on dyadic conversational data consisting of pose and audio of both participants, confirming the importance of adaptive attention between monadic and dyadic dynamics when predicting avatar pose. We also conduct a user study to analyze judgments of human observers. Our results confirm that the generated body pose is more natural, models intrapersonal dynamics and interpersonal dynamics better than non-adaptive monadic/dyadic models.

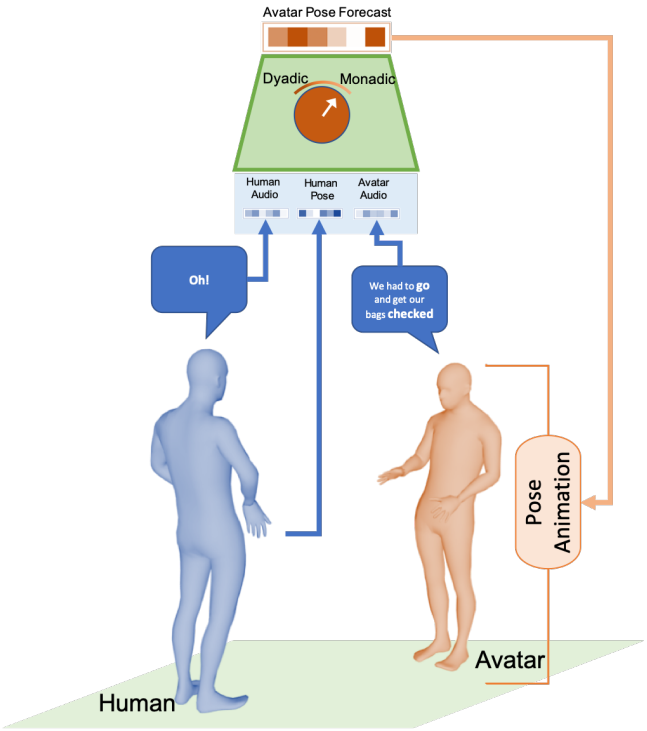

Figure 1: Overview of the visual pose forecasting task, which takes avatar's audio and predicted pose history along with human's audio and pose to forecast the avatar's future pose and generate a natural looking avatar animation. The model dynamically decides which of monadic or dyadic dynamics to focus to make the prediction.

CCS CONCEPTS

• **Computing methodologies** Procedural animation; Motion processing; *Neural networks*; • **Human-centered computing** *Virtual reality*; Auditory feedback.

KEYWORDS

dyadic interactions, pose forecasting, multimodal fusion

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '19, October 14–18, 2019, Suzhou, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353725

ACM Reference Format:
Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3340555.3353725

1 INTRODUCTION

Telepresence has the potential to evolve the way people communicate. With the application of immersion theory, stereoscopic vision and spatial audio, a 3D virtual space has characteristics inspired from the real world [36]. Communicating in a virtual world poses some interesting challenges. A person sitting thousands of miles away, where only speech signals are available, an avatar will need to represent not only his/her facial expressions [30], but also produce realistic non verbal body cues.

Non verbal behaviours such as hand gestures, head nods, body posture and para-linguistic cues play a crucial role in human communication [42]. These can range from simple actions like pointing at objects, head nod agreement, to body pose mirroring. Consider a person giving a monologue. Barring the minimal reaction of the audience (e.g. laughing on a joke he/she made), the speaker relies on his/her audio and his/her hand gestures, head motions and body posture to convey a message to the audience. These behaviours can be combined under the umbrella term of *intrapersonal* behaviours. Realism in *intrapersonal* behaviours is crucial to communication in the virtual world [5]. People can display different kind of gesture patterns, hence there is a need of driving the body pose of personalized avatars using the audio as input.

During dyadic interaction, behaviours of a person will be influenced by the behaviour of the interlocutor [37]. In other words, forecasting an avatar's pose should take *interpersonal* dynamics into consideration. This brings an interesting challenge on how to integrate back channel feedback [43] and other interpersonal dynamics while animating the avatar's behaviour. Examples of such behaviour can be seen in situations where people mimic head nods in agreement [8] or mirroring a posture shift at the end of conversation turn. Modeling such interpersonal behaviour can aid in building a more realistic avatar.

Speaker and listener roles, in a dyadic conversation, can change multiple times during the course of the conversation. A speaker's behaviour is affected by a combination of their non verbal signatures and interpersonal feedback from the listener. Similarly, a listener's behaviour is affected by a combination of some non verbal signatures and mostly providing feedback to the speaker in form of head nods, pose changes and short utterances (like 'yes', 'ya', 'ah' and so on).

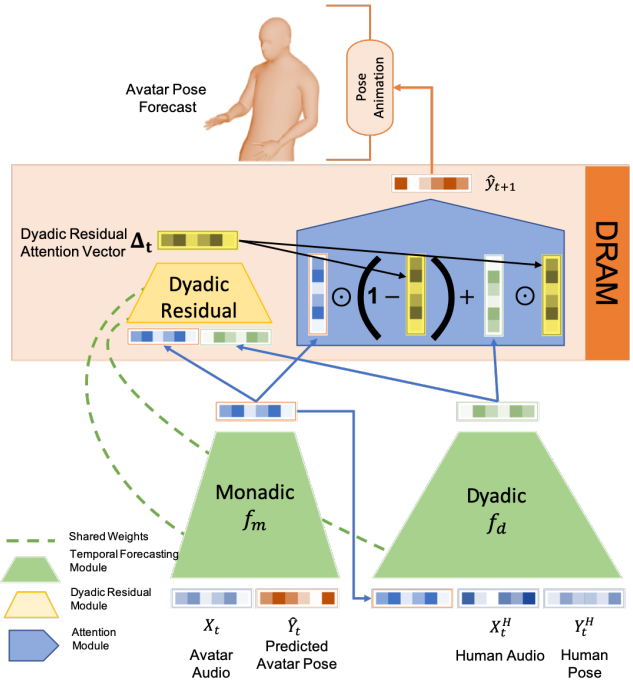

Figure 2: Overview of the proposed model Dyadic Residual-Attention Model (DRAM) designed to model the end-to-end visual pose forecasting task. Avatar's *monadic pose* forecast along with human's audio and pose history forecasts the next *dyadic pose* conditioned on dyadic (or interpersonal) dynamics. Avatar's *monadic* and *dyadic* pose predictions are inputs to DRAM which first calculates the dyadic residual attention vector (Δ_t) followed by an attention layer over monadic and dyadic pose predictions to make the final forecast $\hat{y}_t$.

Hence, to produce avatars capable of dyadic interactions with a human interlocutor, pose forecasting models need to anthropomorphise the character based on two facets of a conversation: interpersonal and intrapersonal dynamics.

In this paper, we learn to predict non verbal behaviours (i.e body pose) of an avatar[1] conditioned on the para-linguistic cues extracted from input audio and behaviours of the interlocutor as described in Figure 1. Central to our approach is a dynamic attention module that can toggle between monadic-focused (e.g. speaking with limited input from the listener) and dyadic-focused (e.g. interacting with the interlocutor) where interpersonal dynamics are also integrated. Our model Dyadic Residual-Attention Model (or DRAM) allows us to dynamically integrate intrapersonal (a.k.a monadic) and interpersonal (a.k.a dyadic) dynamics by attending to the interlocutor as and when needed. We present two variants of

[1]Project webpage: http://chahuja.com/trontr/

our model based on recurrent neural networks and temporal convolutional networks. These models are trained on a dataset consisting of conversations between two people. We study the avatar pose forecasting of one participant generated by these models on three challenges (1) Naturalness, (2) Intrapersonal Dynamics, and (3) Interpersonal Dynamics, by analyzing the effects of missing audio or pose information. Finally, we conduct a user study to get an overall human evaluation of the generated avatar pose sequences.

2 RELATED WORK

Pose forecasting has been previously studied with approaches ranging from goal conditioned forecasting [1, 32], image [10] and video [19] conditioned forecasting to pose synthesis using high-level control parameters [22, 31]. These vision-only approaches do not make use of audio signals from the speech.

It has been shown that fusing audio and visual information can give more robust predictions hence leading to improved performance [6, 26] especially for emotion modeling [44, 45]. Emotions are correlated to body motions [34] implying that audio is also correlated to body pose. Earlier work directly studied rhythm relationships of audio with body pose [15], correlation of head motion and speech disfluencies [23] and influence of audio on gestures in a dyadic setting [42].

In context of audio conditioned generation of facial expression and head pose, previous work includes creating voice driven puppets [7] and more recently deep learning approaches have improved the quality of lip-movement generation [38], facial expression generation [18, 28, 40] and facial expression generation in a conversation setting [13]. A related topic is generating speech by measuring vibrations in a video [14]. Follow up works include separating input audio signals into a set of components that corresponds to different objects in the given video [20], and separating audio corresponding to each pixel [46].

Cassell et al. [9] created the Behavior Expression Animation Toolkit (BEAT), which takes text as input to generate synthesized speech along with gestures and other nonverbal behaviors such as gaze and facial expression. The assignment is done on the linguistic and contextual analysis of the input text, relying on rules predefined based on evidence from previous research on human conventional behavior. Scherer et al. [33] proposes a markup language for generalizing perceptual features and show its effectiveness by integrating it into an automated virtual agent. Non verbal behaviours generated in this approach constructs a fixed set of body gestures[11, 12, 29], hence posing it as a classification problem. Fixed set of gestures cannot generalize to new behaviours, which is a drawback to this approach.

Parameterizing avatars with joint angles instead can alleviate this shortcoming. Extending this idea to audio conditioned pose forecasting, Takeuchi et al. [39] use linguistic features extracted from audio to predict future body poses using a bi-directional LSTM. As this method uses audio information from the future, it cannot be used for pose forecasting in real-time. In comparison, our models are auto-regressive in nature, using only information from the past. We note that our focus is on scenarios where manual text transcription may not be available, so our focus stays on the non-linguistic components of audio signals.

To our knowledge, our proposed work is the first to integrate both intrapersonal and interpersonal dynamics for body pose forecasting. Our aim is to generate natural looking sequence of body poses which correlate with audio signals driving the avatar as well as paralinguistic cues and behaviour of the interlocutor.

3 PROBLEM STATEMENT

Consider a conversation between two human participants, one of which is interacting remotely (henceforth referred as avatar) and only the audio is available. For the local participant (henceforth referred as human/interlocutor), we have pose and audio information. The goal of the forecasting task is to model future body pose of the avatar. Formally, given sequence of local human audio features $X_t^H = \left[x_t^H, x_{t-1}^H, \ldots, x_{t-k-1}^H\right]$, human pose $Y_t^H = \left[y_t^H, y_{t-1}^H, \ldots, y_{t-k-1}^H\right]$, avatar's audio features $X_t = [x_t, x_{t-1}, \ldots, x_{t-k-1}]$ and history of avatar's pose $\hat{Y}_t = [\hat{y}_t, \hat{y}_{t-1}, \ldots, y_{t-k-1}]$, we want to predict avatar's next pose $\hat{y}_{t+1}$. Let x_t, x_t^H be vectors of dimension a, and y_t, y_t^H be vectors of dimension p for all t. k is the size of feature history used by the model. Hence, $X_t^H, X_t \in \mathcal{R}^{a\times k}$ and $Y_t^H, \hat{Y}_t \in \mathcal{R}^{p\times k}$ are matrices.

This is equivalent to modeling the probability distribution $P(\hat{y}_{t+1}|X_t^H, Y_t^H, X_t, \hat{Y}_t)$. Concatenating all input features $X_t^H, X_t, Y_t^H, \hat{Y}_t$, we define a joint model $f : \mathcal{R}^{2(a+p)\times k} \rightarrow \mathcal{R}^p$ which predicts the future pose $\hat{y}_{t+1}$

$$\widehat{y}_{t+1} = f\left(X_t^H, Y_t^H, X_t, \hat{Y}_t; \theta\right) \tag{1}$$

where θ are trainable parameters of function f. As the history of avatar's previously predicted pose sequence $\hat{Y}_t$ is used to predict the future pose, f is an autoregressive model.

In this section we discuss challenges of jointly modeling interpersonal and intrapersonal dynamics. We propose our model Dyadic Residual-Attention Model (DRAM) to tackle these challenges in the next Section while maintaining the generalizability of the model.

Challenges of Jointly Modeling Interpersonal and Intrapersonal Dynamics

Interpersonal dynamics are important in providing a realistic social experience in a virtual environment. Ignoring verbal or non-verbal cues from the interlocutor may result in generating avatar body-poses which are not synchronous with

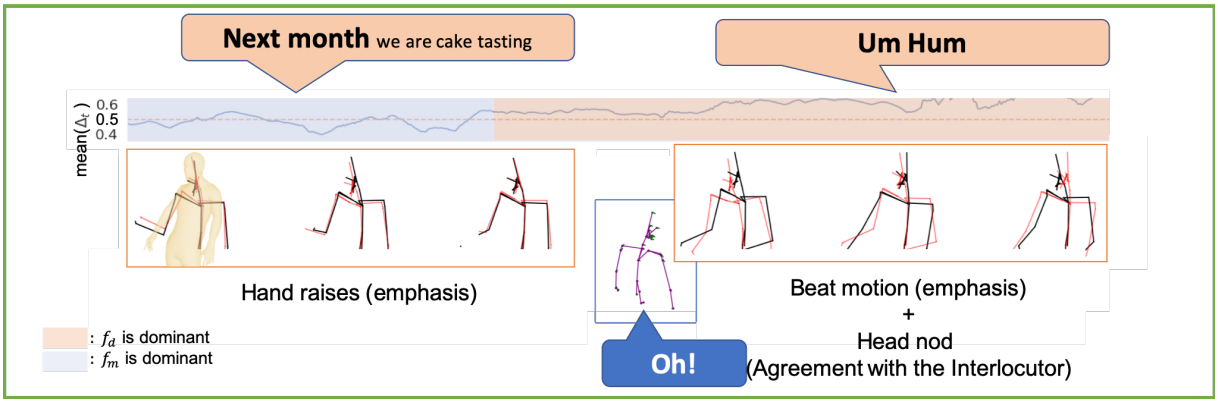

Figure 3: An example demonstrating a hybrid of Intrapersonal and Intrapersonal dynamics in predictions made by DRAM. For the avatar, the black skeleton is the current pose and the red skeleton is the pose from one second in the past. Similarly for the interlocutor, the black skeleton is the current pose and the purple skeleton is the pose from one second in the past. For the first 3 seconds, the avatar focusses on words 'Next Month' and DRAM forecasts hand raises denoting emphasis.mean(Δ_t) is moslty less than 0.5. As soon as the interlocuter chimes in with an exclaimation 'Oh!', mean(Δ_t) rises up implying more focus on the interlocutor. DRAM forecasts head nods denoting agreement with the interlocuter. Beat motions are also predicted by the model which is probably due to emphasis on the words 'Um Hum'.

the interlocutor[27]. Given enough dyadic conversation data, the function in Equation 1 has the capacity to jointly model interpersonal and intrapersonal dynamics.

However, an imbalance between interpersonal and intrapersonal dynamics in dyadic conversations is common, with generally more instances of the intrapersonal dynamics (e.g. between speech and gestures of the same person). This naturally occurring imbalance ends up treating X_t and Y_t from intrapersonal dynamics as the stronger prior than signals from the interlocutor (X_t^H and Y_t^H) while solving the optimization problem,

$$min_\theta \sum_t \|f\left(X_t^H, Y_t^H, X_t, Y_t; \theta\right) - \hat{y}_{t+1}\|_2^2 \tag{2}$$

where $\|.\|_2$ is L2 Norm.

Hence, interpersonal dynamics could end up getting ignored, leaving the pose generation largely monadic and intrapersonal.

4 DYADIC RESIDUAL-ATTENTION MODEL

To combat the imbalance in dyadic conversations, our proposed approach shown in Figure 2 decomposes pose generation of the avatar to monadic function (f_m) and dyadic function (f_d), which model Intrapersonal and Interpersonal dynamics respectively.

We propose a model Dyadic Residual-Attention Model (DRAM) with $\Delta_t \in \mathcal{R}^p$ as a time-continuous vector. Δ_t allows for smooth transitions between monadic and dyadic models and it is a vector with the same dimensions as the pose of the avatar/interlocutor. We use Δ_t like an attention vector which attends to different joints at different points in time. Hence, our proposed model $f_{DRAM}\left(X_{t-1}, Y_{t-1}, X_{t-1}^H, Y_{t-1}^H; \theta_{DRAM}\right)$ can be written as

$$f_{DRAM} = (1 - \Delta_t) \odot f_m + \Delta_t \odot f_d \tag{3}$$

where $f_{DRAM} : \mathcal{R}^{2(a+p)\times k} \rightarrow \mathcal{R}^p$ and Δ_t is the Dyadic Residual Vector. Δ_t can be used as a trainable parameter which enables the model to implicitly learn attention weights for each joint at each time-step.

In this section, we describe the formulation of the Dyadic Residual Vector (Δ_t) which is further used as soft-attention weights on Dyadic (f_d) and Monadic (f_m) models resulting in the formation of DRAM. We end the section by explaining the loss function and the training curriculum for the model.

Dyadic Residual Vector

Monadic model $f_m : \mathcal{R}^{(a+p)\times k} \rightarrow \mathcal{R}^p$ learns the intrapersonal dynamics (i.e. demonstrating emphasis using head nods or hand gestures etc.) of the avatar. This is equivalent to the avatar giving a monologue without any ineterlocutor,

$$\hat{z}_t^m = f_m\left(X_{t-1}, Y_{t-1}; \theta_m\right) \tag{4}$$

Dyadic model $f_d : \mathcal{R}^{(a+2p)\times k} \rightarrow \mathcal{R}^p$ can be written as,

$$\hat{z}_t^d = f_d\left(X_{t-1}^H, Y_{t-1}^H, \hat{Z}_t^m; \theta_d\right) \tag{5}$$

where $\hat{Z}_{t-1}^m = \left[\hat{z}_{t-1}^m, \hat{z}_{t-2}^m \ldots, \hat{z}_{t-k-1}^m\right]$.

Dyadic model f_d depends on the Monadic dynamic $\hat{Z}_t^m$ and features of the interlocutor[2]. Hence, it learns to model

[2]For computational efficiency, we use $\hat{Z}_t^m$ as an input for the dyadic network. An alternative equivalent model is where raw avatar features (X_{t-1}, Y_{t-1}) are used as one of the inputs.

the interpersonal dynamics (i.e. head nods, pose switches, interruptions etc.) between the avatar and interlocutor.

The absolute difference between Monadic and Dyadic Models $|f_d - f_m|$, or the residual, is representative of the joints that were affected by interpersonal dynamics[3]. If, at a given time t, the residual for some joints is high, the interlocutor is influencing those joints, while if the residual is low, the avatar's audio and pose-history is dominating the avatar's current pose behaviour.

The residual for all joints is always positive, so to compress all dimensions of the residual vector Δ_t between 0 and 1, we use tanh non-linearity to get the dyadic residual vector Δ_t,

$$\Delta_t = \tanh |f_d \left(X^H_{t-1}, Y^H_{t-1}, \widehat{Z}^m_{t-1}; \theta_d\right) - f_m \left(X_{t-1}, Y_{t-1}; \theta_m\right)| \tag{6}$$

The interpretability of the Dyadic Residual Vector is an added advantage. We know which joint has the maximum influence at every time t, and hence can estimate whether the non-verbal behaviours of the avatar are due to interpersonal or intrapersonal dynamics.

Using Equations 3, 4, 5 and 6, the predicted pose of our proposed DRAM can be re-written as:

$$\widehat{y}_t = (1 - \Delta_t) \odot \widehat{Z}^m_t + \Delta_t \odot \widehat{Z}^d_t \tag{7}$$

Loss Function

Pose is a continuous variable, hence we use Mean Squared Error (or L2) loss. Based on predicted pose in Equation 7, the loss function is

$$\mathcal{L} = \sum_t \|\hat{y}_t - y_t\|_2^2 \tag{8}$$

$$= \sum_t \| (1 - \Delta_t) \odot \widehat{Z}^m_t + \Delta_t \odot \widehat{Z}^d_t - y_t\|_2^2 \tag{9}$$

Design Choices for f_d and f_m

The prediction functions f_m (for monadic) and f_d (for dyadic) of our proposed model can work with any autoregressive temporal network that depends only on features from the past. This gives our model the flexibilty of incorporating temporal models that may be domain dependent or pre-trained on some other dataset.

Recurrent neural architectures have been shown to perform very well in such autoregressive tasks [2, 21, 25], especially for pose forecasting algorithms [10, 31]. Recent work demonstrates the utility of bi-directional LSTMs to model speech signals to forecast body pose [39]. One weakness of this approach is the dependency of the pose prediction on future speech input, hence making it unusable for real-time applications. A uni-directional LSTM model is used as a baseline model [39].

Temporal convolutional networks (TCNs) work just as well in many practical applications [4]. It was shown that adding residual connections and dilation layers [41] can boost the empirical performance of TCNs equal to, if not better than, LSTM and GRU based models.

In our experiments, both TCNs and LSTMs are used as temporal models for f_d and f_m, which demonstrates the versatility of our proposed approach.

5 EXPERIMENTS

Visual pose-forecasting of an avatar during dyadic conversations can be broken down into three core challenges,

(1) **Naturalness**: How natural is the flow of poses and how close are they to the ground truth?
(2) **Intrapersonal Dynamics**: How correlated is the generated pose sequence with avatar's speech?
(3) **Interpersonal Dynamics**: Is the generated pose sequence reacting realistically to the interlocutor's behaviour and speech?

Experiments, both subjective and objective are designed to evaluate there 3 aspects of pose forecasting.

In the following subsections, we describe the dataset and have a short discussion on pre-processing of audio and pose features. This is followed by constructing a set of competitive baseline models to compare our own proposed DRAM model.

Dataset

Our models are trained and evaluated on a previously recorded dataset of dyadic face-to-face interactions. The dataset contains one person who is the same across all conversations. This person interacts with 11 different participants for around 1 hour each. The participants were given different topics (like sports, school, hobbies etc.) to choose from and the conversation was guided by these topics. No script were given to either of the participants. The capture system included 24 OptiTrack Prime 17W cameras surrounding a capture area of approximately 3 m × 3 m and two directional microphone headsets. Twelve of the 24 cameras were placed at 1.6 m height. Participants wear marker-attached suits and gloves. The standard marker arrangement provided by OptiTrack for body capture and glove marker arrangement suggested in Han et al. [24] was followed.

For each conversation, there are separate channels of audio signals for each person with a sampling rate of 44.1 kHz. Body pose was collected at a frequency of 90 Hz using a Motion-Capture (MoCap) setup, and gives 12 joint positions

[3]It may be possible to use a separate network to model the attention vector. Our proposed network Dyadic Residual-Attention Model, in a manner of speaking, shares weights with existing networks f_m and f_d to estimate the attention vector Δ_t. This helps us limit the number of trainable parameters.

Table 1: Objective Metric Average Position Error (APE) for DRAM is compared with all baseline models. Lower values are better. The first row networks, Human Audio Only and Human Monadic Only, model Intrapersonal dynamics, while the second row networks, Avatar Audio Only and Avatar Monadic only, model Intrapersonal Dynamics. The third row networks, Early Fusion and DRAM w/o attention, non-adaptively model Interpersonal and Intrapersonal dynamics jointly. Fourth row networks, DRAM, adaptively choose from Interpersonal and Intrapersonal dynamics. Two-tailed pairwise t-test between all TCN models and DRAM-TCN where $**$- p<0.01, and $*$- p<0.05

Dynamics	Models		Average Position Error (APE) in cms							
			Avg.	Torso	Head	Neck	RArm	LArm	RWrist	LWrist
Interpersonal	Human Audio Only (f_m)	LSTM	4.9	0.2	1.1	0.4	0.2	0.2	14.4	21.3
		TCN	3.3**	0.2	1.2	0.5	0.2	0.3	9.5	13.8
	Human Monadic Only (f_m)	LSTM	3.6	0.2	1.1	0.4	0.2	0.2	13.2	12.5
		TCN	3.3*	0.2	1.1	0.4	0.2	0.2	9.3	13.8
Intrapersonal	Avatar Audio Only (f_m)	LSTM	3.9	0.6	1.5	1.4	1.0	0.5	10.7	15.0
		TCN	3.5**	0.2	1.1	0.5	0.3	0.3	9.3	15.5
	Avatar Monadic Only (f_m)	LSTM	3.4	0.2	1.3	0.4	0.3	0.2	9.5	14.4
		TCN	3.0*	0.2	1.4	0.5	0.2	0.3	8.8	12.1
Interpersonal and Intrapersonal	Early Fusion (f)	LSTM	3.0	0.2	1.1	0.5	0.3	0.2	9.3	11.7
		TCN	3.2*	0.2	1.1	0.4	0.2	0.3	10.5	12.2
	DRAM w/o Attention (f_d)	LSTM	3.1	0.2	1.8	0.4	0.2	0.2	9.7	12.1
		TCN	3.0	0.1	1.0	0.4	0.2	0.2	8.8	12.6
Adaptive Interpersonal and Intrapersonal	DRAM (f_{DRAM})	LSTM	**2.8**	**0.1**	**1.0**	**0.4**	**0.2**	**0.2**	9.0	**10.8**
		TCN	**2.8**	0.2	1.1	**0.4**	**0.2**	**0.2**	**8.8**	11.1

of the upper body including which can be grouped[4] into Torso, Head, Neck, RArm (Right Arm), LArm (Left Arm), RWrist (Right Wrist) and LWrist (Left Wrist).

Feature Representation

Body pose is shown to have correlation with affect and emotion. GeMAPS is a minimalist set of low level descriptors for audio including prosodic, excitation, vocal tract, and spectral descriptors which increase the accuracy of affect recognition. OpenSmile [17] is used to extract GeMAPS [16] features sub-sampled rate of 90 Hz to match the body pose frequency.

In this work, translation of the body is not considered, as it is largely absent in the data. Instead rotation angles are modeled, which form the crux of dyadic interactions in a conversation setting. In our experiments we use pose features that are 3-dimensional joint coordinates are converted to local rotation vectors and parameterized as quaternions [31].

Baselines

There has been limited work in the domain of gesture generation from audio signals using neural architectures. The model only take into account monadic behaviours of a person using a bidirectional-LSTM [39]. Bidirectional-LSTMs depend on the future time-steps, hence they are unusable in real-time applications. An adapted version of this network (referred as **Avatar Audio only- LSTM** in Table 1) and TCNs are used as temporal models for Dyadic f_d and Monadic f_m models.

To gauge the naturalness of our proposed model **DRAM**s, they are compared with **Early Fusion** (f from Equation 1) and **DRAM w/o Attention** (f_d from Equation 5)

To demonstrate presence of Intrapersonal Dynamics in a dyadic conversation, a reasonable baseline is Monadic Models (f_m from Equation 4) with inputs as the avatar's audio (refer as **Avatar Audio Only**) and avatar's audio+pose history (refer as **Avatar Monadic Only**). Both of these models forecast the pose of the avatar.

To demonstrate presence of Interpersonal Dynamics in a dyadic conversation, a reasonable baseline is Monadic Models (f_m from Equation 4) with inputs as the human's audio (refer as **Human Audio Only**) and human's audio+pose history (refer as **Human Monadic Only**). Both of these models forecast the pose of the avatar.

Objective Evaluation Metrics

We evaluate all models on the held-out set with a metric Average Position Error (APE). Given a particular keypoint p,

[4]Our modeling is performed for all 12 joints, but we group them in our results to help with interpretability

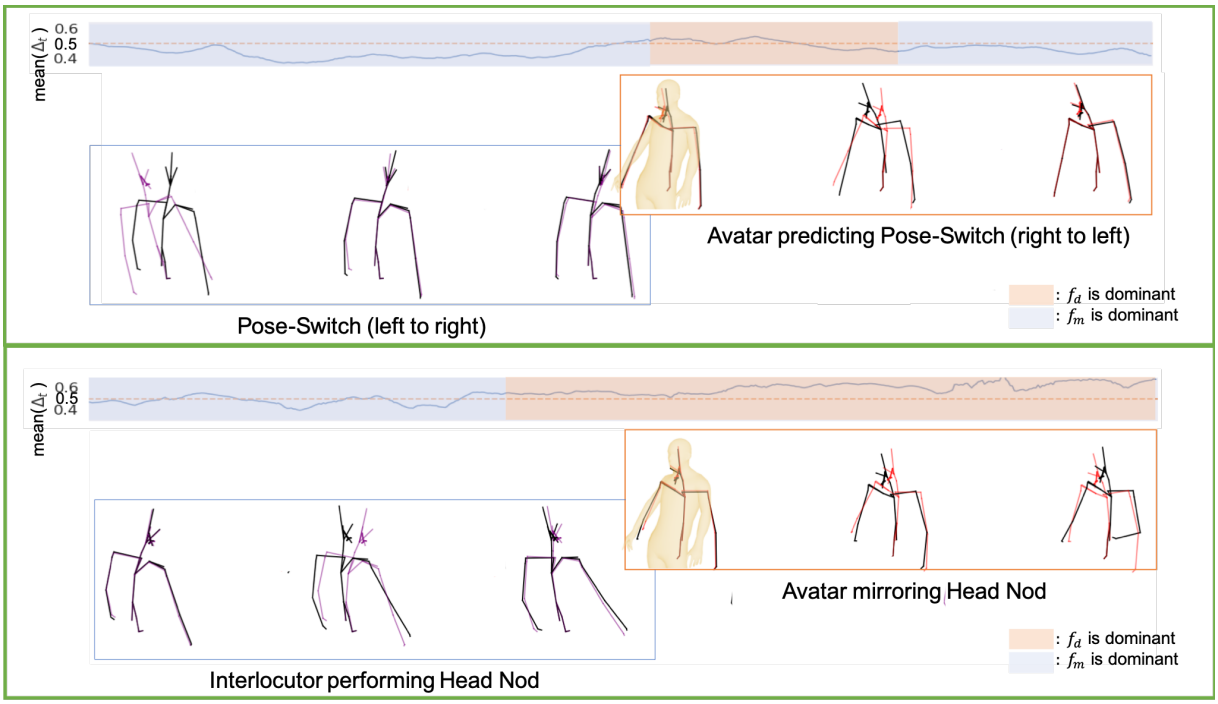

Figure 4: Two examples demonstrating Interpersonal Dynamics in predictions made by DRAM. For the first example, interlocutor performs a pose switch which is followed by predicted pose switch by the avatar. Mean of Dyadic Residual Attention vector (mean(Δ_t)) is moslty below 0.5 till the interlocutor performs a pose switch. DRAM estimates the need to focus on the human (i.e. interpersonal dynamics) with the increase in value of Δ_t and predicts a pose-switch for the avatar. Similarly, the second example has the interlocutor performing a head nod, which is followed by a forecast of head nod by the avatar. mean(Δ_t) values rise to values above 0.5 just after the interlocuter's head nod implying the need for interpersonal dynamics.

it can be denoted as APE(p),

$$\text{APE}(p) = \frac{1}{\mathcal{Y}} \sum_{\mathcal{Y}} \|\hat{y}_t(p) - y_t(p)\|_2 \tag{10}$$

where $y_t(p)$ is the true location and $\hat{y}_t(p) \in \mathcal{Y}$ is the predicted location of keypoint p

Another metric, Probability of Correct Keypoints (PCK) [3, 35], is also used to evaluate all models. If a predicted keypoint lies inside a sphere (of radius σ) around the ground truth, the prediction is deemed correct. Given a particular keypoint p, PCK$_\sigma(p)$ is defined as follows,

$$\text{PCK}_\sigma(p) = \frac{1}{\mathcal{Y}} \sum_{\mathcal{Y}} \delta\left(\|\hat{y}_t(p) - y_t(p)\|_2 \leq \sigma\right) \tag{11}$$

where δ is an indicator function.

User Study: Subjective Evaluation Metrics

Pose generation during dyadic interactions can be a subjective task as reactions of the *avatar* depend on its own audio, and the human's audio and pose. A human's subjective judgement on the quality of prediction is an important metric for this task.

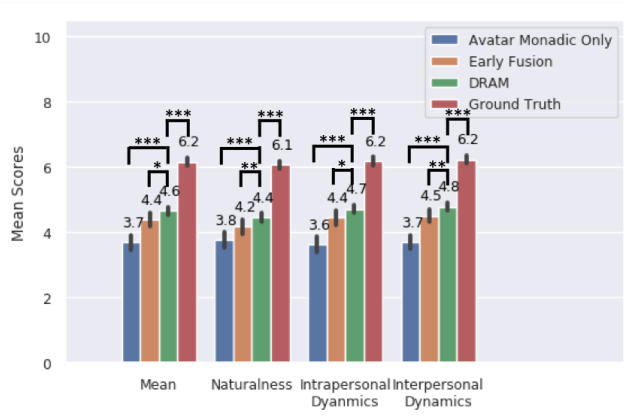

Figure 5: Histograms of Subjective Scores for Naturalness, Intrapersonal Dynamics, Interpersonal Dynamics and the mean across all criteria. Higher scores are better. Two-tailed pairwise t-test between Avatar Monadic only, Early Fusion TCN models and DRAM-TCN where ∗- p<0.05, ∗∗- p<0.01, and ∗ ∗ ∗- p<0.001

To achieve this, we design a user study of the generated videos from a held-out set. During the study, an acclimation phase is performed by showing reference clips (ground truth poses taken from the training set) to annotators to get them acquainted with the natural motion of the avatar. The main part of the study consists of showing annotators multiple one minute clips from the test set. Each video contains predicted avatar pose, avatar's audio and the ground truth audio and poses for the human. Pose is represented in form of a stick figure (Refer to 7). The avatar predicted poses can come from one of these models in Figure 5 or the ground truth. Annotators do not know which model was used to animate the avatar. They are instructed to judge the animation based on the following statements:

- S1 **Naturalness**:The motion of *avatar* looks natural and match his/her audio
- S2 **Intrapersonal Dynamics**:*Avatar* behaves like themself (recall the reference video)
- S3 **Interpersonal Dynamics**:*Avatar* reacts realistically to the interlocutor (in terms of interlocutor's audio and motion)

At the end of each clip they give a score for all the statements following a 1 to 7 on the likert scale where,

1	2	3	4	5	6	7
Disagree		Somewhat Disagree		Somewhat Agree		Agree

A fourth quesiton is asked to know how confident annotators were in scoring each video based on all input modalities. Each video is rated by a minimum of 2 human annotators where the final score is a weighted average with the weights as the confidence rating for each video.

6 RESULTS AND DISCUSSION

Objective Evaluation

Average Position Error (APE): Models with only Interpersonal Dynamics achieve the best APE of around 3.3 which is slightly worse than the best APE of 3.0 on models with Intrapersonal dynamics (Table 1). **Early Fusion** and **DRAM w/o Attention** are models with both interpersonal and intrapersonal dynamics as input, but they are not able to surpass the **Avatar Monadic Only** model. This is not surprising as avatar's speech is highly correlated with its body pose. Models with Adaptive Interpersonal and Intrapersonal Dynamics (e.g. **DRAM**), which achieved an APE of 2.8, are able to exploit changing dynamics in a conversation unlike non-adaptive methods such as Early Fusion and DRAM w/o Attention.

Each joint has different characteristics in a conversation setting. Some of them, like *Torso*, and *Neck*, do not move a lot during the course of the conversation. It is clear from the low APE values for these joints in Table 1 that modeling them is easier when compared to frequently moving joints like *Wrists*. It is also evident from the table that forecasting wrists had a much higher APE across all joints across all models. **Dyadic Residual-Attention Model** gives almost a 1.0 reduction in APE values when compared to the best non-adaptive model.

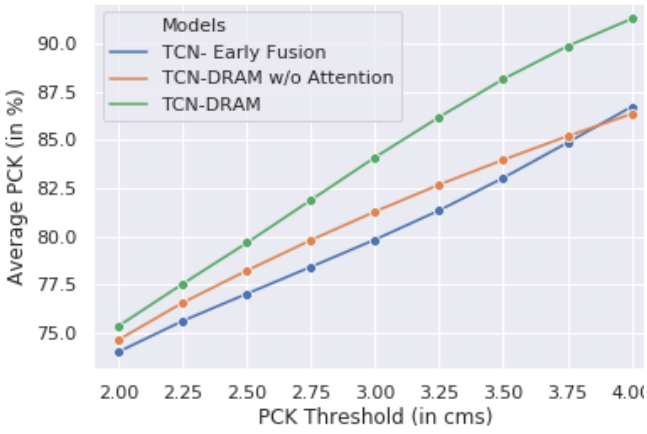

Figure 6: Plots of average Probability of Correct Keypoint (PCK) values over multiple values of PCK threshold (σ) for Early Fusion, DRAM w/o Attention, and DRAM models with TCNs. Higher values are better.

The joint, *Head*, shows some interesting characteristics. It can be fairly hard to predict head motion with just monadic data of the avatar as it sometimes mirrors head nods coming from the interlocutor (or Human). Dyadic information becomes crucial in this scenario. It is interesting to see that head predictions are around 1.0 value of APE worse for models with only intrapersonal dynamics. The monadic model conditioned on only Human features ends up performing reasonably well, probably because it can learn to map Avatar's sporadic head nods to those of the Human.

Probability of Correct Key-points (PCK): The gap between PCK values of DRAM and other baselines increase with the value of PCK threshold (Figure 6), which implies that variance of erroneous predictions by DRAM is lesser than other baselines, making our proposed model more robust.

Subjective Evaluation

Based on the user study we conducted on the generated Avatar Pose (see Figure 5), humans find that **DRAM** generates more natural pose sequences which correlate better with the audio signals (i.e intrapersonal) and the Human's body pose (i.e. interpersonal) than other models. **DRAM** gets an average score of 4.6 which implies that annotators 'somewhat agree' with all the statements in Section 5. On the other

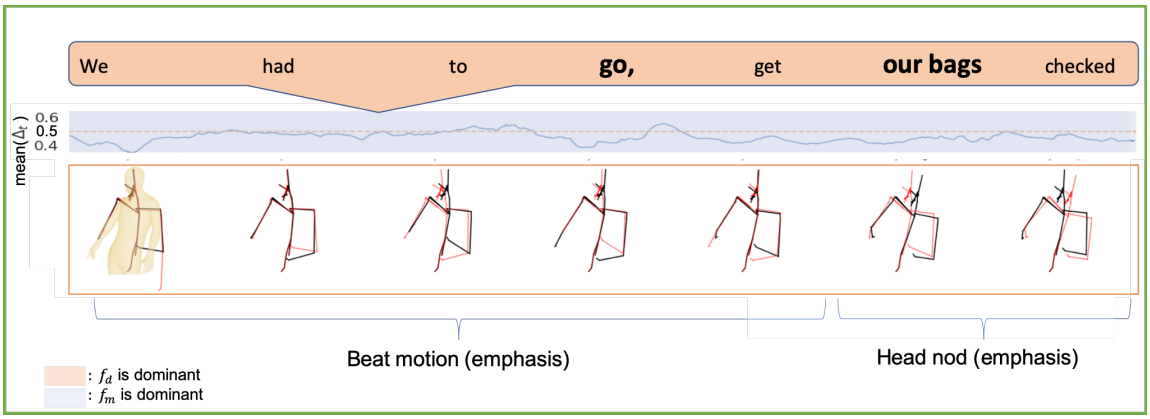

Figure 7: An example demonstrating Intrapersonal Dynamics in predictions made by DRAM. The black skeleton is the current pose, while the red skeleton is the pose from one second in the past. Paralinuguistic cues of emphasis are denoted by a larger font in the spoken sentence. It can be seen that the avatar performs a beat motion for the first five seconds to emphasize the word *go*. For the next two seconds, the avatar nod's its head also to denote emphasis while speaking the words *our bags*. Mean of Dyadic Residual Attention vector (mean(Δ_t)) is moslty below 0.5 which implies that f_m is dominant for the final avatar body pose prediction.

hand, annotators are neutral towards the pose generated by Avatar conditioned Monadic only model.

Qualitative Analysis

Conversations in a dyad contain non-verbal social cues, which might go unnoticed to us humans, but play an important role in maintaining the naturalness of the interaction. Head nod mirroring and Torso Pose switching are two of the most common cues. We pick out 2 cases in Figure 4 with such cues existing in the conversation. Our model detects the head nod and pose switch in the human's pose and is successfully able to react to it.

Another aspect in social cues is hand gestures during conversation. Utterances that are emphasized usually lead to a switch in position of hands almost instantaneously. Our model is able to detect emphasis and moves their hand(s) up and down (Figure 7) to sync with the speech.

Real conversations are a mixture of changing interpersonal and intrapersonal dynamics. Our model is able to detect these changes and react appropriately. In Figure 3, the avatar conducts hand raises which are due to emphasis in the avatar's speech, but when the interlocutor interrupts in with an exclamation, the avatar starts head nodding in agreement while still performing beat gestures to accompany emphasis in its audio signal.

The mean value of Dyadic Residual Vector Δ_t is plotted along the animation to analyze its effects and correlations with changing dynamics in the conversation. First, Δ_t's mean value seems to correlate with changing interpersonal and intrapersonal dynamics. In Figure 3, mean(Δ_t) rises as soon as the interlocuter interrupts the avatar. In Figure 7, mean(Δ_t) remains almost constant as the interlocutor does not seem to have a huge role in that part of the sequence. Second, even though the value of Δ_t correlate with changing roles in a conversation, its absolute value is not extreme. (i.e. at an average it is closer to 0.5 than to 0 or 1). This is not surprising as the contribution of interpersonal and intrapersonal dynamics can often overlap hence requiring both monadic f_m and dyadic f_d models.

7 CONCLUSIONS

In this paper we introduce a new model for the task of generating body pose in a conversation setting conditioned on an audio signal, and interlocutor's audio and body pose. This person specific model, Dyadic Residual-Attention Model , learns to selectively attend to interpersonal and intrapersonal dynamics. The attention mechanism is successfully able to capture social cues such as head nods and pose switches while generating a sequence of poses which appear natural to the human eye. It is a first step towards an avatar for remote communication which is anthropomorphised with non-verbal cues.

8 ACKNOWLEDGEMENTS

This material is based upon work partially supported by the National Science Foundation (Award #1722822). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, and no official endorsement should be inferred.

REFERENCES

[1] Shailen Agrawal and Michiel van de Panne. 2016. Task-based locomotion. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 82.

[2] Chaitanya Ahuja and Louis-Philippe Morency. 2018. Lattice Recurrent Unit: Improving Convergence and Statistical Efficiency for Sequence Modeling. In *AAAI-18*. 4996–5003. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17394

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 3686–3693.

[4] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[5] Jeremy N Bailenson, Nick Yee, Dan Merget, and Ralph Schroeder. 2006. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments* 15, 4 (2006), 359–372.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

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

[8] Justine Cassell and Kristinn R Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 519–538.

[9] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[10] Yu-Wei Chao, Jimei Yang, Brian L Price, Scott Cohen, and Jia Deng. [n. d.]. Forecasting Human Dynamics from Static Images.

[11] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 127–140.

[12] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*. Springer, 152–166.

[13] Hang Chu, Daiqing Li, and Sanja Fidler. 2018. A Face-to-Face Neural Conversation Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7113–7121.

[14] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. 2014. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 33, 4 (2014), 79:1–79:10.

[15] Allen T Dittmann. 1972. The body movement-speech rhythm relationship as a cue to speech encoding. *Studies in dyadic communication* (1972), 135–152.

[16] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[17] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[18] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. 2002. *Trainable videorealistic speech animation*. Vol. 21. ACM.

[19] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*. 4346–4354.

[20] Ruohan Gao, Rogerio Feris, and Kristen Grauman. 2018. Learning to separate object sounds by watching unlabeled video. *arXiv preprint arXiv:1804.01665* (2018).

[21] Alex Graves. 2012. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*. Springer, 5–13.

[22] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. 2017. A Recurrent Variational Autoencoder for Human Motion Synthesis. *BMVC17* (2017).

[23] Uri Hadar, TJ Steiner, and F Clifford Rose. 1984. The relationship between head movements and speech dysfluencies. *Language and Speech* 27, 4 (1984), 333–342.

[24] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. 2018. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 166.

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding* 108, 1-2 (2007), 116–134.

[27] Stanley E Jones and Curtis D LeBaron. 2002. Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of Communication* 52, 3 (2002), 499–521.

[28] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 94.

[29] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 243–255.

[30] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 68.

[31] Dario Pavllo, David Grangier, and Michael Auli. 2018. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv preprint arXiv:1805.06485* (2018).

[32] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. *ACM Trans. Graph.* 37, 4, Article 143 (July 2018), 14 pages. https://doi.org/10.1145/3197517.3201311

[33] Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Louis-Philippe Morency, et al. 2012. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *International Conference on Intelligent Virtual Agents*. Springer, 455–463.

[34] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. 2008. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks* 21, 9 (2008), 1238–1246.

[35] Tomas Simon, Hanbyul Joo, Iain A Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping.. In *CVPR*, Vol. 1. 2.

[36] Namrata Singh and Sarvpal Singh. 2017. Virtual reality: A brief survey. In *Information Communication and Embedded Systems (ICICES), 2017 International Conference on*. IEEE, 1–6.

[37] Anthony Steed and Ralph Schroeder. 2015. Collaboration in Immersive and Non-immersive Virtual Environments. In *Immersed in Media*. Springer, 263–282.

[38] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 95.

[39] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM, 365–369.

[40] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 93.

[41] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. [n. d.]. WaveNet: A generative model for raw audio.

[42] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

[43] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[44] Martin Wöllmer, Moritz Kaiser, Florian Eyben, BjöRn Schuller, and Gerhard Rigoll. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing* 31, 2 (2013), 153–163.

[45] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923* (2018).

[46] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. *arXiv preprint arXiv:1804.03160* (2018).