Digital Humans: Automatic Character Animation

Jonathan Windle

University of East Anglia

This thesis is submitted for the degree of

Doctor of Philosophy

School of Computing Sciences March 2025

Acknowledgements

I would like to express my gratitude to Dr. Sarah Taylor for providing me with the opportunity to study such an exciting topic and for her continuous and invaluable support throughout my research. I also extend my thanks to Dr. Iain Matthews for his guidance and support during the project. Additionally, I greatly value the supervision and assistance of Dr. Ben Milner and Dr. David Greenwood throughout this endeavour.

The Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA) challenges and workshops have been invaluable resources for my learning experience. I want to extend my thanks to all the organizers and participants for these opportunities.

Finally, I would like to acknowledge my family, especially Georgie, for their unwavering support and patience.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Jonathan Windle

March 2025

Publications

- Taylor, S., Windle, J., Greenwood, D., and Matthews, I. (2021). Speech-driven conversational agents using conditional flow-vaes. In *Proceedings of the 18th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '21, New York, NY, USA. Association for Computing Machinery

- Windle, J., Greenwood, D., and Taylor, S. (2022a). Uea digital humans entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 771–777

- Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022b). Arm motion symmetry in conversation. *Speech Communication*, 144:75–88

- Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022c). Pose augmentation: mirror the right way. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA. Association for Computing Machinery

- Windle, J., Matthews, I., Milner, B., and Taylor, S. (2023). The uea digital humans entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 802–810, New York, NY, USA. Association for Computing Machinery

- (2024). Llanimation: Llama driven gesture animation. In *Computer Graphics Forum*, volume 43, page e15167. Wiley Online Library

Abstract

This thesis covers topics related to automatic generation of co-speech gesture animation. This vast field traditionally employs automatic rule-based, statistical, and machine learning approaches. This thesis expands on machine learning approaches, applying new methods to co-speech gesture generation. Initially, one of the most extensive co-speech gesture datasets is examined to provide insight into gesture production and lateral symmetry in gestures. The thesis then focuses on the application of four machine learning generative modelling approaches. Each proposed method answers a specific research question while simultaneously striving for the best performance in automatic gesture animation.

First, the common data augmentation technique of lateral mirroring is shown to be problematic through dataset analysis, which also introduces new gesture analysis methods and statistically derived gesture spaces. The effect of using multiple, body-part-specific decoders is compared to a single decoder that predicts the whole body. This experiment finds that leg motion is negatively impacted while the arms and hands benefit. A novel style-controlled diffusion model focusing on the impact of long-term historical knowledge is introduced. This sheds light on the importance of historical memory, finding performance improved when extended, producing smooth, contextually correct animation with emotive style control. Conversational speech often occurs in a dyadic setting, where the other person's response influences communication, such as through back-channel communication. A model including the second speaker's speech as a feature shows minor improvements, particularly in head nods and gesture turn-taking. An experiment using Large-Language Models (LLMs) as

a feature extractor is performed and evaluated to determine their effectiveness in isolation and combination with audio features. Using LLAMA2 features enables well-timed, contextually rich gestures without an audio embedding demonstrating that Large-Language Model (LLM) features contribute more to the perceived quality of the results than audio features. These findings offer valuable insights for improving automatic co-speech gesture generation.

Table of contents

List of figures

List of tables

Listings

Acronyms

D_{KL} Kullback–Leibler Divergence. 18

AI Artificial Intelligence. 4

BA Beat Alignment. 25, 26

BLSTM Bi-Directional Long Short Term Memory. xix, xx, xxvii, 16, 17, 76, 80–82, 84–93, 95, 175

BPE Byte-Pair Encoding. 152, 155

BVH Biovision hierarchical data. 40

CLIP Contrastive Language-Image Pre-Training. 19

CNN Convolutional Neural Network. 18

CSMP-Diff Contrastive Speech and Motion Pretraining Diffusion. 170–172

ELBO Evidence Lower Bound. 18

FGD Frèchet Gesture Distance. 24, 25

GAN Generative Adversarial Network. 17, 18, 24

GENEA Generation and Evaluation of Non-verbal Behaviour for Embodied Agent. ix, xvi, 5, 6, 8, 9, 20, 21, 23, 26, 40, 41, 75, 76, 89, 90, 124–126, 135, 140, 147, 157, 160, 165, 168, 170, 175, 176

HEMVIP Human Evaluation of Multiple Videos in Parallel. 90, 140

LLM Large-Language Model. vi, vii, 5–7, 150, 152–154, 156, 157, 170, 172, 176–179

LSTM Long Short Term Memory. 16, 17, 76, 178

MFCC Mel-frequency Cepstral Coefficient. 4, 17–19, 33, 34, 105, 151

MPJPE Mean Per Joint Position Error. 24

NMI Normalised Mutual Information. 63

PASE+ Problem Agnostic Speech Encoding. xxvi, 4, 34, 65, 78, 105, 150, 151, 154, 159, 172, 175–177

RNN Recurrent Neural Network. 16

STT Speech-To-Text. 33

TTS Text-To-Speech. 4, 172

VAE Variational Autoencoder. 17, 18

Chapter 1

Introduction

1.1 Motivation

Digital humans are virtually embodied human agents with a vast number of applications regularly associated with computer games and animated television and film productions. However, digital human solutions are now becoming common in several newer areas, including digital assistants, healthcare, and telecommunications, where human-computer interaction is pivotal to their application. Within these areas, the current level of realism and appropriateness of motion is inconsistent with varying effects and impacts in the end use case. For example, some computer games or television and film productions make a stylistic choice to use certain animation styles that are less human-like than is possible with current technology. These forms of media are still considered engaging and well-received. Despite the stylistic option, many computer games, television, and film productions strive for true realism. Current approaches will often rely on motion capture which is expensive and slow. This thesis aims to improve science in the realism of speech-driven gesture methods for digital humans making the generation of digital humans cheap, easy, automated and quick.

This chapter introduces the thesis topic of automated co-speech gesture generation. First, co-speech gesture is defined and described. This provides a baseline knowledge of how

gesture and speech coincide to produce effective communication. Then, a broad overview of automatic gesture generation is described to provide background for where this research fits in. The aims of the research presented in this thesis are then defined, including research questions that should be answered during later chapters. The key academic contributions are then described before discussing the limitations and explicitly stating what the research is not aiming to address. Lastly, an outline of the thesis structure is provided.

1.2 Co-Speech Gesture

Co-speech gesture refers to body movements that occur alongside speech. This thesis focuses on co-speech animation and gesture generation from audio and text speech representations. Gesturing complements speech, playing a pivotal role in human communication and can form both communicative and non-communicative roles. Communicative gestures provide semantic content of utterances, emotion, and emphasis on what's being said. Around 90% of gestures occur during speech articulation [92, 105]. Humans often use arms and hands to represent literal and metaphorical objects, for example, a steep hill or changing hands from one side to the other to represent 'before' and 'after'. These are examples of gestures that rely on semantic context and are directly related to the lexical understanding of the speech. Another common human trait is using 'beats', which are gestures used to emphasise spoken words. These gestures are less related to lexical understanding but commonly coincide with speech timing and intensity. Conversely, non-communicative gestures are self- and object-touching, such as scratching or holding a glass of water. While these do not play a direct communicative role, they are common in communicative settings such as conversation.

Co-speech gestures are greatly important in communication. Iverson et al. [59] found that gesture had a tight relation to childrens' lexical and syntactic development, and children often use gestures to communicate before they use words. While gestures alone may not be an effective method of communication, when paired with speech, they enhance communicative

understanding. Further, removing natural speech gestures reduces the communicative extent of conversational agents, and perceived realism is reduced [34]. Given the importance of gesture in communication and the perceived realism of embodied conversational agents, this serves as motivation to improve the field of automatic co-speech gesture generation.

1.3 Automatic Gesture Generation

Automatic gesture generation involves computationally deriving co-speech gestures from a speech source, typically audio. This work will focus on using deep learning methods to produce smooth, human-like, and appropriate gesture motion that should coincide with a speech signal to make it appear natural and timely.

There are many approaches for automatic gesture generation. Most follow a similar generation pipeline shown in Figure 1.1, which overviews a typical speech-to-gesture animation pipeline. In each case, the method and feature extraction portions are where most contributions lie and provide the most significant impact on performance.

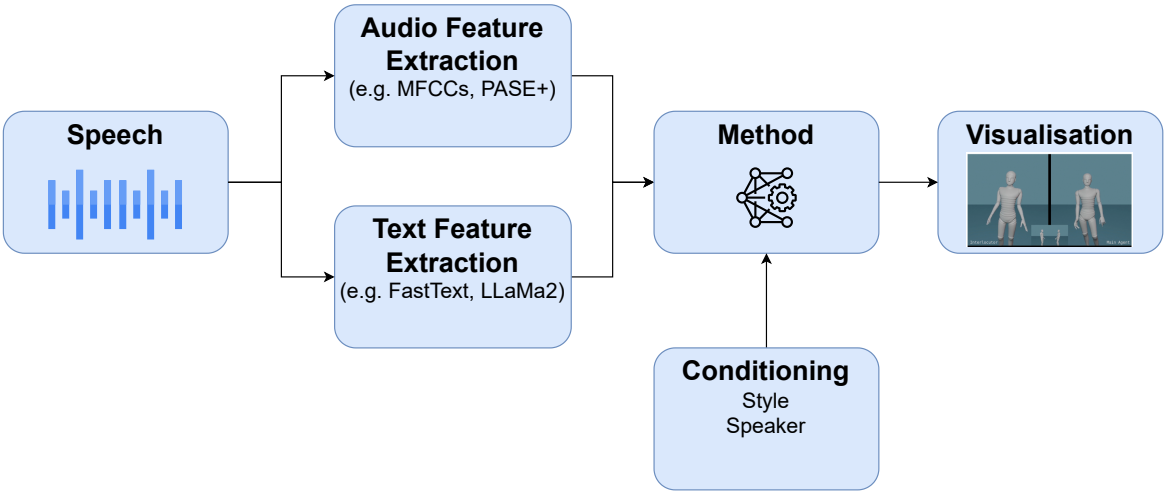


Fig. 1.1 Overview pipeline of automatic gesture generation. This shows the steps involved in going from a speech signal to rendered animation.

Initially, a speech signal is provided. Ideally, this signal will be a high quality, recorded human voice; however, this may also be generated from a Text-To-Speech (TTS) method. Given this speech signal, speech features are extracted. Most commonly, audio features such as Mel-frequency Cepstral Coefficients (MFCCs) or Problem Agnostic Speech Encoding (PASE+) [120] features are extracted from the raw audio signal. Increasingly, text-based features are extracted to incorporate semantic understanding. A text transcript with timings may be manually generated from the speech audio, or automated tools may be used to generate these automatically. Using this transcript, text features such as FastText [16] or LLAMA2 [136] features can be extracted. A conditioning variable, such as the speaker's identity or speech style, may also be provided to influence style control. A method may use audio, text, or a combination of both and optionally a conditioning variable to generate parameters, often joint rotations, to represent human motion. These parameters are then visualised depending on the use case for deployment or evaluation.

1.4 Research Aims and Objectives

This thesis aims to develop generative Artificial Intelligence (AI) methods for co-speech gesture generation. This work will focus on data analysis and generative modelling methods, each with a subset of research questions to answer. Data analysis aims to clarify data formats and data augmentation methods. To determine the validity of data augmentation methods, it is first essential to analyse the current literature regarding the relationship between speech and gesture. The main aim is to develop generative AI models which should strive to perform best and simultaneously answer a particular research question. The work in this thesis should specifically answer the following for co-speech gesture:

- Is lateral mirroring a valid data augmentation technique?
- Is a single decoder or a group of body part-specific decoders preferable?

- Does the amount of historical context of generated gestures influence generative performance?
- Does introducing a second, dyadic speaker influence the gestures generated?
- Can Large-Language Models (LLMs) be used as effective speech feature extractors, and how do they compare to audio features?

.

1.5 Research Contributions

This thesis strives to advance the field of automatic speech-to-gesture generation for 3D character animation. There are five key contributions described throughout this thesis.

Arm Motion Symmetry Analysis: An analysis is performed on a large dyadic co-speech gesture dataset and is published in the Speech Communication journal 2022 [147]. This analysis focuses on arm motion symmetry to determine whether lateral mirroring of a pose is a valid data augmentation technique. Analysis is performed to investigate per-frame and temporal symmetry of motion, and information theory is used to examine the information gain by mirroring the data. This contribution also introduces statistically derived gesture spaces and provides suggestions for including mirrored data in gesture generation methods.

Single vs. Multiple Decoder Comparison: A comparison between using a single decoder to predict the full body and the use of a number of body-part-specific decoders is performed and published in the International Conference on Multimodal Interaction 2022 [145]. Performance was analysed using both objective and subjective measures in the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Challenge 2022 [154], producing a competitive solution.

Gesture Diffusion Network: A novel diffusion model architecture is introduced for speech-to-gesture currently submitted for peer review. This Gesture Diffusion Network is defined as a diffusion model using an underlying Transformer-XL architecture, which can generate smoothly varying animations for any given input size. The model is conditioned on speech as well as speaker style, enabling controllable style animation at inference time. Results are compared to state-of-the-art methods using both objective and subjective measures.

Dyadic Cross-Attention Model: A model that adapts the Transformer-XL architecture for a dyadic setting is introduced and published in the International Conference on Multimodal Interaction 2023 [146]. This contribution alters the underlying Transformer-XL self-attention mechanism to introduce cross-attention. This cross-attention includes the second speaker, known as the interlocutor, during prediction to provide dyadic influence to the prediction. This method was evaluated as part of the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Challenge 2023 [76] and performs competitively when compared to other methods.

LLAniMAtion : Large-Language Model (LLM) driven model: A method of generating gestures using LLAMA2 [136] LLM features is introduced and will be published in the Computer Graphics Forum as part of the Symposium on Computer Graphics [2]. This method compares the use of LLM features and audio features both in isolation and in combination to determine the impact of semantic and prosodic elements on gesture generation. Results are compared to state-of-the-art methods using both objective and subjective measures.

1.6 Limitations

While it is helpful to highlight the contributions and aim of the thesis, it is also valuable to highlight research areas that are not expected to be covered. This thesis covers only

co-speech gesturing. The data used includes conversation while standing; therefore, the leg motion is included. While the conversational agent can walk or sway, this is typically limited to mimicking the context of a standing conversation rather than conversation during an activity such as walking. Activity generation is a separate research area to co-speech gesture generation. For example, activity generation may generate explicit activities such as jumping or strafing. This work focuses on generating the natural motion of co-speech gestures. Some work includes a style aspect regarding age and affective state; however, the focus remains on co-speech gesture generation in these stylistic settings. For example, if scared, the speaker may mimic crouching to hide while speaking. However, the goal is not necessarily to generate crouching explicitly.

While facial motion is an essential aspect of a digital human, the work in this thesis will include head motion, e.g., nodding and shaking. Still, it will not include the animation of facial features such as frowning or lip-syncing.

1.7 Thesis Structure

The following two chapters introduce the gesture generation problem. Chapter 2 introduces the gesture generation landscape, defining the types of gestures and methods of gesture generation and evaluating gesture generation methods. Chapter 3 describes gesture datasets, introducing the modalities involved, data capture methods, and also describing the relevant dataset releases. Each subsequent chapter is then focused on each contribution described in Section 1.5. Chapter 4 performs arm motion symmetry analysis, Chapter 5 performs comparison of single and multiple decoders, Chapter 6 introduces the novel Gesture Diffusion Network, Chapter 7 explores the introduction of the dyadic interlocutor into a model and Chapter 8 explores the use of LLM features for gesture generation. Finally, Chapter 9 concludes the thesis and outlines further work that has been identified from this work.

Chapter 2

Gesture Generation Landscape

2.1 Introduction

This chapter reviews the related work regarding gesture generation literature, focusing on co-speech body gestures. Given the array of gesture types, it's crucial first to understand when each type is performed and whether their production is related to speech prosody or semantic understanding. This chapter defines gesture types and discusses their relationship to speech. This provides context for which gestures are appropriate based on speech content and introduces the variety of co-speech gestures that much of this work aims to predict.

As a large portion of this thesis's contributions involves generative modelling, it's useful to understand the current landscape of generative modelling for gesture generation. An overview of current methods is described with a small introduction to each technique. These methods introduce the trends and improvements of gesture generation over time and provide the context for where the contributions of this work expand from historically and why they should be considered in the future.

The Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) workshops [74, 154, 76] and challenges are a key driving factor in some of this work. Therefore, this workshop is introduced in this chapter and described to provide context for how and

why two contributions in this thesis are driven and evaluated by these GENEA challenges. The challenge aims to overcome some difficulties in assessing gesture generation methods. This chapter also introduces common subjective and objective methods of evaluation, highlighting the benefits and problems of each.

2.2 Co-Speech Gesture Types

Given the importance of co-speech gestures in communication, understanding how these gestures play that role is critical to examining co-speech data. For this thesis, although gesture, as a generic definition, refers to any body motion, the focus is on co-speech body motion, with a particular focus on arm and hand motion.

Gestures can form both *communicative* and *non-communicative* roles. *Non-communicative* gestures are referred to as *adaptor* movements while *communicative* gestures can be classified under five definitions from McNeill [94]: *emblematic, deictic, iconic, metaphoric* and *beat*. To ensure realism, an automatic speech-to-gesture method should produce a natural and appropriate mix of all communicative and non-communicative motions.

2.2.1 Adaptor

The only non-communicative gesture definition is adaptor motions. These motions are touches with objects or oneself, such as holding a glass or scratching. Figure 2.1 shows an example adaptor gesture where a person scratches their head while thinking. While these gestures bear no meaning in the conversation, these movements can portray helpful information regarding the speaker.

Adaptors may indicate a speaker's state. Freedman [39] suggests that self-adaptors can signal the speaker's need to focus and concentrate. Waxer [142] found that individuals with low emotional stability, i.e. anxiety and depression, produced more self-adaptor movements

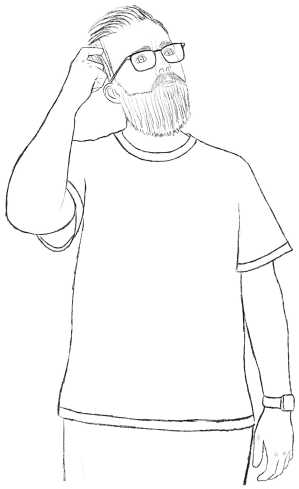


Fig. 2.1 Example adaptor gesture - scratching head while in thought. Illustration inspired by McNeill [92]

during speech. Neuroticism was found to be associated with self-touching, pausing during a conversation, and an absence of expressive gesture [19]. Neff et al. [102] provide guidance on embodied conversation agent design, indicating that it is essential for particular personality types to produce non-communicative gestures, such as self-adaptors.

Given the relationship between adaptor motion and personality, particularly neuroticism, these motions are valuable studies in gesture generation, mainly when the speaker's speech style or emotion is to be preserved.

2.2.2 Emblematic

Gestures that can replace speech and bear conversational meaning are considered *emblematic* gestures. These gestures are commonly associated with exact words or phrases uttered during speech and provide semantic meaning. Each emblematic gesture may convey a different meaning depending on context or some pre-defined rules.

Figure 2.2 shows three examples. The finger ring shown in Figure 2.2a is often used to denote "O.K" or "affirmative". A thumbs up shown in Figure 2.2b can typically denote "good" or "yes'. Emblematic gestures can also be culture-specific; for example, Figure 2.2c

(a) Finger Ring - “O.K.”, “Affirmative”

(b) Thumbs up - “Yes”, “Good”

(c) Purse Hand - “Che vuoi?” “What do you want?”

Fig. 2.2 Example emblematic gestures. Each sub-caption contains the format gesture description - A list of examples of associated words or phrases. Illustrations inspired by Morris [99]

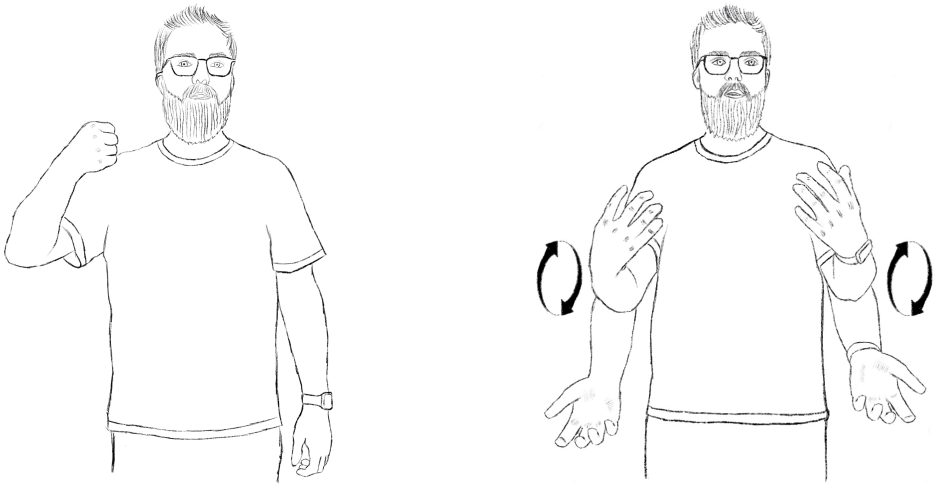
shows a purse hand or finger pinch gesture. This is a gesture commonly found in Italian culture [63] that is used to denote the phrase “Che vuoi?” and translates to “What do you want?”.

Emblematic gestures are not particularly common during conversation and are often invoked when channels of vocal communication are limited. Vocal channels may be restricted due to noise, distance, or external factors, such as a building site or during a deep sea dive. The thumbs-up gesture shown in Figure 2.2b is an example of an emblem with multiple meanings in situations. In most situations, this means “good” or “yes”; however, when diving, this gesture indicates that a diver is about to ascend and should not be used to suggest the conventional “good” sentiment.

These types of gestures are rare during unimpaired speech conversations. They are not typically used as semantic, communicative gestures but have a more pragmatic function. Therefore, these emblematic gestures can be re-categorised during co-speech gestures to be *metaphoric* [143].

2.2.3 Iconic and Metaphoric

Gestures that resemble a particular physical aspect of the conveyed information are known as iconic gestures. These aim to illustrate some properties of the speech and relate to the semantics of the speech. For example, when describing the shape or size of an object,



(a) Iconic Gesture - Mimic the use of an umbrella

(b) Metaphoric Gesture - Circling arms to mimic time passing

Fig. 2.3 Example Iconic and Metaphoric gestures. Illustrations inspired by McNeill [92]

a speaker may use their hands to resemble this. Figure 2.3a shows an example speaker mimicking the use of an umbrella during the speech; while no umbrella is present, the gesture shows the use of one. These are common in co-speech gestures and enhance understanding by illustrating the act or object being discussed for additional context.

Like iconic gestures, metaphoric gestures illustrate speech content; however, rather than directly illustrating the object/act, the gesture illustrates this through a metaphoric third element. McNeill [92] gives an example of a metaphoric gesture as when moving their arms in a circular motion to resemble a speaker saying “And now we get into the story proper”. Figure 2.3b shows an example of a speaker moving their arms in a circular motion to represent time passing.

2.2.4 Deictic

Gestures often involve pointing movements, which are known as deictic gestures. These pointing motions may indicate a tangible person or object, a physical location or direction, but they may also be used metaphorically for abstract or imaginary things. These gestures

are particularly important when shifting focus; for example, a person may point at a specific area of an object and wish to shift attention to another, in which case the pointing gesture moves. Each hand may perform its own pointing gesture, where two points of focus relate to each other. For example, when describing the layout of a building, each hand may be used to indicate landmarks in relation to each other. Figure 2.4 illustrates an example Deictic gesture. This person is seen pointing to an object above their head while uttering the words “If you look at this up here” as a means to shift focus to a particular object.

Fig. 2.4 Example deictic gesture - Person pointing up at an object uttering “If you look at this up here”.

2.2.5 Beat

Beat gestures are common in co-speech gestures. These gestures are rhythmic motions commonly synchronised with prosodic events in speech [114]. These movements are simple, typically small and repetitive motions that do not relate to the semantic information of the speech. Figure 2.5 shows an example beat gesture where the right arm takes two positions. Each position would represent a motion beat, i.e. a sharp change in velocity of arm motion that is in time with an audio beat. Beat gestures are commonly associated with lexically stressed syllables in words and are often used to emphasise certain areas of speech.

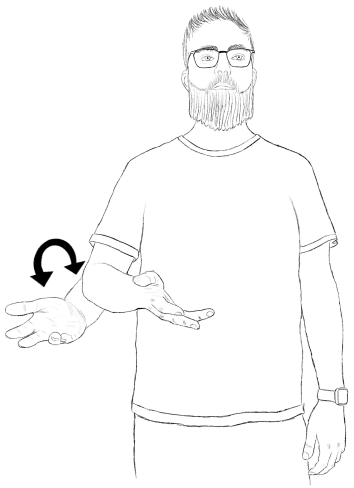


Fig. 2.5 Example beat gesture - small repetitive motion with the right arm, each illustrated right arm position represents a motion beat that would be synchronised with an audio beat.

2.3 Automated Co-Speech Gesture Generation

There are many existing approaches to automated co-speech gesture generation. Early techniques employed rule-based generation and statistical methods. Most modern techniques use a learning-based approach, which will be the focus of this thesis. This section describes the earlier rule-based and statistical approaches before reviewing deterministic and stochastic learning-based approaches.

2.3.1 Rule-based Approaches

Research first approached the automated speech-to-gesture generation problem through the use of rule-based approaches [50, 23, 21, 90, 69, 110, 127, 22, 68, 111, 79, 135]. Rule-based techniques mainly dealt with semantic aspects of human gesturing, using carefully designed heuristics to select an appropriate gesture for speech. In 2001, Cassell et al. [23] introduced the Behavior Expression Animation Toolkit. This toolkit allowed animators to produce nonverbal behaviours assigned based on actual linguistic and contextual analysis of the typed text. While rule-based methods have scarcely been developed in recent years, they

are still occasionally released as in 2022, Zhou et al. [158] presented a graph system that embeds style and rhythmic information for an audio clip and selects the closest matching sequences from a database of gestures. Chiu et al. [27] used Deep Conditional Neural Fields for estimating a *gestural sign* from the utterance, parts-of-speech tags and prosodic features. However, rule-based approaches typically lack diversity in the generated motions and are limited in capturing the nuances and variations of natural human gestures.

2.3.2 Statistical Approaches

Early data-driven approaches were based on probabilistic modelling. For example, in 2008, Neff et al. [101] computed the probability that a body gesture from a fixed library of gestures is to be generated, conditioned on context. In 2014, Chiu and Marsella [26] used Gaussian process latent variable models to learn a mapping from speech to hand gestures through an intermediate representation of gesture annotation. Yang et al. [151] constructed a motion graph that preserves the statistics of a database of recorded conversations. This graph is then searched for a motion sequence that closely respects coordination to the phonemic clause, listener response, and partner's hesitation pause. As early as 2005, Kipp [66] proposed a statistical system that models an individual's gesture profile, such as handedness, timing and communicative function, by analysing an annotated co-speech gesture dataset. Gesture profiles were then modelled using statistical models from speech recognition and dialogue act recognition [122]. Levine et al. [81] trained a hidden Markov model on prosody features. This idea was later integrated into a reinforcement learning framework [80]. Exposing the underlying probability distribution of body motion conditioned on speech is desirable as it allows for non-deterministic sampling. However, the Gaussian assumptions of these prior works are somewhat limiting.

2.3.3 Learning Approaches

As motion capture data became more available and deep learning techniques developed, the focus shifted to learned data-driven approaches. These approaches vary drastically and can be both deterministic and stochastic in generation.

2.3.3.1 Deterministic Approaches

Autoregressive Methods: Autoregressive methods are popular generative deep learning models. These models continually predict the next element in a sequence using historic context from previous inputs and outputs. These models tend to allow for generative sequences of non-fixed length.

Yoon et al. [153] developed a text-driven system based on a recurrent encoder-decoder network for robot co-speech gesture generation. Kucherenko et al. [73] instead regressed pose from text and audio features using an autoregressive, sliding window feed-forward neural network architecture.

Bi-Directional Long Short Term Memory (BLSTM): BLSTM models are a popular technique found in gesture generation [131, 51]. Long Short Term Memory (LSTM) models are a form of Recurrent Neural Network (RNN) capable of learning long-term dependencies introduced by Hochreiter et al. [56]. An LSTM [56] comprises a chain of repeating cells, each producing an output and hidden state. It is these hidden states that provide long-term dependencies through the use of three gates. The first, *forget* gate determines whether the information from the previous timestep should be kept. The second, *input* gate, is used to quantify the importance of the new information. The final, *output* gate is the output from the LSTM cell. A BLSTM [44] contains two cells that process the data in both the forward and backward directions.

Takeuchi et al. [131] used phonemic features from speech audio data as input to output time sequence data of rotations of bone joints. Hasegawa et al. [51] used a BLSTM to learn the speech-gesture relationships with both backward and forward consistencies over a long period of time. This model used perceptual features extracted from audio using Mel-frequency Cepstral Coefficients (MFCCs).

Gated Recurrent Unit (GRU): GRU models have also been used [37, 72, 152]. GRUs are similar to LSTMs but contain only two gates. The *update* gate determines how much of the past information is preserved, and the *reset* gate decides how much of the past information to forget.

Ferstl et al. [37] use a GRU-based network with an encoder-decoder structure that takes in prosodic speech features and generates a short sequence of gesture motion. Kucherenko et al. [72] utilised GRU networks to form a motion encoder *MotionE* and a motion decoder *MotionD*, as well as a speech to corresponding motion representation encoder *SpeechE*. At inference time, *SpeechE* predicts the latent motion representations based on a given speech signal, and *MotionD* then decodes these representations to produce motion sequences.

2.3.3.2 Probabilistic Approaches

Recently, probabilistic models such as Generative Adversarial Networks (GANs) [48, 152, 47], Variational Autoencoders (VAEs) [9, 41, 42, 82] and Flow-based models [133, 6, 53] have gained popularity for gesture generation since they better model the ambiguities between speech and human motion over deterministic approaches.

Generative Adversarial Network (GAN): A GAN consists of two key components, a *generator* and *discriminator*. The *generator* aims to produce samples that closely resemble the ground truth data, and the *discriminator* reinforces the *generator* by categorising the output as real or fake.

Yoon et al. [152] incorporate the multimodal context of speech text, audio, and speaker identity as input to the model, trained as a GAN. Habibie et al. [48] utilised a Convolutional Neural Network (CNN) architecture within a GAN to produce gestures from MFCCs extracted from speech. Habibie et al. [47] passed the nearest neighbour from a gesture database to a conditional Generative Adversarial Network to refine the motion.

Variational Autoencoders (VAEs): Autoencoders are encoder-decoder models that learn a parametrised mapping from data through a lower-dimensional latent space and predict it back to be the same data. The encoder encodes data to the lower dimension, and the decoder performs the mapping from latent space to data space. Variational Autoencoders (VAEs) instead describe an observation in latent space in a probabilistic manner. The encoder, $q_\phi(\mathbf{h}|\mathbf{x})$, learns the parameters of a Gaussian distribution that is used to approximate the posterior distribution $p(\mathbf{h}|\mathbf{x})$; where ϕ are the encoder network parameters, $\mathbf{x}$ is the input and $\mathbf{h}$ is the latent space. The decoder, $d_\theta(\mathbf{x}|\mathbf{h})$, maps samples from the variational distributions back to the input domain, with network parameters θ. VAEs are trained by maximising the Evidence Lower Bound (ELBO), which consists of two terms; a reconstruction error and the Kullback–Leibler Divergence (D_{KL}) between the approximate posterior and the prior distribution [15].

Li et al. [82] utilised a conditional VAE to model speech-to-gesture. Ghorbani et al. [41, 42] used a VAE to learn a style embedding, making it easy to modify style through latent space manipulation. Using this latent style embedding and a separate audio embedding, the two latent vectors were used as condition variables to generate stylised gestures. Ao et al. [9] use a Vector Quantised (VQ)-VAE to map rhythm and semantic information to gesture. VQ-VAEs adapt the standard VAE model to use a discrete codebook of latent embeddings.

Flow-based: Flow-based models are also popular [133, 6, 53]. Normalising flows provide a highly flexible method for transforming a simple distribution (E.g. Gaussian) to a more

complex distribution through a series of invertible and differentiable transformations [123]. These transformations can be considered as expansions or contractions of the initial density.

Alexanderson et al. [6] adapted a flow-based locomotion model [53] to produce gestures from MFCC speech encodings. This method produced upper-body motion only and could be conditioned on both speech and arbitrary style parameters such as average hand height, gesture speed and gesture radius.

Diffusion: Diffusion models [55, 104] are currently considered state of the art in probabilistic generative modelling due to their impressive ability to produce realistic and diverse output, particularly in text-to-image generation [32, 124, 119, 126]. Recent research has shown that diffusion models can also handle time-series or sequence-based data, such as generating 3D motion from text [65, 134, 157] or dancing and speech gestures from audio [28, 8, 156]. Sequence-based diffusion models have used Transformer architectures [134, 65] to effectively model temporal information. The diffusion process consists of two steps, *noising* and *denoising*. The *noising* process consists of a Markov Chain $q(\mathbf{x}_k|\mathbf{x}_{k-1})$ for $k \in \{1,...,K\}$ where K denotes the number of diffusion steps. During training, given a ground truth sequence of poses $\mathbf{x}_0$, the Markov Chain adds noise progressively until $q(\mathbf{x}_K|\mathbf{x}_0)$ approximates a standard normal distribution and no longer resembles $\mathbf{x}_0$. The *denoising* process is a parameterised backwards process $p(\mathbf{x}_0|\mathbf{x}_k)$ which gradually reduces the noisy data point $\mathbf{x}_k$ to the original data point $\mathbf{x}_0$. This parameterised denoising step may include additional conditioning information, such as speech audio.

Alexanderson et al. [8] used a Conformer-based diffusion process with control over motion style. This method uses MFCC features to generate realistic, stylised gestures. Ao et al. [10] use a Contrastive Language-Image Pre-Training (CLIP) model that extracts style representations from multiple modalities such as text, motion clip or video. The approach uses a latent diffusion model that utilises the CLIP embeddings, audio acoustic features, and text, T5 [149] embeddings.

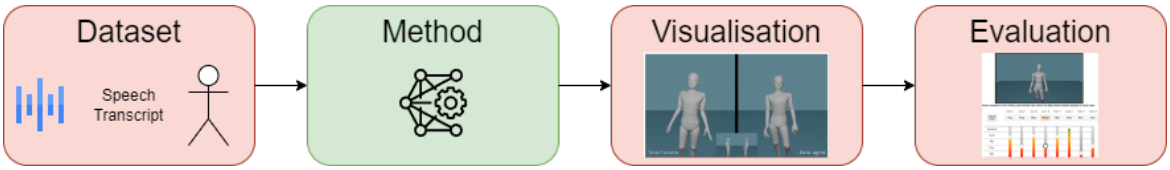


Fig. 2.6 Overview of Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Challenge pipeline. Red box defines a standardised and fixed stage, Green box defines a differing stage between each participant submission.

2.4 Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA)

A particular motivation for this thesis is driven by the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Workshop. This workshop aims to bring together researchers working on generating and evaluating nonverbal behaviour for social robots, virtual agents or the like. This workshop plays an essential role in the literature on co-speech gesture generation, particularly in attempting to standardise comparisons between methods.

The GENEA Challenge is a significant part of this workshop. This is a recurring challenge, taking place in 2020, 2022, and 2023, and it aims to serve as a means of qualitatively analysing co-speech gesture generation performance over time. For this challenge, organisers release a dataset, and participating teams create a speech-to-gesture method and submit synthesised motion for a test set where only the speech information is available at inference time. These synthesised motions are then compared and evaluated in a large-scale evaluation of gesture generation models.

The GENEA Challenge follows the same structure each year and provides standardisation for certain parts of the challenge pipeline. Figure 2.6 gives an overview of this pipeline, showing each stage. The *Dataset* stage is standardised so that each participating team is provided with precisely the same data at the same time. This data has varied across each

recurring year and is explained further in Section 3.4.2. Each year, a dataset of speech, audio and text transcripts and corresponding 3D co-speech gesture motion is provided. The pipeline differs between submissions in the following *Method* stage. This is where participating teams create their methods of gesture synthesis. There are no limits on the techniques participants can use except the synthesis process must be automated, i.e. motion cannot be manually edited after synthesis, but automatic processing such as smoothing is valid. Each participating team provides synthesised output for the test set, where the ground truth motion is not provided. Natural motion and baseline methods are also provided by organisers to aid evaluation and to monitor the performance of gesture generation over time across multiple Challenges. This synthesised motion is then passed to the next step, *Visualisation*. The visualisation is standardised across all methods to ensure fair and standardised evaluation between methods. Once visualised, the *Evaluation* stage is performed. This is again standardised for all methods and consists of a large-scale subjective evaluation utilising a user study.

The GENEA Challenge evolves as the gesture generation literature changes over time. The first challenge in 2020 focused on predicting the upper body only and no hand motion. The second challenge in 2022 evolved to predict the full body and finger motion as discussed in Chapter 5. The third challenge in 2023 encouraged using the second speaker in a dyadic conversation as an additional data source available at inference time, as discussed in Chapter 7.

2.5 Embodied Conversational Agent Evaluation

Evaluating generated gesture quality is an inherently complex problem. This stems from the one-to-many mapping of speech and gesture. Many gestures can be considered accurate and appropriate for the same speech. Gestures are also spontaneous, highly idiosyncratic and non- periodic, which adds further complexity to the evaluation.

Three primary motion characteristics are essential to evaluate: *human-likeness, appropriateness* and *stylistic accuracy*. *Human-likeness* is regarding how natural and close the generated motion is to ground-truth motion. *Appropriateness* pertains to whether the generated motion is appropriate to the speech utterance. *Stylistic accuracy* is an appropriate measure only when there is a particular style conditioning, which is how well the predicted motion fits the expected style. This section describes the standard evaluation techniques for gesture generation, including subjective and objective measures that aim to evaluate the primary motion characteristics.

2.5.1 Subjective

Subjective measures through the use of human perception studies are widely considered the most effective measure of gesture generation performance [106, 154, 76]. These methods typically consist of pairwise or multiple clips being compared side-by-side [106, 154, 76] and asked a corresponding question relating to the comparison of clips. Each question will typically involve rating each clip on a pre-determined scale or a preference test, in which the user indicates a preference between clips. To ensure the methods are compared fairly, each condition should have results rendered in the same environment and using the same avatar. This ensures the effect being measured is the desired question and there is no bias towards a particular environment or avatar.

2.5.1.1 Example Questions

The benefit of human perception studies is the ability to ask questions regarding different aspects of motion. These questions often cover *human-likeness* of the motion, *appropriateness* of the motion to the speech, and *stylistic accuracy* if the style is a feature.

Human-likeness can be evaluated with or without the speech signal, depending on the study setup. Human-likeness is evaluated independently of the appropriateness in the 2022

and 2023 Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Challenges [76, 154]. All clips are rendered using the same avatar and speech signal; however, the speech signal is muted to participants on playback. This ensures the evaluation only examines human-likeness independent of appropriateness. Human-likeness can also be evaluated with appropriateness in mind in a combined comparison with a question such as "Which character's motion do you prefer, taking into account both how natural-looking the motion is and how well it matches the speech rhythm and intonation?" as in Alexanderson et al. [8].

Should a comparison be made to exclusively assess *Appropriateness* with differing human-likeness qualities, the perceived appropriateness or style accuracy may be interfered with by the differing quality of motion [74]. Therefore, the matched vs. mismatched paradigm has been proposed to alleviate this effect [60, 154, 76, 121]. A *matched* sequence is the predicted motion sequence for the particular speech, whereas the *mismatched* clip is motion generated for a different speech sequence using the same method. This ensures the perceived human-likeness independent of speech is equal. Users are asked to choose the most appropriate clip for the speech and assess gesture appropriateness for speech, rhythm, and interlocutor behaviour.

In Alexanderson et al. [8], *Style Accuracy* has been evaluated through a user study. The question was posed as "Based on the body movements alone (disregarding the face), which of the two clips looks most like **{STYLE}**?" where **STYLE** is a phrase that is representative of the conditioning style label. Speech audio was not permitted for this study, as one modality can affect the perception of the other [17, 8, 74, 60]. This allows the users to evaluate the motion style independent of vocal style.

2.5.1.2 Challenges with Subjective Evaluations

Although subjective evaluation is a particularly effective evaluation method in gesture generation, it has challenges. The main challenges are the time-intensive and expensive nature of user studies. There are no particular suggestions on the number of participants; however, it is essential to ensure a sufficient number of participants and questions being asked to reveal any statistical significance.

2.5.2 Objective

Objective measures mitigate the time and expense of subjective user studies. A concern regarding objective measure is due to the one-to-many mapping of speech and gesture, which means that intuitive distance metrics such as Mean Per Joint Position Error (MPJPE) and Percentage Correct Keypoints (PCK) [95] are inappropriate. Instead, a combination of metrics can be used [8, 84, 152].

2.5.2.1 Frèchet Distance

Generative models in many domains are using Frèchet Distance metrics [54, 64, 85, 12, 137]. Namely, the Frèchet Inception Distance, which Heusel et al. [54] designed to evaluate the quality of generated images and the performance of GANs.

Frèchet Gesture Distance (FGD) was introduced by Yoon et al. [152]. When comparing two generation methods, the comparison can be done in feature or world space. For feature space comparison, an autoencoder is trained on the ground-truth data. For each set of poses to be compared, the pre-trained autoencoder aims to extract domain-specific features using the encoder. Given a mean and variance to represent the Gaussian distribution of poses in feature or world space, the FGD score is the Frèchet distance between the two multivariate Gaussian distributions. The FGD between the ground truth gestures, X and the generated gestures, $\hat{X}$, $FGD(X,\hat{X})$ is defined in Yoon et al. [152] as:

$$FGD(X,\hat{X}) = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}} \tag{2.1}$$

where μ_r and Σ_r are the first and seconds moments of the latent feature distribution, Z_r of real human gestures, X and μ_g and Σ_g are the first and second moment of the latent feature distribution Z_g of generated gestures $\hat{X}$.

This measure indicates a similarity between generated and ground truth poses but does not include temporal alignment with the speech in its measure. Kucherenko et al. [77] found FGD has a moderate correlation with gesture human-likeness ratings from a user study where human-likeness was measured independently from speech appropriateness.

The Frèchet distance measures have been used for more than just raw pose representation; they have also been extended to characteristics. Ng et al. [103] uses Frèchet Kinetic Distance (FD_k), which is similar to FGD. However, there is no auto-encoding process. Instead, the first derivative of each joint is used to determine the distribution of velocities for both the ground truth and predicted motion. FD_k is the Frèchet Distance between these two distributions.

The FGD measure in feature space relies on a pre-trained pose auto-encoder, which introduces difficulties when reproducing results from different publications. It is impossible to directly compare results from publication to publication unless the pre-trained autoencoder is released and used and the pose skeleton hierarchy matches that of the autoencoder. This also makes the autoencoder dataset specific, making comparing models trained on different datasets more challenging.

2.5.2.2 Beat Alignment

To mitigate the lack of temporal alignment measure in the FGD score, it is often coupled with Beat Alignment (BA). BA was initially introduced to evaluate dance synthesis by Li et al. [83], measuring how closely a movement matched a beat in the music. Liu et al. [84] since adapted this measure for speech gestures. BA provides a measure of synchrony between

speech and gestures using a Chamfer Distance between audio and gesture beats. In Liu et al. [84], an audio beat is determined using a root mean square onset and a defined threshold. A motion beat is defined using local minimums in velocity for a particular joint, typically a left or right wrist. The BA score can be defined as:

$$BeatAlign = \frac{1}{m}\sum_{i=1}^{m} exp(-\frac{min_{\forall t_j^y \in B^y}||t_i^x - t_j^y||^2}{2\sigma^2}) \tag{2.2}$$

where $B^x = \{t_i^x\}$ is the kinematic beats, $B^y - \{t_j^y\}$ is the music beats and σ is a parameter to normalise sequences with different FPS.

2.6 Discussion

This chapter has reviewed how speech and gesture coincide with an insight into the different roles gestures play in communication. It has also introduced the various approaches to automated co-speech gesture generation, from rule-based, statistical and machine learning approaches. The Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) workshops are a key influence in co-speech gesture generation and this has been introduced and explained, with its pivotal role in driving co-speech gesture evaluation methods. This chapter also explains these evaluation methods and their inherent difficulties.

The rest of the work in this thesis will focus on the generation of co-speech gestures using machine-learning approaches. The approaches will include deterministic and probabilistic approaches and will be evaluated using the recommended methods discussed in this chapter. Two introduced methods are also evaluated as part of the 2022 and 2023 GENEA Challenges.

Chapter 3

Gesture Datasets

Contributing Publications

- Taylor, S., Windle, J., Greenwood, D., and Matthews, I. (2021). Speech-driven conversational agents using conditional flow-vaes. In *Proceedings of the 18th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '21, New York, NY, USA. Association for Computing Machinery

3.1 Introduction

Gestures have historically been primarily researched in psychology. This research has relied on physical observation of humans or video recordings [86, 92, 105, 91] or extracting hand position and head orientation from video [117]. Recent advancements in data capture and the availability of high-quality gesture data have led to advances in data-driven speech-to-gesture models. This chapter addresses gesture data from the perspective of data-driven approaches to automatic character animation, its modalities, processing and augmentation, and analysis of currently released datasets to provide an in-depth understanding of the current dataset and modality options for speech driven gesture generation.

3.2 Modalities

Data-driven speech-to-motion generation is a multimodal task, with each modality playing a pivotal role in ensuring effective communication. This section focuses on three critical modalities for communicative effect during conversational gesture: motion, audio and speech transcription as text. This section introduces the different modality data types.

3.2.1 Motion

The human body follows a pre-defined skeleton with numerous bones. Although every human body differs, the fundamental skeleton structure remains the same or similar between individuals. This is convenient because the underlying motion format for animating digital humans can follow a set structure and graph topology. Motion formats formed from a skeleton follow a set topology of joints (nodes) and bones (vertices). Each node represents either the rotation or position of the respective joint, while the vertices represent a corresponding bone length. Figure 3.1 shows an example skeleton hierarchy used in motion modelling.

Each skeletal structure can contain many kinematic chains, which provide constrained motion to a chain of joints depending on their interactions. Using known joint angles, Forward Kinematics derives each joint's position and movement velocity. Conversely, Inverse Kinematics uses the known position of the kinematic chain end effector. It derives the position and rotations of the rest of the chain to reach the desired position of the end effector. Inverse Kinematics are helpful in character animation; however, they can only look realistic and natural, assuming the underlying kinematic rig is derived from a complete dynamic physical model [125]. Due to this complex physical model requirement and current data capture techniques, the data-driven gesture generation literature favours Forward Kinematics and predicts each joint rotation or position.

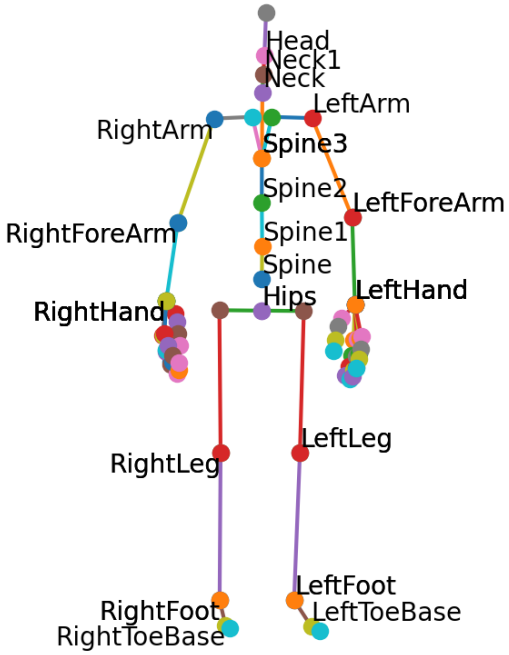


Fig. 3.1 Example skeleton topology, showing joints (nodes) with their respective rotations and bones (vertices) with their respective bone lengths.

Given a pre-defined skeleton, the rotations and bone lengths of each joint, Forward Kinematics produces a skeletal pose for each frame that can have a mesh applied to render the animation. This is commonplace for animating 3D motion material, and numerous pre-existing animation pipelines exist. These pipelines often use a pre-defined file format called Biovision hierarchical data (BVH). The BVH format consists of two parts: a header (hierarchy) section, which describes the hierarchy and initial pose of the skeleton, and a data (motion) section, which defines each frame of motion in a sequence. The hierarchy section describes the skeleton with its root joint and corresponding kinematic chains to multiple end effectors. Listing 3.1 shows an example BVH file. This file only contains two joints, `Hips` and `Spine`, each includes an `OFFSET` parameter which defines the length and direction

to draw the parent segment, for example, in Listing 3.1, the offset defines that `Hips` joint should be drawn at `-20.557875 87.152710 -32.076614` from the origin and the `Spine` joint should be drawn `0.000000 8.814246 -2.080107` from the `Hips` joint, each in the `X`, `Y` and `Z` direction respectively. Each joint in this section also contains a `CHANNELS` keyword followed by value labels. While there is no standard for the number of values possible for each joint or the label value, each label typically corresponds to a position or rotation. The rotation represents an Euler angle rotation change while the position is a second offset, which moves the joint as required; for example, the root joint has an initial offset. However, the global motion of the skeleton is represented as an offset to this original value. It is expected to have the root joint contain six values, Three for position offset and three rotations, while all other joints control three rotations values only. Some data include three position values for each joint to provide a per-joint position offset, which may be helpful to account for bone stretching during the motion capture process. The `MOTION` section of the BVH file contains some information regarding the number of frames in the file and the frame time, which defines how long each frame is representing; for example, in 30fps motion, a frame time will be $0.03\dot{3}$ seconds (33ms). Each subsequent line is space-separated values that follow the exact order of the defined skeleton and their respective `CHANNELS` values. Each line of the BVH file and, therefore, a pose at frame n can be represented mathematically as:

$$\mathbf{p}_n = [x_n, y_n, z_n, r_{j,1,n}, ..., r_{j,3,N}] \tag{3.1}$$

where x, y, z denote the global skeleton position and $r_{j,1:3,n}$ form rotations for each joint j as Euler rotation representation.

```
HIERARCHY
ROOT Hips
{
    OFFSET -20.557875 87.152710 -32.076614
```

```
    CHANNELS 6 Xpos Ypos Zpos Zrot Yrot Xrot
    JOINT Spine
    {
        OFFSET 0.000000 8.814246 -2.080107
        CHANNELS 3 Zrot Yrot Xrot
    }
}
MOTION
Frames: 2
Frame Time: 0.0333333
5.23 88.29 8.71 -0.24 1.33 -17.24 0.82 0.048 10.29
6.11 88.33 7.97 -0.27 0.98 -17.23 0.64 0.069 10.23
```

Listing 3.1 BVH Example

3.2.1.1 Joint Angle Representations

Although Euler angles are an intuitive joint angle representation, they have inherent issues. There is a many-to-one mapping from angles to joint position, this is to say that there are many different combinations of 3 angle values that will result in the same end result. Zhou et al. [159] state that Euler angle representations are discontinuous for SO(3), which is the group of all rotations around an origin in 3D space and can result in large regression errors. As well as this, Euler angles suffer from the Gimbal lock paradigm. This is where two of the three axes align resulting in a loss of one degree of freedom. To mitigate these issues in a deep learning environment, alternative angle representations are often employed such as the exponential map representation and the 6 degrees of freedom representations proposed by Zhou et al. [159].

3.2.2 Speech

High-quality motion and high-quality audio are both required for data-driven gesture generation. Audio capture is typically cheap compared to motion capture but requires several considerations when recording. Recording audio for co-speech gestures should be kept consistent and with noise limited as much as possible. In addition to speech audio, speech text transcriptions can be desirable as an extra modality to provide speech semantics.

It is ideal to ensure this happens in a controlled, consistent environment with the same recording device when recording audio. This environment includes the room or area in which the recording is occurring and the position of the recording device relative to the speaker and its surroundings. Audio is particularly sensitive to noise, such as the background hum of air conditioning or the rustling of clothes. These aspects can be challenging to keep consistent, particularly across multiple recording sessions that can be recorded on separate days. This introduces inconsistencies across audio recordings that require attention before being used in a data-driven approach.

Audio should be normalised as much as possible between recording sessions to ensure the relative loudness is consistent. Humans often use loudness as an emotional response or reaction. It's important to isolate this change in the recording due to emotion rather than environmental changes.

The microphone type is a crucial recording consideration, each with strengths and weaknesses. A Lavalier microphone attaches to the speaker and has the benefit of limiting noise from the rest of the room and keeping the distance from the microphone, hopefully resulting in greater consistency of audio recording between sessions. A disadvantage is the person wearing it can accidentally brush over it, creating a significant noise spike. These are extra cables for a speaker to wear and could be restrictive of motion. A stand-alone microphone is a microphone that is not attached to the speaker and captures audio at a much lower sensitivity. While this method removes the possibility of a speaker touching the

microphone by accident, the greater distance from speaker to microphone means the amount of noise captured is significantly increased.

Depending on the data capture setup, there may be multiple speakers involved. Using Lavalier microphones for each speaker is beneficial in a dyadic, triadic or polyadic recording setting. This allows each speaker to capture their audio independently and produce audio files. Independent audio is helpful as it enables separating audio streams within a data-driven approach. An issue with introducing multiple speakers is the possibility of audio bleed across microphones. Therefore, although quieter, one or more speakers may be caught by another microphone, introducing speaker noise. This is often unavoidable and is something to consider at the data-driven stage.

Raw audio signals and many audio features do not capture the semantic meaning of the spoken word. For data-driven gestures, semantic meaning is a valuable feature; therefore, text transcripts can be desirable. Transcripts can be manually labelled or extracted from audio using Speech-To-Text (STT) systems. While the former should be more accurate, labelling this data can be time-consuming and expensive. STT systems have become highly precise in recent years and are becoming a viable option for text extraction.

3.2.2.1 Audio Embedding

When trying to map audio speech to gesture, the raw audio signal is not a particularly useful data representation. Instead it is common to extract particular audio features that encode features such as prosody, pitch and energy. The most suitable audio representation for speech-motion synthesis is an open research question. One of the most common audio speech representations chosen in previous work is Mel-frequency Cepstral Coefficients (MFCCs) [6, 48, 133]. MFCCs represent the spectral characteristics of sound that are important for human speech perception. These are well performing audio features for speech-to-gesture [6]. While these have provided impressive results, there is scope for more descriptive features.

Recently, learned audio encodings have become available [11, 120, 25], aiming to improve on manually extracted features such as MFCCs. In the majority of this work, PASE+ [120] will be utilised as the audio embedding of choice. PASE+ [120] uses a trained network to extract audio features. The network was trained using a multi-task encoder-decoder method where the same encoding is used as input to multiple decoders. Each decoder has a different task, in PASE+, there are 12 regression tasks, which include estimating MFCCs, FBANKs and other speech-related information, including prosody and speech content.

The extensive pre-training objectives of PASE+ ensures that these features encode particularly useful features for gesture such as prosodic elements, MFCCs and other speech-related features. The prosodic elements of speech have been highlighted in many studies to have a close relationship to gesture movement [36, 71, 35, 45, 58] suggesting that the objectives of PASE+ are of a particular benefit to gesture generation. Other methods such as Wav2Vec[11], WavLM [25], deepspeech [49] and Whisper [118] do not actively encode prosody, nor are they evaluated on prosodic tasks. As well as this, the embedding size is smaller for more efficient training when compared to competitive embeddings such as WavLM [25], with an embedding size 768 over 25ms of audio compared to PASE+ having size 256 over 10ms audio.

3.2.3 Style

Many datasets aimed at gesture generation include a level of style control. This usually occurs through different speaker styles or different emotional performances. The speaker is self-explanatory in that the conditioning should match the style of a specific speaker. Emotion can be either natural or acted. Genuine emotion will typically only be available from in-the-wild video sources, whereas emotion can be acted and exaggerated during controlled data capture. This can be a desirable attribute to data as it presents the opportunity to provide fine-grained control and style transfer across the same speech conditioning.

3.3 Motion Data Capture

Data collection is often challenging, particularly when multiple modalities and human participants are involved. This section explores the advantages and disadvantages of particular data collection techniques and their impact on data-driven approaches.

3.3.1 Video

Video is a widely used source of co-speech gesture data due to the abundance of widely available footage of speech-related videos. Motion is extracted from video using landmark prediction for a pre-defined skeleton representing a subset of joints in the human skeleton. As video is a two-dimensional recording medium, extracting 2-D landmarks from video is a much simpler problem than 3-D. This, in turn, leads to the majority of video-driven datasets being limited to 2D [43, 153, 3]. While the performance of predicting 3D positions from 2D data has increased in recent years, the extraction process can be challenging. Inferring the 3rd dimension from 2D can often lead to noise and inaccurate and inconsistent depth predictions. As discussed in Section 3.2.1, this is a distinct issue as 3D data is desirable for its ability to fit into standard 3D character animation pipelines.

There are a few critical issues regarding using video for motion data extraction, mainly when using in-the-wild video sources: occlusion, camera positioning and camera cuts. Each provides an issue for both 2D and 3D joint prediction, often having a more significant impact on the prediction of the 3rd dimension. This has the most significant effect on the hand, particularly finger motion, which is very fine-grained and often subtle.

Occlusion means that the body joint is not visible in the scene. This is a common problem, mainly as hands or fingers are crossed or a camera is positioned to the speaker's side. In a controlled environment, it is possible to control the camera position to limit this and ask the speakers to be aware of potentially occluding certain areas of the body; however, this

may restrict the motion produced by a speaker and appear unnatural. Occluded joints may be interpolated or given a defined skeleton constrained to conform to the skeleton given a particular bone length and topology knowledge. While this is a feasible approach, it often misses granular details such as hand and finger positions, which can be helpful in communication. Performance issue also scales with the amount of occlusion in a skeleton, particularly if the occlusion is stacked along the kinematic chain.

Cameras are rarely in the same position between videos or even within the same video. This can cause issues with scaling bone lengths and even missing joints altogether; for example, one camera may show the whole body, whereas the next may only show the upper body. Video cuts are often used when editing videos to ensure the content is shortened with only the required information present. When extracting co-speech gestures, it's critical to know when a speech sequence starts and ends, as without this knowledge, there is a risk that the scene may cut and there is a significant, unnatural jump within a sequence.

Where it is possible to cheaply extract a large volume of data from 'in-the-wild' sources, the output would require extensive cleaning. This is because recordings are not taken in a consistent and controlled environment, which can lead to audio issues such as those detailed in 3.2.2. These issues can be limited but likely not eradicated by ensuring a controlled recording environment. This will inherently restrict the volume of data available.

3.3.2 Motion Capture

The current state-of-the-art motion tracking is in the form of large motion capture (mocap) studios. These studios contain cameras at multiple angles, requiring the tracked actor to wear a mocap suit. This suit has retroreflective or LED markers positioned at key joint positions. Each camera is time-synced and tracks each marker to derive the positions of each joint on the suit, which are then fit a virtual skeleton with a defined topology. Motion capture is

typically tracked in a consistent and controlled environment and mitigates many issues from video-tracked data.

Motion capture is a highly effective method for gathering gesture data; however, it is also costly in terms of both financial and processing expenses. Tracking requires a minimum of two cameras; however, the accuracy of tracking scales with the number of cameras. Like the occlusion issue discussed in the previous section, motion capture suffers from the same fate. However, capturing multiple motion views means that others can accurately track an occluded marker in one camera. Literature utilising the OptiTrack motion tracking commonly uses 6, 10 or 12 cameras [100] to mitigate occlusion and ensure accurate tracking. This is particularly effective when comparing hand and finger tracking from motion capture and video capture.

3.3.3 Motion Capture and Video Hybrid

It is possible to accurately capture 3D motion from video feeds using a similar approach to traditional motion capture. This approach does not work for in-the-wild videos and requires a controlled environment with multiple camera views. Taylor et al. use humble, non-specialist hardware and a setup that is easy to replicate for future collaborative growth [133]. Section 3.4.1 discusses this approach in more detail.

3.4 Releases

There are many datasets available for co-speech gesture generation. Table 3.1 shows a complete list with details. This section will focus on the subset of datasets used in studies discussed in this thesis.

Name	Hours	# Speakers	# Session Speakers	Motion Format	Motion Source	Modalities	Fingers
IEMOCAP [18]	12	10	2	mp4 video	MoCap	Ges., Audio, Text, Emotion	
SaGA [89]	1	6	2	mp4 video	MoCap	Ges., Audio, Gest. properties	
Creative-IT [96]	2	16	2	3d joint rot.	MoCap	Ges., Audio, Text, Emotion	
MPI-EBEDB [139]	1.43	8	1	3d joint rot.		Ges., Text	
Gesture-Speech Dataset [132]	5	2	1	3d joint rot.		Ges., Audio	✓
CMU Panoptic [62]	5.5	50	1-8	3d joint rot.	Multi-Video	Ges., Audio, Text	
Trinity Speech-Gesture I [37]	6	1	1	3d joint rot.	MoCap	Ges., Audio	
Speech-Gesture [43]	144	10	1	2d joint pos.	Video	Ges., Audio	
TED Dataset [153]	106	1295	1	2d joint pos.	Video	Ges., Audio	
Talking With Hands [78]	50	50	2	3d joint rot.	MoCap	Ges., Audio	✓
PATS [3]	250	25	1	2d joint pos.	Video	Ges., Audio, Text	
Trinity Speech-Gesture I GENEA Extension [74]	6	1	1	3d joint rot.	MoCap	Ges., Audio, Text	
Trinity Speech-Gesture II [38]	4	1	1	3d joint rot.	MoCap	Ges., Audio, Gest. segment	
Speech-Gesture 3D extension [48]	144	10	1	3d joint pos.	Video	Ges., Audio	
UEA-DH [133]	3.5	1	2	3d joint rot.	Multi-Video	Ges., Audio	
Talking With Hands GENEA22 Extension [154]	20	17	2	3d joint rot.	MoCap	Ges., Audio, Text	✓
SaGA++ [75]	4	25	2	3d joint rot.	MoCap	Ges., Audio, Gest. properties	
ZEGGS Dataset [42]	2	1	1	3d joint rot.	MoCap	Ges., Audio, Emotion	✓
BEAT Dataset [84]	26	30	1-2	3d joint rot.	MoCap	Ges., Audio, Text, Emotion, Gest. properties	✓
Talking With Hands GENEA23 Extension [76]	20	17	2	3d joint rot.	MoCap	Ges., Audio, Text, Interlocutor Alignment	✓
Audio2Photoreal [103]	8	4	2	3d joint rot.	Multi-Video	Ges., Audio, facial codes.	✓

Table 3.1 A comprehensive list of available co-speech gesture datasets. Table adapted from Nyatsanga et al. [106].

3.4.1 UEA Digital Humans

A male speaker (Speaker A) was filmed conversing with a female speaker (Speaker B) who was off-camera. Speaker A was filmed before a green backdrop from three synchronised views (see Figure 3.2). The video was recorded at 25fps and 1080p resolution with 48kHz audio. The dataset contains ≈3.5 hours of dialogue and comprises three parts: Part 1 (1 hour) includes an unscripted conversation between the speakers. Part 2 (1 hour) is a debate on a topic chosen from Speaker A's list. Speaker B argued the opposing view to Speaker A to incite a heated discussion. Part 3 (1.5 hours) is a performance of scripted emotional monologue vignettes, which were included to provoke a broader range of affective states.

2D key points are located in each camera view using the monocular body pose detection system OpenPose [20]. Cameras are calibrated using a checkerboard target, and the 2D key points are projected into 3D world space through triangulation. Occlusion may still occur; the key point is omitted in this case. Due to the speaker's sitting position on a stool, the lower body is not reliably tracked and discarded. As this dataset aims at dyadic gesture generation, the head is considered a rigid object, and consequently, any joints above the neck are reduced

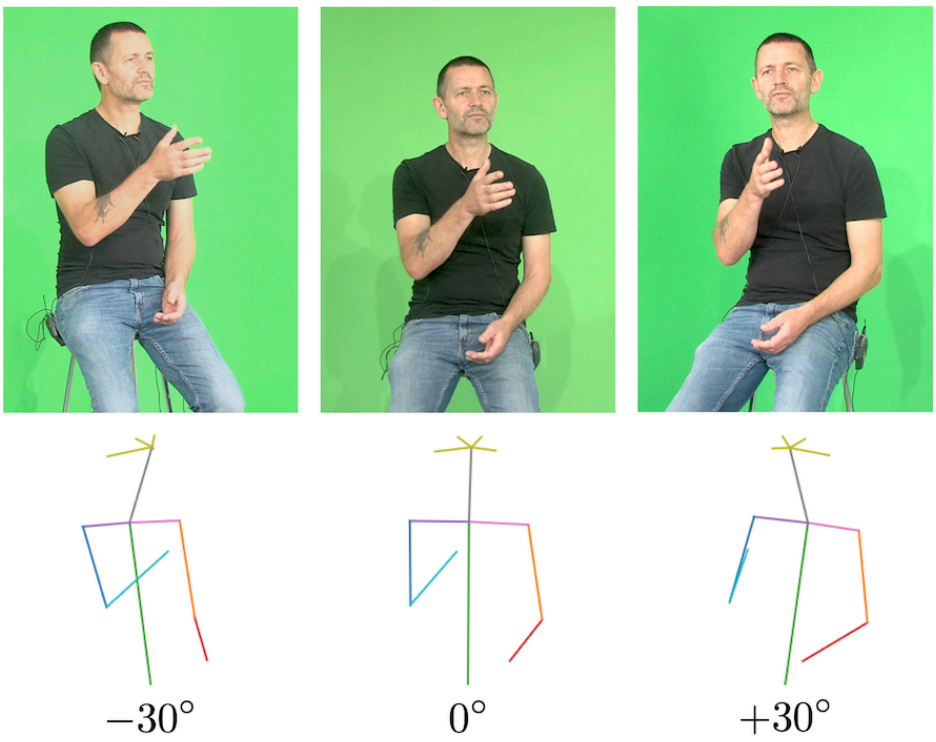


Fig. 3.2 A frame from each camera view (top), and the corresponding pose at 0 and ±30 degrees from frontal pose (bottom) shown on a reference skeleton.

to a single rotation. The body gesture is therefore represented using nine remaining joints as shown in Figure 3.2. Poses are translated so that the base of the spine rests at the world origin, $(0,0,0)$.

While the speaker's position in relation to the camera is kept reasonably consistent between recordings, as these were recorded over multiple sessions, the speaker's global orientation may vary over time. The pose is frontalised using procrustes alignment to compute a rotation, $\mathbf{R}$, minimising the distance between a clip's hip and shoulder landmarks to the corresponding landmarks of a frontal reference skeleton. $\mathbf{R}$ is computed and applied once per clip, and a clip is defined as a natural break in the capture or a 13-minute segment, whichever is shortest.

3.4.2 Talking With Hands

One of the largest and highest quality motion datasets is the Talking With Hands 16.2M from Lee et al. [78]. This motion-captured dataset contains high-quality audio and motion from 50 speakers, each recorded in a dyadic conversational setting.

Each conversation follows a free-talking or video-retelling construct. Free-talking sessions started with a provided, casual topic to spike conversation; however, participants were encouraged to drift from this topic to encourage natural conversation flow. The video-retelling construct consisted of Speaker A watching a video, for which Speaker A must then tell Speaker B what happened in the video. After this, Speaker B must retell the story to Speaker A. This interaction encourages a natural, active speaking and listening paradigm, with interruptions and questions.

Participants wore mocap suits fitted with retroreflective markers. The motion was captured using 24 OptiTrack cameras sampled at 90 frames per second, and the motion capture data was converted to joint angle representations in Biovision hierarchical data (BVH) format. The audio was captured using directional microphone headsets on each speaker sampled at 48kHz.

The dataset comprises 16.2 million motion frames for 50 speakers; however, much has not been publicly released due to anonymity concerns. Instead, a subset of 36 speakers has been released, still making it a substantial amount of data. The anonymity concerns also mean the audio data has been heavily redacted to remove any personally identifiable information. Of the 36 available speakers, only 18 have audio aligned with the motion. Each audio file has been manually edited to mute any personally identifiable information, and therefore, many of the sequences contain missing information and muted elements.

Despite this missing data and muted audio, the GENEA workshop, described in Chapter 2.4, recognised the value of this large corpus of high-quality data and was used in the GENEA challenges. For the 2022 challenge, Yoon et al. [154] released an improved subset of this,

cleaning some of the motion and adding speech text transcripts. The 2023 Challenge enhanced this data further with more data cleaning and the alignment of interlocutor information. In the original Talking With Hands data release, although the motion and audio from both speakers were released, the motion and audio between both may not have been synchronised. The GENEA Challenge 2023 aligns both speaker's motion and audio to allow for the dyad information to influence gesture prediction.

3.4.3 ZeroEGGS

Gesturing style can be determined by an actor's age and affective state [42]. Controlling this style can be a useful tool for speech-to-gesture systems. The ZeroEGGs dataset [42] contains two hours of female-voiced monologue acting in 19 different styles.

Both body motion and high-quality audio were captured, allowing for diverse predictive styles and high-quality motion. Styles performed in the data are *Agreement, Angry, Disagreement, Distracted, Flirty, Happy, Laughing, Oration, Old, Neutral, Pensive, Relaxed, Sad, Sarcastic, Scared, Sneaky, Still, Threatening* and *Tired*. The diverse range of styles aims to cover a variety of postures. For example, *Oration* motion consists of shoulders held in high posture whereas *Old* is a very hunched and low posture. Styles also encourage a variety of head motions; for example, *Agreemment* and *Disagreement* consist of nodding and shaking of the head, respectively. The affective state is also covered in the motion with styles such as *Happy* and *Angry* containing exaggerated movements to reflect the speaker's emotive state.

3.5 Augmentation

Data augmentation is a technique used to increase the amount of data by adding slightly modified copies of real data or creating synthetic data from existing data. The most common technique for this is through *data warping* defined by Perez et al. as an approach to directly

augment the input data to the model in *data space* [113]. Augmentation approaches vary depending on the data type and the problem domain.

When working with image data, simple transformations are commonly applied to each image. These include flipping, scaling, rotating, translating, noise injection and colour space transformation [128]. While flipping, scaling, rotating and translating can apply to a 3D skeleton representation of body motion data, it is not necessarily appropriate. Scaling the skeleton by a different amount in each dimension would alter the identity. If we scale by the same amount and joint angles represent the skeleton pose, this scaling would not provide additional information as the angles would remain identical. Applying a global rotation to the skeleton might introduce unnatural positioning (e.g. losing foot contact with the ground). Translating the skeleton would not effectively augment the data as the speaker would still move in the same way but in a different location. Adding noise to the captured motion would cause unnatural, jittery motion. Flipping (or laterally mirroring) the skeleton is the only data augmentation approach that produces potentially valid human body motion.

3.6 Discussion

This chapter introduces the modalities involved in automatic co-speech gesture generation and provides details on how these multiple modalities are gathered to form the released datasets. It also describes methods of data augmentation to increase the amount of data available however, the next chapter will extend the data augmentation discussion with a particular focus on the lateral mirroring augmentation technique.

This thesis involves a particular focus on the Talking With Hands dataset described in Section 3.4.2, featured in the majority of proceeding chapters. The ZeroEGGS dataset described in Section 3.4.3 is also used in Chapter 6. These are both high-quality and large datasets and, therefore, lend themselves to this work. In each case, the modalities of speech

and style are fundamental to each experiment. The speaker's identity or gesture styles are utilised in each experiment to add an extra level of gesture generation conditioning.

Chapter 4

Speech Gesture Symmetry

Contributing Publications

- Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022b). Arm motion symmetry in conversation. *Speech Communication*, 144:75–88
- Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022c). Pose augmentation: mirror the right way. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA. Association for Computing Machinery

4.1 Introduction

Domain knowledge is required in any machine learning problem, particularly for data processing and performance evaluation. This chapter introduces the topic of gesture and its relationship to speech and explores speech gesture data. This analysis pays particular attention to gesture symmetry and how the common augmentation technique of mirroring a pose may impact communicative understanding and naturalness. This analysis also reveals practical objective measures regarding motion characteristics including introducing two

statistically derived gesture spaces that could be utilised in future work to analyse the performance of data-driven speech-to-gesture generation.

4.2 Arm Gesture and Symmetry

Relying on speech or gesture alone does not allow a speaker to communicate most optimally. Removing either of these modalities leads to a reduction in semiotic versatility [140] and communicative understanding [57]. One reason is that each modality may represent certain information better than the other. For example, hands might better describe shape or direction using visual cues. The gestures that form these cues may or may not be symmetrical, which may, in part, depend on the particular shape or direction being described.

Environmental conditions greatly contribute to the importance of each modality during a conversation. A small and enclosed space may cause a person to be conservative with their gesturing, whereas to communicate the same speech in an expansive, outside environment, they may gesture more actively as they have more space. Proximity and facing direction of the conversational partner within the environment will also affect the extent and gesturing. If conversation occurs while walking alongside their partner, this will prompt different behaviour to a static face-to-face interaction. Similarly, if the partner is far away, gestures may be emphasised to account for the reduction in the received audio volume. It has been found that gesture activity increases during adverse listening conditions, such as acoustic noise and non-native speaking conversational partners [33].

Objects surrounding or colliding with the speaker introduce physical constraints that inhibit or otherwise affect gesturing. For instance, a wall to one side of the speaker will limit their available gesture space, constrain physical activity and likely increase asymmetry. Similarly, a speaker's hand might be occupied with an object, such as a glass of water, which would alter gestural behaviour.

Individuals exhibit gestural idiosyncrasies. Some speakers may commonly perform self-adaptor traits such as self-touching or scratching. Others may have physiological restrictions, making particular gestures impossible and affecting the realisation of others. In each of these cases, asymmetry in the positioning of the arms is likely.

The amount of conversational gesturing during an interaction can be linked to a speaker's personality. It has been found that a speaker's *Big Five* personality traits (extroversion, neuroticism, conscientiousness, agreeableness and openness to experience) are correlated with the amount of gesture production [57]. In particular, extroversion is positively correlated with representational gesture production, possibly due to extroverted people having high energy in social situations and, therefore, gesturing regardless of communicative effect.

McNeill defined a gesture space [94], stating that most gestures happen in the *central gesture space*, which encompasses the area below the neck and between the shoulders and elbows. *Peripheral gesture space* encapsulates gestures performed outside of the central gesture space and can be thought of as the extremes of gesturing. They suggest that the peripheral gestures aim to capture visual attention.

McNeill also defined a classification of the semantic functions of gesture types [94]. As described in Section 2.2 they categorised gestures as either emblematic, iconic metaphoric, deictic or beat:

- ***Emblematic gestures***: bear a conventionalised meaning
- ***Iconic gestures***: resemble a particular physical aspect of the conveyed information
- ***Metaphoric gesture***: is an Iconic gesture resembling abstract content
- ***Deictic gesture***: point out locations in space
- ***Beat gestures***: is simple and fast movements of the hands commonly synchronised with prosodic events in speech [114]

However, in practice, a gesture may perform many semantic functions. Instead, it has been proposed to treat each gesture category as a dimension on which gestures load to differing degrees [93].

A speaker's handedness has been found to impact gesture production, particularly regarding the positioning of the left and right arms. It has been found that beat-style gestures were more commonly performed with a speaker's dominant hand. In contrast, representational gestures in right-handed speakers had a right-handed preference, while left-handed speakers did not have a hand preference [24]. There is an association between gestural handedness and the emotional dimensions of pleasure and arousal. Kipp and Martin [67] found a significant correlation between emotion category and handedness of the gesture, where speakers consistently used their left hands to gesture during a relaxed, positive mood and their right hands to gesture when in a negative, aggressive mood.

These works each analyse gestural symmetry during conversation. However, these works are limited by the data used. Data is often observed manually from video [94] or limited to a few speaker's worth of data [67]. This reveals a limitation in current studies that is addressed in this chapter.

4.3 Data and Pre-processing

This study performs an analysis of the body motion from the Talking with Hands dataset [78] described in Section 3.4.2. This dataset consists of 50 different speakers during dyadic conversation, capturing 90fps motion capture and associated audio. Unfortunately, not all of this data is publicly available; therefore, the available subset of 36 speakers has been used. The majority of speakers were only captured in conversation with one other speaker (*shallow* speakers), while a small number had multiple conversational partners (*deep* speakers). Any non-conversation data segments (e.g. T-Pose sequences) are removed before analysis.

The dataset provides a set of 3D skeleton joint key points for each frame. This study focuses on arm movements and considers only the 3D locations of the left and right shoulder, elbow, forearm and wrist. The skeleton was translated per frame such that the mid-point between each shoulder joint was at the origin. This simplifies the analysis and accounts for large translations of arm positions from motion originating from the spine, such as leaning forward and backwards. This allows the evaluation of translations made by motion generated from the arms independently of the rest of the pose. The coordinate system utilised in this chapter is as follows:

- Y - Height (Up and Down)
- X - Depth (Back and Forth)
- Z - Width (Left and Right)

A consistent colour scheme is used throughout all figures to represent each forearm. Red depicts the right forearm and Blue depicts the left forearm.

4.4 Mean Pose Symmetry

The symmetry of the mean poses for each speaker is first evaluated, aiming to reveal an impression of the per-speaker symmetry across all of their motion. Using all the frames of motion, the per-speaker mean pose is calculated. The right arm is then projected to the space of the left arm by laterally mirroring (along the y-axis). Symmetry can be quantified using the Euclidean distance between all joints in the left arm and the projected right arm. The lower this distance, the closer the two arms are to each other, which indicates a more symmetrical pose where a distance of 0 would represent complete symmetry.

The symmetry range is shown in Figure 4.1. This presents that a person's mean pose is not always symmetrical. Shallow3 is found to have the most symmetrical mean pose, whereas Deep3 has the most asymmetric pose according to the Euclidean distance.

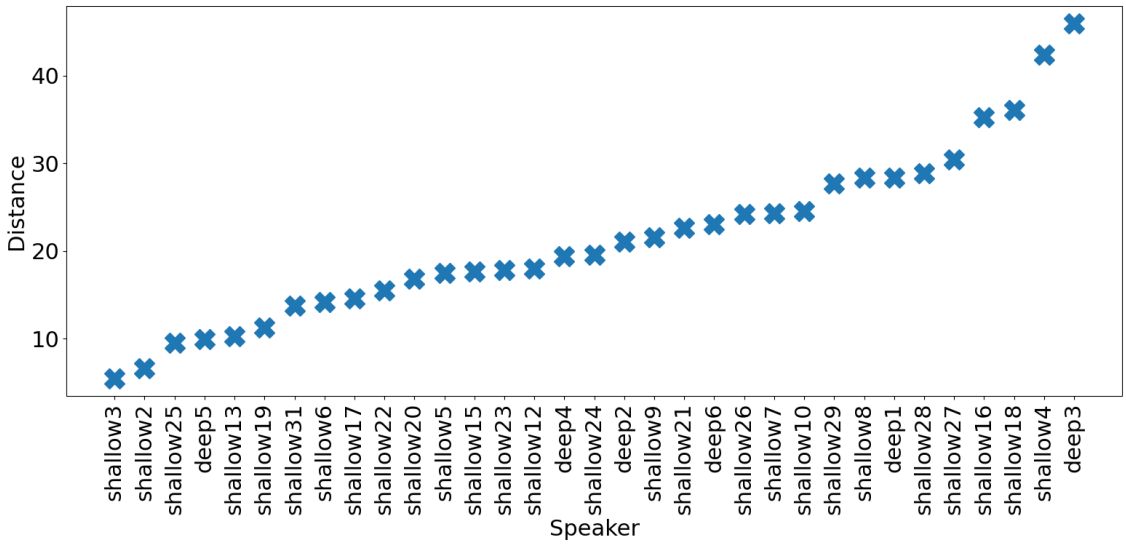


Fig. 4.1 Symmetry distance between each speaker's mirrored right arm and the left.

From the 36 speakers, the two with the highest and the two with the lowest Euclidean distance are chosen for further examination, representing the subjects exhibiting the least and most arm symmetry in their mean pose. It is possible to visualise the level of symmetry by overlaying a perspective projection of the mirrored right arm onto the left arm. Figure 4.2 shows this projection from both a frontal and side view for each of the four speakers. There is an apparent asymmetry in the mean arm pose of Deep3 and Shallow4 (columns one and two). The left arm of Deep3 shows itself angled towards the right side of the body, whereas the right arm points away from the body, towards the camera. Shallow4 orients their right wrist away from their body while their left wrist points towards them. At the other extreme, Shallow3 and Shallow2 are symmetrical (columns three and four). In these examples, the mirrored right arm overlaps the left arm from the shoulder to the elbow with a slight divergence from the elbow to the wrist.

The most considerable differences between the arm positions are observed in the side view, whereby each left arm is positioned further forward than the right. While this observation is

more prominent on the two most asymmetric speakers, it holds for each speaker in Figure 4.2.

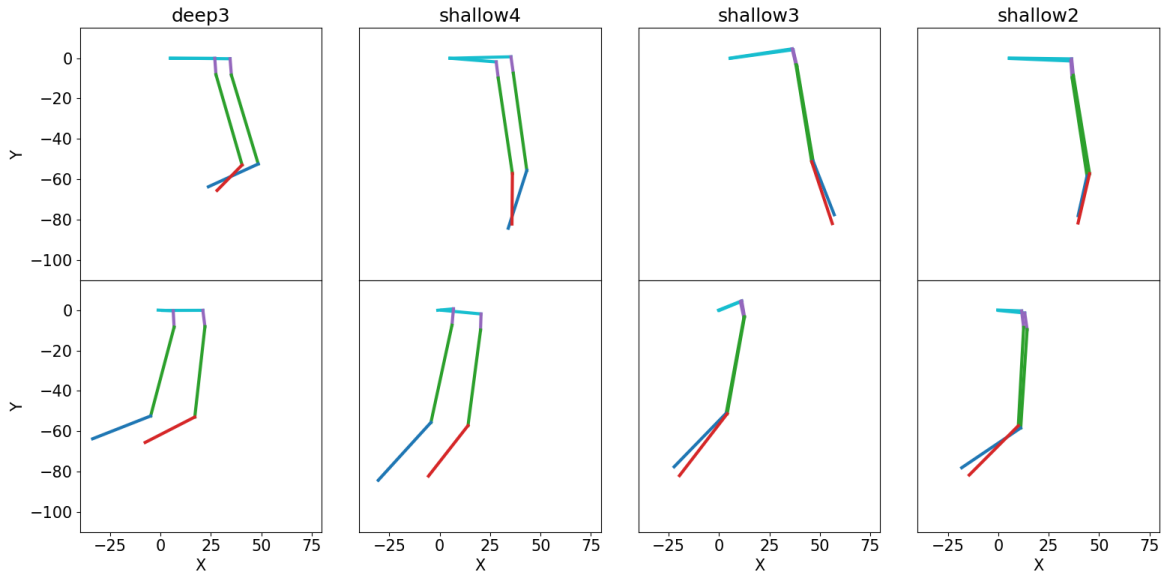


Fig. 4.2 A projection of the mean pose for four speakers. In each case, the right arm (red forearm) has been mirrored and overlaid onto the left arm (blue forearm). Top row: front view. Bottom row: side view.

4.5 Spatial Symmetry

The mean pose analysis in Section 4.4 indicates the symmetry of a speaker's average (or neutral) arm positions. However, it does not explain whether the *motion* of the arms is similar or symmetrical. This section investigates whether the observed asymmetry affects a speaker's tendency to gesture more on one side than the other and whether the arms occupy symmetrical gesture spaces. 3D key points are used to gather statistics regarding the arm motion of each speaker, describe the speakers' motion ranges and traits, and define their data-driven gesture spaces.

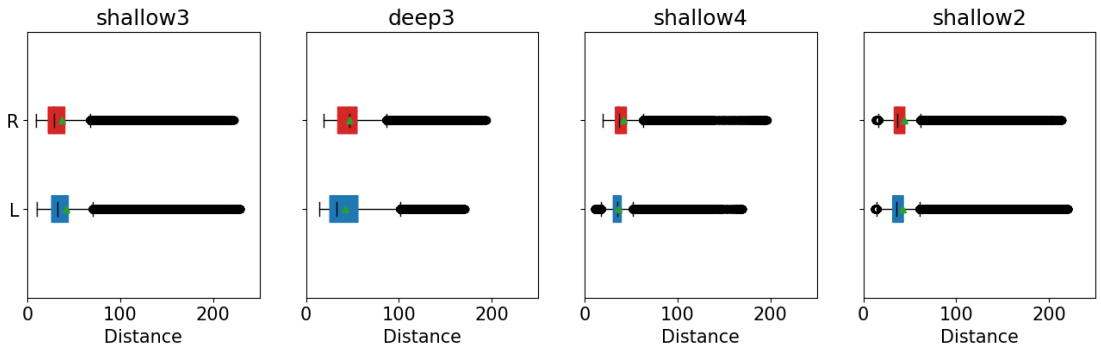


Fig. 4.3 Per-frame Euclidean distance from the mean of each arm for four speakers. L=Left arm, R=Right arm.

4.5.1 Full Arm Motion Range

To reveal whether a similar amount of movement is performed by the left and right arms, the deviation from the mean pose is measured. Frame-wise Euclidean distances are computed independently from each arm to its respective mean pose. These statistics are calculated over all arm joints.

Figure 4.3 shows box plots indicating the distribution of the amount of deviation from the mean pose of each arm. The results are for the four speakers identified as exhibiting the least and most symmetry in their mean pose in Section 4.4. The deviation from the mean pose in the left and right arms is not considerably different when assessing the poses within the whiskers, representing those within 1.5 $\times$ the interquartile range beyond the first and third quartiles. However, the outliers appear somewhat asymmetrical for speakers Deep3 and Shallow4, each displaying more significant divergence from the mean with the right arm compared to the left. Shallow3 and Shallow2 exhibit more symmetrical outliers, indicating that both arms encompass a similar amount of space during these infrequent, more significant gestures. The maximum and minimum values for each speaker follow the same trend, with larger maximum values recorded for the right arm in the former two speakers and similar values for both arms in the latter two.

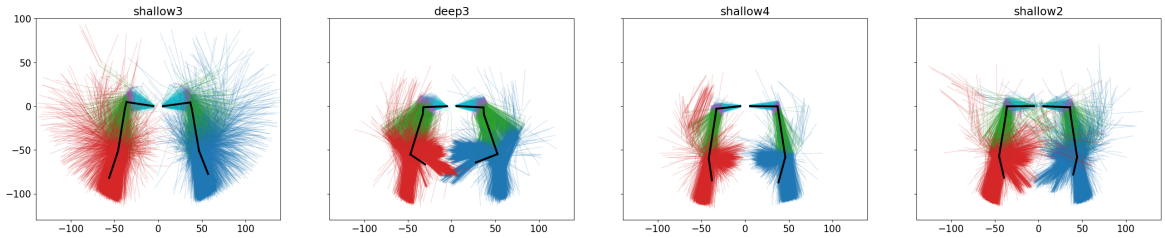


Fig. 4.4 A frontal perspective projection of all poses per speaker, taken at one-second intervals with the mean pose overlaid in black.

Figure 4.4 shows a frontal perspective projection of each speaker's arm pose taken over their respective conversations at 1-second intervals. Variability is observed in the gestural symmetry and the amount of gesturing per speaker. Shallow3 appears the most symmetrical, with a wide range of positions produced by both arms. Despite having a highly symmetrical mean pose, Shallow2 exhibits a high degree of asymmetry in the peripheral poses, whereby the right arm reaches more expansive poses than the left. Still, the left arm produces higher gestures than the right. Deep3 and Shallow4 raise their right hands more frequently than their left, suggesting increased expressiveness in that dominant hand. These plots show that asymmetry is most apparent in the peripheral gesture space where the extreme gestures are performed. Although relatively infrequent, these extreme gestures capture visual attention and are perceptually significant [94].

4.5.2 Gesture Spaces

McNeill defines the *central gesture space* as the area below the neck and between the shoulders and elbows, and the *peripheral gesture space* as any gestures performed outside of the *central gesture space* [94]. Given the variability between the spaces occupied by each speaker's arm and the frequency in which they extend into their respective peripheral spaces, a data-driven approach to defining speaker-specific gesture spaces is defined. Statistics are used to define a speaker's *common gesture space* and *extreme gesture space*. The *common*

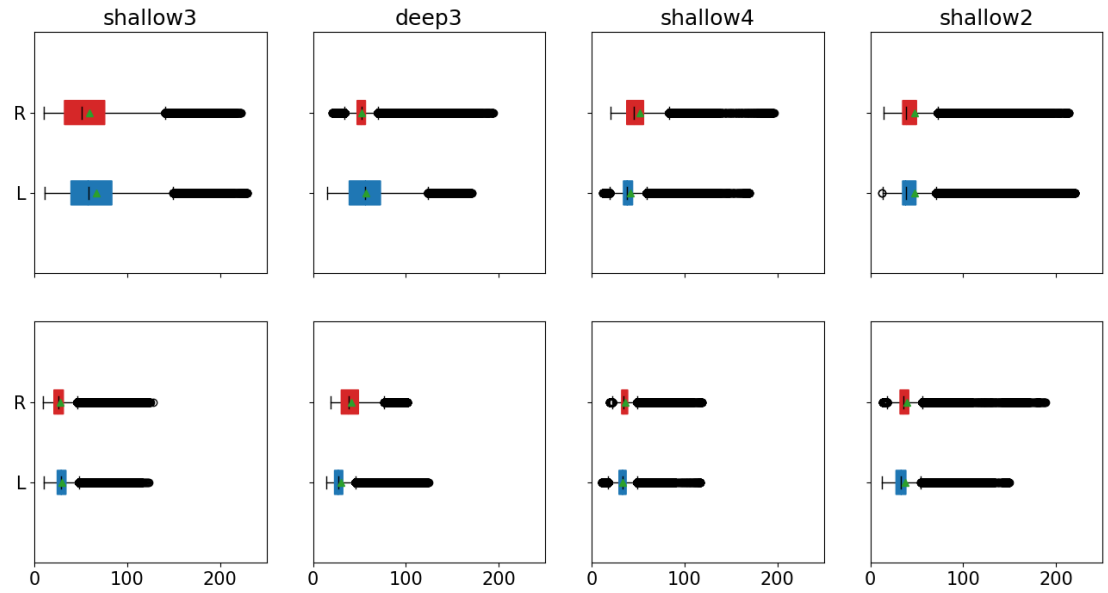


Fig. 4.5 Per-frame Euclidean distance from the mean of each arm, split into *Extreme Gesture Space* (Top) and *Common Gesture Space* (Bottom). L=Left Arm. R=Right Arm.

gesture space is the region within a single standard deviation of the respective speaker's mean arm pose. The *extreme gesture space* is the space outside a single standard deviation of the mean pose, away from the body.

Using this new definition, data is partitioned into two sections. The *extreme* partition contains all poses with at least one arm in the *extreme gesture space*, and the *common* partition contains the remaining data. Per-speaker distance from the mean pose is again computed for each partition, and the results can be seen in Figure 4.5. For most speakers, the distances from the mean for gestures within the common gesture space are similar for both left and right arms (Figure 4.5, bottom row). An exception is the speaker Deep3, whose range is more extensive for the right hand. The most significant differences between the left and right arms are observed in the extreme gesture space (Figure 4.5, top row), particularly for the asymmetric speakers Deep3 and Shallow4. In each case, one hand diverges further from the mean than the other.

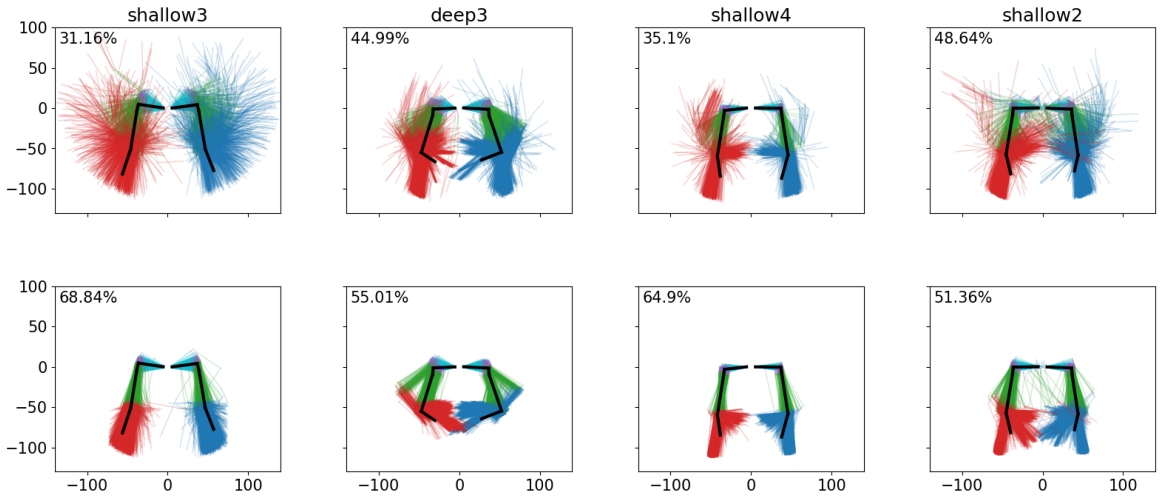


Fig. 4.6 Frontal projections of all poses from four speakers at one-second intervals, split into *Extreme Gesture Space* (Top) and *Common Gesture Space* (Bottom) with the mean pose overlaid in black. The percentage in the corner denotes the percentage of poses belonging to the respective gesture space for the respective speaker.

For Deep3, the left arm is more active in the extreme gesture space than the right, and the reverse is true in the common gesture space. The perspective projection of all poses is plotted corresponding to the extreme and common gesture spaces in Figure 4.6 for each speaker to visualise these differences. The top row reveals that the right arm of Deep3 contributes to gesturing in the extreme gesture space, but the poses of the left arm are more expansive, taller, and further from the mean pose. In contrast, the bottom row shows more movement in the right arm than the left in the common gesture space, but not significantly.

Figure 4.6 highlights that the positioning of the arms in common gesture space appears to be more symmetrical than in extreme space across all speakers. Each speaker exhibits different types of asymmetry in the extreme gesture space. Shallow4 lowers its left arm and raises the right, and shallow2 extends its right arm wider than the left. Shallow3 has highly mobile arms but holds symmetry in both spaces reasonably well, consistent with the findings in Section 4.5.1. The percentage of poses within each gesture space as shown in Figure 4.6 impacts the effect of mirroring. Given more symmetry in the common gesture space, if a

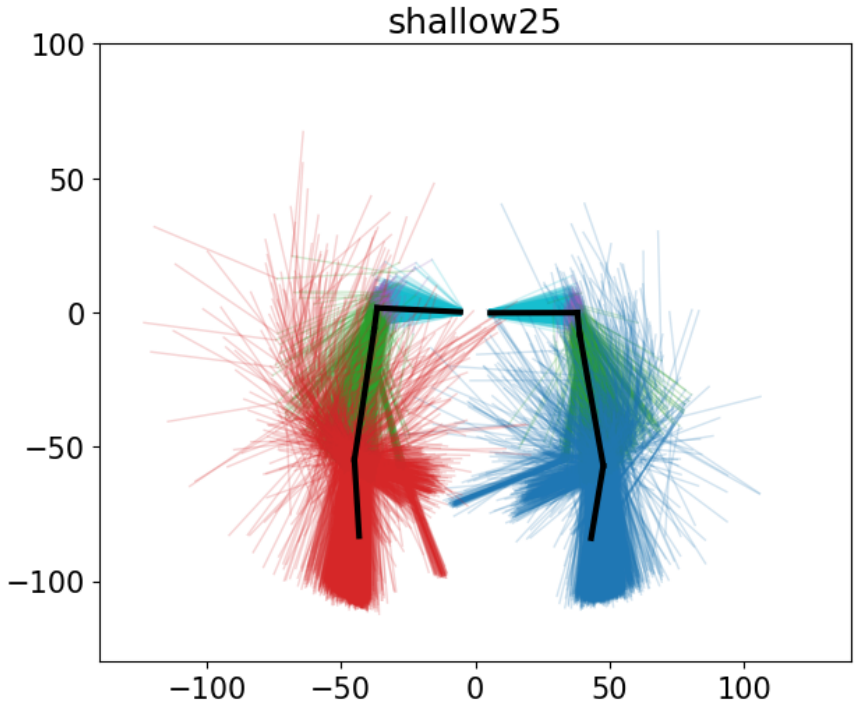


Fig. 4.7 Shallow25 poses taken at one-second intervals with the mean pose overlaid in black. This speaker exhibits self-adaptor movements whereby the left hand frequently touches the right forearm.

speaker uses the extreme gesture space less, the potential negative impact of mirroring is reduced.

4.5.3 Self-adaptor Traits

Self-adaptors are movements that co-occur with speech gesturing and typically include self-touch, such as scratching the neck, clasping at an elbow, adjusting hair or interlocking fingers. These traits tend to be realised asymmetrically.

Figure 4.7 shows the poses of speaker Shallow25, who frequently touches their left hand to their right forearm. The reverse, the right hand touching the left forearm, is absent in any motion. If laterally mirrored, this self-adaptor movement would not accurately represent a valid pose from that speaker. The presence and degree of self-adaptor traits have been found

to significantly impact the perceived level of neuroticism of a speaker [102], and the effect of reversing the handedness of the behaviour is not well established.

4.6 Symmetry in Gesture Types

The type of gesture being performed may be necessary when considering the impact of symmetry. By reviewing several speech-motion pairs, it's possible to determine what impact may occur from the gesture being mirrored. While it is not possible to generalise from these few examples, it should be helpful to consider specific aspects of gestures suitable when mirrored.

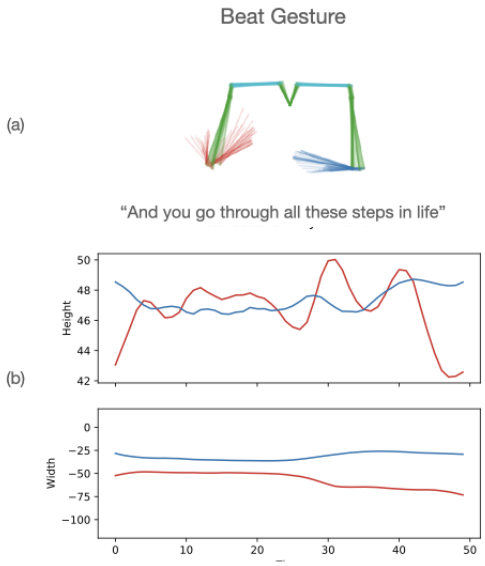


Fig. 4.8 A speaker performing a beat gesture. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left-right) and height (up-down) directions.

The study supports that beat gestures are often performed by a single hand. Figure 4.8 shows a pose plot of a beat gesture and the values of each wrist position over time. While the pose plot appears fairly symmetrical with both arms raised, it is clear that the right arm

is moving up and down while the left stays fairly static. While knowledge of the dominant hand of this speaker is unknown, some trends similar to those of Çatak et al. [24] where one hand is performing the gesture are observed.

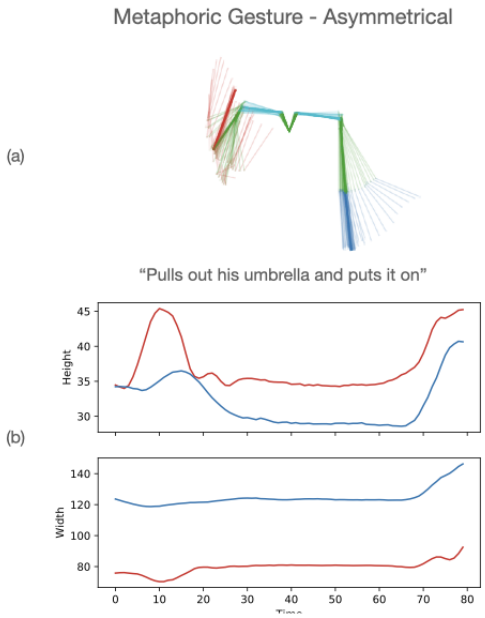


Fig. 4.9 A speaker performing a metaphoric gesture. In this case, the gesture is **asymmetric** due to context. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left-right) and height (up-down) directions.

Çatak et al. [24] suggest that representational gestures are performed by a dominant hand for right-handed speakers, but no dominant hand was found in left-handed speakers. Handedness cannot be compared in this work, but instead consider that the context of the gesture can determine the symmetry of the gesture performed. Figure 4.9 shows a metaphoric gesture mimicking using an umbrella. It is typical for a person only to use a single hand while using an umbrella; therefore, a single hand is used to depict this. Should this pose be mirrored, it may still make logical sense as a single hand will be used, but the handedness of the speaker may not be maintained. Figure 4.10 is a gesture performed by another speaker. However, they are referring to moving a heavy object onto a table. Typically, moving heavy

objects as outlined in the speech would require two hands; therefore, two hands have been used to depict this. In this instance, there are high degrees of symmetry between each arm movement, both arms moving and seemingly at the same or similar time.

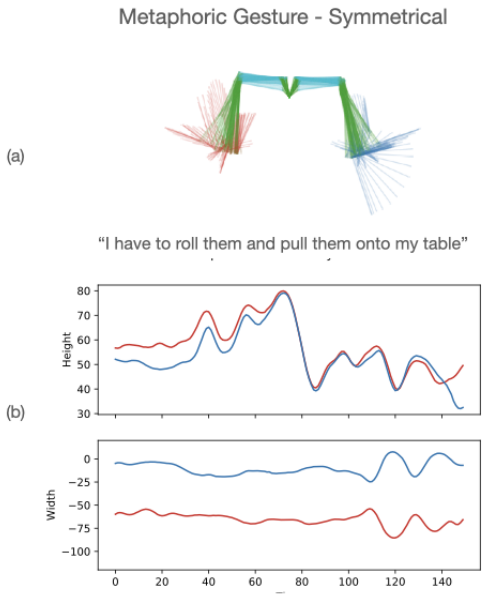


Fig. 4.10 A speaker performing a metaphoric gesture. In this case, the gesture is **symmetric** due to context. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left-right) and height (up-down) directions.

Regarding directional Deictic gestures, the hand closest to the direction was often used. The time plot in Figure 4.11 shows a gesture referring to each end of a building. "That end of the building" is referred to using the right arm, pointing towards the same direction to depict an area far away. "this end of the building" is seemingly where they stood, and a slight movement of the left arm refers to this. Figure 4.11 time plot shows a clear spike as the right arm moves to the peak directional gesture; the left arm lowers, suggesting asymmetry.

Some examples of symmetrical and asymmetrical poses and their associated gesture type have been described. Sometimes, a mirrored, symmetrical pose may still portray the same meaning. An excellent example is when an inconic action requires both hands to lift

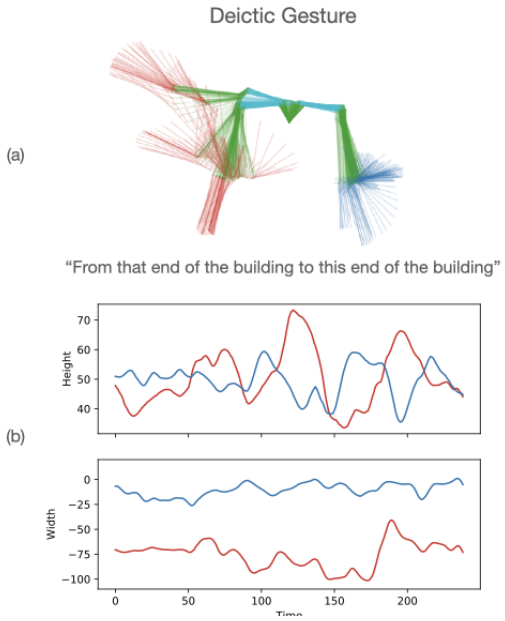


Fig. 4.11 A speaker performing a Deictic gesture. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left-right) and height (up-down) directions.

something. However, in the example of the Deictic gesture, this would not continue to make sense when performed in the exact location.

4.7 Mirrored Pose Validity

For some machine learning approaches, laterally mirroring body pose aims to generate further valid examples of the same speaker. In these cases, validity only holds if the mirrored poses fall within the gesture space of the original data belonging to that speaker. This section visualises and quantifies mirrored pose validity using this definition.

A nearest neighbour search was performed for each mirrored pose in the original motion data per speaker. The distance metric used is the Euclidean distance, which is computed over the joint locations in both arms. Poses within the extreme gesture space are the focus, defined as any pose outside one standard deviation away from the mean pose (Section 4.5.2). Figure

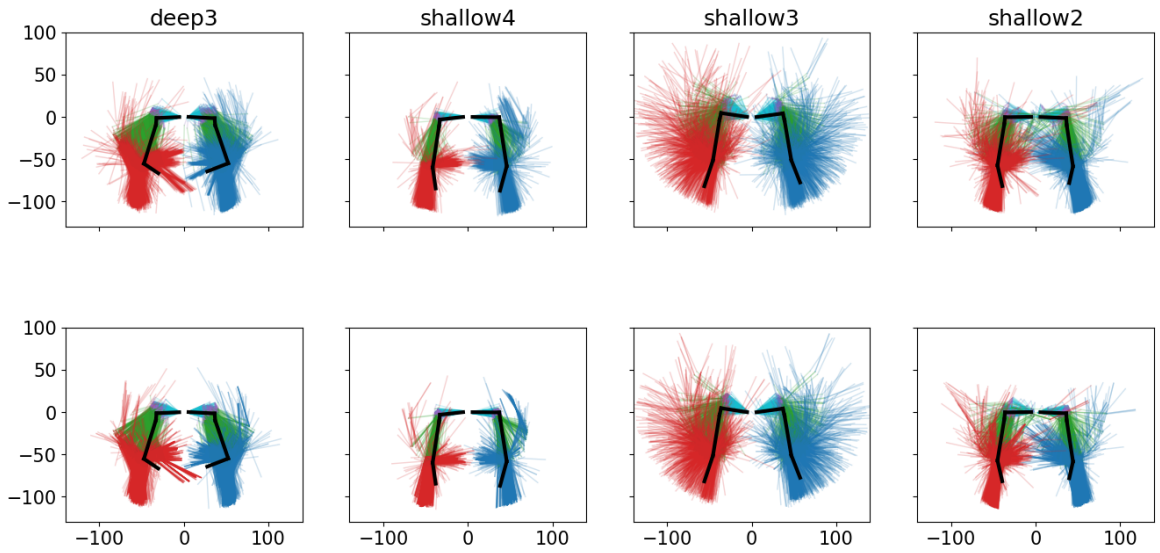


Fig. 4.12 The frontal 2D projections with mean pose overlaid of the mirrored poses that are at least one standard deviation away from their mean pose (top) and the closest respective mean poses from the original data (bottom).

4.12 presents a visualisation of the nearest neighbours. In this plot, the top row shows a subset of the mirrored poses for each speaker, and the bottom row shows the nearest neighbours from the original motion data. It is evident from this figure that it is not possible to cover the full range of motion found in the mirrored poses in the original data. For each speaker, there are areas in world space for which the arm does not reach in the original data.

In the rightmost column of Figure 4.12, it shows that, with speaker Shallow2, for the left arm to reach out as wide as it does in the mirrored poses, in the original data, the right arm also has to extend. This suggests that in the original data, it is characteristic for either both arms to move to a wide position together or for the right arm to move out wide independently. It is uncharacteristic for the left arm to reach out alone from the right arm. For Deep3 and Shallow4 (leftmost columns), when the mirrored poses are at their most extreme poses (i.e., the arms elevated to their highest and widest positions), it is impossible to match these in the original data.

Figure 4.13 shows mean distances between the mirrored poses and the closest match in the original data. Although Deep3 was associated with the least symmetrical mean pose from the dataset (Section 4.5), it shows that, in the extreme gesture space, they produce similar gestures with both left and right hands.

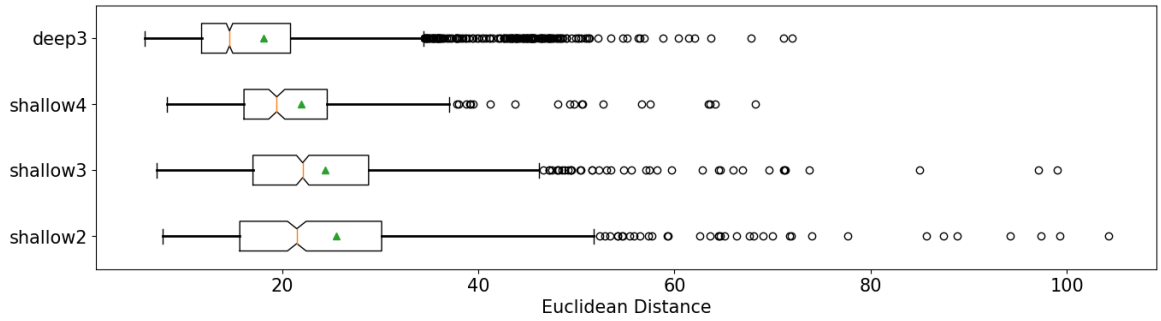


Fig. 4.13 Euclidean distance between mirrored arm position and the closest pose from the original data for poses in the extreme gesture space.

4.8 Temporal Symmetry

Analysis so far has considered only frame-wise statistics, which do not account for differences in the dynamics of each arm. Lateral mirroring for body data augmentation swaps the positions of the arms on a frame-by-frame basis, so the dynamics of the respective arms are inherently swapped. In practice, an asynchrony, or a temporal shift, may exist between the motion of the two arms, particularly if the speaker gestures with a dominant hand. In this section, a cross-correlation analysis is performed to reveal any temporal lag between left and right hands.

Correlation between the left and right-hand positions is computed over a 401-frame window ($\approx$ 4.5s), centred at frame t. For each windowed frame in the left-hand data, $t = 0, \ldots, T$, a window slides over the right arm data from frames $t - 200$ to $t + 200$ and computes the correlation coefficient between the segments. A larger window size was not

used since a lag longer than 2.2 seconds was more commonly due to a rhythmic motion than an asynchrony caused by a leading hand. The cross-correlation analysis is performed for each motion sequence on a per-speaker basis. The analysis is run independently on each directional axis, and the Euclidean distance to the mean pose of each hand, and the results can be seen in Figure 4.14.

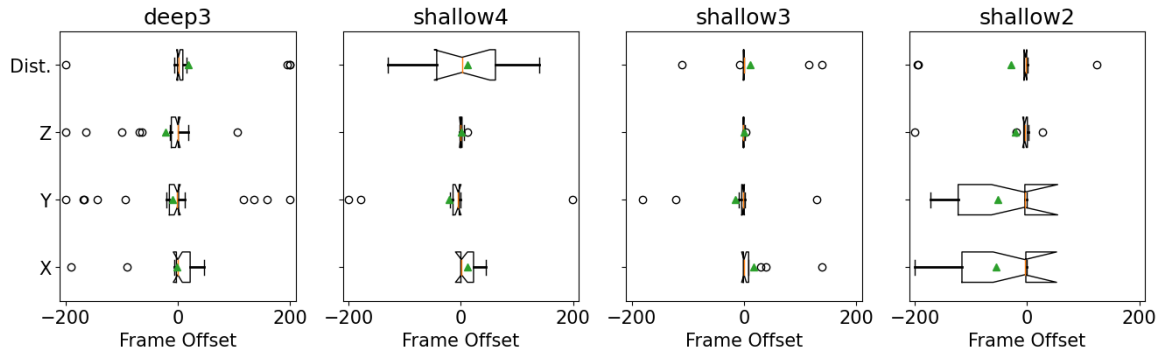


Fig. 4.14 Cross correlation analysis between left and right-hand position for each directional axis and Euclidean distance from the mean. Dist. denotes the overall distance from the mean pose, and X,Y, and Z are joint depth, height and width, respectively.

Although Shallow2 has a relatively symmetrical gesture space (Figure 4.4), Figure 4.14 clearly shows a dominant hand in the temporal domain. This Figure shows that this speaker leads with their right hand with a mean offset of 28 frames ($\approx 0.31s$) when considering the distance from the mean pose. When considering the individual axes, it is evident that the right-hand leads in all cases, and the X and Y axes, the offset is greater than $0.5s$. This suggests that, although a symmetrical pose is formed, a temporal offset exists between hands achieving this pose.

Other speakers' motions are more symmetrical, and minimal temporal offsets were found. Shallow3 in Figure 4.14 is an example where the mean offset does not exceed a mean of 17 frames ($0.19s$) in any axis.

4.9 Mutual Information

This section explores mirroring for data augmentation from an information theory perspective. Specifically, whilst mirroring effectively doubles the amount of data, how much additional *information* does it introduce? Mutual information is computed between the original data and its mirrored counterpart to reveal the dependence between the two distributions.

Normalised Mutual Information (NMI) [130] is measured on a per-speaker, per-axis basis at the wrist joint. NMI is computed using the following:

$$NMI(X,\tilde{X}) = \frac{I(X,\tilde{X})}{\sqrt{H(X)H(\tilde{X})}} \tag{4.1}$$

where $I(X,\tilde{X})$ is the mutual information between the original data X and mirrored data $\hat{X}$, and $H(X)$ and $H(\tilde{X})$ is the entropy of the original and mirrored data respectively. The entropy is calculated using the nearest neighbour approach [70].

Normalising the Mutual Information allows for easy comparison between speakers and axis, producing a value between 0-1. This NMI value describes the dependence of the two variables. At zero NMI, the variables are completely independent, and as the NMI increases to 1, it indicates a reduction in uncertainty and largely dependent variables.

The NMI for each speaker is shown in Figure 4.15. This shows that the amount of mutual information in the wrists is speaker-dependent. However, when considering the relative mutual information between axes, the Y-axis (movement of the wrist in the vertical axis) consistently has higher values. Therefore, this analysis suggests that more information will be gained in the movement along the X-axis (forward-back) and the Z-axis (left-right) from augmenting the dataset with mirrored poses. Information symmetry is revealed from NMI. Low levels of NMI and, therefore, low information symmetry indicate the importance of both wrists to predictive models. This is particularly important when regarding motion datasets gathered from video. As occlusion is common, arms are often interpolated or

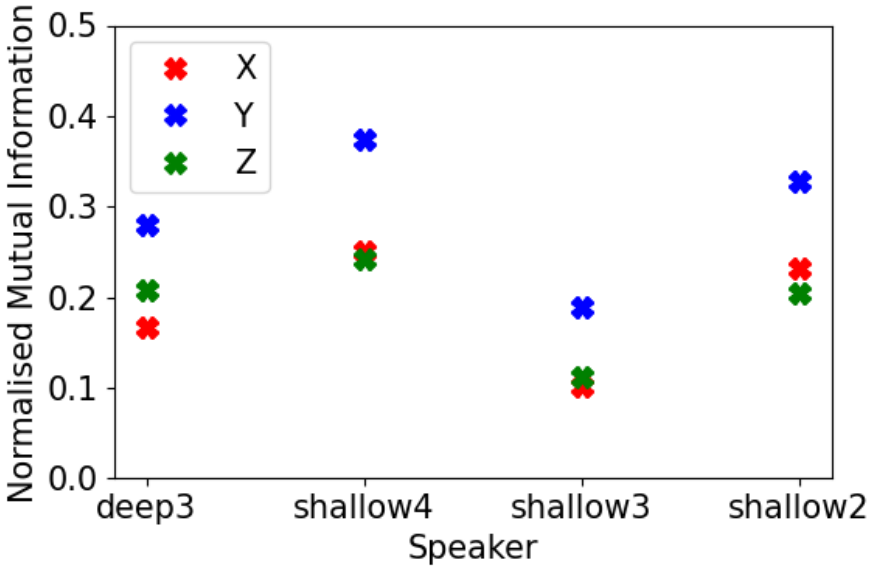


Fig. 4.15 Normalised Mutual Information per-speaker, per-axis measured between the original and mirrored wrist joints. Lower values represent a higher degree of independence.

missing from the data. By removing or including potentially incorrect arm movement on one side, important information is lost or large amounts of uncharacteristic information are introduced.

4.10 Mirroring Effect on Generative Modelling

To further support these findings, a Long Short-Term Memory (LSTM) model is trained on different data splits and use various augmentation settings to map from speech to body pose. This aims to determine the impact of including the potentially uncharacteristic mirrored motion for a speaker and whether including the mirrored speaker as a new *virtual identity* improves results.

4.10.1 Motion Representation

Of the 36 speakers released, only 18 have audio and motion capture available; therefore, this subset is used. The motion was down-sampled to 30fps to maintain realistic motion, but training time was reduced. A test sequence is randomly held out for each speaker, and the remaining data, 20%, is held out for validation, and 80% is used for training. Each speaker's global position is inconsistent; therefore, the respective mean global root position is removed from each frame on a per-sequence basis. 3D positions in world space are the target values, standardised by subtracting the mean pose and dividing by the standard deviation computed over all speakers across all training sequences.

4.10.2 Audio Representation

Mel Spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) are often used in speech-to-motion pipelines [48, 6, 133]. Instead, a model trained using a multi-task learning framework that is comprised of 12 regression tasks is used. PASE+ [120] features encode an audio waveform and should implicitly encode MFCCs and other speech-related information, including prosody and speech content. Speech is downsampled using a band-sinc filtering method from 44.1KHz to 16KHz.

4.10.3 Generative Model

An LSTM-based model trained to predict a frame of motion from a motion frame's worth of audio (33ms) is used. The prediction is also conditioned using a learned feature vector that encodes a speaker's identity to ensure the motion is speaker-specific. This learned feature vector should adequately associate the speaker and their gesturing style. This learned feature vector should allow the introduction of a speaker's potentially uncharacteristic mirrored motion to the model without affecting the gesturing style of the original speaker.

The LSTM model contains 4 bi-directional layers, each with 1024 hidden units and a 40% dropout, followed by a ReLU non-linearity layer and a fully connected layer. The output from the fully connected layer is the estimated (standardised) body pose at that frame.

4.10.4 Training Procedure

Models are trained using the Adam optimiser with a learning rate 0.0001 and batch size of 256. Not all sequences contain hand motion; where this is the case, the loss is computed against all joints in the body except the hands. 30-frame long sequences are used to train, with a 25-second overlap on each window.

A multi-term loss function is employed minimising the position values as an L_2 loss on joint positions and an L_2 loss on joint velocity and acceleration. Introducing the velocity and acceleration allows the model to produce smoother and more realistic transitions. On observation of some bone stretching artefacts due to positions not having any constraint on distance apart, an L_1 loss on bone length is included. The final loss L_c is computed as:

$$\begin{aligned} L_p &= L_2(y, \hat{y}) \\ L_v &= L_2(f'(y), f'(\hat{y})) \\ L_a &= L_2(f''(y), f''(\hat{y})) \\ L_b &= L_1(y_{lengths}, \hat{y}_{lengths}) \\ L_c &= L_p + L_v + L_a + L_b \end{aligned} \tag{4.2}$$

where y and $\hat{y}$ is the ground truth and predicted motion, and $y_{lengths}$ and $\hat{y}_{lengths}$ are Euclidean distances between each joint and its parent in the skeleton hierarchy for the ground truth and predicted motion respectively. L_p represents positional accuracy, L_v velocity accuracy, L_a acceleration accuracy, L_b bone length accuracy, and L_c is the combined loss. L_1 and L_2 represent Mean Absolute Error and Mean Squared Error, respectively.

4.10.5 Experimental Setup

The same model architecture is trained on each of the settings defined as follows:

All Data. A baseline formed using all available training data with no augmentation.

Half Data. A random subsample of the training data reduces the number of samples by approximately 50%. A model trained using this reduced data is helpful to compare the effect of doubling the size of the training set by augmentation versus adding additional ground truth data.

Mirrored Same Identity. Created by augmenting the *Half Data* training set by laterally mirroring the pose at each frame. Mirrored data is assigned the **same** identity label as the original speaker. This determines the impact of introducing uncharacteristic motion for a specific speaker.

Mirrored Virtual Identity. The *Half Data* training set augmented by laterally mirroring the pose at each frame. A **new** virtual identity label is assigned to the mirrored data during training. This determines if adding motion that could be considered characteristic for a different speaker aids or hinders performance.

All Data Mirrored Virtual Identity The model is also trained on all available training data and the laterally mirrored augmentation. The augmented sequences are assigned new virtual identity labels as in the *Mirrored Virtual Identity* setting. This represents the most optimal setting.

4.10.6 Results

Motion characteristics are used to evaluate performance. These include positional pose plots, distances from the mean pose and temporal handedness. This analysis should indicate how characteristic the predicted motion is and whether the introduction of motion has had an impact on performance. The same processing procedure as in Section 4.3 was followed,

and poses were translated per frame so that the midpoint of the left and right shoulders are centred on the origin. Overlaying pose projections at 1s intervals provides an overview of the predicted motion range and shows the type of gestures performed. Distributions showing the distance from the mean pose indicate a characteristic of how active a speaker is, which will indicate how similar the predicted motion of a speaker moves to the ground truth. Temporal offsets are analysed using cross-correlation lags between the onset of left wrist motion given right wrist motion in the Z (left-right) and Y (up-down) axes, which should show if the temporal characteristics are preserved across speakers.

4.10.6.1 Using the same identity

Two key findings can be observed: the mirrored data produced far more muted and symmetrical motion than desired.

The movement generated was found to be positionally symmetrical over the whole pose, particularly with arm movements. Figure 4.16a shows each of the arms consistently raising simultaneously when using mirrored data as the same identity. While using just half of the data and no mirror augmentation, more asymmetrical poses are closer to the characteristics performed in the ground truth.

Figure 4.16b indicates the time and distance away from the mean pose. It is a common trend across speakers that the distance from the mean pose was lower in the mirrored with the same identity split compared to motion generated from half of the data and the ground truth. This indicates the muted motion observed, producing slow and small movements.

Temporal symmetry is notably present when using the same identity. When the left-hand moves, the right hand also moves at the same time, producing unnatural motion. Figure 4.16 shows a strong correlation between the left and right wrists moving at a temporal lag offset of ± 1 frame. Compared to the ground truth, this high temporal symmetry is very uncharacteristic of the speaker.

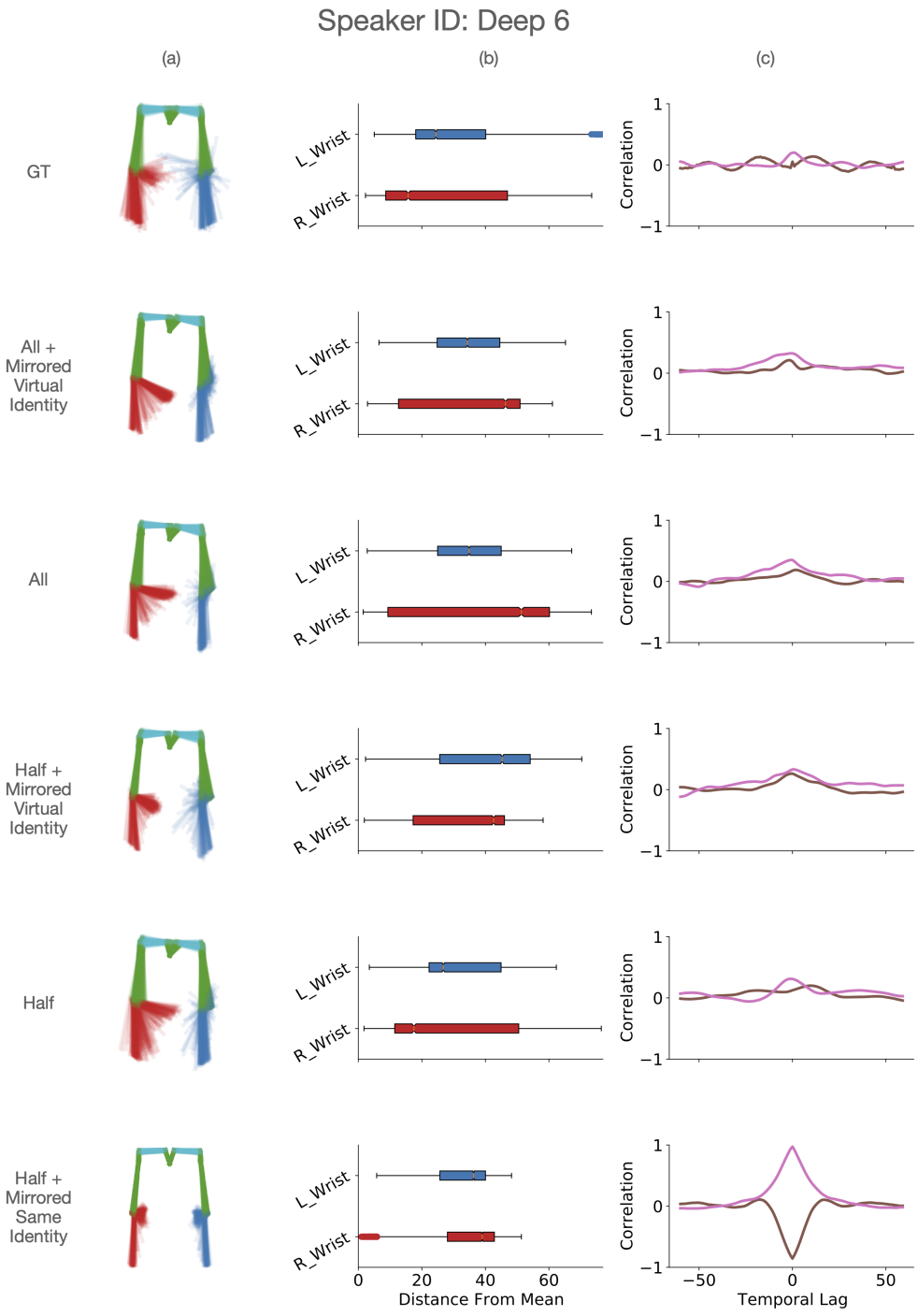


Fig. 4.16 A comparison for a single speaker's generated motion showing the detrimental impact of including mirrored motion under the same identity. Each row corresponds to a different data split used. Column (a) contains the orthographic projection of a pose at every second in the sequence. Column (b) shows the distribution of distances from the mean arm pose. Column (c) shows the cross-correlation lags between the onset of left wrist motion given right wrist motion in the Z (left-right) and Y (up-down) shown in brown and pink respectively.

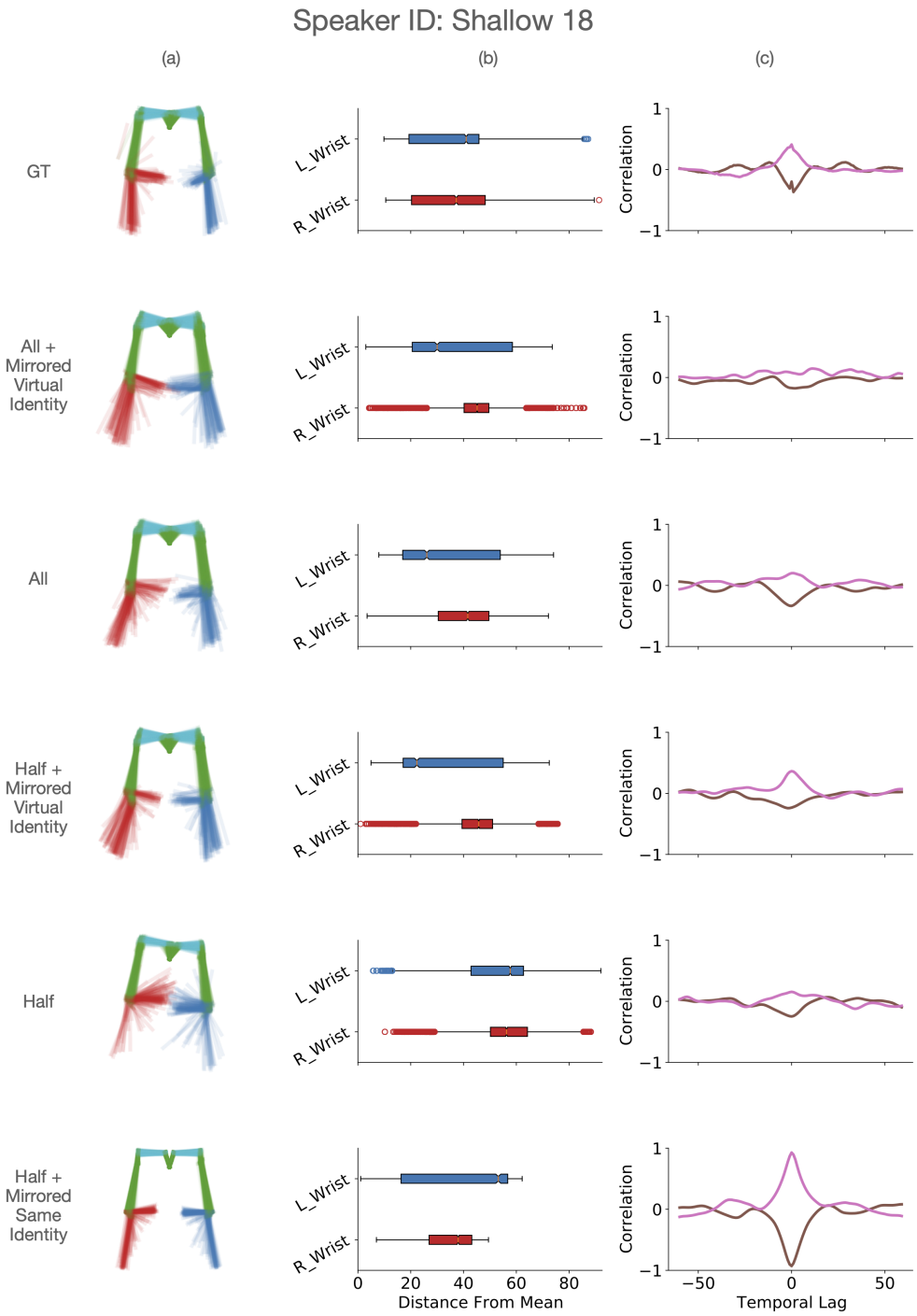


Fig. 4.17 A comparison for a single speaker's generated motion showing the detrimental impact of including mirrored motion under the same identity. Each row corresponds to a different data split used. Column (a) contains the orthographic projection of a pose at every second in the sequence. Column (b) shows the distribution of distances from the mean arm pose. Column (c) shows the cross-correlation lags between the onset of left wrist motion given right wrist motion in the Z (left-right) and Y (up-down) shown in brown and pink respectively.

4.10.6.2 Augmenting With a Virtual Identity

With the detrimental effect of including mirrored data under the same identity, further examination of the impact of including mirrored data under a virtual identity (*Mirrored Virtual Identity*) is required.

Improvements are identified in generated motion quality, which varies between speakers. However, a negative impact on performance was not found. Mirroring with a virtual identity was found to be competitive with a model trained with all the available data, often improving positioning, adding more movement that closely resembles the ground truth and generating motion from all of the data.

An example of improvement from including lateral mirrored data is shown in Figure 4.17. The distribution of distances from the mean pose shown in Figure 4.17b decreases from half of the data and half mirrored as a virtual identity. Poses in Figure 4.17a appear closer to the predictions using all of the data and ground truth. Lowering the arms more often than the generated motion using half of the data supports the hypothesis that adding mirrored data as a virtual identity can be competitive with a model including all data.

4.11 Discussion

In this chapter arm symmetry during dyadic conversation and its impact on lateral mirroring for body motion data augmentation were analysed. This presents the potential issues that could arise and when it would not be a suitable data augmentation approach.

If lateral mirroring is used for pose data augmentation, caution should be taken if the gesturing style and handedness of the speaker are to be preserved. From this analysis, it is clear that mirroring can result in valid poses and dynamics for specific speakers who move with a high degree of arm symmetry. Statistical analysis can be performed on a per-speaker basis to ensure that this is the case. However, the information gained from mirroring the arm

motion might be minimal for these highly symmetrical speakers. In most cases, the speakers did not move symmetrically, and the mirrored data would not reflect the actual characteristics of a speaker's gesturing style. While mirroring could produce a physically valid pose for a speaker, it may not fit with their motion style or handedness.

From the generative modelling discussed, a naive mirroring implementation did not predict characteristic or plausible motion and was found to be detrimental to model performance. Instead, the suggestion is for using a new *virtual identity* for the mirrored poses. It was found that the amount of improvement was speaker-dependent. This may be due to the non-uniform data distribution across the speakers. The dataset used has *shallow* and *deep* speakers, so the amount of data available per speaker varies. Although the models appeared to capture the speaker identities well, there is a chance that with small amounts of data for some speakers, the motion characteristics required to describe this speaker's motion are not present in the training data. The improvement may be due to increased generalised characteristics common across all speakers. Suppose the aim is to preserve the gesturing style and handedness of the original speaker. In that case, lateral mirroring should instead be used to increase the number of speakers in a dataset by treating the mirrored data as its own *virtual* identity. Care must still be taken to account for directional cues in the training data speech that could lead to a multi-modal disparity.

Shallow25 in Figure 4.7 is an example of an asymmetrical self-adaptor trait characteristic of that speaker. The left arm touching the right arm is common in their data, but the right arm does not appear to touch the left arm similarly. If this stylistic motion were maintained, simply mirroring the body pose would not suffice.

Mirroring the data has the potential to cancel out temporal offset characteristics. It is evident that specific speakers gesture with a leading hand. The generative model trained on original and augmented motion data with the same identity removes any temporal offsets and

produces temporally symmetrical motion. This synthetic motion would not be faithful to the original speaker.

Given the speaker-dependent nature of the amount of symmetry, it can be expected that the inclusion of a symmetry statistic to aid in numerous evaluation tasks. The use of statistics for synthetic motion evaluation is discussed in Section 4.11.1; however, it is also suggested to use these statistics for identity classification. Motion symmetry could be critical to the classification of speaker identity. It is expected that a discriminatory model (i.e. "Does this motion resemble the expected speaker?") could be successful when classifying using symmetry motion characteristics. More work is required to determine what degree of success classifying a speaker's identity using motion symmetry alone could provide.

The mutual dependence between the mirrored poses and the original is speaker-dependent, and it's evident that some information is gained through lateral mirroring. *More information* may be enticing. However, this measure does not inform on appropriateness, and the added information may introduce uncharacteristic motions.

Previous work by Çatak et al. [24] has considered the impact of handedness on the beat and representational gestures. They found that beat gestures preferred the dominant hand of the speaker, whereas representational gestures varied. There was no preference for left-handed speakers, but for right-handed speakers, there was a right-handed preference. This suggests that, although arm positions could be reflectively similar, the types of gesturing could be varied. When training a generative body motion model using mirrored motion, there is a risk that both hands will produce beat gestures in the synthesised animation, which may reduce realism or even understanding.

After analysing a few gesture types and their relationship to symmetry, it's not possible to generalise from this small analysis alone but it would be sensible to consider when certain gesture types could be adequately mirrored. Handedness must be maintained during directional or positional gestures, such as pointing, to communicate a direction. If a speaker

uses a gesture to signify to the left and the augmented version points to the right without adapting the corresponding audio speech, this would lead to a disparity in the multi-modal context. When building gesture-generation systems, keeping the handedness of gestures produced consistent would be beneficial.

Further study is required to determine the impact of modifying positional and temporal symmetry on realism and understanding. However, this chapter's findings suggest that care should be taken when augmenting data using lateral mirroring. There is a risk that with this augmented data, the motion could lose speaker-dependent characteristics.

4.11.1 Evaluating Synthetic Motion

Evaluating synthesised body animation is a significant challenge in data-driven embodied agent synthesis. It is common to evaluate the performance of generative models using a user study [6]. Assuming the synthesised data is to represent that of a particular speaker, the analysis from this study could also be considered as a performance evaluation method.

If the goal is to generate animated body motion that is faithful to the style of a particular speaker, it would be expected that the animation possesses the same positional and temporal characteristics as the speaker's ground truth motion. Statistical analyses based on the work presented in this chapter would provide good indicators of these qualities. The per-speaker percentage of time spent in the extreme gesture space, distributions of velocity, degree of spatial symmetry, and temporal lag of the animated result compared to the ground truth motion would indicate the similarities in both gesturing style and handedness.

Chapter 5

Body-Part-Specific Decoding

Contributing Publications

- Windle, J., Greenwood, D., and Taylor, S. (2022a). Uea digital humans entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 771–777

5.1 Introduction

Gesture generation from speech has evolved from previously focusing on the upper-body motion only [133, 6, 74] to now predicting the full body [154], including global position. Given this advancement, how the lower body is predicted could be crucial to gesture generation performance. One particular research question is whether using a single decoder to predict all joints in the whole body or a multi-decoder approach to produce body-section-specific experts could be preferable.

This chapter describes the UEA Digital Humans entry to the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) challenge 2022. The challenge aimed to further the scientific knowledge using a large-scale, joint subjective evaluation of

many gesture generation systems. Two models were presented to the challenge to evaluate the part-specific decoding method. A Bi-Directional Long Short Term Memory (BLSTM) to full body and a BLSTM multi-decoder to produce body-section-specific experts. Each system uses audio and word embeddings to predict a sequence of 6D rotation [159] values for each body joint, producing appropriate gesture animation. The GENEA Challenge compares these methods against other competitive systems in a large subjective user study.

A BLSTM is chosen as a baseline model for comparison due to their strength in modelling sequential data. Many speech-to-motion deep learning techniques are built upon BLSTMs [37, 131, 51] and LSTM-based models are a commonly used baseline in pose generation work [6, 133, 53]. Inspired by the multiple decoders used in Habibie et al. [48], a model is presented that uses BLSTMs to encode audio and text features and multiple BLSTM-based decoders that model specific areas of the body. The full body is divided into four sections: head, upper body (including arms), hands and legs.

This chapter compares the performance of generating gestures using body-part-specific decoding and predicting the full body. The data is first introduced, followed by a full description of each approach and its training methodology. The results are then evaluated and compared against each other using both objective and subjective measures.

5.2 GENEA22 Data

Each of these models is trained on the GENEA 2022 data [154] derived from the Talking With Hands dataset [78] as described in Chapter 3.4.2. This data consists of high-quality 30fps mocap data in Biovision Hierarchical (BVH) format, with corresponding speech audio and text transcripts. Talking With Hands recorded dyadic conversations with the mocap and audio, separated by each speaker and, in this challenge, treated independently. Each model uses the pre-trained PASE+ [120] speech audio encoder and pre-trained FastText [97] word encoder for multi-modal representations.

5.2.1 Motion Representation

Both models in this experiment predict global skeleton position and joint rotations as their output. A body pose at frame n is defined as:

$$\mathbf{p}_n = [x_n, y_n, z_n, r_{j,1,n}, ..., r_{j,6,N}] \tag{5.1}$$

where x, y, z denote the global skeleton position and $r_{j,1:6,n}$ form rotations for each joint j in the 6D rotation representation presented by [159]. $\mathbf{p}_n$ is a vector of length 501, representing the 83 joint rotations in the skeleton. These rotation representations have gained traction in 3D pose estimation [46, 141] due to Zhou et al. [159] finding these are more suitable for learning applications. Rotations can then be converted to 3D keypoint positions in world space via forward Kinematics.

BVH file formats contain two types of offset to consider, as discussed in Chapter 3.2.1. Global joint offsets and per-frame joint offsets. In BVH format, it is common to have a joint offset for each joint that represents each bone length. A per-frame joint offset is typically only present in the joint representing world position; in the case of Talking With Hands format, the joint is labelled `body-world`. However, Talking With Hands is different as each joint has a per-frame offset, too, possibly to account for bone-stretching in the data capture.

Talking With Hands contains multi-modal data of multiple speakers and, therefore, different physical attributes. For each speaker identity, a slight difference in bone lengths is observed between BVH files corresponding to the same speaker. This is likely due to the recording setup. However, the differences were minimal. For rendering purposes, a single random BVH file for each speaker is chosen from the training dataset, and these values are used across all outputs for the respective speaker.

Regarding the per-frame offsets found in the Talking With Hands dataset, low variance is observed in these values. Through visual inspection of the ground truth data, removing or

keeping these values static throughout all frames did not impact visual performance. While local playback of predicted motion was fine with the removed offsets, a static offset to each frame was added to ensure the BVH format was correctly formatted for the challenge. This static offset was chosen from the same random BVH file per speaker as the joint offsets, but only the first frame offset was used and repeated across all frames in each BVH file. Keeping the bone lengths and per-frame offsets static should allow the model to focus on representing the motion characteristics rather than physical attributes.

5.2.2 Audio Representation

The most suitable audio representation for speech-motion synthesis is an open research question. One of the most common audio speech representations chosen in previous work is Mel Frequency Cepstral Coefficients (MFCCs) [6, 48, 133]. While these have provided impressive results, there is scope for more descriptive features. Through empirical evidence, the Problem Agnostic Speech Encoding (PASE+) [120] outperformed MFCCs. PASE+ adequately encodes an audio waveform to represent features required for 12 regression tasks. These 12 tasks include estimating MFCCs, FBANKs and other speech-related information, including prosody and speech content. Therefore, MFCCs and other useful speech-related features are implicitly encoded in these features. PASE+ features are extracted before training. The PASE+ model expects audio waveforms to be sampled at 16kHz. Therefore, the audio was downsampled using a band-sinc filtering method from 44.1kHz to 16kHz. Most speech signals contain frequency components of only up to 8kHz. Therefore, 16 kHz is an optimal sampling frequency for speech, and downsampling should not hinder speech quality. The released, pre-trained PASE+ model extracts an audio feature embedding of size 768 for each 33ms motion frame.

5.2.3 Text Representation

Text embedding is included in the model to provide explicit word-based context to gestures. The FastText word embedding described by Bojanowski et al. [16] is extracted using the pre-trained model released by Mikolov et al. [97]. This word embedding has been used in multi-modal gesture generation before [152], suggesting it is known to produce effective word embeddings for gesture generation. Each word embedding is extracted at a size of 300 per word aligned with its respective time frame in the context of the audio waveform. For each frame of motion, the word embedding of the word being spoken at the time of the frame is included. A vector of zero values is passed if no word is spoken at a given frame. When a word is spoken across multiple frames, the vector is repeated for the appropriate number of frames.

5.2.4 Data Presentation

The speaker's identity is provided as a unique ID passed to an embedding layer. This layer contains a lookup table that stores a fixed vector embedding representative of the speaker. The layer contains trainable weights, meaning vector representations of speakers that move similarly should be close in vector space. This embedding acts as a style conditioning variable and produces motion that closely represents the style of the speaker ID provided. For this dataset, as there are only 17 different speaker identities, a small embedding size of 2 is adequate to represent the different speaker styles.

Before training, speech audio and text transcripts are pre-processed as described in Section 5.2. For both PASE+ and FastText models, these weights are frozen and not updated during training. Each data modality is then concatenated to a flat vector of size 1070 per motion frame, ready to be passed through the rest of the network.

5.3 Decoding Methods

There are two decoding methods of interest in this chapter. The ***BLSTM-Full*** baseline system represents a high-performing, simple, but effective method. A second model, ***BLSTM-Parts***, uses a BLSTM encoder, followed by BLSTM body-section-specific decoders. The encoder aims to represent the motion so that the decoders can each be specialists in predicting their respective body sections. Hyperparameters for both models have been fine-tuned using a hyperparameter sweep, and final parameters were chosen using a combination of low loss and objective measure scores and empirical observation from researchers.

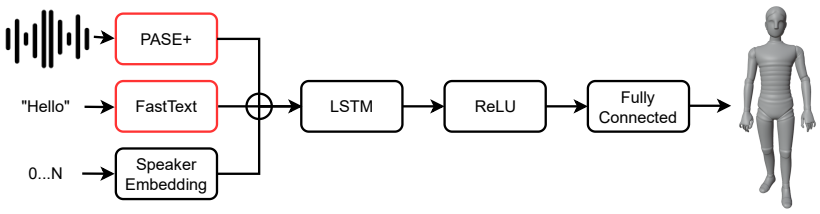


Fig. 5.1 Outline of the BLSTM-Full baseline model used for full body speech-to-motion prediction. This model takes as input speech audio, text transcript and speaker encoding. Outputs are the joint rotation values. A pre-trained model is used to extract the audio and text inputs. Red box defines frozen weights.

5.3.1 Bi-Directional Long Short Term Memory Baseline

A Bi-Directional Long Short Term Memory (BLSTM) baseline system is first trained, which is referred to as ***BLSTM-Full***. Figure 5.1 gives an end-to-end model overview. Data is processed and concatenated using the method described in Section 5.2 and then passed to a BLSTM network. The BLSTM output is then linearly projected using a fully connected layer to predict the pose $\mathbf{p}_n$ as defined in Equation 5.1.

This model consists of 4 bi-directional layers, each with 1024 hidden units and a 40% dropout, followed by a ReLU non-linearity layer and a fully connected layer. The output

from the fully connected layer estimates the 6D rotations of each joint and the global position of the `body-world` joint.

5.3.2 Body Part-Specific Decoders

A second architecture with part-specific expert decoders is introduced and referred to as ***BLSTM-Parts***. Figure 5.2 shows an end-to-end view of this model. Each decoder is responsible for a subset of joints representing the head, upper body (including arms), legs and hands. This should be a sufficient grouping of subsets, as each subset should consist of closely related joints. For example, Chapter 4 found that there is a close relationship between how each arm moves and, therefore, should be predicted together to ensure that any prediction made for one is also aware of the prediction of the other to keep this relationship.

The number of joints in each body part section varies. Table 5.1 summarises the number of joints predicted by each specific decoder. Although the hand joints contain the shortest bones, they also contain the largest number. This is due to each finger also containing between 3 and 4 joints. The head section only contains two joints, while the original skeleton contains very granular details of the face, including joints for the eyes, nose, and tongue. Despite being present in the skeleton, the joints are not tracked and, therefore, have a static rotation value and do not need to be included in the prediction.

Body Section	**Number Joints**	**Number of Values**
Head	2	12
Body	16	96
Hands	36	216
Legs	9	57

Table 5.1 Number of joints predicted by each body part-specific decoders.

The encoder consists of 4 bi-directional layers, each with 768 hidden units and a 40% dropout followed by a ReLU non-linearity layer. This follows a similar architecture as the baseline and provides a reliable encoding of motion from the input. Each body section is

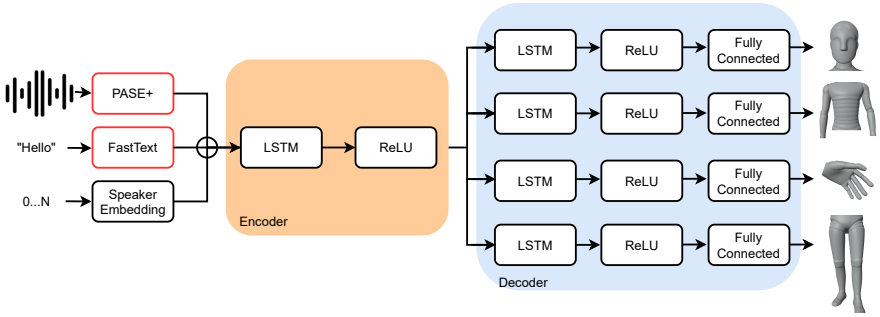


Fig. 5.2 Outline of BLSTM-Parts model used for speech-to-motion prediction with part-specific decoders. Red box defines frozen weights

predicted using a different decoder that follows the same architecture. A decoder in the architecture consists of 2 bi-directional layers, each with 768 hidden units and a 40% dropout, followed by a ReLU non-linearity layer and a fully connected layer. The output from each fully connected layer is the 6D rotations of representative joints. The decoder responsible for the legs also predicts the `body-world` position as the leg movement should have the most significant impact on the global position of the speaker.

5.3.3 Training Procedure

Each model is trained using the same procedure. The loss function contains multiple terms and weights. While 6D rotation values are learned, positions are also included when computing the loss. The loss comprises an L_2 loss on the rotations, positions, acceleration and movement velocity. By adding these terms, empirical findings found motion became smoother and expanded the range of motion performed compared to a rotation loss alone.

The final loss L_c is computed as:

$$
\begin{aligned}
L_p &= \lambda_p L_2(\mathbf{y}_p, \hat{\mathbf{y}}_p) \\
L_v &= L_2(f'(\mathbf{y}_p), f'(\hat{\mathbf{y}}_p)) \\
L_a &= L_2(f''(\mathbf{y}_p), f''(\hat{\mathbf{y}}_p)) \\
L_r &= \lambda_r L_2(\mathbf{y}_r, \hat{\mathbf{y}}_r) \\
L_o &= \lambda_o L_2(\mathbf{y}_o, \hat{\mathbf{y}}_o) \\
L_c &= L_p + L_v + L_a + L_r + L_o
\end{aligned}
\tag{5.2}
$$

where $\mathbf{y_r}$ and $\hat{\mathbf{y}}_\mathbf{r}$ are ground truth and predicted 6D rotations respectively, $\mathbf{y_p}$ and $\hat{\mathbf{y}}_\mathbf{p}$ are the world positions derived from the 6D rotations via Forward Kinematics and $\mathbf{y_o}$ and $\hat{\mathbf{y}}_\mathbf{o}$ are the global offsets for the root joint. f' and f'' are the first and second derivatives respectively. L_p is representative of positional distance, L_v similarity in velocity, L_a similarity in acceleration, L_r is the similarity in 6D rotations and L_o is how close the root offset is. λ_p is the weighting of positions, λ_r is the weighting of rotations and λ_o is the weighting of offsets. These weights are applied to bring all terms into the same order of magnitude and increase the importance of some terms. L_2 represents the Mean Squared Error between the two sets of data. A small parameter search was used to find the optimal term weights. Setting $\lambda_p = 0.1$, $\lambda_o = 0.01$ and $\lambda_r = 20$ produced the best motion from observation.

The Adam optimiser is used during training with a learning rate of 0.0001 and a batch size of 256. Where hand motion is absent from the dataset, the hand motion is excluded during the loss calculation. This encourages the model to learn effective finger movements and avoid learning a static hand position. To balance training time and data samples, motion is split into 30-frame chunks with the corresponding audio with a 25-frame overlap. Each model predicts a 30-frame sequence of motion, one frame at a time. Only the training data is used during training, and the validation data is reserved for model selection purposes only.

The BLSTM baseline is trained for 300 epochs, and the part-specific decoder, 240 epochs, which are determined by observed motion quality.

5.4 Evaluation

The performance of both the BLSTM-Full baseline and BLSTM-Parts models are first evaluated by objective measures and empirical observation. As discussed in Chapter 2.5.2, no single metric can effectively evaluate the quality of generated gestures. Instead, a combination of Frèchet Gesture Distance (FGD) [152, 13] and Beat Alignment (BA) [83, 84] scores have been used for their ability to reflect perceived realism and the alignment of the motion to the speech [7, 84, 152].

Further to the objective evaluation, empirical observation identifies two key issues from both models. Rotations were sometimes predicted to unnatural values, particularly in the shoulders and arms. In addition to this, foot contact and natural leg movement were not always guaranteed.

5.4.1 Objective Results

The BLSTM baseline is compared against the BLSTM-Parts method with respect to the ground truth motion. Results are summarised in Table 5.2.

The part-specific decoder method outperforms the BLSTM-Full baseline in both objective measures. The FGD score is much lower for the BLSTM-Parts method. This suggests that the distribution of motion performed during these predictions more closely resembles that of the ground truth test motion than the baseline BLSTM-Full. Beat Alignment scores suggest that both models' gestures align closely with audio beats. Again, the BLSTM-Parts model outperforms the BLSTM-Full baseline model, however, only slightly.

	FGD↓	BA↑
BLSTM-Full	189.263	0.822
BLSTM-Parts	**107.949**	**0.836**

Table 5.2 Frèchet Gesture Distance (FGD, lower is better) [152] and Beat Alignment (BA, higher is better) [84] scores for each system calculated with respect to the ground truth test dataset.

5.4.2 Ground Truth Comparison

The results of both models are shown in Figure 5.3, where six different test sequences are compared to the Ground Truth test sequences. Each sequence of poses sampled at 1-second intervals is plotted and overlayed to show the extent and types of gesturing that are generated. Additionally, the magnitude of the velocity for the head, right and left wrists and feet are calculated, and the distribution for both ground truth and predicted sequences is displayed. The velocity magnitude aims to describe the movement characteristics, which should closely match that of the ground truth. Each distribution should also indicate the amount of motion from each joint. Arm and head motion are important factors in communicative gestures. The joints are chosen as the joints are at the end of the kinematic chain and, therefore, display the most velocity change in relation to others.

For each sequence shown in Figure 5.3, the plotted predicted poses show similar levels of motion to the ground truth. The range of arm motion predicted by both BLSTM-Full and BLSTM-Parts appears to be more extensive than the ground truth. The magnitude of velocity distributions for the arms show that while these sequences are more active, the motion characteristics are similar, as the velocities seen in the predicted sequences closely match those of the ground truth. A common trend among the arm motion is the BLSTM-Parts distributions tend to show slightly higher velocities more often than the BLSTM-Full. These velocity distributions are marginally closer to the Ground Truth than the BLSTM-Full baseline.

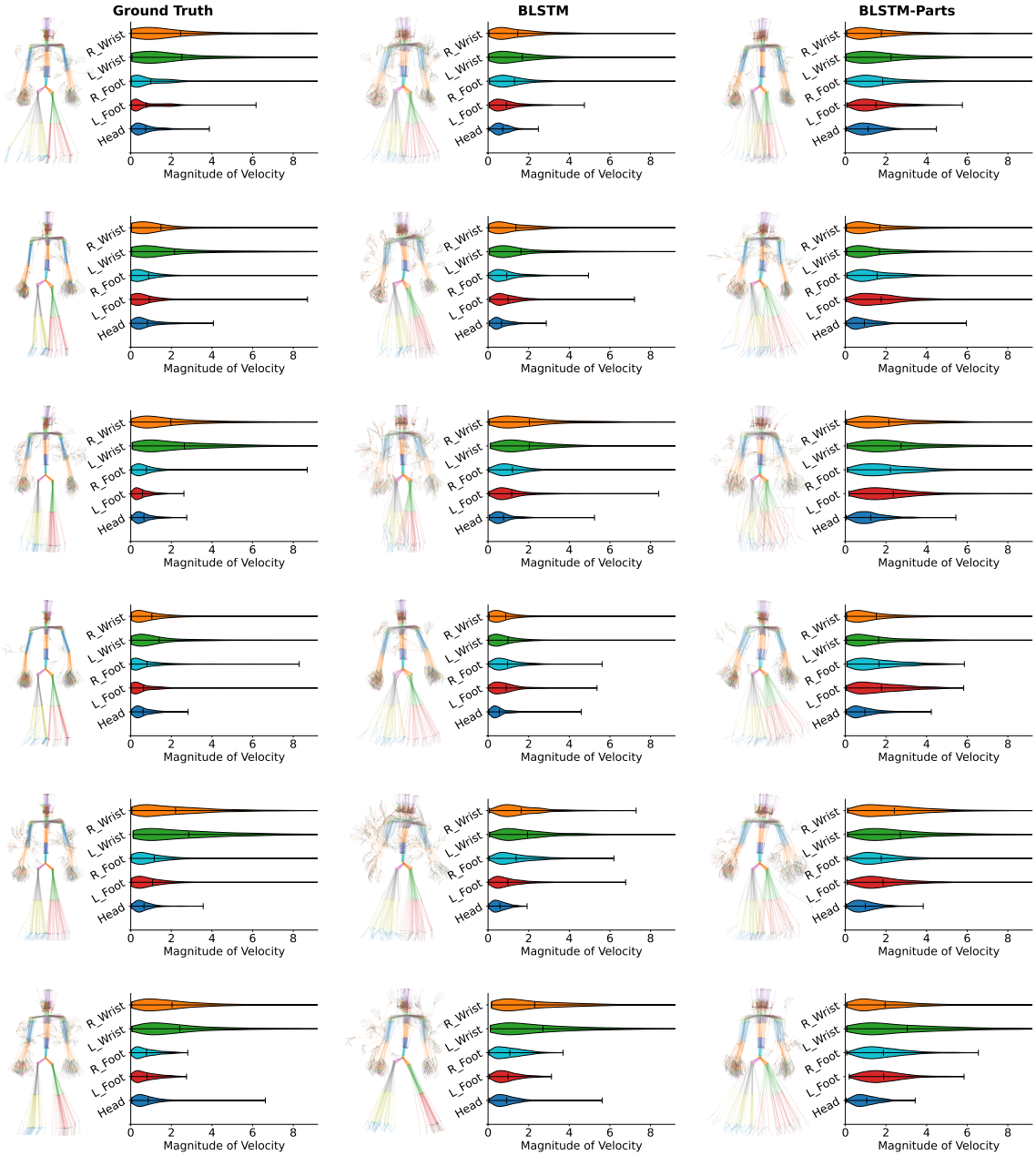


Fig. 5.3 A test sequence from 6 different style categories shown for ground truth (left) and sequences predicted from the Bi-Directional Long Short Term Memory (BLSTM)-Full Network (middle) and BLSTM-Parts (right) for the same audio sampled every 1 second. Each example is shown with the distribution of the velocity magnitude for the wrists, feet and head to indicate the amount and speed of motion across the sequence.

The most considerable difference in motion between predicted sequences and Ground Truth is regarding leg motion. In the BLSTM-Full and BLSTM-Parts predicted poses, the legs appear to form wider stances and move substantially more than the ground truth. These leg motions appear unnatural and may be distracting when visualising results. The BLSTM-Parts method also seems to include a number of erroneous leg poses that are persistently off the ground, removing foot contact. When examining the magnitude of the velocity of leg motion, the BLSTM-Full baseline appears relatively similar to the ground truth motion. The BLSTM-Parts method has a distribution that favours a much higher velocity than the ground truth data.

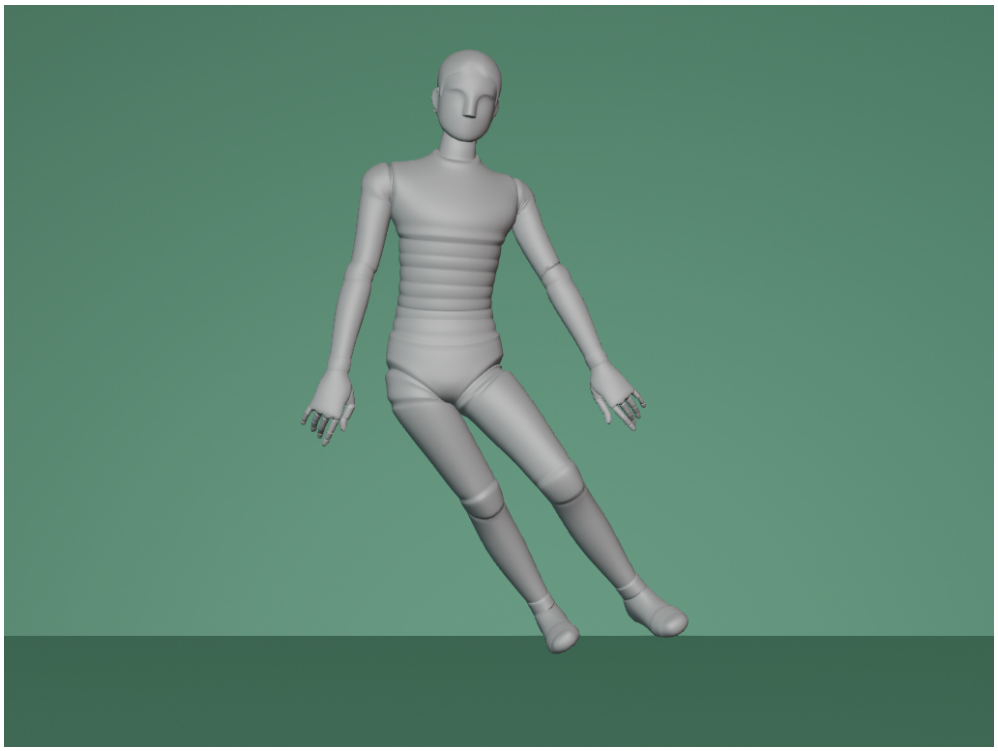


Fig. 5.4 Failure case for the BLSTM-Parts part-specific decoder model incorrectly predicting leg motion. This shows a pose where both legs are visibly raised from the ground in an unnatural position for the legs.

5.4.3 Foot Contact

Given these findings in the ground truth comparison, further investigation into foot contact is needed. The baseline BLSTM-Full model achieved some level of plausible leg movement and foot contact. However, the part-specific decoder model struggled to predict valid leg motion and foot contact. While some sequences of leg motion were realistic and appropriate, the predicted leg motion often involved large errors of foot contact where both feet were far from the ground.

Figure 5.4 shows an example of both legs raised unnaturally. While the results suggest that the part-specific decoders produce better arm, head and hand movement, this leg motion is distracting and essentially negates the better motion from the rest of the body.

5.4.4 Unconstrained Rotation

Although the inclusion of positions in the loss function was found to be beneficial, it introduced the issue of extreme rotations. If no weighting is applied to L_p in Equation 5.2, this term dominates the loss and causes unnatural rotations to be formed. This is due to solving inverse kinematics, as many solutions exist to form a particular pose. The model tended to produce impossible rotations. For example, rotations exceed a typical value range for a particular joint. Despite these physically impossible rotations, absolute positions of end-effectors relative to the over-rotated joint in world space appeared to be accurate.

Introducing a weight to constrain the positional influence allows a balance of valid rotation values and positive position influence. Despite the weight inclusion, there are still some issues regarding unnatural rotations. When viewing rendered sequences, unnatural poses can sometimes be observed being formed. Figure 5.5a shows an example of a pose where the right shoulder has a rotation value outside of the typical range and causes an unnatural pose. This issue would remain for several frames before recovering to a well-

formed pose. Figure 5.5b shows a recovered pose from the same sequence as Figure 5.5a. This issue is common in both proposed models, albeit slightly more prominent in the BLSTM

(a) Effect of the shoulder joint exceeding its typical range.

(b) Recovered pose

Fig. 5.5 An example of a sequence where a joint rotation exceeds a typical range of motion. In this case, the shoulder joint produces a rotation value, which pushes the right arm back into an unnatural position. These unnatural poses resolve themselves after a while, as shown by a pose from the same sequence once the rotation has returned to a normal range.

baseline. The motion predicted during these phases of over-rotation is still appropriate, and gesturing still appears to be as correct to the speech as in other phases. This issue could cause a negative effect when evaluating the human likeness of the predicted motion. However, the appropriateness of gestures should be less affected.

5.5 User Study Results

Each model was evaluated in the GENEA user study, comparing both models to other systems. The GENEA Challenge 2022 consisted of two user studies, full-body and upper-body, where only one system could be submitted to each. Due to the severe leg motion and foot contact issues mentioned previously, for the subjective user study, the BLSTM-Parts model is compared to other systems using only the upper body. This is to determine whether the other part-specific experts may aid performance without being unfairly compared when

including the distracting leg movement, which detracts from the performance of the rest of the body.

To effectively compare synthesised gesture, the GENEA Challenge compared the submitted synthesised motion to both natural motion and pre-existing baseline systems. Each motion condition was assigned a three-letter ID following a set structure:

U/F - Representing Upper/Full body respectively.

N/B/S - Representing Natural/Baseline/Submission respectively.

A-Q - Representing a unique ID.

For example, the full body natural motion is assigned the ID **FNA** and full body baseline, **FBT**. The BLSTM-Full baseline is entered into the full-body tier with the ID **FSG**, and the part-specific decoder, BLSTM-Parts, is entered into the upper-body tier with the ID **USM**. Table 5.3 provides results of the user-study from the main challenge paper [154].

5.5.1 Human-likeness

Evaluating human-likeness determines whether the synthesised motion looks like the motion of an actual human, controlling for the effect of the speech. The video stimuli audio is removed to ensure the measure is for human-likeness alone and decoupled from appropriateness. A Human Evaluation of Multiple Videos in Parallel (HEMVIP) [61] method was used to measure this. For this question, multiple motion examples are presented in parallel, and the subject is asked to assign a rating to each one. Each question asked "*How human-like does the gesture motion appear?*" while being presented with eight video stimuli to be rated on a scale from 0 (worst) to 100 (best) by adjusting an individual GUI slider for each video. This rating allows for pairwise statistical tests and produces a median rating for each condition.

Both models ranked fourth in human-likeness compared to all other systems except natural motion. While not performing best, this is still higher than four other systems, including the baseline models, suggesting the models successfully capture a reasonable level

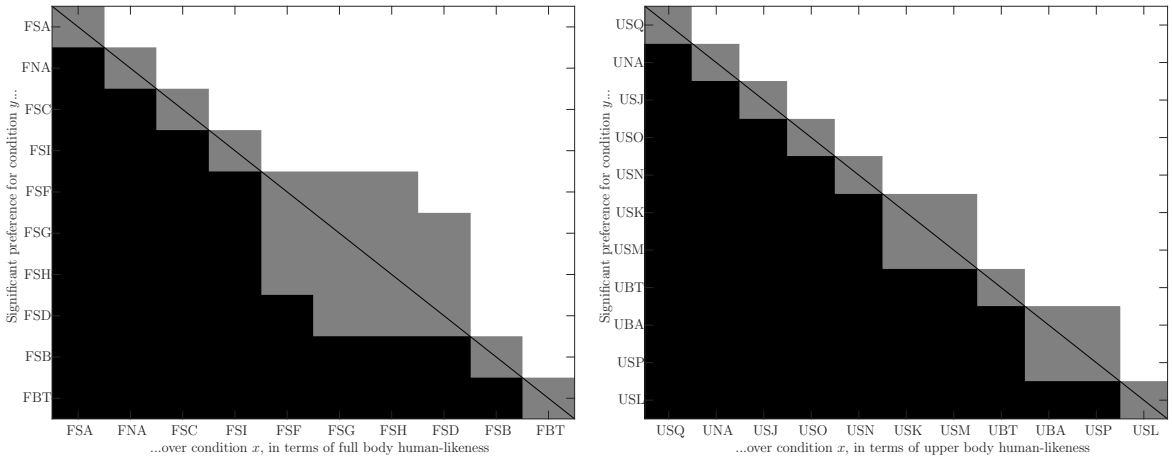
ID	Human-likeness		Appropriateness			
			Number of responses			Percent matched
	Median	Mean	Match.	Equal	Mismatch.	(splitting ties)
FNA	$70 \in [69, 71]$	66.7 ± 1.2	590	138	163	$74.0 \in [70.9, 76.9]$
FBT	$27.5 \in [25, 30]$	30.5 ± 1.4	278	362	250	$51.6 \in [48.2, 55.0]$
FSA	$71 \in [70, 73]$	68.1 ± 1.4	393	216	269	$57.1 \in [53.7, 60.4]$
FSB	$30 \in [28, 31]$	32.5 ± 1.5	397	163	330	$53.8 \in [50.4, 57.1]$
FSC	$53 \in [51, 55]$	52.3 ± 1.4	347	237	295	$53.0 \in [49.5, 56.3]$
FSD	$34 \in [32, 36]$	35.1 ± 1.4	329	256	302	$51.5 \in [48.1, 54.9]$
FSF	$38 \in [35, 40]$	38.3 ± 1.6	388	130	359	$51.7 \in [48.2, 55.1]$
FSG	$38 \in [35, 40]$	38.6 ± 1.6	406	184	319	$54.8 \in [51.4, 58.1]$
FSH	$36 \in [33, 38]$	36.6 ± 1.4	445	166	262	$60.5 \in [57.1, 63.8]$
FSI	$46 \in [45, 48]$	46.2 ± 1.3	403	178	312	$55.1 \in [51.7, 58.4]$

(a) Full Body Results

ID	Human-likeness		Appropriateness			
			Number of responses			Percent matched
	Median	Mean	Match.	Equal	Mismatch.	(splitting ties)
UNA	$63 \in [61, 65]$	59.9 ± 1.3	691	107	189	$75.4 \in [72.5, 78.1]$
UBA	$33 \in [31, 34]$	34.6 ± 1.4	424	264	303	$56.1 \in [52.9, 59.3]$
UBT	$36 \in [34, 39]$	37.0 ± 1.4	341	367	287	$52.7 \in [49.5, 55.9]$
USJ	$53 \in [52, 55]$	53.6 ± 1.3	461	164	365	$54.8 \in [51.6, 58.0]$
USK	$41 \in [40, 44]$	41.5 ± 1.4	454	185	353	$55.1 \in [51.9, 58.3]$
USL	$22 \in [20, 25]$	27.2 ± 1.3	282	548	159	$56.2 \in [53.0, 59.4]$
USM	$41 \in [40, 42]$	41.9 ± 1.4	503	175	328	$58.7 \in [55.5, 61.8]$
USN	$44 \in [41, 45]$	44.2 ± 1.4	443	190	352	$54.6 \in [51.4, 57.8]$
USO	$48 \in [47, 50]$	47.3 ± 1.4	439	209	335	$55.3 \in [52.1, 58.5]$
USP	$29.5 \in [28, 31]$	32.4 ± 1.4	440	180	376	$53.2 \in [50.0, 56.4]$
USQ	$69 \in [68, 70]$	67.5 ± 1.2	504	182	310	$59.7 \in [56.6, 62.9]$

(b) Upper Body Results

Table 5.3 Table of results from main challenge paper [154]. Summary statistics of user-study ratings from all user studies, with confidence intervals at the level $\alpha = 0.05$. "Percent matched" identifies how often participants preferred matched over mismatched motion regarding appropriateness. The part-specific decoder BLSTM model results are highlighted in pink. Higher is better for Median, Mean, Match and Percent Matched columns. For Mismatch, lower is better, and for Equal, lower is preferable.



(a) Full-body study

(b) Upper-body study

Fig. 5.6 Figure from main challenge paper [154]. Significance of pairwise differences between conditions. White means the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction.

of human-likeness. This limited performance of both models is likely due to the over-rotation issues described in Section 5.4.4.

While it is not possible to compare the results of each model directly in this study, it is possible to compare each performance with their respective ground truth ratings. Although the upper-body median is only 3 points higher, it is interesting to compare this against the median of the ground truth. The median rating of the BLSTM-Full baseline in the full-body study is 32 points lower than the ground truth. However, a lower median value of the upper-body ground truth means the gap between the BLSTM-Parts model and ground truth is 22. This suggests the BLSTM-Parts model may produce motion that is closer in human-likeness to the ground truth than the BLSTM-Full baseline when ignoring the lower body.

Challenge organisers also included baseline systems in the challenge. These use the IDs **FBT/UBT** for text-only baselines and **UBA** for the audio-only baselines. Figure 5.6 shows that the proposed models are significantly better in both challenge tiers than all baselines.

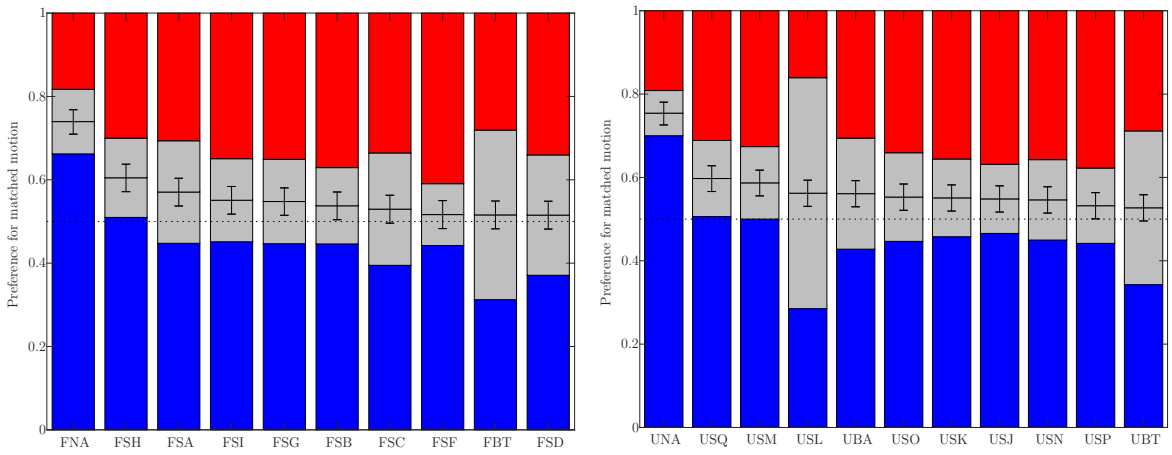
5.5.2 Appropriateness

How appropriate the motion is to speech is one of the most critical performance characteristics for gesture generation. The appropriateness evaluation is designed to assess the appropriateness separate from the intrinsic human-likeness of the motion. To evaluate this, subjects are presented with a pair of videos containing the same speech audio. One video is the motion associated with that audio, and the other is a mismatched motion for the audio, i.e. the motion is related to another speech audio stimulus. Both motion stimuli are from the same condition system to ensure human-likeness does not influence the choice. Subjects were asked to "*Please indicate which character's motion best matches the speech, both in terms of rhythm and intonation and in terms of meaning.*" with answers gathered by selecting the character on the left, on the right, or indicate that the two were equally well matched. This allows calculating the percentage number of times subjects prefer the matched over mismatched conditions.

Both BLSTM models performed well in the appropriateness of gesture to speech. BLSTM-Full (**FSG**) ranked third, and the BLSTM-Parts (**USM**) ranked second against all other systems excluding natural motion. Figure 5.7 visualises the distribution in responses from the appropriateness study. The full-body model remained in the middle of the pack but can still be considered significantly more appropriate than random chance as the confidence interval does not overlap with the 0.5 value of random chance.

While it is not possible to draw a statistical significance against any other submissions, the fact that the upper-body submission went from the middle of the pack in human likeness to gaining the second-highest appropriateness score in the submissions is promising.

It is difficult to derive the reason for this performance gain from the user study alone. However, it is possible to speculate based on visual observation. Observed gestures produced from this model would start at the expected time in relation to speech. The gesture intensity was also an expected value, particularly in the arm motion. The timing of beat gestures



(a) Full-body study

(b) Upper-body study

Fig. 5.7 Figure from main challenge paper [154]. Bar plots visualising the response distribution in the appropriateness studies. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. The black horizontal line bisecting the light grey bar shows the proportion of matched responses after splitting ties, each with a 0.05 confidence interval. The dashed black line indicates chance-level performance. Conditions are ordered by descending preference for matched after splitting ties.

can be related to prosodic characteristics of speech [17]. The observed accurate timing and intensity may come from using PASE+ features that adequately encode many speech features, including prosody. As the arm-specific decoder only has to focus on predicting arms, it is possible this decoder can more effectively use these features.

5.6 Discussion

While both models performed well in general, particularly given the simplicity of the architecture, there are still many things to consider going forward. Leg movement is a limiting factor in the predicted motion, particularly with the multiple-decoder model. This may be due to a weak correlation between speech and leg motion. Gestures are rarely made by legs alone; instead, the leg motion likely depends on the motion of the rest of the body.

There appears to be a disparity between the leg movement and the rest of the body. Qualitative observation and objective measures suggest that the addition of independent decoders for separate body parts works well and has been shown to work effectively in Habibie et al. [48]. Motion in the fingers, arms and head appears to improve over the BLSTM baseline. However, with the severe limitation of leg motion disparity, decoding motion in separate body sections cannot be recommended in its current state against predicting all joints using a single decoder. Decoding the legs with the core body may help with the disparity in leg movement.

Both models have room for improvement regarding human-likeness. This may be due to the occasional extreme rotation described in Section 5.4.4. This may be solved by further cleaning of the data or in future work, it may be useful to include constraints on joints. For example, setting hard limits on how far a joint can rotate. These could be learned from data or hand-crafted limits on a per-joint, per-speaker basis.

Chapter 6

Style Conditioned Speech-To-Gesture Generation With Long-Term Context

6.1 Introduction

This chapter introduces a novel transformer diffusion architecture for speech-to-gesture generation. The diffusion mechanism generates latent samples that are then transformed by a Transformer-XL architecture into a sequence of gestures. The proposed Gesture Diffusion Network has expanding memory and can generate smoothly varying animations for any input size. Additionally, the model can be conditioned on speaking style, enabling stylised animation that can be controlled at inference time. The audio is encoded using PASE+ features [120], which allows the model to generalise across out-of-domain audio.

Speech-to-gesture generation is a challenging problem due to the inherent ambiguity in mapping speech-to-body poses. Traditional supervised techniques, which model the problem as deterministic, often struggle to achieve good results. Probabilistic models, such as diffusion models, are more suitable for this task as they can handle the many-to-many relationship between speech and gestures. Furthermore, gestures can be temporally sparse or extended over long durations, which standard recurrent techniques do not handle well.

Transformer architectures can be effective but are typically limited to fixed input contexts and do not scale well with increasing input size. The Transformer-XL architecture [29] addresses this limitation by providing an extended attention mechanism and segment-level recurrence, allowing for larger input context and efficient training and inference.

Experiments show this model can generate diverse, natural-looking gesture animations for different voices and speaking styles. A user study supports that this approach produces appropriate and realistic gesturing and outperforms state-of-the-art methods. Rendered animations from the proposed model are available[1].

This chapter will first motivate style-controlled animation and the use of probabilistic, recurrent methods. The Gesture Diffusion Network is then defined before an experimental setup is described to produce stylised co-speech gestures. These produced gestures are then evaluated using empirical and objective measures to determine the efficacy of gesture production and style conditioning against the ground truth, as well as the ability to interpolate styles and produce motion for out-of-domain audio. An extensive subjective and objective evaluation is performed to compare the Gesture Diffusion Network to other state-of-the-art methods. The amount of historical memory available to the model is an important aspect of the proposed approach, and therefore, an ablation study looks at the effect of length of memory on both training and inference time.

6.2 Style Controlled Diffusion Motivation

It is desirable for the animator to be able to control the style of the generated motion at a high level, using conditioning variables, for instance, [4, 42, 41, 152, 6, 7]. Style can refer to characteristics of the overall motion, such as hand height, speed, radius, and symmetry [6] or to a speaker's individual traits [152]. Gesturing style can also be determined by an actor's

[1] https://youtu.be/x6B8rAffUJ0

age and affective state [42] or by back-channel inputs such as whether the actor is shouting or whispering, speaking or listening [133]. Prior work has incorporated style by concatenating a one-hot vector to the input [133] or learning a style embedding space [152, 116]. The method introduced in this chapter takes the latter approach since it allows interpolation within the embedding space and gives continuous style control.

Probabilistic models such as Variational Autoencoders [9, 41, 42, 82] and Flow-based models [133, 6, 53] have gained popularity for gesture generation since they better model the ambiguities between speech and human motion. Diffusion models [55, 104] are considered state of the art in probabilistic generative modelling due to their impressive ability to produce realistic and diverse output, particularly in text-to-image generation [32, 124, 119, 126]. These models work by sampling noise from a distribution and gradually refining it to produce a realistic target sample, typically incorporating conditioning values to guide the denoising process. One of the key features of diffusion models is their ability to generate multiple results for the same conditioning value, as the random sample of noise also influences the output.

6.3 Unlimited Sequence Length Prediction

Recent research has shown that diffusion models can also handle time-series or sequence-based data, such as generating 3D motion from text [65, 134, 157] or dancing and speech gestures from audio [28, 8, 156]. Sequence-based diffusion models have made use of Transformer architectures [134, 65] for their ability to model temporal information effectively. However, a limitation of vanilla Transformer architectures is that they do not scale well with the length of the sequence. As the sequence length increases, the size of the self-attention mechanism also grows exponentially, leading to memory and computational limitations. Also, diffusion models necessitate passing data through the model multiple times, further exacerbating these limitations. As a result, architectures have to limit the maximum generative

length to a fixed number of frames [65]. Speech sequences may naturally exceed the limits imposed by length-limited models. A naive method to process them is to divide the speech into segments and process each segment separately before combining the results. However, this approach preserves very little context of previous predictions and introduces a discontinuity between segments.

The Transformer-XL [29] architecture allows for extended attention beyond a fixed length. It introduces a recurrence mechanism to the Transformer architecture using segment-level recurrence with state reuse and a learned positional encoding scheme. The memory length can be adjusted after training, allowing more context to be included during inference than during training. As changing memory length can alter the positional context of historical events, the Transformer-XL introduces a learned, *relative* positional encoding scheme to accommodate the context length change. The Transformer-XL can, therefore, be trained more efficiently on narrower segments than a vanilla Transformer without compromising on the length of historical influence.

6.4 Gesture Diffusion Network

This section introduces a *Gesture Diffusion Network* that combines a diffusion model with a Transformer-XL for predicting a sequence of poses, $\mathbf{X}$, from a stream of audio, $\mathbf{a}$, and a given style. The diffusion process consists of two steps, *noising* and *denoising*. The *noising* process gradually adds a small amount of noise to a sequence of poses. The Gesture Diffusion Network then gradually *denoises* the noisy sequence and generates a sequence of poses conditioned on speech and style.

This model is the first to integrate diffusion modelling with a Transformer-XL architecture to model the longer-term relationship between speech and gesture for input of any duration. A similar concurrent architecture by Alexanderson et al. [7] combined diffusion models with Conformers for audio-driven gesture and dance generation. This method is not theoretically

constrained by sequence length due to the inclusion of TiSA (Time Interval aware Self-Attention for popularity prediction) [144] for positional embeddings, however, due to the nature of Transformer-based approaches, the complexity for this model scales quadratically with sequence length. This means the model would often be inhibited due to memory constraints being met due to long sequence lengths. By instead using a Transformer-XL based architecture, this memory usage can be split into smaller segments and allow for more efficient inference for long sequences while still retaining historical context from past predictions.

This section provides an end-to-end description of conditioned speech-to-gesture generation with long-term context. It starts by describing the diffusion process and feature extraction before introducing the *Gesture Diffusion Network* and describing how this model handles long-term context.

6.4.1 Diffusion Noising Process

The *noising* step consists of a Markov Chain $q(\mathbf{x}_k|\mathbf{x}_{k-1})$ for $k \in \{1,...,K\}$ where K denotes the number of diffusion steps. During training, given a ground truth sequence of poses $\mathbf{x}_0$, the Markov Chain adds noise progressively until $q(\mathbf{x}_K|\mathbf{x}_0)$ approximates a standard normal distribution and no longer resembles $\mathbf{x}_0$. The noise added is sampled from a normal distribution such that

$$q(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k; \alpha_k \mathbf{x}_{k-1}, \beta_k \mathscr{I}) \tag{6.1}$$

where α_k is defined as $\sqrt{1-\beta_k}$ [55, 129]. The value β_k is defined by a variance schedule that determines the intensity of the noise ε_k being added. The cosine schedule used by Nichol et al. [104] progressively adds a small amount of noise each time. For the initial noising step at inference where $k = K$, ε_K is sampled from $\mathcal{N}(0, \mathscr{I})$.

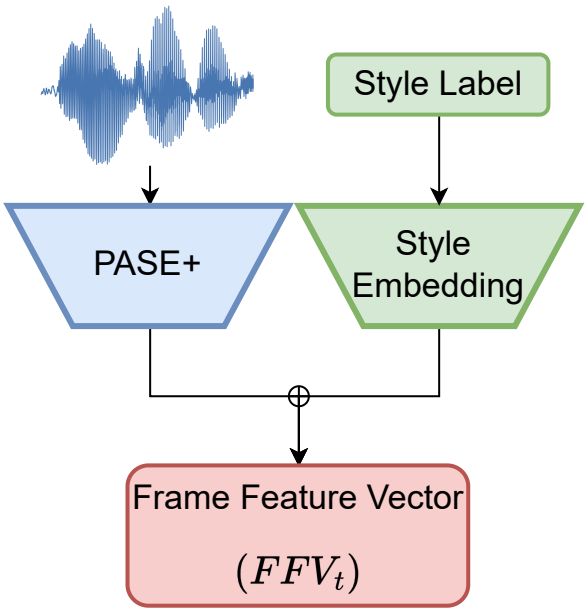


Fig. 6.1 Feature Extraction process. At 30fps, PASE+ features are extracted [120] as well as the learned style embedding and concatenated to a vector representing the input for a single motion frame.

6.4.2 Feature Extraction

Each audio sequence $a_{1:T}$ can vary in length, T. Each sequence is split into non-overlapping segments of length w frames. Splitting into segments is beneficial to avoid the computational limitations described in Section 6.3 and allow prediction without a limit on audio length. The proposed model choice ensures that the transition between predicted windows remains coherent and smooth. Frame Feature Vectors (FFVs) are computed for each window, which encode the speech and style for each frame in a segment using the architecture in Figure 6.1 to produce $\mathbf{FFV}_{t:t+w}$. A frame is defined as a motion frame sampled at 30fps. The Gesture Diffusion Network is trained to predict a sequence of poses $\hat{\mathbf{x}}_{1:T}$ from the series of Frame Feature Vectors $\mathbf{FFV}_{1:T}$ and a *noised* sequence of poses. At inference time, these sequences of poses are a random sample from a normal distribution, and at training time, these are the ground truth poses that have noise added as described in Section 6.4.1.

6.4.3 Gesture Diffusion Network

The *Gesture Diffusion Network* provides the denoising mechanism to reverse the *noising* process using a parameterised backwards process, $p(\mathbf{x}_0|\mathbf{FFV})$. The Transformer-XL model [29] gradually reduces $\mathbf{x}_K$ to $\mathbf{x}_0$ conditioned on the associated **FFV** given a noise sample ε_k. In diffusion models, the noise ε is commonly predicted at every diffusion step. However, inspired by other works [134, 119], the estimated pose $\hat{\mathbf{x}}_k$ is predicted at each step, which enables the direct use of geometric terms in the loss function. The Transformer-XL includes memory from the previous context, introducing segment-level recurrence with state reuse.

Figure 6.2 shows an overview of the proposed Gesture Diffusion Network, which is repeated K times for each segment. At each noise step k, its corresponding sinusoidal time embedding is used as input to a Diffusion Step Embedding. The embedding is from a network comprising two linear layers with a Sigmoid Linear Unit activation function between the two. The output of the Diffusion Step Embedding is the first sequence input to the Transformer-XL. Each $\mathbf{FFV}_{t:t+w}$ is concatenated with the corresponding noise sample $\varepsilon_{k,t}$, which has passed through a single linear layer. These combined features are then transformed using another single linear layer before being passed to the Transformer-XL as the rest of the sequence.

The Transformer-XL contains attention mechanisms much like a vanilla Transformer; however, instead of computing attention only on the current sequence, it also has access to the past context. This knowledge is given in the form of reusable states from earlier segments, defined as $\mathbf{W}_{t-m}$ where m is the memory length. Memory length can extend beyond a single segment in the past, and due to the learned relative positional encoding, it can differ during training and inference.

The sequence output from the Transformer-XL architecture is passed to a decoding layer. The decoder layer outputs the predicted pose $\hat{\mathbf{x}}_{k,t}$. The last denoising step $(k = 1)$ predicts the final pose sequence. For all values of $k > 1$ noise is added to $\hat{\mathbf{x}}_k$ according to the noise schedule to produce $\varepsilon_{k-1,t}$.

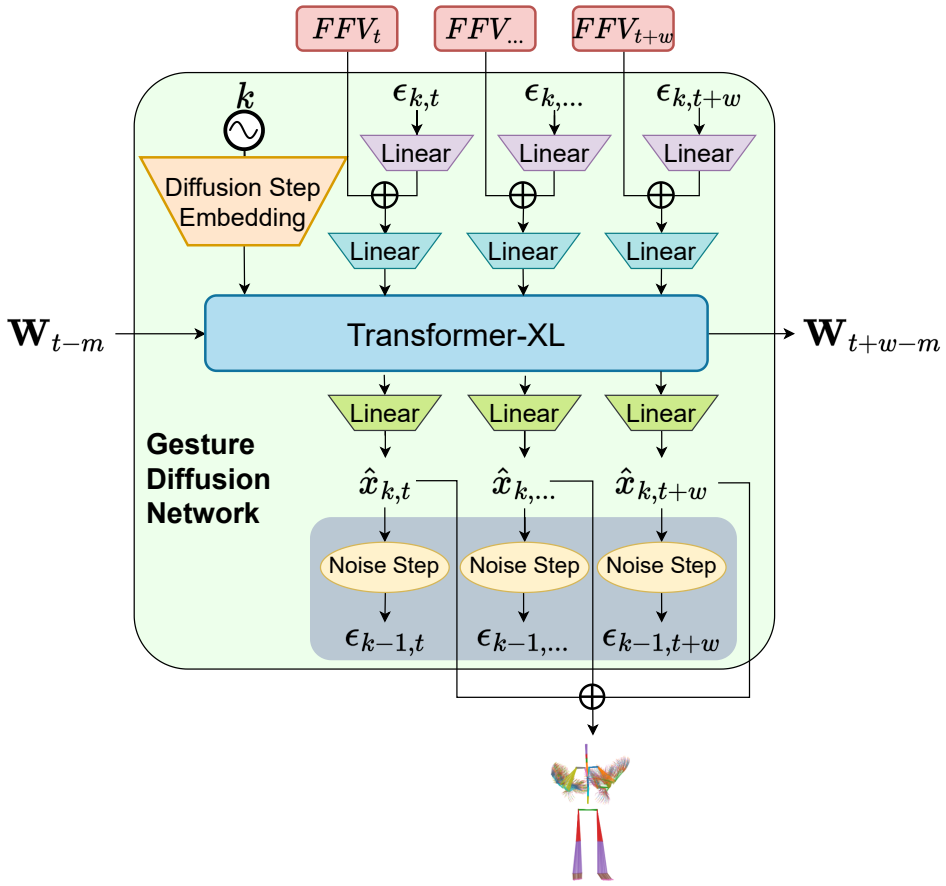


Fig. 6.2 Gesture Diffusion Network. The diffusion process runs for $k = 1000 : 1$ steps. Given a sequence of feature vectors, **FFV**, and noisy pose vectors, ε, of length w, each step predicts the corresponding denoised pose sequence $\hat{\mathbf{x}}$. For all steps $k > 1$, the prediction $\hat{\mathbf{x}}$ is subsequently noised and fed to the next denoising step, concatenated with the same **FFV**. Colours indicate the same layer being used for each input when applicable.

6.4.4 Extended Context

One of the contributions of the proposed architecture is the model's ability to remember the long-term context of variable length that can extend beyond a single segment. Figure 6.3 shows an example of predicting $\hat{\mathbf{x}}_{1:360}$ from $\mathbf{FFV}_{1:360}$ using a segment size of $w = 90$ frames and memory length of $m = 180$. For the first input segment, $\mathbf{FFV}_{1:90}$, the Gesture Diffusion Network returns the predicted pose sequence $\hat{\mathbf{x}}_{1:90}$. The Transformer-XL model also *caches* the *fixed* hidden states $\mathbf{W}_{1:90}$. For the subsequent input segment, $\mathbf{FFV}_{91:180}$, the Gesture Diffusion Network also has this additional context of the *cached* previous segment.

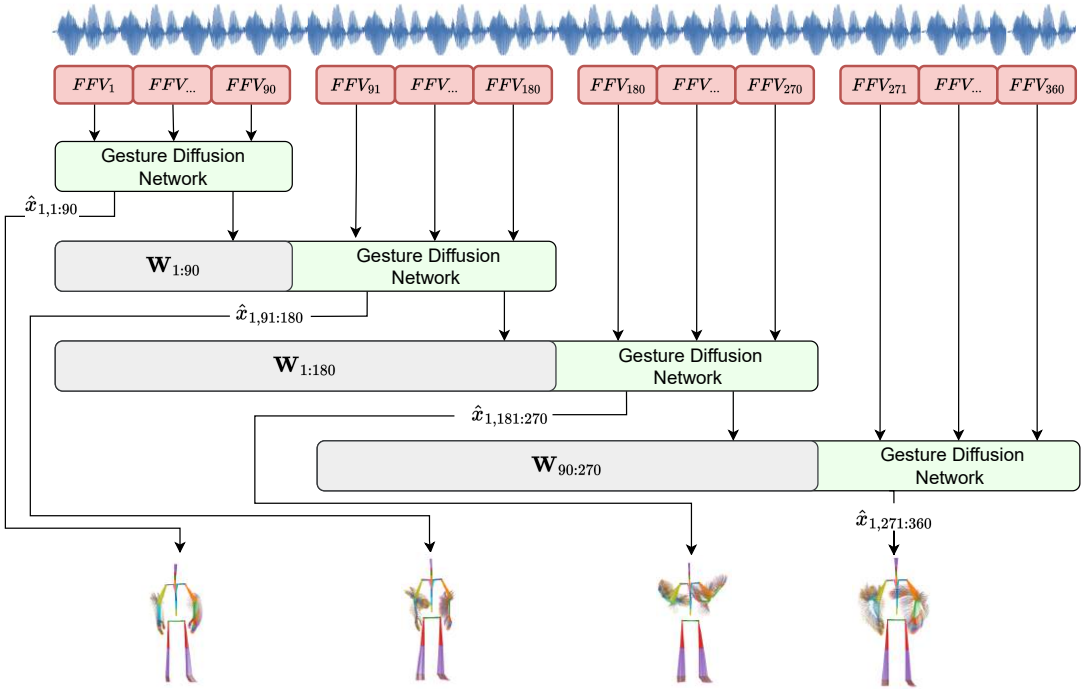


Fig. 6.3 Overview of the long-term context model at inference time. The audio is split into segments of length 90, where each frame corresponds to a motion frame window (sampled at 30fps). Frame Feature Vectors (**FFV**) are derived from the audio segment as defined in Figure 6.1. Each segment of **FFV** values is passed to a Gesture Diffusion Network as described in Figure 6.2. This model will output 90 frames of motion and the previous states from the Transformer-XL model as $\mathbf{W}$, which stores the long-term context of up to 180 previous frames.

The process is repeated for each following segment, each time *caching* the previous 180 hidden states. For example, when processing $\mathbf{FFV}_{271:360}$ with a memory length $m = 180$, the Gesture Diffusion Network only has the *cached* context of $\mathbf{W}_{90:260}$, and no longer retains context from predictions $\hat{\mathbf{x}}_{1:90}$. The length of cached context can change, even after training.

6.5 Experimental Setup

The Gesture Diffusion Network is trained to generate style-conditioned animation using the ZeroEGGs speech and motion dataset [42], described in Chapter 3.4.3. The data is partitioned into a training and test set as provided, containing 94 and 40 minutes of motion, respectively.

This section introduces the motion, speech and style conditioning representations before providing an overview of the training procedure and post-processing.

6.5.1 Motion Representation

The ZeroEGGs data is downsampled from 60fps to 30fps, and skeleton joint rotations are extracted to encode the pose, $\mathbf{x}$, at each frame. Joint rotations are a convenient pose representation as these fit into pre-existing 3D animation pipelines and can be re-targeted to a character mesh. Euler angles from ZeroEGGs are converted to the 6DOF rotations defined by Zhou et al. [159] for their suitability to deep learning tasks. The body pose at time n is defined as:

$$\mathbf{x}_n = [x_n, y_n, z_n, r_{j,1,n}, ..., r_{j,6,N}] \tag{6.2}$$

where x, y, z denote the global skeleton position and $r_{j,1:6,n}$ form rotations for each joint j in the 6D rotation representation. These values are standardised by removing the mean and scaling to unit variance calculated using the training data to bring these values to the same scale as the rotations. While joint positions are not predicted, they are used in the loss function defined in Section 6.5.4 and are calculated using Euler angles and pre-defined skeletal offsets.

6.5.2 Speech Representation

Audio features are extracted using the Problem Agnostic Speech Encoding (PASE+) [120]. PASE+ implicitly encodes multiple features, including MFCC, FBANKs and other speech-related information, including prosody and speech content. The released, pre-trained PASE+ model extracts an audio feature embedding of size 768 for each frame of motion. The weights of this model are frozen and not updated during training.

6.5.3 Style Conditioning Representation

Each motion sequence in the training data is acted according to a labelled style. The model is trained with these labels to generate gesture animation in a particular style at inference time. Style is represented using a learned vector embedding layer. The perceptually similar styles should be close in the latent space, and those dissimilar ones should be far apart. Interpolation can be performed in the latent space to transition between styles smoothly.

6.5.4 Training Procedure

At each denoising step the pose, represented by a vector of joint rotations $\hat{x}$ is predicted. Joint positions are evaluated from the rotations using Forward Kinematics to provide additional geometric constraints during training. The loss function comprises multiple terms including a L_1 loss on the rotations (L_r), positions (L_p), acceleration (L_a), velocity (L_v) and kinetic energy (L_{v^2}) of each joint. $\mathbf{x}_r$ and $\hat{\mathbf{x}}_r$ represent ground truth and predicted 6D rotations, respectively; $\mathbf{x}_p$ and $\hat{\mathbf{x}}_p$ to be positions in world space, the following loss function is used:

$$\begin{aligned} L_r &= \lambda_r L_1(\mathbf{x}_r, \hat{\mathbf{x}}_r) \\ L_p &= \lambda_p L_1(\mathbf{x}_p, \hat{\mathbf{x}}_p) \\ L_v &= \lambda_v L_1(f'(\mathbf{x}_p), f'(\hat{\mathbf{x}}_p)) \\ L_{v^2} &= \lambda_{v^2} L_1(f'(\mathbf{x}_p)^2, f'(\hat{\mathbf{x}}_p)^2) \\ L_a &= \lambda_a L_1(f''(\mathbf{x}_p), f''(\hat{\mathbf{x}}_p)) \\ L_c &= L_p + L_v + L_a + L_r + L_{v^2} \end{aligned} \tag{6.3}$$

Where f' and f'' are the first and second derivatives respectively. Each term has a λ weighting to ensure they are all within the same order of magnitude and to indicate importance.

The optimal parameter configuration was identified through a thorough parameter sweep (shown in Table 6.1). The Gesture Diffusion Network was trained for 3350 epochs using

Hyperparameter		Value
TransformerXL	Head Dimension	32
	Inner Dimension	4096
	Number Heads	32
	Number Layers	8
Embeddings	Feature Embedding	1024
	Pose Embedding	256
	Style Embedding	8
	Diffusion Step Embedding	1024
Training	Dropout	0.2
	Batch Size	32
	Learning Rate	0.00001
	λ_r	1
	λ_p	0.01
	λ_v, λ_a	0.5
	λ_{v^2}	0.2
Context	Segment Length	90 frames
	Memory Length	180 frames

Table 6.1 Training hyperparameters.

the AdamW [87] optimiser. The most successful results were achieved with a sequence segment size of 90 frames (3 seconds) and a memory length of 180 frames (6 seconds). This was determined by a combination of low Frèchet Gesture Distance (FGD) scores and the observed quality of the rendered predicted sequences.

6.5.5 Post-Processing

The raw model output can contain low levels of high-frequency noise that detracts from the overall realism of the motion. Following other work on motion synthesis [156, 158], a Savitzky-Golay Smoothing filter is applied to mitigate this using a window length of 9 and polynomial order of 2. The small window size and low polynomial order mean this filter provides minimal smoothing while retaining accurate beat gestures. Comparatively, the proposed method uses much less smoothing than another diffusion model from Zhang et al. [156], which requires a window size of 32 and polynomial order 4.

6.6 Evaluation

Fig. 6.4 Gesture Diffusion Network uses a novel transformer diffusion architecture for generating gestures from Speech. The model has a variable length context, and the Animation can be conditioned on style. The resulting Animation can be retargeted to rigs such as the MetaHuman [40] shown.

A comprehensive evaluation in which motion predicted from the Gesture Diffusion Network is compared against ground truth motion for the held-out test set. The effect of varying and interpolating styles and the model's ability to generalise to out-of-domain audio is also explored. A rendered example is shown in Figure 6.4, and further results can be found online[2].

6.6.1 Ground Truth Comparison

An attribute of diffusion models is their ability to generate multiple results for the same conditioning values. This is a desired attribute, but the predicted motion will not necessarily resemble the ground truth motion closely at every inference step. This is due to the one-to-many relationship between speech and gesture. Therefore, the *characteristics* of generated motion are evaluated and compared against the characteristics of corresponding ground truth sequences. Motion characteristics are beneficial to evaluate gesture as discussed in Chapter 4.11.1, particularly when comparing a stylistic aspect such as speaker identity or emotion.

[2] https://youtu.be/x6B8rAffUJ0

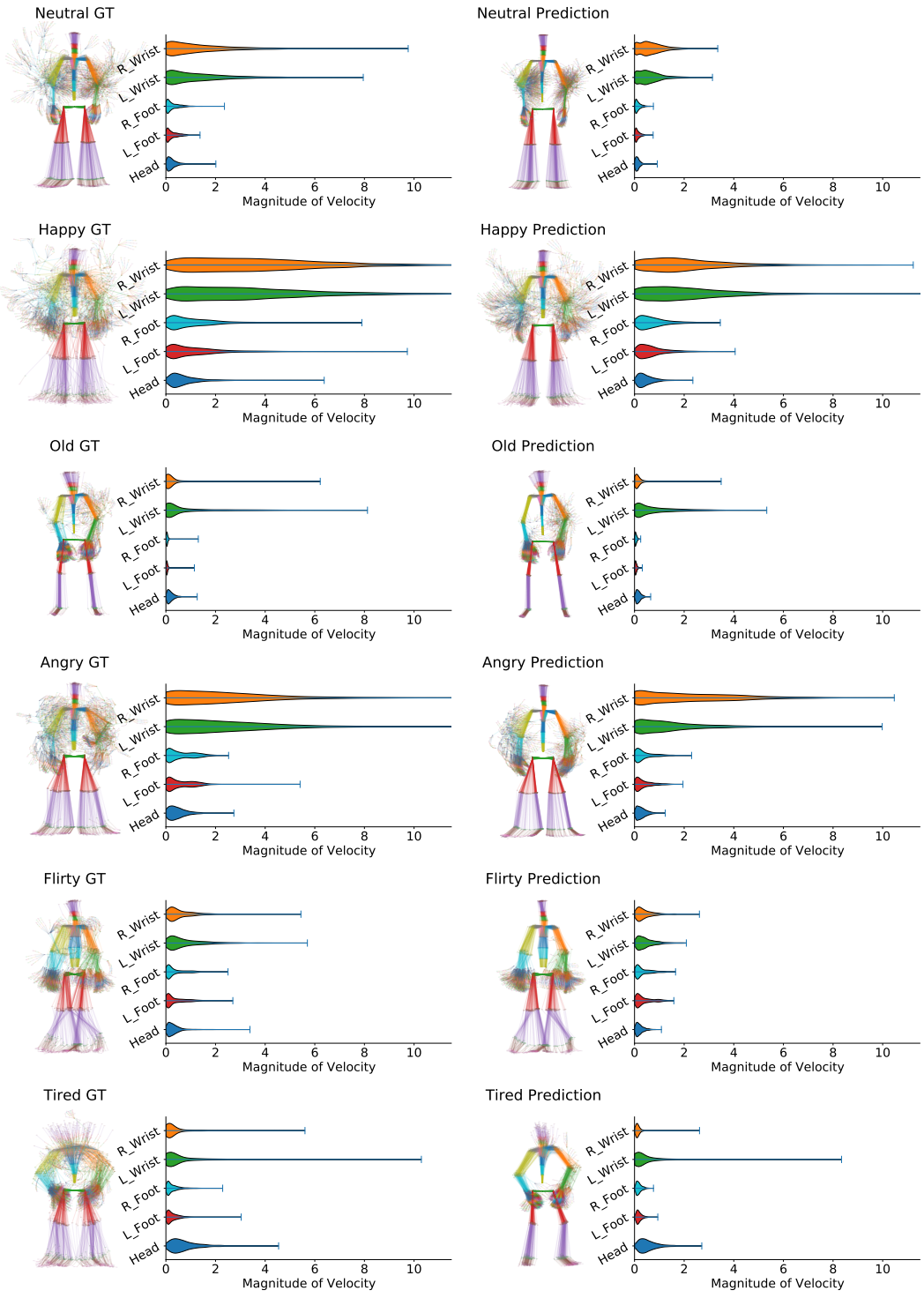


Fig. 6.5 A test sequence from 6 different style categories shown for ground truth (left) and sequences predicted from the Gesture Diffusion Network for the same audio (right) sampled every 1 second. Each example is shown with the distribution of the velocity magnitude for the wrists, feet and head to indicate the amount and speed of motion across the sequence.

The results of the GDN prediction are shown in Figure 6.5, where six different test sequences are compared. The style conditioning label is set to match the style of the test audio for each sequence. The full sequence of poses is displayed, sampled at 1-second intervals to show the extent and types of gesturing that are generated. Additionally, the magnitude of the velocity is computed for the head, the right and left wrists, and the feet, and it displays the distribution for both ground truth and predicted sequences. The poses in the predicted sequences closely resemble those of the respective ground truth values. The poses formed in each prediction are also particularly representative of their respective styles. For example, in the *Happy* sequence, the arms are highly active with a broader span compared to the *Old* or *Tired* sequences. The Gesture Diffusion Network can capture the nuances of each style, even if they are subtle. For example, during the *Old* sequence, the model generates animation with the actor's hunched back posture; in the *Tired* style, the model captures the flexing at the knees and hips.

The rate of movement is an essential characteristic of gesture, particularly when considering style. Observed from the plots in Figure 6.5, styles such as *Happy* and *Angry* are associated with faster motion with joint velocity values centred higher than other styles such as *Flirty* and *Tired*. This trend persists across the motion generated from the Gesture Diffusion Model, indicating that the dynamics of the generated Animation closely match that of the ground truth.

6.6.2 Style Conditioning

The impact of style on gesture generation is evaluated using the same audio input and sampling noise while varying the conditioning style. Figure 6.6 shows orthographic projections of the pose at each second of a sequence conditioned on different styles and a corresponding frame from the rendered sequence.

Fig. 6.6 Animation generated from the same audio and noise sample but conditioned on different styles. An orthographic projection of the pose at every second of the generated sequence is shown with a rendered frame from the same sequence below.

Evident characteristics of each style can be observed in Figure 6.6. *Happy* is particularly expressive, resembling motion similar to a *Neutral* sequence but with exaggerated, wider arm movement. *Angry* motion is also predicted with active arms but less low body motion. Additionally, the arms maintain a wide pose and do not cross the body, which is consistent with the ground truth. *Old* motion maintains a hunched-over posture with the hands resting on the actor's legs, only raising occasionally. *Speech* also has active arm motion. However, this is consistently raised instead of often lowering as in other styles. The results indicate that the Gesture Diffusion Network can produce gestural motion appropriate to the given style.

6.6.3 Generalisation to Out of Domain Audio

PASE+ features effectively encode the content of the Speech while being agnostic to the speaker and language [120]. Given the agnostic aspect of the features, audio from sources other than the ZeroEGGs dataset can be used to generate gesture. This is demonstrated using audio from the GENEA challenge 2022 data [154] and Multilingual LibriSpeech [115]. Rendered results are shown in the supplementary video.

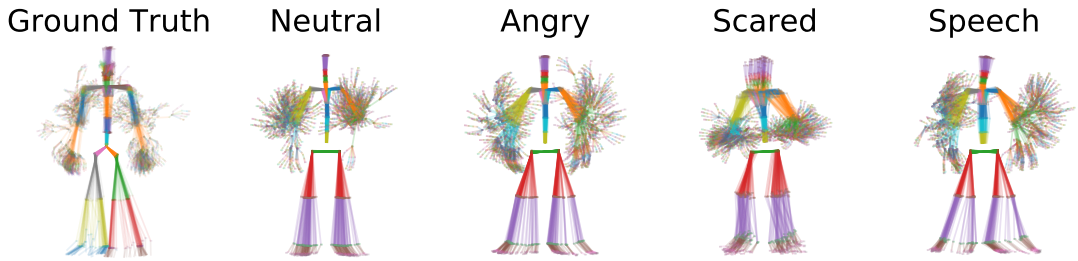


Fig. 6.7 Gesture generation using out-of-domain audio, conditioned on different styles. The ground truth sequence is shown on the left.

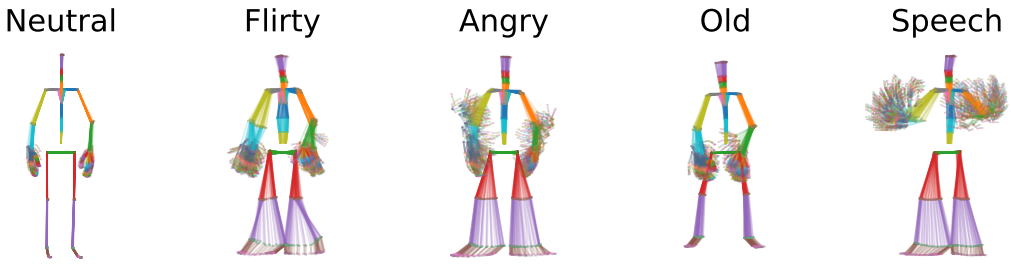


Fig. 6.8 Generated Animation for out-of-domain Polish Speech, conditioned on different styles.

An example of a GENEA sequence can be seen in Figure 6.7 with the ground truth motion of the actor. The GENEA audio contains muted sections where the authors remove periods of speech to preserve anonymity. Despite this, the Gesture Diffusion Network generates natural motion that preserves the conditioning style.

The model's ability to generate realistic animation for speech in different languages is demonstrated by generating animation for Polish speech, as shown in Figure 6.8. The method is agnostic to language, generates realistic animation appropriate to the given style, and synchronises well with the audio.

6.6.4 Style Interpolation

Linear interpolation can be performed between styles in the latent embedding space, which shows how the animation changes. Figure 6.9 shows animation for the same audio sequence, with the same noise for consistency between predictions. The style embedding value is varied

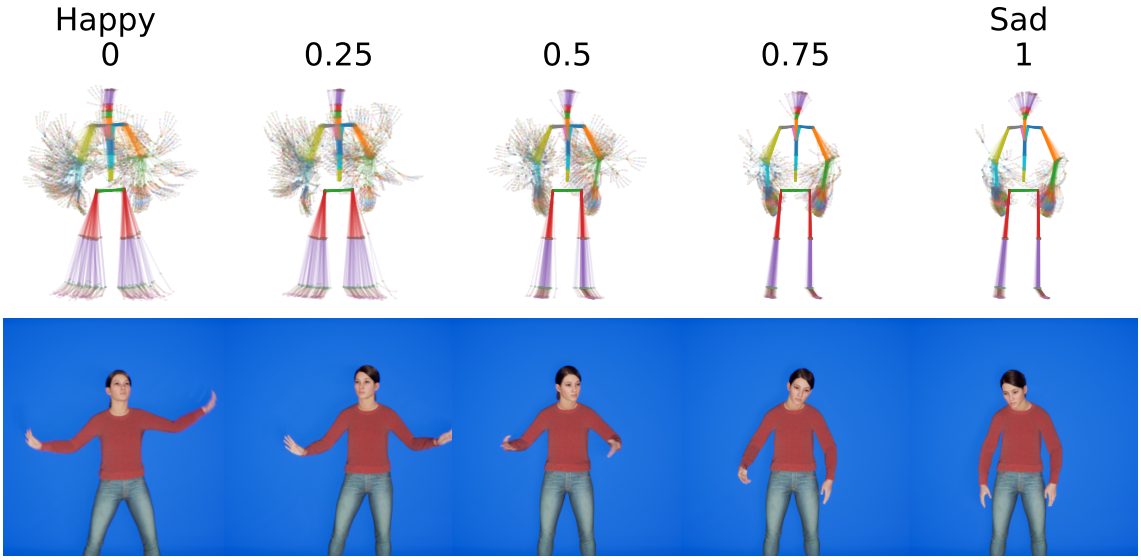


Fig. 6.9 Linearly interpolating between *Happy* and *Sad* in the embedding space generates Animation that gradually becomes more expressive.

by linearly interpolating between the *Happy* and *Sad* styles. The gradual shift from a highly active *Happy* sequence to a *Sad* sequence is evident. As the weight of the interpolation shifts towards *Sad*, the movement decreases an appropriate amount. Further examples can be found in the supplementary video[3].

6.7 Comparison to other methods

Results are compared using both objective and subjective measures against the baseline ZeroEGGs (ZE) [42] and StyleGestures (SG) [6] approaches as these are both state-of-the-art methods for each dataset used.

Gesture motion is computed for the test audio from the ZE system using the released source code and pre-trained model checkpoint[4]. To compare against StyleGestures, the Gesture Diffusion Network is retrained on the Trinity dataset [37] using the GENEA 2020

[3] https://youtu.be/x6B8rAffUJ0

[4] https://github.com/ubisoft/ubisoft-laforge-ZeroEGGS

challenge [74] train/test splits and the same hyperparameters as identified in Section 6.5.4 except for number of epochs which is reduced to 2400 to avoid overfitting the smaller dataset. The Trinity dataset does not have style labels, so the model is conditioned on audio alone for this task. The released source code[5] was used to train the SG model according to the parameters provided by the authors.

All test samples are used in each respective dataset for computing the objective measures and a random subset in the user study.

6.7.1 Objective Results

No single metric can effectively evaluate the quality of generated gestures. Instead, a combination of Frèchet Gesture Distance (FGD) [152, 13] and Beat Alignment (BA) [83, 84] scores have been used for their ability to reflect perceived realism and the alignment of the motion to the speech [7, 84, 152].

Frèchet Gesture Distance is a measure based on the Frèchet Inception Distance (FID) [54] and is commonly used to evaluate generative models. This measure indicates the similarity between the generated and ground truth poses but does not capture how well the generated examples temporally align with the audio. To address this, the Beat Alignment score is also reported. Originally introduced for dance synthesis [83], the Beat Alignment score has been adapted for evaluating speech gestures [84]. BA measures synchrony between gestures and audio using a Chamfer Distance between audio and gesture beats. Chapter 2.5.2 provides further details regarding these measures.

The Gesture Diffusion Network is compared both with and without smoothing (denoted as *GDN* and *GDN-NS* respectively) against ZE, SG and Ground Truth (GT) motion for each dataset. Results are summarised in Table 6.2.

[5] https://github.com/simonalexanderson/StyleGestures

Dataset	Trinity		ZeroEGGs	
Evaluation	FGD↓	BA↑	FGD↓	BA↑
GT	-	0.848	-	0.811
GDN-NS	**64.45**	**0.897**	35.23	**0.864**
GDN	64.72	0.895	**33.49**	0.855
SG	218.32	0.890	-	-
ZE	-	-	56.046	0.797

Table 6.2 Frèchet Gesture Distance (FGD, lower is better) [152] and Beat Alignment (BA, higher is better) [84] scores for each system calculated with respect to the ground truth test dataset. Computed over both the Trinity and ZeroEGGs datasets.

The GDN method, both with and without smoothing, outperforms the ZeroEGGs approach on FGD and BA. On the Trinity dataset, the BA scores are very similar across all systems, and the GDN and GDN-NS methods outperform SG according to FGD. Note that the output from the SG model is post-processed to reverse the pre-processing that was applied to the data before training.

One concern with applying a smoothing filter is that gestures can become less pronounced. Given the minimal amount of smoothing applied in the method, this did not have much of an impact. This is supported by the Beat Alignment scores, which only fell by 0.09 on the ZeroEGGs dataset and 0.02 on the Trinity dataset. The FGD score improved slightly when smoothing was applied when trained on ZeroEGGs, and a minimal change was observed on the Trinity dataset.

The Beat Alignment is higher for the GDN method and SG than for GT. This might be due to the model's lack of semantic understanding and reliance on the prosody embedded in the audio for generating gestures. These prosodic features relate to acoustic energy and pitch, which are commonly synchronised with beat gestures [147, 114].

6.7.2 User Study

A user study is presented to evaluate the perceived human likeness and appropriateness of the gestures generated by the Gesture Diffusion Network compared with other methods. The

relationships between motion, speech and style are analysed. Participants were hired through the prolific[6] platform with 30 participants in each experiment after removing any participants that failed attention checks. Participants were filtered to be fluent in English. This study uses a similar methodology to [7].

All test sequences for each method were rendered on a MetaHuman [40] with static facial animation as shown in Figure 6.10. A different MetaHuman character is used for the Trinity and ZeroEGGs studies to match the actor's voice appropriately.

To evaluate the Gesture Diffusion Network on the ZeroEGGs dataset, 35 random 10-second test clips are rendered spanning five examples of 7 different styles (*neutral, happy, old, sad, angry, pensive and scared*) for each of GT, ZE and GDN method. For the Trinity dataset, 35 random 10-second test clips are rendered for each GT, SG and GDN method. Methods are compared in a pairwise manner by presenting the participant with two side-by-side videos that were each generated for the same audio but with different systems. For each dataset, there are three system combinations to compare (*GT vs. GDN, GDN vs. ZE/SG* and *GT vs. ZE/SG*) and the order is flipped so that the videos are seen on both the left and right sides of the screen (35 clips $\times$ 3 combinations $\times$ 2 sides = 210 comparisons).

Each participant was presented with a subset of 30 pairwise comparisons consisting of 10 random pairs for each of the 3 system combinations, and they were asked to watch both of the two 10-second video clips. Figure 6.10 shows an example of the user study interface. The scoring methodology uses a merit system [109] where an answer is given a value of 2, 1 or 0 for clear preference, slight preference and no preference, respectively. A one-way ANOVA test with a post-hoc Tukey test was subsequently used for significance testing. Preference testing allows a win rate calculation where a win is assigned when there is an identified preference for a system, not including ties.

[6]https://www.prolific.co/

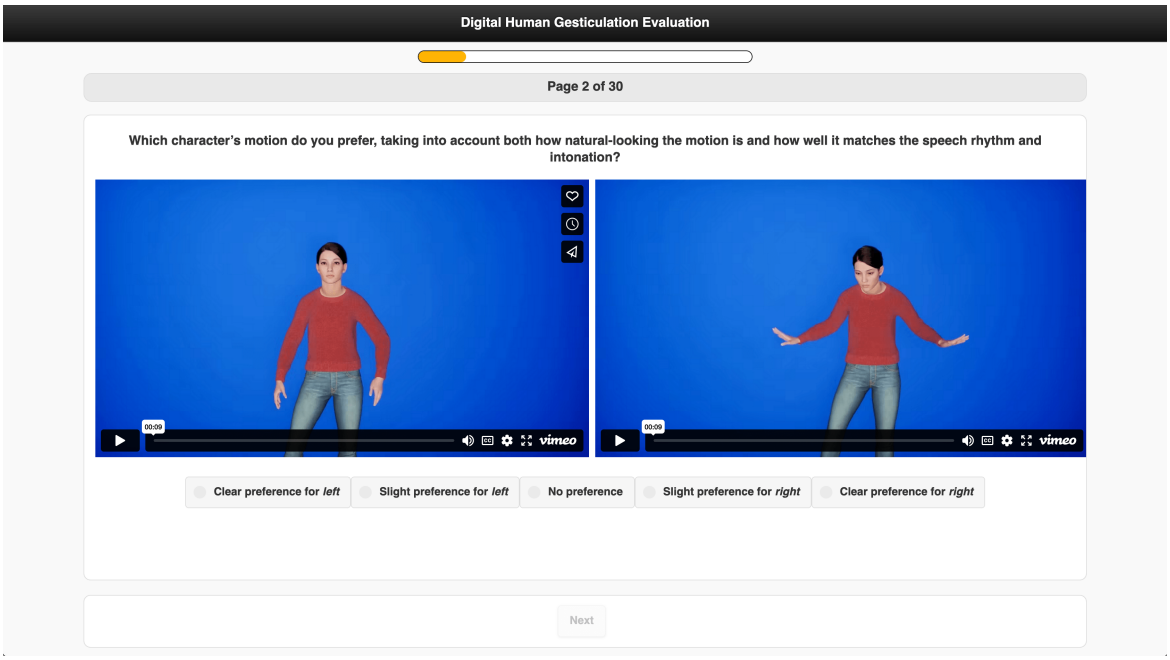


Fig. 6.10 Example user study question with answer options.

6.7.2.1 Perceived Motion Realism and Appropriateness to Speech

The overall motion preference and how well the gestures correspond to the speech are first analysed. This test uses the same question-and-answer options as Alexanderson et al. [7]. The question was posed as "Which character's motion do you prefer, taking into account both how natural-looking the motion is and how well it matches the speech rhythm and intonation?". The participants were asked to choose from the options {**Clear preference for *left***, **Slight preference for *left***, **No Preference**, **Slight preference for *right*** and **Clear preference for *right***}.

Table 6.3 summarises the results. Average merit scores are presented to indicate the relative performance score of each method. Win and tie rates are also provided to indicate model performance vs. the *GDN* method. This provides and insight into how often a method is preferred to the *GDN* and if there are many occasions when the performance is tied.

On the ZeroEGGs dataset (right column), the *GDN* model outperforms ZE in both merit score and win rate. The merit score indicates that animation generated using the Gesture

Dataset	Trinity			ZeroEGGs					
Evaluation	Realism/Appropriateness			Realism/Appropriateness			Style		
Measure	Merit Score	Win Rate	Tie Rate	Merit Score	Win Rate	Tie Rate	Merit Score	Win Rate	Tie Rate
GT	1.34±0.07	80.0%	8.4%	1.37± 0.06	72.8%	14.1%	0.89±0.07	50.7%	15.6%
GDN	**0.36±0.05**	-	-	**0.50 ± 0.06**	-	-	**0.68 ± 0.04**	-	-
SG	0.29±0.05	33.2%	25.6%	-	-	-	-	-	-
ZE	-	-	-	0.18 ±0.04	18.8%	27.8%	0.24 ± 0.06	14.4%	19.6%

Table 6.3 User study results. Merit scores [109] with 95% confidence intervals and win and tie rates for each method vs. the Gesture Diffusion Network (GDN). The highest merit scores for each experiment are written in bold. With the exception of the GT condition, the animation from the GDN was preferred in all cases, outperforming SG and ZE. Notably, on the ZeroEGGs dataset, animation from the GDN model is preferred over or tied with ground truth 27.2% and 49.4% of the time for realism/appropriateness and style matching, respectively.

Diffusion Network was preferred over ZeroEGGs ($p < 0.001$), while GT was considered better than both methods ($p < 0.001$). While the *GDN* method falls short of Ground Truth, the GT only had a win rate of 72.8%. This means that animation from the *GDN* model is preferred over or tied with ground truth 27.2% of the time. Comparatively, ZE had a win rate of 18.8% and a tie rate of 27.8% when compared to *GDN*, meaning that the *GDN* model has a winning preference over ZE 53.4% of the time. These results support the objective measures described in Section 6.7.1.

Next, the performance on the Trinity dataset (Table 6.3, left column) is evaluated. Ground Truth performs significantly better than both systems ($p < 0.001$). Although the GDN is rated higher than SG, this difference is not statistically significant ($p < 0.19$). However, SG is only preferred 33% of the time when compared to the GDN system. With a tie rate of 25.6%, this means that the GDN method has a winning preference 41.2% of the time. Note that the model hyperparameters were tuned on the ZeroEGGs dataset, and increased performance may be observed by re-configuring them using the Trinity data.

6.7.2.2 Perceived Style Appropriateness

To measure style preference participants are asked a similar question to Alexanderson et al. [7], posed as “Based on the body movements alone (disregarding the face), which of the two

clips looks most like **{STYLE}**?" where **STYLE** is a phrase that is representative of the conditioning style label, such as "an old person" or "happy". Audio was not permitted for this study, as one modality can affect the perception of the other [17, 7, 74, 60].

Results are shown in Table 6.3. Based on the merit score, GT animations were perceived as matching the style more closely than the GDN method ($p<0.001$). However, the Gesture Diffusion Network results were considered significantly better than ZE ($p<0.001$) at this task. Notably, participants only awarded a preference to GT 50.6% of the time, which means that for 49.4% of the test sequences, GDN was either indistinguishable or preferable to GT.

6.8 Effect of Memory Length

Memory length has a minimal effect over short isolated speech sequences, but as the sequence length grows, the importance of retaining contextual information also increases. Over a sequence of gesturing, there is a clear relationship between present and historic motion. For example, a behavioural or comfort motion, such as a weight shift from the left to right leg, rarely occurs multiple times over a period of 10 seconds in the data. However, when the Gesture Diffusion Network's historical context is limited to a few seconds, a weight shift occurred repeatedly in quick succession as the model had no knowledge of the previous activations.

An ablation study is presented to determine the effect of memory length on the realism of the generated animation. First, the impact of training models with varying memory lengths is evaluated and then the effect of varying the memory length after training. See the supplementary video[7] for rendered examples.

[7] https://youtu.be/x6B8rAffUJ0

6.8.1 Training Memory

Five model variants are trained using the hyperparameters shown in Table 6.1 and set the memory length to 1, 30, 90, 180 and 270 frames, respectively, so that the impact of memory length on performance can be isolated. When the memory length was set to 1, the generated animations lacked continuity and were not human-like. However, as the memory length increased, the animations became more natural and appropriate to the speech. A memory length of 180 frames (6 seconds) was determined to produce motion that was most contextually correct over long sequences.

For shorter memory lengths, gestures were observed to become out of phase with the speech. For example, arms lower to a rest position prematurely and raise unnaturally quickly to compensate and regain phase with the speech. This was noticeable in longer audio clips with short pauses in speech, for example, when the speaker is in thought between sentences. With only a small amount of previous context, it is difficult to know if the speaker is in an inactive state or if there is just a small break in the speech. With the memory set to 180 frames (6 seconds), the previous context retains the knowledge of speech occurring and keeps the arms raised appropriately. The ability of the GDN model to capture long-term context generates more natural animation.

Predicted sequences are compared against Ground Truth test sequences using the measures described in Section 6.7.1. A canonical correlation analysis (CCA) is additionally performed, which produces a score that indicates how correlated the two sequences are, with 1 being the most correlated. Table 6.4 summarises the results. Both CCA and FGD indicate that a single frame's concise historical context is insufficient. This is also supported by subjective analysis, where motion was very noisy and did not predict any natural human-like motion. Performance starts to plateau around the 90-180 frame memory length with only minor improvements gained between 90 and 180. There was a noticeable drop off in performance when memory was extended beyond 180, which may be due to the model complexity

reaching its limit for the data available. When increasing beyond 180, motion is still natural and smooth; however, the gesturing becomes less expressive.

Beat Alignment does not change much as the memory length is varied. However, an outlier is noted when the model is trained with a single frame of historical context. This is likely due to the noisy generated motion with sharp velocity changes, identified as *beats*, producing false positive gesture beat activations. This suggests that this measure should only be used in combination with other metrics, such as FGD or subjective human evaluations.

Memory Length (# frames)	**CCA ↑**	**FGD ↓**	**Beat Align ↑**
1	0.978	132.00	**0.915**
30	0.992	45.85	0.862
90	0.993	36.67	0.869
180	**0.995**	**35.23**	0.864
270	0.973	40.29	0.859

Table 6.4 Measuring the effect of varying the training memory length. Predicted sequences are compared against ground truth over the held-out test sequences. The best results in each column are written in bold. A training memory length of 180 frames (6 seconds) is optimal.

6.8.2 Memory at Test Time

The training memory length is fixed to 180 frames, and the context is altered up to 270 frames (9 seconds) to test the impact of changing context length after training. The increase in memory length at inference time from training is possible due to the Transformer-XL architecture using a learned relative positional encoding that is robust to a change in context memory length. Differences are compared between memory lengths of 1, 90, 180, 210 and 270.

Measures from Section 6.7.1 are again calculated together with CCA and present the results in Table 6.5. Reducing the memory length to 1 frame maximises the Beat Alignment score, but perceptually, the sequences appear noisy and temporally unstable. FGD scores indicate that reducing the number of memory frames from 180 to 90 at test time has little

Memory Length (# frames)	CCA ↑	FGD ↓	Beat Align ↑
1	0.915	139.89	**0.908**
90	0.995	**35.16**	0.859
180	0.995	35.23	0.864
210	**0.996**	37.66	0.861
270	**0.996**	40.65	0.866

Table 6.5 The effect of varying memory length at test time. Predicted sequences are compared against ground truth sequences on the held-out test set. The best results in each column are written in bold.

impact on performance. Although the CCA values continued to increase monotonically as the memory length increased, the improvement was marginal. It concludes that a memory length between 90 and 180 is reasonable for the data.

6.9 Discussion

This chapter presented the Gesture Diffusion Network for generating stylised gestures from speech using a Transformer-XL + Diffusion model that leverages long-term context from previously predicted poses. This approach Demonstrates effective style control and shows how styles can be interpolated and varied over a single animated sequence. Provides evidence of the effectiveness of using PASE+ features for speech-driven gesture generation and how they enable the model to *generalise* across different voices and languages. A particular benefit to this method is the ability to predict motion seamlessly without a limit on audio length.

Through recurrent state reuse, the model overcomes the common sequence length limitation with Transformer based models while retaining knowledge of a long-term context. Long-term context is important for gesture generation, enabling smooth and plausible transitions across segment boundaries. It also suppresses unnatural repetitive behaviours (e.g., quickly shifting weight from one leg to another). It allows the model to disambiguate be-

tween periods of inactivity and short pauses in the speech. An ablation study identified that a Gesture Diffusion Network trained with a memory length of 6 seconds was optimal for the ZeroEGGs dataset.

The new approach's effectiveness is proven through objective and subjective measures, outperforming the ZeroEGGs approach in both measures. By embedding style in a latent representation, generating animation sequences for the same audio input in multiple styles is possible. This latent representation also allows smooth interpolation between styles. Using PASE+ feature embeddings for audio allows the trained model to effectively generalise to unseen speakers and languages.

Chapter 7

Towards Dyadic Contribution

Contributing Publications

- Windle, J., Matthews, I., Milner, B., and Taylor, S. (2023). The uea digital humans entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 802–810, New York, NY, USA. Association for Computing Machinery

7.1 Introduction

Speech-driven gesture generation has predominantly focused on estimating motion for monadic speech input of a main-agent, with no knowledge of interlocutor speech and no concept of interaction. This chapter instead focuses on generating gestures in a dyadic setting – predicting a main-agent's motion from the speech of both the main-agent itself and also the speech of the interlocutor.

Most speech-to-gesture approaches focus on monadic, non-verbal communication during speech. This is sometimes due to the dataset being a monologue of a single actor, however, many approaches such as those participating in the Generation and Evaluation of Non-verbal

Behaviour for Embodied Agent (GENEA) Challenge 2022 [154] use a single speakers speech, even when a second speaker is present in the data capture process. While these non-verbal roles are an important aspect of communication, these approaches ignore the effectiveness and importance of back-channel communication. Back-channel communication is the act of incorporating vocalisations, facial expressions, gaze, and gestures—involving responsive feedback to the speaker [98]. This is typically performed by the member of the conversation in a *listening* state to provide feedback information to the main *speaker*, for example, agreement, disagreement, or confusion. These back-channels aid conversation; for example, should the *listener* provide a back-channel that suggests confusion, the *speaker* may wish to expand or explain a concept again, which influences conversational flow.

This chapter explores the use of a model where the motion of a *main-agent* is predicted using the speech of both the main-agent itself and also the speech of the interlocutor. This model is an adapted Transformer-XL [29] architecture to generate smooth, contextually and temporally coherent motion that can adapt to varying lengths of historical context. Specifically, the Transformer-XL model is extended to provide cross-attention with the interlocutor's speech to impart knowledge of both speakers into the prediction.

The Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) Challenge is a particular motivation for this chapter which is described in more detail in Section 2.4. In this chapter, the GENEA Challenge 2023 dataset is described to introduce the dyadic, *main-agent* and *interlocutor* interaction. The Transformer-XL [29] model is described, and the dyadic contribution adaptation is introduced for the interlocutor to influence the generation of gestures. The method was evaluated subjectively as part of the GENEA 2023 Challenge and compared to other methods developed using the same dataset, two baselines, and the natural ground truth motion. These results and findings are then discussed. Video examples and code are available [1].

[1] github.com/JonathanPWindle/uea-dh-genea23

7.2 GENEA23 Data

The GENEA 2023 challenge data [76] is derived from the Talking With Hands dataset [78] which is described in more detail in Chapter 4.3. This data includes dyadic conversations between a main-agent and interlocutor and consists of high-quality 30fps mocap data in Biovision Hierarchical (BVH) format, with corresponding speech audio and text transcripts. This chapter aims to generate the main-agent motion conditioned on both main-agent and interlocutor speech. Main-agent and interlocutor speech data is processed using the same approach, using all available modalities; motion, speech, transcription and speaker identity.

7.2.1 Motion

Euler angles are required for evaluation purposes within the GENEA Challenge and are a convenient representation supported by many available 3D animation pipelines. Despite this, Euler angles are discontinuous and difficult for neural networks to learn [159]. Rotations are converted to the 6D rotation representation presented by Zhou et al. [159] for their suitability to deep learning tasks. Global skeleton position is encoded using three x, y, z values. All values are standardised by subtracting the mean and dividing by the variance computed from the training data.

Each speaker identity in the dataset has a skeleton with differing bone lengths. Additionally, per-frame joint offsets are present in the data, possibly to account for bone stretching in the data capture. However, analysis of these joint offset values revealed very low variance, and setting them to a pre-defined fixed value for all frames did not impact visual performance. One set of bone lengths and offsets per speaker is used to simplify the training pipeline. A sample is randomly selected, corresponding to each identity, and the bone lengths and offsets are fixed accordingly using the first data frame. Joint positions can then be computed using

the joint angles (measured or predicted) and pre-defined speaker-specific bone measurements using Forward Kinematics.

7.2.2 Speech

Audio features are extracted using the problem-agnostic speech encoder (PASE+) [120]. PASE+ is a feature embedding learned using a multi-task learning approach to solve 12 regression tasks to encode important speech characteristics. These 12 tasks include estimating MFCCs, FBANKs and other speech-related information, including prosody and speech content. The particular benefits of these features are described in Section 3.2.2.1.

PASE+ requires audio to be sampled at 16KHz, so band-sinc filtering is used to reduce the audio sample rate from 42KHz to 16KHz. The released, pre-trained PASE+ model is used to extract audio feature embeddings of size 768, which represents a 33ms window of audio, to align with the 30 fps motion. The weights for this model are not updated during training.

Word-level features are also extracted from the text transcriptions using the FastText word embedding described by Bojanowski et al. [16] using the pre-trained model released by Mikolov et al. [97]. A word embedding is extracted for each spoken word, and each embedding is aligned to a 33ms motion frame. A vector of zero values is passed if no word is spoken at a given frame. When a word is spoken across multiple frames, the vector is repeated for the appropriate number of frames.

7.3 Transformer-XL Architecture

Many speech-to-motion deep learning techniques are built upon recurrent models, such as bi-directional Long Short-Term Memory models (LSTMs) [37, 131, 51]. Transformer architectures are growing traction in favour of LSTM models in sequence-based AI, with

sequence-based motion prediction models already making use of them [134, 65, 14, 88]. Transformer models do not have a concept of temporal position but can effectively model temporal information, often using a sinusoidal position embedding, which is added to the input.

Transformers rely on attention mechanisms which inform the network which parts of data to focus on [138]. In self-attention, the mechanism is applied to the input sequence to find which elements within the same sequence may relate to each other and which are key to focus on. Conversely, cross-attention is computed for one input source in relation to a separate input source, calculating which elements from one sequence may relate and be important to focus on in another sequence.

To perform sequence-to-sequence generation using a vanilla transformer as defined in Vaswani et al. [138], a sequence is processed over a sliding window with a one-frame stride. For each window of input, one frame of output is generated. This is computationally expensive, and window size is limited by the longest input sequence seen during training. As the sequence length increases, the size of the self-attention mechanism also grows exponentially, leading to memory and computational limitations.

The Transformer-XL architecture [29] differs from the traditional transformer architecture in two key ways: 1) Attention is calculated conditioned on the previous context, and 2) the positional encoding uses a learned relative embedding. The Transformer-XL architecture allows for extended attention beyond a fixed length by using segment-level recurrence with state reuse, allowing the alteration of context length. Therefore, the Transformer-XL can be trained efficiently on small segment lengths while retaining historical influence through state reuse. The historic context length can vary, so the Transformer-XL introduces a learned, *relative* positional encoding scheme. Due to its improved ability for modelling sequences, this work adapts the Transformer-XL architecture for dyadic gesture generation.

7.4 Dyadic Contribution Method

The Transformer-XL [29] architecture is adapted for dyadic, speech-driven gesture generation. Specifically, this architecture is modified to use both self and cross-attention. This model is referred to as ***X-Att-XL***. The advantage of the Transformer-XL architecture in this task is that it models the longer-term relationship between speech and gesture for input of any duration.

The feature extraction process shown in Figure 7.1, is used to generate a feature vector $\mathbf{X}$ of length w for both the main-agent and interlocutor to get $\mathbf{X}^{\mathbf{ma}}$ and $\mathbf{X}^{\mathbf{in}}$. These features are then passed to the *X-Att-XL* model as shown in the overview Figure 7.2 where they are processed using a number of *Self-Attention Blocks* and *Cross-Attention Blocks*.

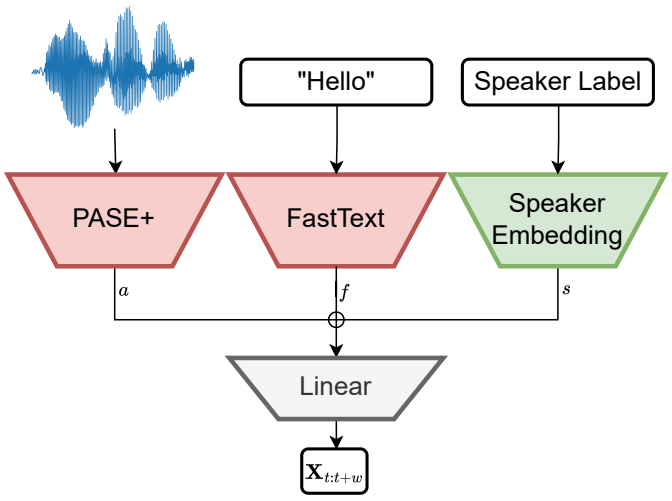


Fig. 7.1 Outline of the data processing pipeline. The process takes as input w frames starting at frame t of speech audio, text transcript and a speaker identity label to generate a feature vector $\mathbf{X}$. Pre-trained models are used for the audio and text inputs. Red box defines frozen weights.

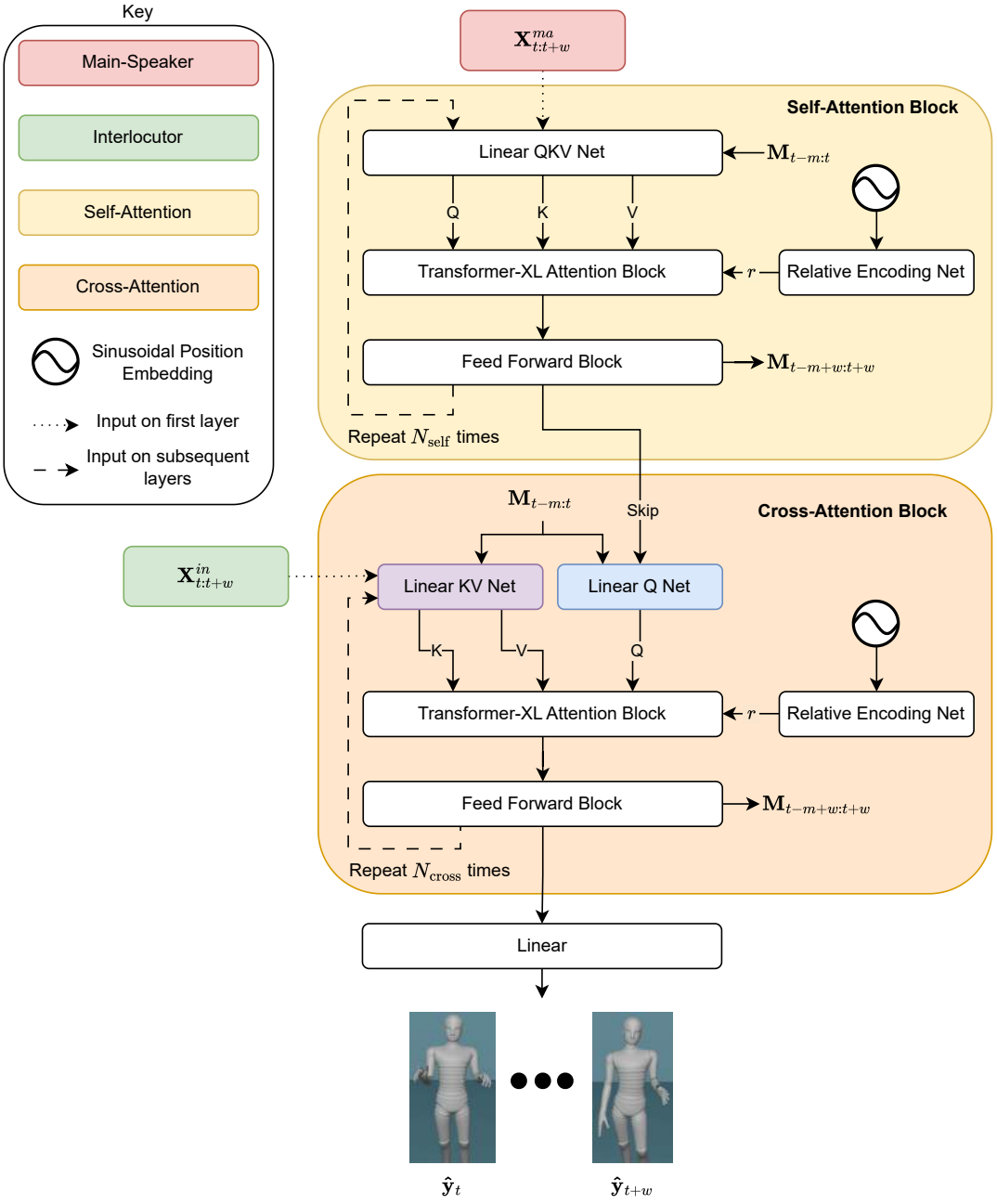


Fig. 7.2 Outline of the proposed *X-Att-XL* dyadic prediction model which takes as input, w motion frames worth of encoded conditioning information starting at time t and predicts w frames of body motion. This shows a self-attention block and cross-attention block, where Q, K, V vectors are extracted using main-agent or interlocutor speech according to the attention type conditioned on previous m number of hidden states $\mathbf{M}$. These vectors are passed to the Transformer-XL attention block to calculate attention before being fed into a feed-forward block. A final linear layer predicts w poses $\hat{\mathbf{y}}_{t:t+w}$.

7.4.1 Data Representation

Input is segmented into non-overlapping segments of length w frames. For each segment, an input feature vector $\mathbf{X}$ is generated and used to predict $\mathbf{Y}$, a sequence of poses of length w. The model is called for each non-overlapping w-length frame feature vector $\mathbf{X}$. Therefore, a speech sequence of length T is called $\lceil \frac{T}{w} \rceil$ times.

For each segment, audio (PASE+) features $\mathbf{a}_{t:t+w}$, and text (FastText) features $\mathbf{f}_{t:t+w}$ are extracted as described in Section 7.2.2, where t represents the start frame of a segment of length w. For each utterance, a speaker label is also provided. This is a unique ID which is passed to a learned embedding layer. The embedding layer acts as a lookup table for learned feature embeddings that are representative of each speaker style. The trainable weights ensure that two speakers with similar gesture styles are close in the latent embedding space, and conversely, those with different gesturing styles are far apart.

Each modality is extracted and concatenated into a single feature vector $\mathbf{X}$ as shown in Figure 7.1. Feature vectors for both the main-agent and the interlocutor are extracted in the same way using the same learned weights. This is because a speaker may appear as the main-agent in some sequences and the interlocutor in others.

7.4.2 Self-Attention

As shown in Figure 7.2, features from the main-agent are processed using a self-attention block. The attention score is defined in Vaswani et al. [138] as:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{7.1}$$

where Query Q, Key K, and Value V are all vectors, queries and keys of dimension d_k, and values of dimension d_v. These vectors are often linear projections of an input vector into their respective dimensions d.

When calculating attention scores in the Transformer-XL model, historic context is included using segment-level recurrence with state reuse. This is achieved by caching previous hidden state sequences, which can be used when processing future segments. When no historic context is present at the start of the speech sequence, *X-Att-XL* extracts Q, K and V vectors from the *main-agent* inputs alone. The historic context from processed segments $\mathbf{M}$ of length m is cached as each segment is processed. Q, K and V vectors are then extracted from the subsequent inputs, conditioned on previous context. This process is completed using a Linear QKV Net shown in Figure 7.2 which is a single linear layer.

Transformer models do not have inherent knowledge of positional order. To ensure temporal coherency, a positional encoding is often added to the input vectors to inject some position context to the model. As the Transformer-XL architecture can have varying lengths of historic context and is not constrained to a maximum length, a learned relative position encoding r is instead utilised. The learned relative encoding is from a single linear layer and takes a sinusoidal position embedding for the full length of context, that is the sum of both memory length available and the query length. Rather than injecting the temporal information to the input before calculating Q, K and V, which is the approach used in Vaswani et al. [138], the Transformer-XL inputs this information after these vectors have been extracted at the time of calculating the attention score.

Using Q, K and V in conjunction with the relative position encoding r, the *Transformer-XL attention block* is used to calculate attention vectors. As Figure 7.2 shows, these attention vectors are then passed to a Feed Forward Block which comprises of two Linear layers, with a ReLU activation on the first output and dropout applied to both.

Each self-attention block has multiple attention heads, each aiming to extract different attention features and a self-attention block is repeated N_{self} times, with each layer feeding its output to the next. Memory values $\mathbf{M}$ are persisted on a per-layer basis and therefore hidden

states are specific to each self-attention block. The length of this memory m can be altered during training and evaluation.

7.4.3 Cross-Attention

While it is reasonable to assume the main-agent speech is driving the majority of the gestures, the interlocutor can also influence the motion of the agent indicating turn taking and backchannel communication. For example, the main-agent might nod to show agreement or understanding when the interlocutor is speaking. Therefore, the main source of information driving the motion is from the main agent's speech, but also includes the interlocutor's speech. The Transformer-XL is adapted to not only compute self-attention over the main-agent inputs, but to also utilise cross-attention from the *interlocutor* while maintaining segment-level recurrence and relative position encoding. This cross-attention block is shown in Figure 7.2.

Cross-attention is an attention mechanism where the Query Q is extracted from the input source and the Key K and Value V are extracted from an external input element. The introduced cross-attention block uses a similar approach as the *self-attention block* defined in Section 7.4.2, but instead has two separate networks to process the inputs; one to extract Q from the *main-agent* self-attention encoding and one to extract K and V derived from the *interlocutor* speech. For each layer of cross-attention blocks, the input to the Q net is a skip connection from the output of the self-attention encoder and therefore remains the same input for all *cross-attention blocks*. The input to the KV net in the first iteration is the interlocutor feature vectors (described in Section 7.4.1), and the output from a cross-attention block thereafter.

The output from the cross-attention block is then passed to a single linear layer which predicts $\mathbf{Y}$, the standardised 6D rotations of each joint and the global position of the skeleton.

7.4.4 Training Procedure

For each segment of speech of length w, the pose represented by a vector of joint rotations $\hat{\mathbf{Y}}$ of length w is predicted. In motion synthesis it is common to include both geometric and temporal constraints in the loss function to ensure that the model generates output that is both geometrically and dynamically plausible [134, 145, 42]. The loss function L_c comprises multiple terms including a L_1 loss on the rotations (L_r), positions (L_p), velocity (L_v), acceleration (L_a) and kinetic energy (L_{v^2}) of each joint. $\mathbf{y_r}$ and $\hat{\mathbf{y}}_\mathbf{r}$ denote natural mocap and predicted 6D rotations, respectively and $\mathbf{y_p}$ and $\hat{\mathbf{y}}_\mathbf{p}$ to be positions in world space computed using Forward Kinematics given the predicted joint angles and the pre-defined speaker-specific bone lengths, the following loss function is used:

$$
\begin{aligned}
L_r &= L_1(\mathbf{y_r}, \hat{\mathbf{y}}_\mathbf{r}) \\
L_p &= L_1(\mathbf{y_p}, \hat{\mathbf{y}}_\mathbf{p}) \\
L_v &= L_1(f'(\mathbf{y_p}), f'(\hat{\mathbf{y}}_\mathbf{p})) \\
L_{v^2} &= L_1(f'(\mathbf{y_p})^2, f'(\hat{\mathbf{y}}_\mathbf{p})^2) \\
L_a &= L_1(f''(\mathbf{y_p}), f''(\hat{\mathbf{y}}_\mathbf{p})) \\
L_c &= \lambda_p L_p + \lambda_v L_v + \lambda_a L_a + \lambda_r L_r + \lambda_{v^2} L_{v^2}
\end{aligned}
\tag{7.2}
$$

Where f' and f'' are the first and second derivatives respectively. Each term has a λ weighting to control the importance of each term in the loss. These values vary in orders of magnitude due to positions and rotation values being in differing orders themselves. This ensures that all terms are within the same order of magnitude.

Table 7.1 summarises the parameters used, optimised using a random grid search parameter sweep. These settings were chosen using a combination of low validation loss values and observed quality of the predicted validation sequences. The *X-Att-XL* model is trained

Hyperparameter		Value
TransformerXL	Head Dimension	32
	Number Heads	32
	Self-Attention Layers (N_{self})	6
	Cross-Attention Layers (N_{cross})	2
Feed Forward Block	Dropout	0.2
	Hidden Size	4096
Embeddings	Feature Embedding	1024
	Speaker Embedding	8
Training	Batch Size	32
	Learning Rate	0.00001
	λ_r	1
	λ_p	0.01
	λ_v, λ_a	0.5
	λ_{v^2}	0.2
Context	Segment Length (w)	90 frames
	Memory Length (m)	180 frames

Table 7.1 Training hyperparameters.

for 1770 epochs using the AdamW [87] optimiser and found that a segment length w of 90 frames and memory length m of 180 frames was optimal. The *Feed Forward Blocks* used in both self and cross-attention layers are comprised using the same topology and size.

7.5 Results

The animation generated from the *X-Att-XL* model is smooth and temporally coherent without jitter or sudden shifts in motion while maintaining gesture beats in time with speech. Performance is first evaluated empirically and then subjectively. First, the model performance of beat gestures is described, followed by a comparison to the natural motion that is performed. Performance is then evaluated subjectively as part of the GENEA Challenge.

7.5.1 Beat Gestures

The *X-Att-XL* model appears to reliably and realistically animate beat gestures. Beat gestures are simple and fast movements of the hands and have a close relationship to prosodic activity, such as acoustic energy and pitch [147, 114]. The PASE+ model used for encoding audio in

the *X-Att-XL* system was trained to estimate prosodic features as one of its downstream tasks, making the derived audio features particularly suitable for animating beat gestures.

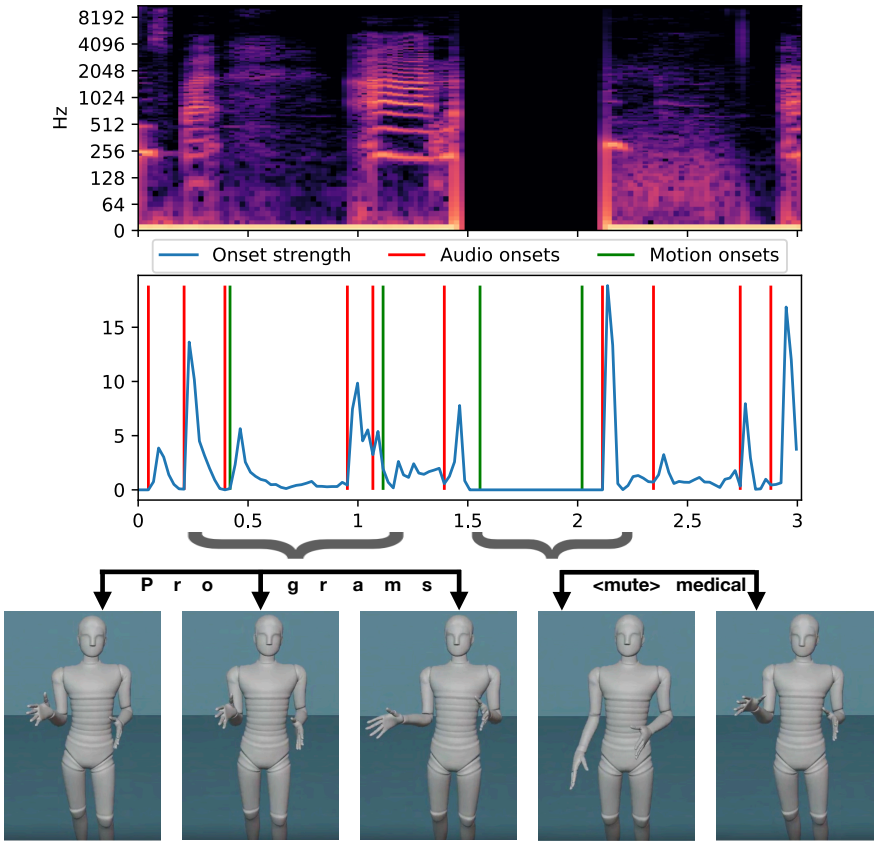


Fig. 7.3 Generated gestures for given audio beats. Using a 3s audio clip from the test dataset, the audio spectrogram is shown, as well as aligned audio beat onsets and their corresponding onset strengths as well as motion gesture onset detection of the right wrist using the method of beat detection defined in Liu et al. [84]. During the syllable utterance "pro", it shows the speaker moves their right hand hand from right to left, and as the stressed syllable "grams" is spoken, the hand begins to move left to right. When there is silence, the arms begin to rest and again gesture in the next utterance.

Gestures are not expected to occur during every audio beat, but they should synchronise with the speech when they happen. Using the motion and audio beat extraction method used in the beat align score calculation presented in Liu et al. [84] and discussed in Section 2.5.2.2, the onset of audio beats and motion gestures over time can be visualised. Figure 7.3 shows two well-timed gestures for a 3-second audio clip. The utterance of "programs" shows a beat gesture where during the syllable utterance "pro", the speaker moves their right hand

from right to left, and as the stressed syllable “grams” is spoken, the hand begins to change velocity and move from left to right. An example of muted speech is also shown where the *X-Att-XL* model continues to perform well. As there is no speech, there is little to inform gesture; it shows the right arm drops to the side, and the left arm lowers slightly. However, as the speech begins again, both arms raise in time with the speech.

7.5.2 Natural Motion Comparison

Despite the one-to-many mapping of speech and gesture, the characteristics of motion will be similar. Comparing predicted motion to natural motion determines whether the generative model captures the general motion characteristics and speaker-specific traits. Reviewing the poses formed and velocities provides an overview of how the model performs against natural motion.

Results are compared between six predicted sequences and the corresponding ground truth natural motion in Figure 7.4. Each sequence of poses sampled at 1-second intervals is plotted and overlayed to show the extent and types of generated gesturing. Additionally, the magnitude of the velocity for the head and the right and left wrists and feet are calculated, and the distribution for both ground truth and predicted sequences is displayed. The velocity magnitude describes the movement characteristics, which should closely match the ground truth. Each distribution should also indicate the amount of motion from each joint. Both the wrist and head joints are important factors in gesture. These joints are chosen because they are at the end of the kinematic chain and, therefore, display the most velocity change in relation to others.

Figure 7.4 shows that the predicted sequences closely match the style of the ground truth motion. The span of motion is slightly dampened compared to some ground truth sequences; however, these are in the rare, extreme gesture space, which does not occur often in the

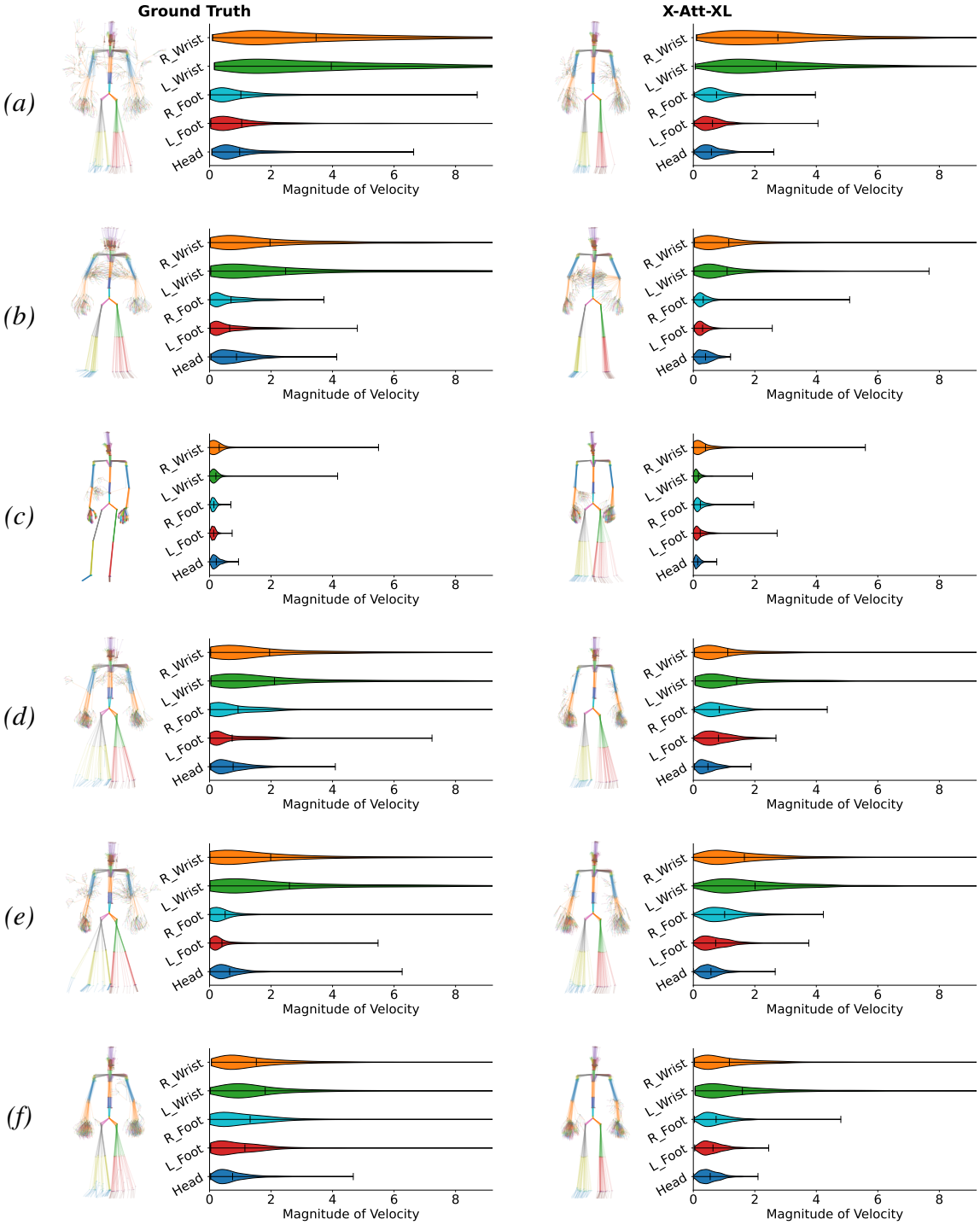


Fig. 7.4 A test sequence from 6 different style categories shown for ground truth (left) and *X-Att-XL* (right) for the same audio sampled every 1 second. Each example is shown with the distribution of the velocity magnitude for the wrists, feet and head to indicate the amount and speed of motion across the sequence.

ground truth data. In each sequence, the magnitude of velocities is closely related to the ground truth, with similar distributions present for each joint.

One aspect of the *X-Att-XL* model is the ability to model each speaker's characteristics. Rows *a*, *b* and *c* in Figure 7.4 represent different speakers in the dataset. It is clear from both the speaker's stance and range of motion, as well as the distribution of velocities, that the speaker's style has been adequately learned. Row *b* shows a speaker with wide arms and legs, with both arms moving up and down a similar amount on each side, which is also present in the predicted motion. Row *c* shows a speaker performing very little motion in the ground truth, with only the right arm making any gestures. While the predicted motion shows slightly more leg movement, the rest of the body is very similar, with only the right arm used for gestures.

A difference between natural mocap motion and the *X-Att-XL* generated animation is that the latter does not exhibit sporadic, non-speech related motion such as self-adaptor traits. Self-adaptors are movements that typically include self-touch, such as scratching of the neck, clasping at an elbow, adjusting hair or interlocking fingers [102]. Despite the indirect relationship between these behaviours and speech, these traits are linked to the perceived emotional stability of an agent [102] and may influence perceived human-likeness.

7.5.3 Back-Channelling

Subtle but important influences were observed in the results. There is evidence of back-channel communication, such as head nodding and gesture turn-taking. These are sporadic and difficult to show in static figures. However, these are evidenced in the publicly available video[2] associated with this work.

[2] github.com/JonathanPWindle/uea-dh-genea23

7.5.4 User Study Results

The *X-Att-XL* approach is evaluated in conjunction with the GENEA Challenge 2023 [76] which is discussed in Section 2.4. Each challenge participant submitted 70 BVH files for the main-agent motion generated using the main-agent's and interlocutor's speech for each interaction. Motion is rendered on the same character for comparison using these submitted BVH files. There are three studies of interest in this challenge; human likeness, appropriateness to speech and appropriate to interlocutor.

Similar to Section 5.5, to effectively compare synthesised gesture, the GENEA Challenge compared the submitted synthesised motion to both natural motion and pre-existing baseline systems. Each motion condition was assigned a three-letter ID following a set structure:

N/B/S - Representing Natural/Baseline/Submission respectively.

A-L - Representing a unique ID.

For example, **NA** denotes the natural motion of the mocap sequences, and the ID which will be used in Figures and Tables throughout to represent the proposed *X-Att-XL* is **SJ**. Baselines slightly change the format where **BD** and **BM** are baseline systems in a dyadic and monadic setting, respectively.

7.5.5 Human Likeness

This user study aims to evaluate whether the synthesised motion looks like the motion of an actual human, independent of the speech. The evaluation method is similar to Section 5.5.1, slightly changing the question presentation. The video stimuli audio is removed to ensure the measure is for human-likeness alone and decoupled from appropriateness. A Human Evaluation of Multiple Videos in Parallel (HEMVIP) [61] method was used to measure this. For this question, multiple motion examples are presented in parallel, and the subject is asked to assign a rating to each one. Each question asked *"Please indicate on a sliding scale how*

human-like the gesture motion appears" while being presented with eight video stimuli to be rated on a scale from 0 (worst) to 100 (best) by adjusting an individual GUI slider for each video. This rating allows for pairwise statistical tests and produces a median rating for each condition.

Summary statistics (median, mean) are shown in Table 7.2 and significance comparisons are provided in Figure 7.5. The proposed *X-Att-XL* system (**SJ**) was evaluated to be the third highest ranking out of the 14 generative methods with regard to mean and median human likeness score. Figure 7.5 shows only **NA**, **SG**, and **SF** are significantly better than the *X-Att-XL*. The *X-Att-XL* system scores significantly higher than nine other systems, including both baseline systems.

Condition	Human-likeness	
	Median	Mean
NA	$71 \in [70, 71]$	68.4±1.0
SG	$69 \in [67, 70]$	65.6±1.4
SF	$65 \in [64, 67]$	63.6±1.3
SJ	$51 \in [50, 53]$	51.8±1.3
SL	$51 \in [50, 51]$	50.6±1.3
SE	$50 \in [49, 51]$	50.9±1.3
SH	$46 \in [44, 49]$	45.1±1.5
BD	$46 \in [43, 47]$	45.3±1.4
SD	$45 \in [43, 47]$	44.7±1.3
BM	$43 \in [42, 45]$	42.9±1.3
SI	$40 \in [39, 43]$	41.4±1.4
SK	$37 \in [35, 40]$	40.2±1.5
SA	$30 \in [29, 31]$	32.0±1.3
SB	$24 \in [23, 27]$	27.4±1.3
SC	$9 \in [9, 9]$	11.6±0.9

Table 7.2 Summary statistics of user-study ratings from the human-likeness study, with confidence intervals at the level $\alpha = 0.05$. Conditions are ordered by decreasing sample median rating. *X-Att-XL* model results are highlighted in pink. Table and caption from [76].

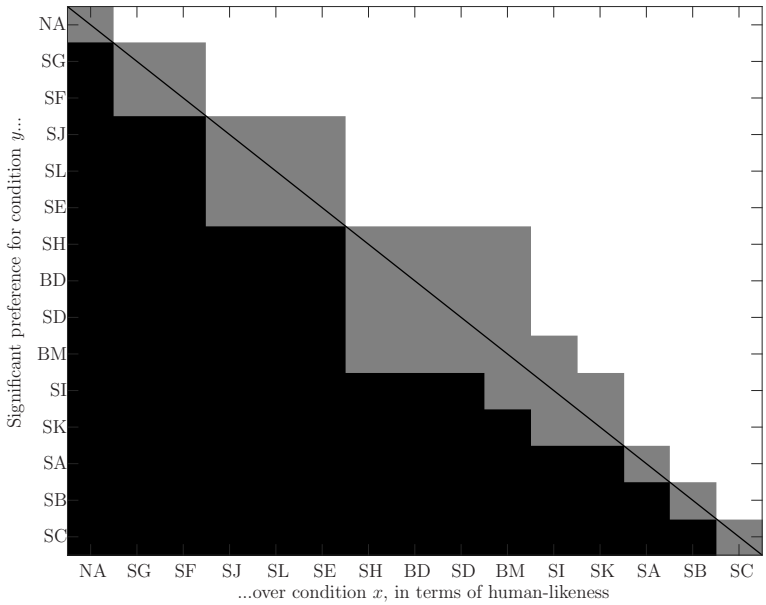


Fig. 7.5 Significance of pairwise differences between conditions in human-likeness study. White means that the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions are listed in the same order as in Table 7.2. Figure and caption from [76].

7.5.6 Speech Appropriateness

The appropriateness evaluation is designed to assess the appropriateness separate from the motion's intrinsic human likeness. It aims to determine whether the gestures produced are appropriate to the speech, given the timing and meaning of both. To measure the appropriateness of gestures to speech, participants were asked to view two videos and answer, "Which character's motion matches the speech better, both in terms of rhythm and intonation and in terms of meaning?". Both video stimuli are from the same condition and thus ensure the same motion quality, but one matches the speech, and the other is mismatched, generated from an unrelated speech sequence. Five response options were available, namely "Left is clearly better", "Left is slightly better", "They are equal", "Right is slightly better", and "Right is clearly better". Each answer is assigned a value of -2, -1, 0, 1, 2 where a negative

value is given for a preference for mismatched motion and a positive value for a preference for matched motion.

Condition	MAS	Pref. matched	Raw response count 2	1	0	−1	−2	Sum
NA	0.81±0.06	73.6%	755	452	185	217	157	1766
SG	0.39±0.07	61.8%	531	486	201	330	259	1807
SJ	0.27±0.06	58.4%	338	521	391	401	155	1806
BM	0.20±0.05	56.6%	269	559	390	451	139	1808
SF	0.20±0.06	55.8%	397	483	261	421	249	1811
SK	0.18±0.06	55.6%	370	491	283	406	252	1802
SI	0.16±0.06	55.5%	283	547	342	428	202	1802
SE	0.16±0.05	54.9%	221	525	489	453	117	1805
BD	0.14±0.06	54.8%	310	505	357	422	220	1814
SD	0.14±0.06	55.0%	252	561	350	459	175	1797
SB	0.13±0.06	55.0%	320	508	339	386	262	1815
SA	0.11±0.06	53.6%	238	495	438	444	162	1777
SH	0.09±0.07	52.9%	384	438	258	393	325	1798
SL	0.05±0.05	51.7%	200	522	432	491	170	1815
SC	−0.02±0.04	49.1%	72	284	1057	314	76	1803

Table 7.3 Summary statistics of user-study responses from the appropriateness to speech study, with confidence intervals for the mean appropriateness score (MAS) at the level $\alpha = 0.05$. "Pref. matched" identifies how often test-takers preferred matched motion in terms of appropriateness, ignoring ties. The *X-Att-XL* model results are highlighted in pink. Table and caption from [76].

Table 7.3 provides summary statistics and win rates, Figure 7.6 visualises the response distribution and Figure 7.7 shows significance comparisons. The *X-Att-XL* approach (**SJ**) ranked second in the submitted systems. Figure 7.7 shows few significant differences between pairwise systems. Only **SG** and the natural mocap (**NA**) rank significantly better than the *X-Att-XL* system. Again, the *X-Att-XL* system ranks significantly better than nine other conditions including the dyadic baseline system.

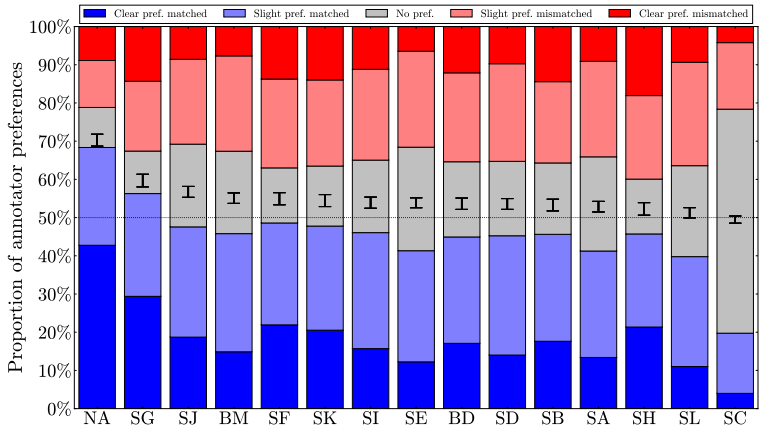


Fig. 7.6 Bar plots visualising the response distribution in the appropriateness to speech study. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. Lighter colours correspond to slight preference, and darker colours to clear preference. On top of each bar is also a confidence interval for the mean appropriateness score, scaled to fit the current axes. The dotted black line indicates chance-level performance. Conditions are ordered by mean appropriateness score. Figure and caption from [76].

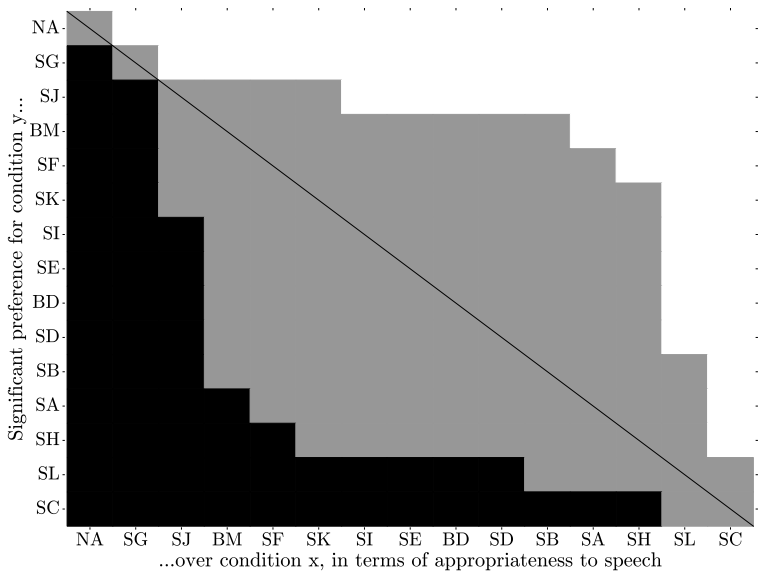


Fig. 7.7 Significance of pairwise differences between conditions in the appropriateness to speech evaluation. White means that the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions are listed in the same order as in Table 7.3. Figure and caption from [76].

7.5.7 Interlocutor Appropriateness

As the model includes awareness of the interlocutor's speech and motion, the appropriateness of the generated main-agent motion to the interlocutor's speech is also evaluated. The study used a similar technique for measuring speech appropriateness but differed in several important aspects. The test data contained pairs of interactions, one with matched main-agent and interlocutor interactions and another with the same main-agent speech but mismatched interlocutor speech. Preference can be quantified for generated motion with matched over mismatched interlocutor behaviour and, therefore, assess how interlocutor behaviour affects the motion.

Cond-ition	MAS	Pref. matched	Raw response count					
			2	1	0	−1	−2	Sum
NA	0.63±0.08	67.9%	367	272	98	189	88	1014
SA	0.09±0.06	53.5%	77	243	444	194	55	1013
BD	0.07±0.06	53.0%	74	274	374	229	59	1010
SB	0.07±0.08	51.8%	156	262	206	263	119	1006
SL	0.07±0.06	53.4%	52	267	439	204	47	1009
SE	0.05±0.07	51.8%	89	305	263	284	73	1014
SF	0.04±0.06	50.9%	94	208	419	208	76	1005
SI	0.04±0.08	50.9%	147	269	193	269	129	1007
SD	0.02±0.07	52.2%	85	307	278	241	106	1017
BM	−0.01±0.06	49.9%	55	212	470	206	63	1006
SJ	−0.03±0.05	49.1%	31	157	617	168	39	1012
SC	−0.03±0.05	49.1%	34	183	541	190	45	993
SK	−0.06±0.09	47.4%	200	227	111	276	205	1019
SG	−0.09±0.08	46.7%	140	252	163	293	167	1015
SH	−0.21±0.07	44.0%	55	237	308	270	144	1014

Table 7.4 Summary statistics of user-study responses from the appropriateness to interlocutor study, with confidence intervals for the mean appropriateness score (MAS) at the level $\alpha = 0.05$. "Pref. matched" identifies how often test-takers preferred matched motion in terms of appropriateness, ignoring ties. The *X-Att-XL* model results are highlighted in pink. Table and caption from [76].

Table 7.4 provides summary statistics and win rates. The *X-Att-XL* system ranked 8th in this study, but only natural mocap (**NA**), **SA**, **BD**, and **SL** are rated significantly higher than it as shown in Figure 7.8. This shows there is no significant difference to any other system, except **SH** where the *X-Att-XL* was significantly better. Statistics in Figure 7.9 show that the *X-Att-XL* system had the lowest number of negative scores (preference for the mismatched dyadic interaction) and a large number of no preference scores.

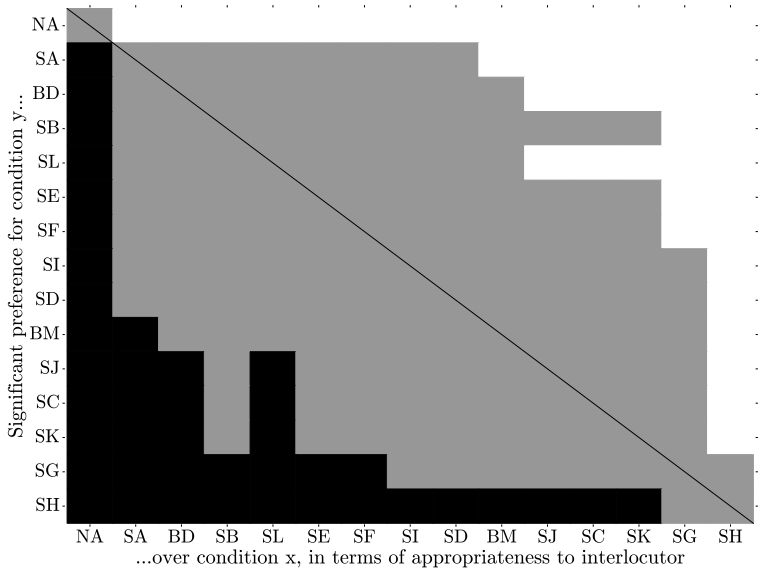


Fig. 7.8 Significance of pairwise differences between conditions in the appropriateness to interlocutor study. White means that the condition listed on the *y*-axis rated significantly above the condition on the *x*-axis, black means the opposite (*y* rated below *x*), and grey means no statistically significant difference at the level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions are listed in the same order as in Figure 7.4. Figure and caption from [76].

7.6 Discussion

The *X-Att-XL* approach performed well with regard to human likeness and appropriateness to speech. The *X-Att-XL* model performed comparably to 10 of the other systems with regards to appropriateness to the interlocutor's speech, but clearly, it can be improved in this area. Figure 7.9 and Table 7.4 show that, for the *X-Att-XL* system, participants preferred

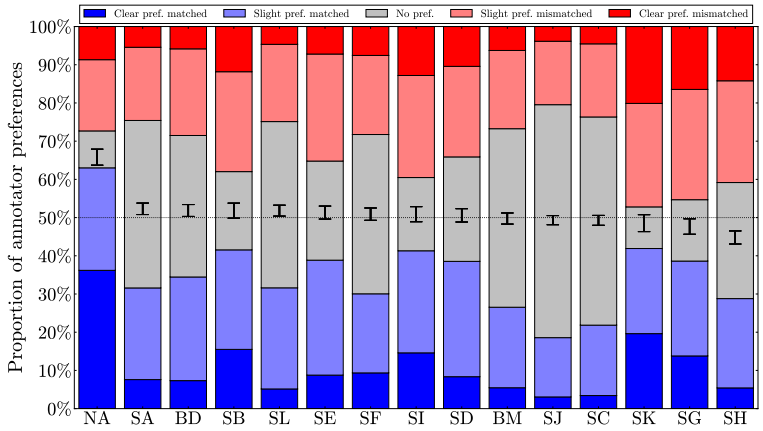


Fig. 7.9 Bar plots visualising the response distribution in the appropriateness to interlocutor study. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. Lighter colours correspond to slight preference, and darker colours to clear preference. On top of each bar is also a confidence interval for the mean appropriateness score, scaled to fit the current axes. The dotted black line indicates chance-level performance. Conditions are ordered by mean appropriateness score. Figure and caption from [76].

the mismatched stimuli least compared to all other systems (including natural mocap). The majority of responses were tied, meaning that they considered the mismatched stimuli to be of equal appropriateness as the matched animation. It is unclear where this uncertainty stems from and more work is required to evaluate this cause. This may be due to the subtle and sporadic nature of the interlocutor influence.

The overall impact of including the interlocutor's speech is difficult to evaluate. Although this chapter has not shown any statistical improvement, there are some observational improvements. The appropriate back-channel communication, such as head nodding and turn-taking through gesture, is evident, although sporadic. The sporadic nature of this means it may be overlooked during subjective evaluation, or many evaluation clips may not include these movements. The GENEA Challenge included two baselines, monadic and dyadic. While there is no significant difference between these two methods, the lack of significant difference and slight improvement from monadic to dyadic suggests that the inclusion of the

interlocutor should not hamper performance and, therefore, is a worthwhile inclusion for generative models in the future.

Although the proposed *X-Att-XL* method is deterministic, i.e. the same inputs will always produce the same outputs, it could be possible to incorporate this design into a probabilistic model. For example, this approach could be adjusted to incorporate probabilistic diffusion [55, 104] methods.

Chapter 8

Large Language Model Driven Gesture Animation

Contributing Publications

- (2024). Llanimation: Llama driven gesture animation. In *Computer Graphics Forum*, volume 43, page e15167. Wiley Online Library

8.1 Introduction

Speech and gesture are codependent, and gesture production is a complex function of audio and text speech content, including semantics and prosody. For instance, beat gestures synchronise with the timing of the speech audio dynamics, while iconic gestures convey the shape of the discussed topic [17]. Historically, methods for automatically generating gestures were predominantly audio-driven, exploiting the prosodic and speech-related content that is encoded in the audio signal. While audio features effectively encode prosody, they may also not capture semantics. Conversely, text features capture the content but may lack prosodic information. It becomes apparent that a combination of features may yield optimal results.

Large-Language Models (LLMs) are exposed to large natural language corpora, making them exceptional in language and content understanding. This chapter explores the integration of LLM embeddings into a gesture generation model to improve the semantic accuracy of co-speech gestures. Experiments comparing methodologies for combining LLM embeddings with audio features are defined, and results are reported using the objective and perceptual performance to determine the contribution of each feature. The introduced approach, named LLAniMAtion, utilises LLAMA2 language embeddings [136] and optionally combines them with PASE+ audio features [120] in a Transformer-XL architecture [146]. Surprisingly, the results show that LLM features perform significantly better than audio features, and no significant difference is recorded when these two modalities are combined. This experiment also demonstrates that LLM features contribute more to the perceived quality of the resulting gesture animations than audio features.

This chapter describes the motivation for using LLM features. The LLAniMAtion approach is introduced and described with an experimental setup of how features can be combined. An ablation study is performed to determine the effectiveness of LLM feature in isolation and in combination with audio features through objective and subjective measures. The methods are then further compared against the current state-of-the-art using objective and subjective measures.

8.2 Large-Language Model Motivation

Speech-driven gesture generation has historically relied on audio features as its primary input. While text-based features have gained momentum in recent research, the utilisation of LLM features remains limited. This section reviews audio features to understand the common features used in gesture generation and why these are useful features. The section also reviews LLM models used in gesture and related fields, describing why they are particularly suited to gesture generation models.

8.2.1 Speech Features for Gesture Generation

Gesture generation systems widely adopt audio-based features. In the co-speech gesture generation review by Nyatsanga et al. [106], out of 40 methods reviewed, 35 used audio as an input feature. In contrast, only 17 methods involved text as an input. Audio features can be embedded using various methods. Perhaps most common is the use of MFCC [51, 6, 116, 47, 108, 8], sometimes combined with other prosodic features such as pitch (F0) and energy [72]. Other latent representations such as Wav2Vec 2.0 [11] and PASE+ [120] have grown in popularity as these can also effectively encode important speech-related information as well as prosodic features [145, 146, 103], while improving speaker independence of the representation. Audio features are advantageous with regard to beat gesture performance as these have a close relationship to prosodic activity, such as acoustic energy and pitch [147, 114].

Numerous approaches leverage a combination of both audio and text features, with different methods for incorporating textual information. Word rhythm was used by Zhou et al. [158] where words are encoded in a binary fashion, taking the value 1 if a word is spoken and 0 if not. Other works, such as those by Windle et al. [145, 146], discussed in previous chapters and Yoon et al. [152] integrate FastText embeddings and Bojanowski et al. [16] which extend the Word2Vec approach [1] exploiting sub-word information. BERT features [31] have been successfully used in conjunction with audio in the work of [9]. BERT, originally designed for language modelling and next-sentence prediction, is composed of transformer encoder layers.

Using text as the exclusive input for gesture generation is infrequent, and performance is often limited when used. Yoon et al. [153] and Battacharya et al. [14] employ word embedding vectors [112] to facilitate gesture generation.

Text-based features are particularly advantageous in gesture generation as these inject a level of lexical and semantic understanding. This should allow a generative model to generate

more nuanced and varied gestures than those relying on prosodic elements only. Despite the recognised advantages of text-based features, LLMs have not been used in the context of gesture generation, whether in isolation or in combination with audio inputs. This highlights a gap in the current research landscape that is explored in this chapter.

8.2.2 Large Language Models

Given the close relationship between language and gesture, the recent advances in LLM performance present a promising avenue for advancing gesture generation.

LLM approaches fall into two categories: Encoder-Decoder/ Encoder only, often referred to as Bidirectional Encoder Representations from Transformers (BERT) [31] and and Decoder only, known as Generative Pre-trained Transformer (GPT) styles. These models typically exhibit a task-agnostic architecture. The primary focus in this chapter is on GPT-style models, which currently stand as leaders in LLM performance.

LLMs are typically trained as a text generation model, the input being the preceding text and the output being proceeding text. GPT models typically consist of multiple Transformer [138] layers followed by a linear layer, which is referred to as the head layer. The transformer layers effectively encode a sequence into a latent embedding and the linear head is trained to perform a specific task, such as sequence generation or classification, using these latent values. Figure 8.1 shows a typical GPT architecture overview.

The text is initially encoded, often using Byte-Pair Encoding (BPE). This process can break relatively rare words into subwords. For example, “thinking” breaks down into “think” and “ing”; this introduces the knowledge that “think” can be used in multiple contexts, often with a similar base understanding, but with additional nuances such as “ing” of “s”. These encodings are tokenised to produce a mapping from word-space to a numeric representation. This is a reversible process, and therefore, encoding and decoding this tokenised representation must be possible. These tokenised values are then passed to a multi

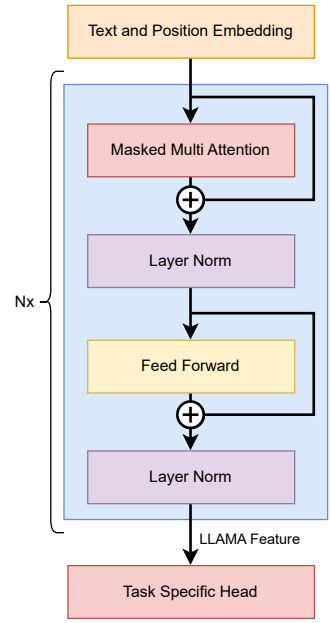


Fig. 8.1 A typical Generative Pre-trained Transformer (GPT) architecture overview. Multiple Transformer [138] layers followed by a linear layer, which is referred to as the task specific head layer.

layer Transformer [138] architecture, typically consisting of large and multiple layers. The output from these multiple Transformer layers is then passed to a Task-specific head, in the LLM training, this is to predict the next word, however, it is common to replace or fine-tune this last layer after the base model is trained. This is because the Transformer layers produce an embedded representation and understanding of the input. It is, therefore, possible to use this as a text embedding value.

Numerous GPT-style models have been introduced, and among them, GPT-4 from OpenAI [107] has emerged as a top performer across various language-based tasks. However, GPT-4 is a closed-source solution. The leading open-source alternative is currently LLAMA2 [136], which surpasses other open-source LLMs in tasks related to commonsense reasoning, world knowledge, and reading comprehension.

LLMs have begun to garner attention in gesture-based tasks but have not been used as a means of gesture generation. For instance, Hensel et al. [52] uses ChatGPT [107] for the selection and analysis of gestures, while Zeng et al. [155] uses ChatGPT to analyse and comprehend performed gestures. There are currently no established methods for generating gestures directly from LLMs.

8.3 LLAniMAtion

In the exploration of using LLMs as a primary feature for co-speech gesture generation, LLAniMAtion is introduced. LLAniMAtion utilises LLAMA2 text embeddings, which can be used as an independent feature or in conjunction with PASE+ [120] audio features. The generative model is based on the adapted Transformer-XL architecture presented in Chapter 7.

8.3.1 Speech Features

The LLAniMAtion method can leverage both audio and text-based features. Each modality has differing sample rates, with audio sample values updating at a faster pace than text tokens. Features are extracted at their original sample rates and aligned to fit the timing of a motion frame at 30fps. N represents the number of $\approx$33ms motion frames in an input sequence. The PASE+ and LLAMA2 model weights are frozen and not updated during training.

8.3.1.1 Audio

Audio features are extracted using the PASE+ model as these have been proven effective for gesture generation [147, 145, 146]. PASE+ was trained by solving 12 regression tasks to learn important speech characteristics using a multi-task learning approach. These tasks include estimating MFCCs, FBANKs and other speech-related information, including prosody and

speech content. Using this model, audio feature embeddings of size 768 for each 33ms audio window are extracted to align with the 30fps motion. Consequently, audio feature vectors, A, with a shape of $(N \times 768)$ are generated for each audio clip.

8.3.1.2 Text

Word-level features are extracted using the pre-trained, 7-billion parameter LLAMA2 model [136]. LLAMA2 adopts a Transformer architecture and has been trained on a corpus of 2 trillion tokens sourced from publicly available materials.

For each speech sequence, the released transcript of the audio clip is tokenised and processed by the LLAMA2 model. A sequence of embeddings is extracted using the transformer layers of the LLAMA2 model. The tokenised input is fed to the model and passed forward through all transformer layers but is not fed through any task-specific linear head. In Figure 8.1, the location of these features is shown with the "LLAMA Feature" label. Therefore, an associated latent vector is extracted from the output of the last transformer layer for each word in the utterance and these are used as the text embedding. For each word in the utterance, an output embedding is assigned and frame-wise alignment is performed to ensure that each embedding is synchronised with its corresponding motion frame timing at 30fps. The process generates text-embedding vectors T of shape $(N \times 4096)$.

Alignment is achieved by repeating text embeddings as needed to synchronise with the audio timing. In instances where a word spans multiple frames, the vector is duplicated for the corresponding number of frames, and a zero-value vector is employed when no word is spoken at a specific frame. Figure 8.2 provides an overview of the alignment process.

The input utterance is tokenised using a BPE method, meaning a single word may be broken into multiple constituent parts. For example, the word "thinking" will be divided into two tokens, "think" and "ing". In such cases, only the embedding for the last token is retained, and the embeddings for the preceding parts are discarded. For example, the

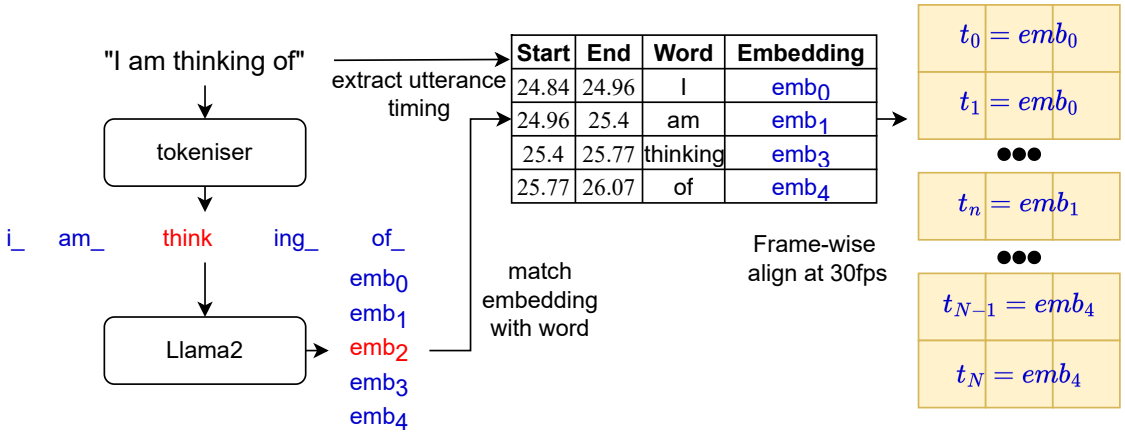


Fig. 8.2 Extracting text features using LLAMA2. The text is BPE-tokenised, and a LLAMA2 embedding is computed for each token. These embeddings are aligned with audio at 30fps by repeating frames as necessary.

embedding associated with "ing" is used rather than "think". This is common practice when using LLMs as the final embedding is expected to encapsulate information about preceding tokens.

8.3.1.3 Speaker style

For each utterance, a speaker label is additionally provided as input. This is a unique ID per speaker which is passed through a learned embedding layer. The trainable weights of this layer ensure that speakers with similar gesture styles are positioned closely in the latent embedding space, while speakers with distinct gesturing styles are situated further apart. An 8-dimensional embedding is used to generate speaker vectors S with a shape of $(N \times 8)$.

8.3.2 Body Pose Representation

The body pose at time n is defined as:

$$\mathbf{y}_n = [x_n, y_n, z_n, r_{j,1,n}, ..., r_{j,6,N}] \tag{8.1}$$

where x, y, z denote the global skeleton position and $r_{j,1:6,n}$ form rotations for each joint j in the 6D rotation representation presented by Zhou et al. [159]. These values are standardised by subtracting the mean and dividing by the standard deviation computed from the training data.

8.3.3 Model Architecture

In this chapter, the primary objective is to evaluate the impact of LLM features on the animation of co-speech gestures. To accurately measure this effect, an established model and training method are employed. Specifically, a model based on the Cross-Attentive Transformer-XL, which demonstrated effectiveness in the GENEA challenge 2023 [146], also described in Chapter 7. This approach is built on the Transformer-XL model architecture [29] which uses segment-level recurrence with state reuse and a learned positional encoding scheme to ensure temporally cohesive boundaries between segments. Chapter 7 extends this architecture using cross-attention to incorporate the second speaker's speech into the prediction when used in a dyadic setting. Notably, this architecture delivers high-quality results without the need for more involved training techniques such as diffusion.

Either a single modality or a combination of features are used to form the input feature matrices $\mathbf{X} \in \{X_a, X_t, X_+, X_\times\}$. Where X_a is audio only, X_t is text only, X_+ is both concatenated, and $X_\times$ is both combined using cross-attention. Please refer to Section 8.4.2 for more details on the construction of this matrix. The model is trained on a dyadic conversation between a main agent and an interlocutor. Specifically, the main-agent's gesturing is predicted, conditioned on both main-agent and interlocutor speech. Consequently, a set of input features are extracted for each speaker, X^{ma} and X^{in}, and a set of target poses for the main-agent, Y. These extracted features are segmented into non-overlapping segments of length W frames.

Given an input feature vector X of length W, the Transformer-XL predicts $\hat{Y}$ of length W using a sliding window technique with no overlap. Consequently, for a speech sequence of length N, the model is invoked $\lceil \frac{N}{W} \rceil$ times. Figure 8.3 shows an overview of this approach.

8.3.4 Training Procedure

Training LLAniMAtion methods use the same methodology as in Chapter 7, and the same geometric and temporal constraints are used in the loss function. The loss function L_c comprises multiple terms including a L_1 loss on the rotations (L_r), positions (L_p), velocity (L_v), acceleration (L_a) and kinetic energy (L_{v^2}) of each joint. The loss is computed as:

$$\begin{aligned} L_r &= L_1(\mathbf{y_r}, \hat{\mathbf{y}}_\mathbf{r}) \\ L_p &= L_1(\mathbf{y_p}, \hat{\mathbf{y}}_\mathbf{p}) \\ L_v &= L_1(f'(\mathbf{y_p}), f'(\hat{\mathbf{y}}_\mathbf{p})) \\ L_{v^2} &= L_1(f'(\mathbf{y_p})^2, f'(\hat{\mathbf{y}}_\mathbf{p})^2) \\ L_a &= L_1(f''(\mathbf{y_p}), f''(\hat{\mathbf{y}}_\mathbf{p})) \\ L_c &= \lambda_p L_p + \lambda_v L_v + \lambda_a L_a + \lambda_r L_r + \lambda_{v^2} L_{v^2} \end{aligned} \tag{8.2}$$

where f' and f'' are the first and second derivatives, $\mathbf{y_r}$ and $\hat{\mathbf{y}}_\mathbf{r}$ are ground truth and predicted 6D rotations and $\mathbf{y_p}$ and $\hat{\mathbf{y}}_\mathbf{p}$ are positions in world space computed using Forward Kinematics given the predicted joint angles and the pre-defined speaker-specific bone lengths. Each term has a λ weighting to control the importance of each term in the loss.

All training parameters were kept the same as in Chapter 7. However, the Cross Attentive Transformer-XL included an additional two self-attention layers. These additional layers were chosen based on validation loss values and the quality of the predicted validation sequences. Models are trained for 1300 epochs using the AdamW optimiser [87].

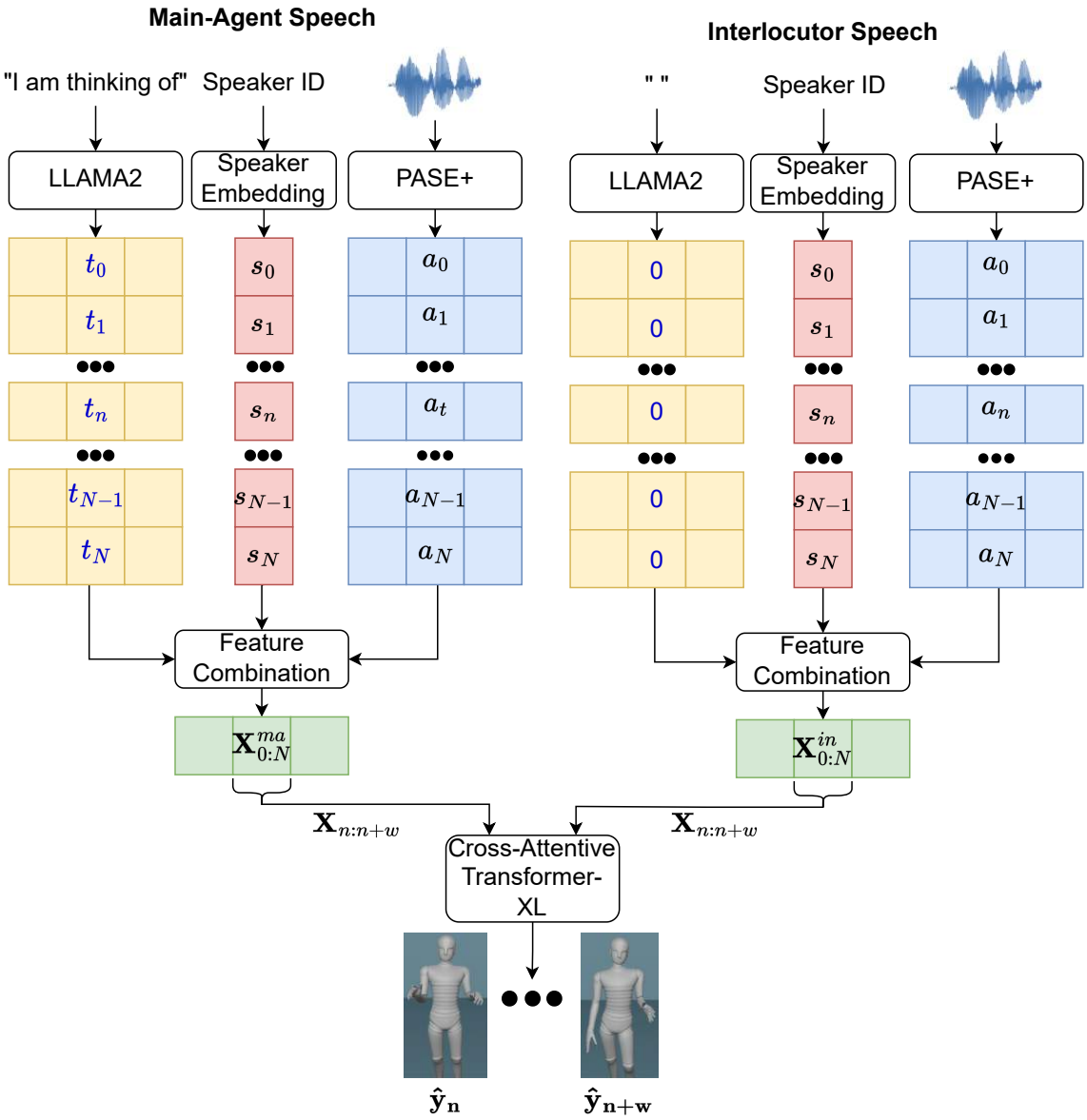


Fig. 8.3 Overview of LLANIMATION method. The model takes LLAMA2 features as input, along with a speaker embedding and optional PASE+ features that encode the speech of a main-agent and an interlocutor. The features are combined and processed through a cross-attentive Transformer-XL model that produces gesture animation for the main-agent.

8.3.5 Smoothing

The raw model output can contain low levels of high-frequency noise. Following other work on motion synthesis [156, 158], a Savitzky-Golay Smoothing filter is applied to mitigate this. A window length of 9 and a polynomial order of 2 is used. The small window size and low polynomial mean this filter provides a very small amount of localised smoothing while retaining accurate beat gestures.

8.4 Experimental Setup

Four distinct models are trained, each with a different set of features: 1) PASE+: An audio-only model, 2) LLAniMAtion: A LLAMA2 text-only model, 3) LLAniMAtion-+: A LLAMA2 and PASE+ concatenated model and 4) LLAniMAtion-×: A LLAMA2 and PASE+ cross-attention model. In this section the data and details of the model configurations are described.

8.4.1 Data

The data used in this study is from the GENEA challenge 2023 [76], discussed in Section 7.2. This dataset is derived from the Talking With Hands dataset [78], containing dyadic conversations between a main-agent and interlocutor. It comprises high-quality 30fps motion capture data in Biovision Hierarchical (BVH) format. The dataset includes both speech audio and text transcripts derived from both speakers in the conversations. The dataset is divided into three splits: 1) train, 2) validation, and 3) test. The validation set is employed for model tuning and refinement, while the test set is exclusively reserved for evaluation.

8.4.2 Feature Combinations

The presented experiments use audio and text modalities in isolation and additionally investigate two approaches for combining the two modalities: 1) post-extraction concatenation and 2) cross-attention, respectively referred to as LLAniMAtion-+ and LLAniMAtion-×. In Figure 8.3, this is shown as a "Feature Combination" box and decides which modalities are used and how they are combined, depending on the LLAniMAtion setting.

8.4.2.1 Single Modalities

To use each modality individually, the speaker S matrices are concatenated with the audio A or text T along the feature dimension to form X_a and X_t, respectively. The concatenated matrix is then passed through a linear layer, giving:

$$\begin{aligned} \mathbf{X}_a &= \mathbf{W}_a(A,S)^\top + \mathbf{B}_a \\ \mathbf{X}_t &= \mathbf{W}_t(T,S)^\top + \mathbf{B}_t \end{aligned} \tag{8.3}$$

where W_a, W_t, $\mathbf{B}_a$ and $\mathbf{B}_t$ are learned parameters. $\mathbf{X}_a$ and $\mathbf{X}_t$ are used as inputs for training the single modality audio and text-based models respectively.

8.4.2.2 Concatenation

To combine modalities A, T and S matrices are concatenated along the feature dimension. The concatenated matrix is then passed through a linear layer, giving:

$$\mathbf{X}_+ = \mathbf{W}(A,T,S)^\top + \mathbf{B} \tag{8.4}$$

where $\mathbf{W}$ and $\mathbf{B}$ are learned parameters. This results in $\mathbf{X}_+$ which are the concatenated audio and text features and serve as the input to LLAniMAtion-+.

8.4.2.3 Cross-attention

Additionally, cross-attention is used to experiment with the method of combining audio and text features. Cross-attention has been shown to be an effective method of combining modalities, as evidenced in Ng et al. [103]. In this approach, the style embedding is first concatenated to both audio and text features. The two concatenated matrices are then linearly projected into the same feature dimension size, d, following Equation 8.3. Cross-attention is performed on the feature dimension, such that the projected audio features, X_a, serve as the query, while the projected text features, X_t are set as the key and value [138]:

$$\mathbf{X}_{\times} = \text{softmax}(\frac{\mathbf{X}_a X_t^{\top}}{\sqrt{d}})\mathbf{X}_t \tag{8.5}$$

giving the cross attention combined audio and text features $X_{\times}$ which are used as input for training LLAniMAtion-$\times$.

8.5 Evaluation

An evaluation is presented to determine the efficacy of LLAMA2 features for gesture generation, in isolation and in combination with audio PASE+ features. This section presents the observations and reports the associated performance metrics. Finally, a user study is described that measures the differences in perceived quality.

8.5.1 Observations

Noticeable differences are observed between the animation produced by the PASE+-based model and the LLAniMAtion method. The PASE+ version primarily generates beat gestures, whereas the LLAniMAtion model exhibits more varied motions, encompassing both beat and semantic gestures. The animation from LLAniMAtion appears to be more expansive

and confident. Video examples and comparisons showing this effect can be seen in the supplementary material[1].

8.5.1.1 Beat Gestures

Beat gestures are characterised by simple and fast movements of the hands, serving to emphasise prominent aspects of the speech [17]. These gestures have a close relationship with the timing of prosodic activity, such as acoustic energy and pitch [147, 114]. Given that these prosodic features can be directly derived from the audio signal, an audio-based model can be very effective at generating beat gestures. A beat gesture is not necessarily expected for every audio beat, but when performed, it is likely to be well-timed with the audio beats.

Using the motion and audio beat extraction method defined in the beat align score calculation proposed by [84], the onset of audio beats and motion gestures over time can be visualised. Remarkably, LLAniMAtion with LLAMA2 and no audio features consistently executes beat gestures in synchronisation with audio beats despite lacking explicit energy or pitch information. Figure 8.4 shows a 1.5-second clip with the left wrist motion onsets in green and audio beat onsets in red. A speaker can be seen swiftly moving their left hand from left to right in time with audio beats and returning close to their original pose.

Although the LLAMA2 embeddings are temporally aligned, providing the model with awareness of word timings, there is no explicit knowledge of syllable-level timing. Further investigation is needed; however, it is plausible that training with LLAMA2 embeddings may effectively encode information regarding the presence of lexically stressed syllables in context within words.

[1] https://youtu.be/jBXpWocXvZ8

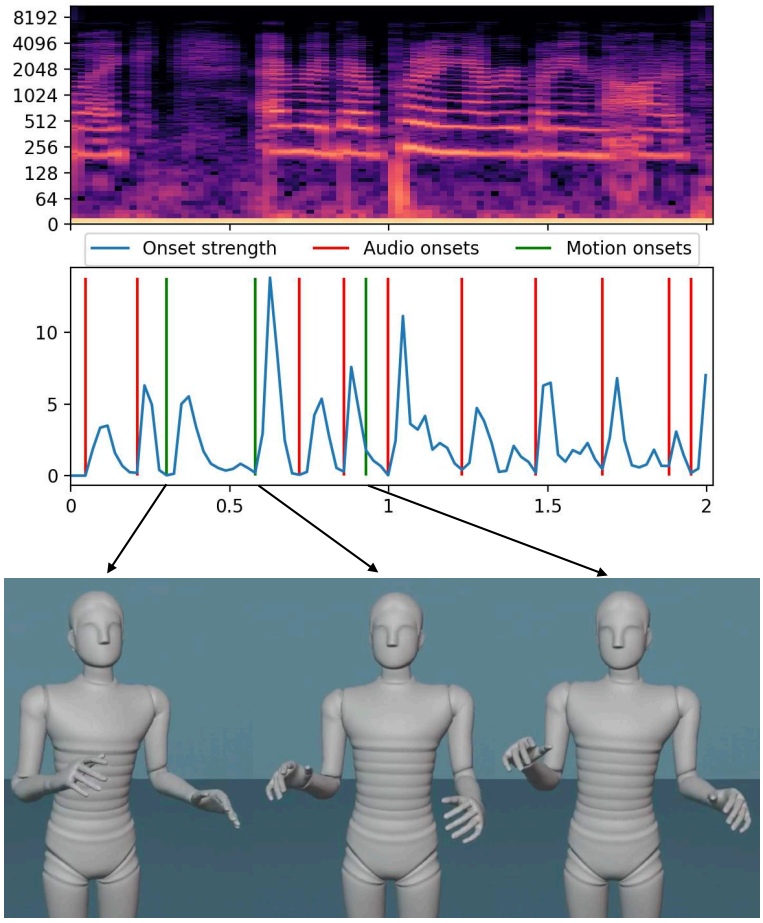


Fig. 8.4 Generated gestures for given audio beats using LLAniMAtion method. Using a 1.5s audio clip from the test dataset, the audio spectrogram is shown, as well as aligned audio beat onsets and their corresponding onset strengths, as well as motion gesture onset detection of the left wrist using the method of beat detection defined in [84]. The speaker moves their left hand from right to left and back again as the syllables are stressed.

8.5.1.2 Semantic gestures

Semantic gestures are often directly linked to speech content, such as mimicking an action or nodding the head in agreement. During empirical observations, the LLAniMAtion method demonstrated superior performance compared to the audio PASE+-based model in generating these types of gestures.

In a test sequence where a speaker is describing the act of eating a crab, the LLAniMAtion gestures exhibit more activity compared to the PASE+ version, particularly when the speaker

uses their hands to illustrate actions. This is exemplified when the hands mimicked sticking a fork in a crab for consumption in time with the verbal description. This sequence can be seen in Figure 8.5 and the supplementary video[2].

LLAniMAtion demonstrates the capacity to adequately encode agreeableness. For example, Figure 8.6 shows a predicted test sequence where the speaker can be seen nodding along with the word yes.

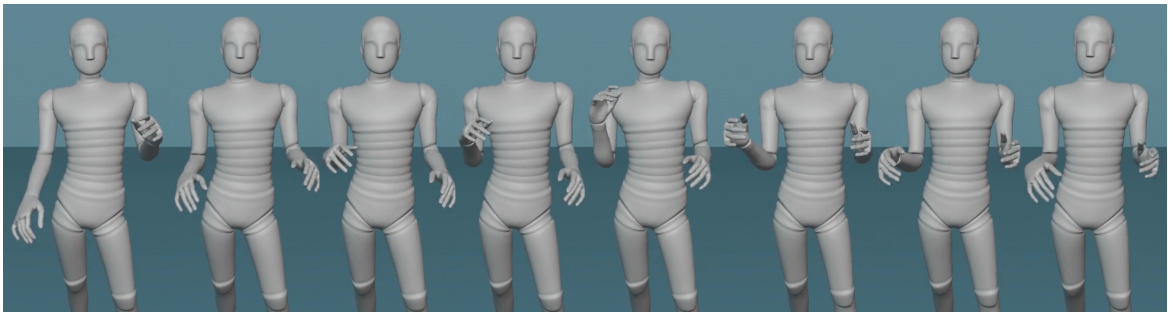


Fig. 8.5 Example sequence showing a speaker mimicking the use of a fork with their right hand while describing eating crab generated using the LLAniMAtion method

Fig. 8.6 Example nod motion temporally aligned with the word "yes" being spoken. from a test sequence generated using the LLAniMAtion

8.5.1.3 Laughter

During the transcription process of the GENEA dataset, laughter without speech was denoted using "###". This representation was directly input to the LLAMA2 model for feature extraction. Although the generated embedding would not encode any semantic meaning,

[2] https://youtu.be/jBXpWocXvZ8

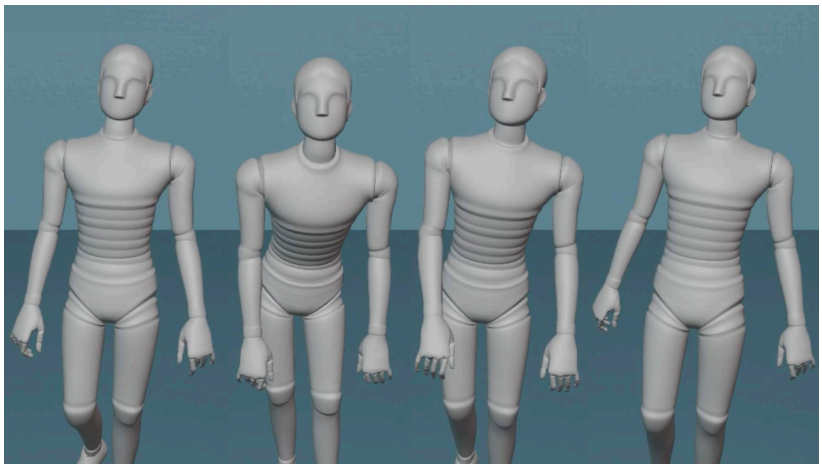


Fig. 8.7 Example laughter sequence generated using the LLAniMAtion method

the model learns to associate these tokens to laughter. The LLAniMAtion method captures moments of laughter as illustrated in Figure 8.7, where the character partially creases over. This specific behaviour is not observed in the gesture animation produced by the PASE+-based model .

8.5.2 Performance Metrics

Evaluating the objective performance of gesture generation poses a challenging research question, primarily due to the many-to-many ambiguous relationship between speech and gesture. No single metric has been developed that correlates with human perception. However, a *combination* of metrics can be used as a means to somewhat evaluate the quality of the generated gesture. Frèchet Gesture Distance (FGD) [152, 13, 103], Frèchet Kinetic Distance (FD_k) [103] and Beat Alignment (BA) [83, 84] are useful metrics for this task. These metrics are indicative of static and dynamic appropriateness and the alignment of motion to speech [8, 84, 152]. Each is discussed in further detail in Section 2.5.2.

Frèchet Gesture Distance is a measure based on the Frèchet Inception Distance (FID) [54], which is commonly used for evaluating generative models. A pre-trained autoencoder extracts domain-specific latent features from both ground truth and predicted motion. The

Model	$FGD \downarrow$	$FD_k \downarrow$	BA$\uparrow$
PASE+	79.90	34.37	**0.871**
LLAniMAtion	61.86	24.23	0.855
LLAniMAtion-+	**47.56**	**23.79**	0.869
LLAniMAtion-$\times$	66.87	25.70	0.865

Table 8.1 Frèchet Gesture Distance (FGD) , Frèchet Kinetic Distance (FD_k) and Beat Alignment (BA) scores for each system calculated with respect to the ground truth test dataset.

FGD score is a Frèchet distance between the two multivariate Gaussian distributions of these features in latent space. This measures the similarity between the generated and ground truth poses but does not necessarily indicate how well the generated examples temporally align with the audio. Frèchet Kinetic Distance is similar; however, there is no auto-encoding process. Instead, the first derivative of each joint is used to determine the distribution of velocities for both the ground truth and predicted motion. FD_k is the Frèchet Distance between these two distributions.

Beat Alignment has been adapted from music synthesis [83] to work with gesture generation [84]. This gives a synchrony measure between audio and gesture beats by using a chamfer distance between them. Beats are detected using the root mean square onset of the audio, and a motion beat is identified by the local minimums of the velocity.

8.5.2.1 Results

The measures presented in Table 8.1 indicate that the FGD and FD_k scores are consistently lower for all LLAMA2-based models than for the model trained on PASE+ features. This suggests that the motion generated by LLAniMAtion may be closer to ground truth, with LLAniMAtion-+ showing the most realistic motion. The BA score suggests that the audio features are the most timely as expected due to the increased prosodic knowledge, however, the differences between this and the LLAniMAtion methods are minimal. Notably, the method with no audio features is competitive in FGD and BA scores.

8.5.3 User Study

A user study is presented to further evaluate perceived human likeness and appropriateness of the animations from the PASE+-based model compared with the LLAMA2-based LLAniMAtion method. Participants were hired through the Prolific[3] platform with 50 participants in each experiment after removing any participants that failed attention checks. Participants were filtered to be fluent in English. For this study, a similar methodology to Alexanderson et al. [8] and the GENEA Challenge 2023 [76] is used.

All test sequences for each method were rendered on the same virtual avatar released by Kucherenko et al. [76], as shown in Figure 6.4. The exact clip timings from the GENEA Challenge are used, comprising 41 clips with an average duration of 10 seconds each. During evaluation, users exclusively heard the audio of the main-agent being animated.

In the pairwise system comparison, participants were presented with two side-by-side videos generated for the same audio but with different systems. To mitigate bias, the question order is randomised and randomly swap the side of the screen that each condition is shown.

The question for all studies was posed as "Which character's motion do you prefer, taking into account both how natural-looking the motion is and how well it matches the speech rhythm and intonation?". The participants were asked to choose from the options {**Clear preference for *left***, **Slight preference for *left***, **No Preference**, **Slight preference for *right*** and **Clear preference for *right***}. The scoring methodology uses a merit system [109] where an answer is given a value of 2, 1 or 0 for clear preference, slight preference and no preference, respectively. Preference testing allows a win rate calculation where a win is assigned when there is an identified preference for a system, not including ties. A one-way ANOVA test with a post-hoc Tukey test was subsequently used for significance testing.

[3] https://www.prolific.co/

		vs PASE+		vs LLAniMAtion		vs LLAniMAtion-+		vs LLAniMAtion-×	
	Merit Score	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate
PASE+	0.37±0.05	-	-	25.4%	11.4%	24.6%	14.8%	28.4%	14.8%
LLAniMAtion	0.68±0.06	63.3%	11.4%	-	-	38.6%	22.3%	44.3%	14.8%
LLAniMAtion-+	**0.69**±0.06	61.0%	14.8%	39.0%	22.3%	-	-	43.2%	20.8%
LLAniMAtion-×	0.64±0.06	56.8%	14.8%	40.9%	14.8%	36.0%	20.8%	-	-

Table 8.2 User study results. Merit scores [109] with 95% confidence intervals, win and tie rates for each comparison.

8.5.3.1 Results

Table 8.2 summarises the results of the user study. These findings validate the objective measure scores in that all LLAniMAtion-based models outperform the PASE+ audio-only method. According to the merit score, all LLAniMAtion methods were significantly preferred over the PASE+ approach ($p < 0.001$). Win and tie rates show that LLAniMAtion methods win or are tied with PASE+ most of the time. Surprisingly, the highest win rate is recorded by the LLAniMAtion method with no PASE+ features included, suggesting that using text as a sole input is sufficient to generate plausible speech gesturing, and that audio features are somewhat redundant in the model

Between each LLAniMAtion method, there is no statistically significant difference in merit scores. The win and tie rates against LLAniMAtion are examined to determine if adding PASE+ features will provide additional preference. It is evident from these rates that the choice between LLAniMAtion settings is almost tied to wins and losses. LLAniMAtion-+ wins 1.9% less than LLAniMAtion-×; however, the tie rate is higher and therefore loses less than LLAniMAtion-×.

This initial study concludes that LLAMA2 features are powerful at encoding information useful to gesture generation and can produce more realistic-looking gestures than a model trained on audio input. Combining modalities also does not make a significant difference, although the concatenation of features performs slightly better than the cross-attention regarding merit scores and win/tie rates.

Model	$FGD \downarrow$	$FD_k \downarrow$	BA↑
LLAniMAtion	61.86	24.23	0.855
LLAniMAtion-+	47.56	23.79	**0.869**
CSMP-Diff	**30.620**	**12.61**	0.866

Table 8.3 Frèchet Gesture Distance (FGD) [152], Frèchet Kinetic Distance (FD_k) and Beat Alignment (BA) [84] scores for each system calculated with respect to the ground truth test dataset.

8.6 Comparison Against State-Of-The-Art

The previous study has shown that a significant performance improvement was achieved by integrating LLM features into gesture-generation models. Additional experiments are performed to compare the best performing LLAniMAtion and LLAniMAtion-+ approaches against both ground truth and the current state-of-the-art method. This broader evaluation aims to assess performance across the field.

LLAniMAtion is compared against the state-of-the-art Contrastive Speech and Motion Pretraining Diffusion (CSMP-Diff) diffusion method [30], which achieved the highest human-likeness and speech appropriateness rating among the entries to the 2023 GENEA challenge. This method uses a CSMP-Diff module, which learns a joint embedding for speech and gesture with the aim of learning a semantic coupling between these modalities. The output from this method is used as a feature in a diffusion model based on the Listen Denoise Action method by Alexanderson et al. [8].

Objective performance metrics are shown in Table 8.3. CSMP-Diff performs better in FGD and FD_k scores. Minimal differences are found in the BA score, with LLAniMAtion marginally outperforming CSMP-Diff. This difference in performance is likely explained by the observed motion from CSMP-Diff looking more human-like and, therefore, scoring higher in FGD and FD_k. However, when considering the appropriate timing of gestures, the timings are very similar, and therefore, the BA scores remain similar.

		vs GT		vs LLAniMAtion		vs LLAniMAtion-+		vs CSMP-Diff	
	Merit Score	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate	Win Rate	Tie Rate
GT	1.16±0.05	-	-	78.6%	8.9%	74.8%	8.9%	68.9%	11.1%
LLAniMAtion	0.34±0.04	12.5%	8.9%	-	-	34.2%	33.6%	31.4%	16.1%
LLAniMAtion-+	0.36±0.04	16.4%	8.9%	32.2%	33.6%	-	-	35.6%	14.4%
CSMP-Diff	**0.58**±0.05	20%	11.1%	52.5%	16.1%	50.0%	14.4%	-	-

Table 8.4 User study results. Merit scores [109] with 95% confidence intervals, win and tie rates for each comparison.

The user study is repeated following the protocol as described in Section 8.5.3, and the results are summarised in Table 8.4. In terms of merit score, the ground truth was perceived as significantly better than any other method ($p < 0.001$), underscoring the current challenge in consistently generating human-realistic gesturing. CSMP-Diff was considered superior to both LLAniMAtion methods ($p < 0.001$). Despite this difference, when examining the win rates against CSMP-Diff, the LLAniMAtion method wins 31.4% of the time and ties 16.1%. Meanwhile, the LLAniMAtion-+ method won 35.6% of the time and ties 14.4%. In each case, LLAniMAtion and LLAniMAtion-+ are rated as good or better than CSMP-Diff 47.5% and 50% of the time, respectively.

CSMP-Diff incorporates both diffusion and contrastive speech and motion pre-training, representing two advanced and complex techniques. Despite these sophisticated methods, the evidence indicates that LLAniMAtion, even in the absence of any audio input, can perform as well as or better than CSMP-Diff nearly half the time. This suggests that LLAMA2 features serve as incredibly valuable feature encodings for gesture animation.

8.7 Discussion

The use of LLAMA2 features for speech-to-gesture generation has been explored using the proposed LLAniMAtion method. With the use of LLAMA2 features it is possible to generate well timed and contextually rich gestures even without the inclusion of any audio feature embedding. The use of combining both audio and text modalities through concatenation and

cross-attention is explored and found that there was no significant difference in the inclusion of PASE+ features when compared to using LLAMA2 features in isolation. The performance improvements when incorporating the LLAMA2 features into a gesture-generation model has been demonstrated through both objective and subjective measures. Given this finding, it can be suggested that human speech-related gesture animation is heavily related to the semantic encoding that is present in the LLAMA2 embeddings and that these embeddings additionally capture a notion of prosody from the language context. This is a somewhat surprising finding, and a result may have great practical impact on future content-aware animation systems.

The LLANIMATION approach is additionally compared to ground truth as well as the state-of-the-art CSMP-Diff approach. The evaluation revealed that both LLANIMATION and CSMP-Diff have areas where improvement is possible as they are unmatched against ground truth. While CSMP-Diff remains state-of-the-art, it is a complex model necessitating both diffusion and contrastive pre-training. Conversely, the introduced LLANIMATION method does not require the computationally expensive diffusion method or any pre-training and was still rated as good or better than CSMP-Diff 50% of the time. Integrating LLM features into state-of-the-art systems will be a step towards bridging the gap between machine-generated and natural gesturing.

The use of only text-based features may prove useful in future as automatic Text-To-Speech (TTS) performance increases. The use of TTS could be considered cheaper and more flexible than audio recording as the speaker voice and style can be directly controlled without complicated recording and actor requirements. This work has also experimented with the use of generating gestures from TTS audio. Speech was generated using Bark [5], with word utterance timings extracted using OpenAI Whisper [118]. Using these automated methods, it is possible to go from text to audio speech and gesture automatically and produce

human-like and appropriate gestures. An example of this can be seen in the supplementary video[4] associated with this work.

[4] https://youtu.be/jBXpWocXvZ8

Chapter 9

Conclusions and Future Work

9.1 Discussion

This work has reviewed the gesture generation landscape, analysed co-speech gesture data and introduced multiple novel approaches for automatic speech-to-gesture animation. Specifically, several models have been developed and evaluated to generate gesture animation from speech audio and text, a combination of both, or each independently. With each development, a key research question underpinned the motivation for each approach.

This work initially delved into a co-speech gesture dataset in Chapter 4, focusing on lateral symmetry to further understand co-speech gestures and whether lateral mirroring is an appropriate data augmentation approach. This work, published in Speech Communication [147], analyses arm motion's positional, temporal and informational symmetry. The discussion of the efficacy of lateral mirroring of the human body for data augmentation concludes that lateral mirroring is unsuitable as a generic approach. Instead, this work suggests that including laterally mirrored poses as a new identity is a suitable data augmentation method. This work also introduces statistically derived gesture spaces, which may be helpful for further analysis and evaluation methods in future.

With complete data analysis and an adequate understanding of the co-speech gesture data, the work focuses on deep-learning approaches for animated gesture generation from speech. The initial approach discussed in Chapter 5 and published in the International Conference on Multimodal Interaction [145] compares a single decoder against a multiple decoder method. The method here explores whether several decoder experts, each focusing on decoding a specific body part, might be better served than decoding the entire body simultaneously. Each model consists of a backbone of a Bi-Directional Long Short Term Memory (BLSTM) model using Problem Agnostic Speech Encoding (PASE+) audio features and FastText text embeddings. Each model performed competitively in the Generation and Evaluation of Non-verbal Behaviour for Embodied Agent (GENEA) challenge 2022, but some issues with the multiple decoder method were found. Notably, there is a disparity in leg motion for the rest of the body, likely due to a weak correlation between leg motion and speech. Despite this, the hand and arm motion did appear to improve when using an expert decoder for each, and therefore, with further work, this method may be effective.

Due to the physical aspect of motion, physiology can influence future motion. For example, what the body does in the past will impact the future motion based on position, acceleration and velocity of motion. Chapter 6 introduces a novel gesture diffusion network which generates style-conditioned gestures from PASE+ audio features with knowledge of an extended historical context using the Transformer-XL [29] architecture. This experiment produced state-of-the-art gesture generation and explored the impact of varying historical context lengths. This method found that increasing the context produced smooth motion, reduced repetitive behaviours such as repeated weight shifting from one leg to another and enabled the model to disambiguate between periods of inactivity and short pauses in the speech. This is an example of a probabilistic approach which is desirable in gesture generation due to the many-to-many mapping of speech and gesture. By using a probabilistic approach, this gives animators more flexibility and diversity in their gesture control, being

able to produce many gesture sequences for the same speech and choose the most appropriate depending on their needs. Probabilistic models are, however, dependent on large datasets that encapsulate the data distribution that is generalised to the problem, which may not always be available and are expensive to capture.

Co-speech gesture often occurs in a dyadic context. While a person is listening and interacting with the second, *interlocutor* speaker, this can often influence motion. The work in Chapter 7 and published in the International Conference on Multimodal Interaction [146] modifies the Transformer-XL attention mechanism to introduce the interlocutor speech via cross-attention. This model was evaluated during the GENEA challenge 2023 and performed competitively, ranking third in human-likeness and second in appropriateness to speech out of 14 other generative methods. Introducing the interlocutor did not drastically improve performance; however, it did subtly improve these results. For example, reactive head nods during listening portions of the interaction and gesture turn-taking are evident during periods where the second speaker is attempting to take the speech turn. These characteristics were not present in a model without the interlocutor included.

Large-Language Models (LLMs) are currently state-of-the-art natural language processing models. These models are being applied to various tasks that include natural language. Given the close relationship between natural language and gesture, Chapter 8 explores using LLM features for gesture generation. This method uses the same model as in Chapter 7 as a solid baseline model and explores using LLAMA2 [136] as a language feature extractor. This introduces the LLAniMAtion model approach and evaluates the performance impact of using LLAMA2 features in combination with PASE+ audio features and in isolation. This work demonstrated that LLM features contribute more to the perceived quality of the resulting gesture animations than audio features. The use of LLAMA2 features was able to generate well-timed and contextually rich gestures even without the inclusion of any audio feature embedding. The work explored combining audio and text modalities through concatenation

and cross-attention and found no significant difference in the inclusion of PASE+ features when compared to using LLAMA2 features in isolation. This work suggests that integrating LLM features into state-of-the-art systems will be a step towards bridging the gap between machine-generated and natural gesturing.

9.2 Application in Industry

Speech-to-gesture is a highly applicable topic in the animation and games industry. Current animation solutions rely on expensive and slow processes such as motion capture or hand animation. The automatic nature of the approaches described in this thesis are explicitly designed to fit into pre-existing animation pipelines. The generated animations may be used as is or provide animators a starting point to control or finely adjust. Non-Playable Characters and background characters are a notable application where these methods may be used. These characters often use pre-recorded lines and therefore the motion can be generated in advance. With the work described in this thesis, these approaches can help companies scale their animation efforts where they cannot afford the number of animations desired. This work can also allow artists to be more flexible with scene composition as many animators will choose camera positions that crop most of the body in order to limit the amount of body animation effort required. With the inclusion of automatic gesture generation in the pipeline, this limitation may be avoided.

When deploying these methods in industry, research and clear expectations need to be considered. In any of the methods mentioned in this thesis, there are limitations, and when applying any of the methods, there is a balance of gesture quality, efficiency in training and inference, latency of approaches and gesture diversity. Diffusion models and LLM based models such as those described in Chapter 6 and 8 respectively produce the overall best gesture quality in this work. However, these are computationally very expensive and may not be appropriate to run client-side. The windowing approaches in most of the introduced

methods mean that there is an inherent latency in these approaches as they rely on some frames of look-ahead to perform effectively. The LSTM approach described in Chapter 5 could be adapted to work without a look-ahead and may be appropriate for a lower latency option at the risk of lower-quality animation. If the goal is to have a diverse range of outputs for a single speech utterance, then a computationally expensive method such as Diffusion would be beneficial, however, at the cost of computational efficiency.

9.3 Further Work

While this thesis covers multiple aspects of automatic co-speech gesture generation, further work is always needed. This section discusses potential further work regarding speech features, particularly text embeddings. It also considers the application of generation in a streaming capability. Finally, additional work on dyadic interaction is discussed.

9.3.1 Speech Features

The extracted speech features can be particularly critical to gesture generation from speech. Chapter 8 found that semantic understanding is as important if not more important than the prosodic elements of speech. Models like that used in Chapter 6 lack speech semantics, so they cannot produce gestures relating directly to the utterance's meaning. More work is required to determine the potential improvement that may be gained by extending the model to utilise additional conditioning features on verbal content to improve semantic understanding [27, 153].

While using LLM features is powerful for generating contextually and semantically correct gestures, more work is required to get performance closer to ground truth gestures. Due to hardware constraints, only the 7-billion parameter release of LLAMA2 has been used

in Chapter 8. The larger 70-billion parameter could be utilised with more resources, which may produce more nuanced and varied gesturing.

LLMs continue to gain greater performance in language tasks; therefore, a complete comparison of the multitude of released LLMs would be helpful. The LLAMA2 model is not fine-tuned for the gesture domain. LLMs are known to perform well with prompting and in-context learning [150] to fine-tune the model. Therefore, there are many opportunities for further performance gain, such as whether fine-tuning these models for gesture generation is beneficial or if the models are fine-tuned for a different conversational downstream task.

9.3.2 Real Time Streaming Models

Each model in this work did not focus on algorithm run-time complexity and was not developed with latent-free streaming in mind. Most approaches described in this thesis rely on a windowing requirement for prediction. This means the model can only be applied in an offline environment and is unsuitable for real-time gesture generation. The window size of the generation defines the latency of these models. In future work, these approaches could be applied to a streaming context with a shorter buffer of frames to be predicted. This, however, leads to a potential increase in the compute required due to the increase in model calls. More work is needed to determine the performance would change when applying these methods to a streaming context. While Chapter 6 describes an increase in historical context as beneficial for gesture generation, it may also be beneficial for a model to predict a window of motion as the model has some future context, too. This future context means the model may be aware of when speech is ending and, therefore, when to reduce the gestures promptly or include turn-taking hints in the motion. This knowledge could be lost in a streaming context, so more work is needed to investigate the repercussions of a real-time streaming method.

9.3.3 Dyadic Interaction

This work has explored the concept of the main speaker and interlocutor in a dyadic interaction, predicting the main speaker from both dyadic speakers' speech. This inclusion of the second speaker may still be improved with future methods of data inclusion. However, future work may also explore the ability to generate both the main-agent and the interlocutor speakers simultaneously. By predicting all agents involved, a model using this approach may be particularly effective at turn-taking and mimicking.

References

[1] (2013). word2vec. Accessed on Jan 2024.

[2] (2024). Llanimation: Llama driven gesture animation. In *Computer Graphics Forum*, volume 43, page e15167. Wiley Online Library.

[3] Ahuja, C., Lee, D. W., Ishii, R., and Morency, L.-P. (2020a). No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1884–1895.

[4] Ahuja, C., Lee, D. W., Nakano, Y. I., and Morency, L.-P. (2020b). Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*, pages 248–265. Springer.

[5] AI, S. (2024). Bark. Accessed on May 2024.

[6] Alexanderson, S., Henter, G. E., Kucherenko, T., and Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, pages 487–496. Wiley Online Library.

[7] Alexanderson, S., Nagy, R., Beskow, J., and Henter, G. E. (2022). Listen, denoise, action! audio-driven motion synthesis with diffusion models. *arXiv preprint arXiv:2211.09707*.

[8] Alexanderson, S., Nagy, R., Beskow, J., and Henter, G. E. (2023). Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):1–20.

[9] Ao, T., Gao, Q., Lou, Y., Chen, B., and Liu, L. (2022). Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19.

[10] Ao, T., Zhang, Z., and Liu, L. (2023). Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18.

[11] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

[12] Balaji, Y., Min, M. R., Bai, B., Chellappa, R., and Graf, H. P. (2019). Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2.

[13] Bhattacharya, U., Childs, E., Rewkowski, N., and Manocha, D. (2021a). Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036.

[14] Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., and Manocha, D. (2021b). Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE.

[15] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

[16] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

[17] Bosker, H. R. and Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, 288(1943):20202419.

[18] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

[19] Campbell, A. and Rushton, J. P. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology*, 17(1):31–36.

[20] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

[21] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994a). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

[22] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994b). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

[23] Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486.

[24] Çatak, E. N., Açık, A., and Göksun, T. (2018). The relationship between handedness and valence: A gesture study. *Quarterly Journal of Experimental Psychology*, 71(12):2615–2626.

[25] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

[26] Chiu, C.-C. and Marsella, S. (2014). Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788.

[27] Chiu, C.-C., Morency, L.-P., and Marsella, S. (2015). Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer.

[28] Dabral, R., Mughal, M. H., Golyanik, V., and Theobalt, C. (2022). Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv preprint arXiv:2212.04495*.

[29] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

[30] Deichler, A., Mehta, S., Alexanderson, S., and Beskow, J. (2023). Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 755–762.

[31] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[32] Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.

[33] Drijvers, L., Özyürek, A., and Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, 39(5):2075–2087.

[34] Ennis, C., McDonnell, R., and O'Sullivan, C. (2010). Seeing is believing: Body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)*, 29(4):1–9.

[35] Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M., and Prieto, P. (2017). The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America*, 141(6):4727–4739.

[36] Esteve-Gibert, N. and Prieto, P. (2013). Prosody signals the emergence of intentional communication in the first year of life: Evidence from catalan-babbling infants. *Journal of child language*, 40(5):919–944.

[37] Ferstl, Y. and McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98.

[38] Ferstl, Y., Neff, M., and McDonnell, R. (2021). Expressgesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, 32(3-4):e2016.

[39] Freedman, N. (1977). Hands, words, and mind: On the structuralization of body movements during discourse and the capacity for verbal representation. In *Communicative structures and psychic structures: A psychoanalytic interpretation of communication*, pages 109–132. Springer.

[40] Games, E. (2023). High-fidelity digital humans made easy. Accessed on Jan 2023.

[41] Ghorbani, S., Ferstl, Y., and Carbonneau, M.-A. (2022). Exemplar-based stylized gesture generation from speech: An entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 778–783.

[42] Ghorbani, S., Ferstl, Y., Holden, D., Troje, N. F., and Carbonneau, M.-A. (2023). Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, volume 42, pages 206–216. Wiley Online Library.

[43] Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019). Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506.

[44] Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer.

[45] Graziano, M. and Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology*, 9:879.

[46] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., and Cheng, L. (2022). Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161.

[47] Habibie, I., Elgharib, M., Sarkar, K., Abdullah, A., Nyatsanga, S., Neff, M., and Theobalt, C. (2022). A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9.

[48] Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., Elgharib, M., and Theobalt, C. (2021). Learning speech-driven 3d conversational gestures from video. In *ACM International Conference on Intelligent Virtual Agents (IVA)*.

[49] Hannun, A. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[50] Hartmann, B., Mancini, M., and Pelachaud, C. (2005). Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*, pages 188–199. Springer.

[51] Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., and Sumi, K. (2018). Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86.

[52] Hensel, L. B., Yongsatianchot, N., Torshizi, P., Minucci, E., and Marsella, S. (2023). Large language models in textual analysis for gesture selection. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 378–387.

[53] Henter, G. E., Alexanderson, S., and Beskow, J. (2020). Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14.

[54] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

[55] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

[56] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[57] Hostetter, A. B. (2011). When do gestures communicate? a meta-analysis. *Psychological bulletin*, 137(2):297.

[58] Hübscher, I. and Prieto, P. (2019). Gestural and prosodic development act as sister systems and jointly pave the way for children's sociopragmatic development. *Frontiers in Psychology*, 10:1259.

[59] Iverson, J. M. and Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, 16(5):367–371.

[60] Jonell, P., Kucherenko, T., Henter, G. E., and Beskow, J. (2020). Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

[61] Jonell, P., Yoon, Y., Wolfert, P., Kucherenko, T., and Henter, G. E. (2021). Hemvip: Human evaluation of multiple videos in parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 707–711, New York, NY, USA. Association for Computing Machinery.

[62] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*.

[63] Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in southern italian conversation. *Journal of Pragmatics*, 23(3):247–279.

[64] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. (2018). Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

[65] Kim, J., Kim, J., and Choi, S. (2022). Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*.

[66] Kipp, M. (2005). *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.

[67] Kipp, M. and Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–8. IEEE.

[68] Kopp, S. and Wachsmuth, I. (2002). Model-based animation of co-verbal gesture. In *Proceedings of Computer Animation 2002 (CA 2002)*, pages 252–257. IEEE.

[69] Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1):39–52.

[70] Kozachenko, L. and Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16.

[71] Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130397.

[72] Kucherenko, T., Hasegawa, D., Henter, G. E., Kaneko, N., and Kjellström, H. (2019). Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104.

[73] Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., and Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 242–250.

[74] Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., and Henter, G. E. (2021a). A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21.

[75] Kucherenko, T., Nagy, R., Neff, M., Kjellström, H., and Henter, G. E. (2021b). Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762*.

[76] Kucherenko, T., Nagy, R., Yoon, Y., Woo, J., Nikolov, T., Tsakov, M., and Henter, G. E. (2023a). The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction*, ICMI '23. ACM.

[77] Kucherenko, T., Wolfert, P., Yoon, Y., Viegas, C., Nikolov, T., Tsakov, M., and Henter, G. E. (2023b). Evaluating gesture-generation in a large-scale open challenge: The genea challenge 2022. *arXiv preprint arXiv:2303.08737*.

[78] Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S. S., and Sheikh, Y. (2019). Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772.

[79] Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer.

[80] Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. *ACM Transactions on Graphics*, 29.

[81] Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 Papers*, SIGGRAPH Asia '09, New York, NY, USA. Association for Computing Machinery.

[82] Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., and Bao, L. (2021a). Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302.

[83] Li, R., Yang, S., Ross, D. A., and Kanazawa, A. (2021b). Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412.

[84] Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., and Zheng, B. (2022). Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer.

[85] Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., and Liao, R. (2024). Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*.

[86] Loehr, D. and Harper, L. (2003). Commonplace tools for studying commonplace interactions: practitioners' notes on entry-level video analysis. *Visual Communication*, 2(2):225–233.

[87] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[88] Lu, S. and Feng, A. (2022). The deepmotion entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 790–796.

[89] Lücking, A., Bergman, K., Hahn, F., Kopp, S., and Rieser, H. (2013). Data-based analysis of speech and gesture: The bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7:5–18.

[90] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35.

[91] McNeill, D. (1985). So you think gestures are nonverbal? *Psychological review*, 92(3):350.

[92] McNeill, D. (1992). Hand and mind. *Advances in Visual Semiotics*, page 351.

[93] McNeill, D. (2008). *Gesture and thought*. University of Chicago press.

[94] McNeill, D. (2011). *Hand and mind*. De Gruyter Mouton.

[95] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017). Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE.

[96] Metallinou, A., Yang, Z., Lee, C.-c., Busso, C., Carnicke, S., and Narayanan, S. (2016). The usc creativeit database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation*, 50:497–521.

[97] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[98] Moran, N., Hadley, L. V., Bader, M., and Keller, P. E. (2015). Perception of 'back-channeling'nonverbal feedback in musical duo improvisation. *PLoS One*, 10(6):e0130070.

[99] Morris, D. (1994). Bodytalk: A world guide to gestures.

[100] Nagymáté, G. and Kiss, R. M. (2018). Application of optitrack motion capture systems in human movement analysis: A systematic literature review. *Recent Innovations in Mechatronics*, 5(1.):1–9.

[101] Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1).

[102] Neff, M., Toothman, N., Bowmani, R., Tree, J. E. F., and Walker, M. A. (2011). Don't scratch! self-adaptors reflect emotional stability. In *International Workshop on Intelligent Virtual Agents*, pages 398–411. Springer.

[103] Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., and Richard, A. (2024). From audio to photoreal embodiment: Synthesizing humans in conversations. *arXiv preprint arXiv:2401.01885*.

[104] Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

[105] Nobe, S. (2000). Where do most spontaneous representational gestures actually occur with respect to speech. *Language and gesture*, 2(186):4.

[106] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., and Neff, M. (2023). A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum*, 42(2):569–596.

[107] OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2023). Gpt-4 technical report.

[108] Pang, K., Komura, T., Joo, H., and Shiratori, T. (2020). Cgvu: Semantics-guided 3d body gesture synthesis. In *Proc. GENEA Workshop. https://doi. org/10.5281/zenodo*, volume 4090879.

[109] Parizet, E., Hamzaoui, N., and Sabatie, G. (2005). Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica*, 91(2):356–364.

[110] Pelachaud, C. and Bilvi, M. (2003). Computational model of believable conversational agents. *Communication in multiagent systems: Agent communication languages and conversation policies*, pages 300–317.

[111] Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., and Poggi, I. (2002). Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 758–765.

[112] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[113] Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

[114] Pouw, W., Harrison, S. J., Esteve-Gibert, N., and Dixon, J. A. (2020). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, 148(3):1231–1247.

[115] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

[116] Qian, S., Tu, Z., Zhi, Y., Liu, W., and Gao, S. (2021). Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086.

[117] Quek, F., Bryll, R., Kirbas, C., Arslan, H., and McNeill, D. (2002). A multimedia system for temporally situated perceptual psycholinguistic analysis. *Multimedia Tools and Applications*, 18:91–114.

[118] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

[119] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

[120] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE.

[121] Rebol, M., Güti, C., and Pietroszek, K. (2021). Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 573–581. IEEE.

[122] Reithinger, N. and Klesen, M. (1997). Dialogue act classification using language models. In *EuroSpeech*. Citeseer.

[123] Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1530–1538.

[124] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

[125] Ruder, H., Ertl, T., Gruber, K., Günther, M., Hospach, F., Ruder, M., Subke, J., and Widmayer, K. (1994). Kinematics and dynamics for computer animation. In *From Object Modelling to Advanced Visual Communication*, pages 76–117. Springer.

[126] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

[127] Salem, M., Kopp, S., Wachsmuth, I., and Joublin, F. (2009). Towards meaningful robot gesture. *Human Centered Robot Systems: Cognition, Interaction, Technology*, pages 173–182.

[128] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

[129] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

[130] Strehl, A. and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

[131] Takeuchi, K., Hasegawa, D., Shirakawa, S., Kaneko, N., Sakuta, H., and Sumi, K. (2017a). 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*, pages 365–369.

[132] Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., and Sakuta, H. (2017b). Creating a gesture-speech dataset for speech-based automatic gesture generation. In *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19*, pages 198–202. Springer.

[133] Taylor, S., Windle, J., Greenwood, D., and Matthews, I. (2021). Speech-driven conversational agents using conditional flow-vaes. In *Proceedings of the 18th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '21, New York, NY, USA. Association for Computing Machinery.

[134] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. (2022). Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

[135] Thiebaux, M., Marsella, S., Marshall, A. N., and Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 151–158.

[136] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[137] Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

[138] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[139] Volkova, E., De La Rosa, S., Bülthoff, H. H., and Mohler, B. (2014). The mpi emotional body expressions database for narrative scenarios. *PloS one*, 9(12):e113647.

[140] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview.

[141] Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., and Dai, B. (2022). Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469.

[142] Waxer, P. H. (1977). Nonverbal cues for anxiety: an examination of emotional leakage. *Journal of abnormal psychology*, 86(3):306.

[143] Webb, R. A. (1997). *Linguistic features of metaphoric gestures*. University of Rochester.

[144] Wennberg, U. and Henter, G. E. (2021). The case for translation-invariant self-attention in transformer-based language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 130–140.

[145] Windle, J., Greenwood, D., and Taylor, S. (2022a). Uea digital humans entry to the genea challenge 2022. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 771–777.

[146] Windle, J., Matthews, I., Milner, B., and Taylor, S. (2023). The uea digital humans entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 802–810, New York, NY, USA. Association for Computing Machinery.

[147] Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022b). Arm motion symmetry in conversation. *Speech Communication*, 144:75–88.

[148] Windle, J., Taylor, S., Greenwood, D., and Matthews, I. (2022c). Pose augmentation: mirror the right way. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA. Association for Computing Machinery.

[149] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

[150] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., and Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

[151] Yang, Y., Yang, J., and Hodgins, J. (2020). Statistics-based motion synthesis for social conversations. In *Computer Graphics Forum*, volume 39, pages 201–212. Wiley Online Library.

[152] Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., and Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16.

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

[154] Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., and Henter, G. E. (2022). The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*, ICMI '22. ACM.

[155] Zeng, X., Wang, X., Zhang, T., Yu, C., Zhao, S., and Chen, Y. (2023). Gesturegpt: Zero-shot interactive gesture understanding and grounding with large language model agents. *arXiv preprint arXiv:2310.12821*.

[156] Zhang, F., Ji, N., Gao, F., and Li, Y. (2023). Diffmotion: Speech-driven gesture synthesis using denoising diffusion model. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, pages 231–242. Springer.

[157] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. (2022). Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

[158] Zhou, C., Bian, T., and Chen, K. (2022). Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 764–770.

[159] Zhou, Y., Barnes, C., Lu, J., Yang, J., and Li, H. (2019). On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.