Speech Communication 144 (2022) 75–88

Contents lists available at ScienceDirect

Speech Communication

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

Arm motion symmetry in conversation

Jonathan Windle [*], Sarah Taylor, David Greenwood, Iain Matthews

University of East Anglia, Norwich, UK

ARTICLE INFO

Keywords:
Speech-driven conversational agents
Motion symmetry
Conversational gesture analysis

ABSTRACT

Data-driven synthesis of human motion during conversational speech is an active research area with applications that include character animation, computer gaming and conversational agents. Natural looking motion is key to both perceived realism and understanding of any synthesised animation. Multi-modal speech and body-motion data is scarce and limited, so it is common to augment real motion data by mirroring the body pose to double the number of training samples. This augmentation is based on the assumption that a person's gesturing is not affected by handedness and that the reflected pose is plausible. In this study, we explore the validity of this assumption by evaluating the reflective symmetry of a speaker's arms during conversational exchanges. We analyse the left and right arm motion of 36 subjects during dyadic conversation and present the per-frame symmetry of the arm gestures. To identify temporal offsets caused by the presence of a leading hand, we compute the time lag between movements of the left and right arms. We perform a nearest neighbour search to test the validity of any mirrored pose. We also consider information theory to examine the information gain from mirroring the data. We implement a speech-to-gesture generative model to determine the efficacy of lateral mirroring techniques for data augmentation. Our findings suggest that both positional symmetry and left–right motion offsets vary from speaker to speaker. We conclude that data augmentation by mirroring is valid in certain cases when considering the mirrored pose as a new *virtual* identity, but that it should be carefully considered as a generic approach if the gesturing style and handedness of the original speaker is to be maintained.

1. Introduction

Co-speech gesturing contributes to language production and perception during conversation. Gesturing provides semantic context, and may be indicative of emotion and emphasis (Kendon, 1994; McNeill, 1985; Studdert-Kennedy, 1994; De Ruiter et al., 2012).

Gesturing in conversational speech serves many purposes including contributing to increased understanding, turn taking and listener feedback. Given the multi-modal nature of conversation, it follows that there is a co-dependency between speech and gesture.

Data-driven approaches for automatically driving body motion from speech is an active research area (Alexanderson et al., 2020a,b; Henter et al., 2020; Korzun et al., 2020; Yoon et al., 2020; Ginosar et al., 2019). Applications for these conversational agents include character animation, computer gaming and codec avatars (Bagautdinov et al., 2021). Such systems require multi-modal data comprised of motion captured body pose with a corresponding audio signal. These datasets are typically time-consuming and both financially and computationally expensive to capture, therefore, availability is scarce. A practised augmentation approach is lateral mirroring (Henter et al., 2020; Alexanderson et al., 2020b; Gong et al., 2021). This is to flip the left and right sided motion with each other.

While lateral mirroring effectively doubles the amount of training data, we raise the question of how natural and appropriate this augmented data is. Asymmetry is known to occur in pose from physical body constraints and gesture style types. We present a study of frame-by-frame position and temporal characteristics to investigate if this mirroring produces natural speaker-dependent movement. This study is not only relevant to gesture generation and data augmentation, it provides an insight into arm symmetry during conversation, providing greater understanding for all relevant fields of research such as gesture recognition and gesture behaviour. Finally we consider the use of this method of analysis as a means to evaluate performance of data-driven synthesised motion.

2. Related work

We present a review on works relating to speech gesturing, body motion datasets, methods for speech-driven body animation, and techniques for data augmentation used by these methods.

[*] Corresponding author.
E-mail address: j.windle@uea.ac.uk (J. Windle).

https://doi.org/10.1016/j.specom.2022.08.001
Received 22 February 2022; Received in revised form 17 June 2022; Accepted 16 August 2022
Available online 29 August 2022
0167-6393/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2.1. Arm gesture and symmetry

Neither speech nor gesture alone allows a speaker to communicate to their full efficiency. Removing either of these modalities leads to a reduction in semiotic versatility (Wagner et al., 2014) and communicative understanding (Hostetter, 2011). One reason for this is that each modality represents certain information better than the other. For example, hands might better describe shape or direction by providing visual cues. The gestures that form these cues may or may not be symmetrical, and this may, in part, depend on the particular shape or direction being described.

Environmental conditions contribute a great deal 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, a person may gesture more actively as they have more space. Proximity and facing direction of the conversational partner within the environment will also effect the extent and type of gesturing. If conversation is taking place 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 (Drijvers et al., 2018).

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 that takes place 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 (Hostetter, 2011). In particular, extroversion is positively correlated with representational gesture production, which might be due to extroverted people having high amounts of energy in social situations and therefore gesturing regardless of communicative effect.

McNeill defined a gesture space (McNeill, 2011), stating that the majority of 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 on the semantic functions of gesture types (McNeill, 2011). They categorised gestures as either emblematic, iconic metaphoric, deictic or beat: *Emblematic gestures* bear a conventionalised meaning; *Iconic gestures* resemble a certain physical aspect of the conveyed information; *Metaphoric gesture* is an Iconic gesture resembling abstract content; *Deictic gesture* point out locations in space; and *Beat gestures* are simple and fast movements of the hands commonly synchronised with prosodic events in speech (Pouw et al., 2020). However, in practice a gesture may perform many semantic functions, and it has instead been proposed to treat each gesture category as a dimension on which gestures load to differing degrees (McNeill, 2008).

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, while representational gestures in right-handed speakers had a right-handed preference while left-handed speakers did not have a hand preference (Çatak et al., 2018). There is an association between gestural handedness and the emotional dimensions of pleasure and arousal. Kipp and Martin (2009) found 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.

We have reviewed works that analyse gestural symmetry during conversation, however, these works are limited by the data used. Data is often observed manually from video (McNeill, 2011) or limited to a few speakers worth of data (Kipp and Martin, 2009). This reveals a limitation in current studies that we aim to address.

2.2. Body motion data and limitations

Conversational body motion data is needed for performing analysis of gestural symmetry, and for training generative speech-to-body animation models. However, the availability of such data is scarce and issues commonly arise during the data collection process resulting in data that is noisy, unnatural or lacking in quantity. Ideally, motion data is recorded using optical motion capture systems that track retroreflective markers on the speaker. The 3D position of each marker is triangulated between multiple cameras. Issues regarding marker jitter, swapping and occlusion often require motion captured landmarks to be manually cleaned. Generally, motion capture is both financially and computationally expensive to collect, but can result in high-quality performance capture. An abundance of body motion data is available if we use video as a data source. However, extracting 3D key points from a single video feed is challenging, often leading to noise and inaccurate depth estimation. This causes a trade off between data quality and quantity.

A dataset that was collected for data-driven synthesis of motion is the Trinity dataset (Ferstl and McDonnell, 2018). It contains 244 min of speech and motion data that was recorded using 20 Vicon cameras, and the motion data is high quality and accurate. However, the Trinity dataset contains only one male speaker producing monologue speech. Gestural motion and symmetry varies across speakers and therefore it is difficult to draw conclusions from a single speaker. Since the speech is monologue, the gesturing that relates to listener understanding and turn taking is also not captured.

Social interaction is not limited to conversation. Joo et al. (2015) presented a dataset that contains social interactions during game scenarios, together with a description of the Panoptic Studio that was used for the capture. The capture system is comprised of a large dome structure containing 480 VGA cameras for video capture, each with calibrated frame timers and positions. Using the known positions of the cameras and 2D pose estimation software, 3D poses are accurately predicted. While this system produces clean motion capture, it is both financially and computationally expensive. With 480 cameras, the datarate is approximately 29.4 Gbps, requiring a large amount of processing power and storage to manage such quantities of data. While this dataset provides multiple speakers' motions, the scenarios recorded are not natural conversations but instead social interactions during games, which will affect the types of gestures that are produced.

There is an abundance of video data available that contains conversational interaction. This is exploited by Ginosar et al. who extracted monologue speech and motion data from videos of talk show hosts, lecturers and televangelists (Ginosar et al., 2019). The videos are shot from a single view and therefore only 2D keypoints were extracted. Further work estimated 3D keypoints for this dataset (Habibie et al., 2021), however the result is noisy and includes errors in depth prediction.

The main limitations of existing motion captured data is the number of identities and lack of natural dyadic conversation. The Talking with Hands dataset presented by Lee et al. mitigates these limitations and is selected for our analysis (Lee et al., 2019). This dataset is described in Section 3.

2.3. Speech-driven body animation

Embodied conversational agents describe both human-like robots and animations that aim to employ human-realistic verbal and non-verbal communicative modalities. Data-driven approaches for automatically driving body motion from speech is an active research area (Alexanderson et al., 2020a,b; Henter et al., 2020; Korzun et al., 2020; Yoon et al., 2020; Ginosar et al., 2019). These approaches aim to estimate a speakers pose, typically represented by a sparse set of skeleton joints, from their corresponding speech audio signal.

Recent approaches for data-driven motion synthesis typically involve deep learning (Alexanderson et al., 2020a,b; Henter et al., 2020; Korzun et al., 2020; Yoon et al., 2020; Ginosar et al., 2019). Their success is highly dependent on the data used to train them. For instance, small datasets or those lacking diversity can lead to models not generalising well or overfitting to training data (Perez and Wang, 2017). Data quality is also important as a model can only learn to be as good as the training data, and inaccurate or poorly labelled data will cause the model to learn incorrect information. To mitigate the limited amount of available body motion data, it is common to augment the dataset. It is key to ensure that the quality of the data is not compromised during augmentation, and the focus of our work is to explore this.

2.4. Data augmentation

Data augmentation are techniques used to increase the amount of data by adding slightly modified copies of real data or created synthetic data from existing data. The most common technique for this is through *data warping* defined in Perez and Wang (2017) as an approach to directly augment the input data to the model in *data space*. Augmentation approaches vary depending on the data type and the problem domain.

When working with image data it is common to apply simple transformations on each image. These include flipping, scaling, rotating, translating, noise injection and colour space transformation (Shorten and Khoshgoftaar, 2019). While flipping, scaling, rotating and translating are all possible to 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 if joint angles are used to 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 of these data augmentation approaches that still produces potentially valid human body motion. It is our goal to determine in what cases this augmentation is a valid approach.

3. Data and pre-processing

This study performs an analysis on the body motion from the Talking with Hands dataset (Lee et al., 2019). The dataset consists of 16.2-million frames of motion at 90 Frames Per Second across 50 different speakers during dyadic conversation. Unfortunately not all of this data is currently publicly available and 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). We removed any non-conversation segments of the data (e.g. T-Pose sequences) prior to performing the analysis.

The dataset provides a set of 3D skeleton joint keypoints for each frame. Our study focuses on the 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 arms from motion originating from the spine such as leaning forwards and backwards. This allows us to evaluate translations made by motion generated from the arms independently of the rest of the pose. The coordinate system utilised in this paper is as follows:

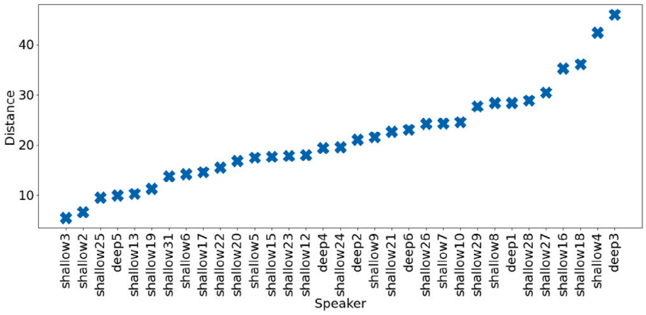

Fig. 1. Euclidean distance between mirrored right arm and the left for each speaker.

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

We also use a consistent colour scheme through all figures to represent each forearm. Cyan depicts the right forearm and Blue depicts the left forearm.

4. Mean pose symmetry

We first evaluate the symmetry of the mean poses for each speaker, 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. We then project the right arm to the space of the left arm by laterally mirroring (along the y-axis). To evaluate the arm symmetry, the Euclidean distance between all joints in the left arm and projected right arm are calculated. The lower this distance, the closer the two arms are to each other, which is indicative of a more symmetrical pose.

We show the range of symmetry in Fig. 1. We observe 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.

From the 36 speakers we select the two with the highest and two with the lowest Euclidean distance, representing the subjects exhibiting the least and most arm symmetry in their mean pose. We visualise the level of symmetry by overlaying a perspective projection of the mirrored right arm onto the left arm. Fig. 2 shows this projection from both a frontal and side view for each of the four speakers. There is clear 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 their body, whereas the right arm is pointing away from their body, towards the camera. Shallow4 orients their right wrist away from their body while their left wrist is pointing towards their body. At the other extreme, Shallow3 and Shallow2 show good symmetry (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 largest differences between the arm positions is observed in the side view, whereby each of the left arms are positioned further forward than the right arms. While this observation is more prominent on the two most asymmetric speakers, it holds for each of the speakers in Fig. 2.

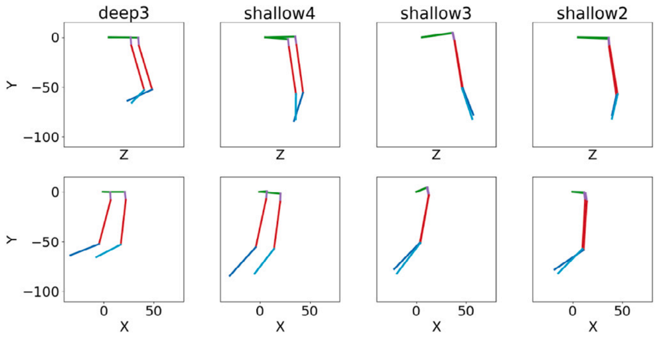

Fig. 2. A projection of the mean pose for four speakers. In each case, the right arm (cyan forearm) has been mirrored and overlaid onto the left arm (blue forearm). Top row: front view. Bottom row: side view. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Spatial symmetry

The mean pose analysis in Section 4 provides an indication of the symmetry of a speaker's most frequent (or neutral) arm positions. However, it does not explain whether the *motion* of the arms is similar or symmetrical. In this section we investigate whether the observed asymmetry is an effect of a speaker's tendency to gesture more on one side than the other, and whether the arms occupy symmetrical gesture spaces. We use 3D keypoints to gather statistics regarding the arm motion of each speaker, discuss the speakers' motion ranges and traits, and define their data-driven gesture spaces.

5.1. Full arm motion range

To reveal whether a similar amount of energy is exerted by the left and right arms, we measure the deviation from the mean pose. We independently compute a frame-wise Euclidean distance from each arm to its respective mean pose. These statistics are calculated over all arm joints.

Fig. 3 shows the results for the four speakers that were identified as exhibiting the least and most symmetry in their mean pose in Section 4. It is evident that the amount of deviation from the mean pose in the left and right arms is not significantly different if we consider the poses that fall within the whiskers, which represent those within $1.5\times$the interquartile range beyond the first and third quartiles. However, the outliers do appear somewhat asymmetrical for speakers Deep3 and Shallow4, each displaying greater divergence from the mean with the right arm compared to the left. Shallow3 and Shallow2 exhibit more symmetrical outliers, indicating that a similar amount of space is encompassed by both arms during these infrequent, larger 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 for the latter two.

Fig. 4 shows a frontal perspective projection of each speaker's arm pose taken over all of their respective conversations at 1 s intervals. We observe variability 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 wider poses than the left, but the left arm produces higher gestures than the right. Deep3 and Shallow4 both raise their right hands more frequently than their left, suggesting increased expressiveness in that dominant hand. From these plots it is evident 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 (McNeill, 2011).

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* (McNeill, 2011). Given the variability between the spaces occupied by each speaker's arm and the frequency in which they extend into their respective peripheral spaces, we propose a data-driven approach to defining speaker-specific gesture spaces. We use statistics to define a speaker's *common gesture space* and *extreme gesture space*. The *common 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 of a single standard deviation of the mean pose, away from the body.

Using our definition, we partition the data 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. We again compute the per-speaker distance from the mean pose for each partition, and the results can be seen in Fig. 5. For the majority of the speakers, the distances from the mean for gestures within the common gesture space are similar for both left and right arms (Fig. 5, bottom row). An exception is the speaker Deep3 in which the range is larger for the right hand. The greatest differences between the left and right arms are observed in the extreme gesture space (Fig. 5, top row), particularly for the asymmetric speakers Deep3 and Shallow4. In each case, one hand diverges further from the mean than the other.

For Deep3, we observe that the left arm is more active in the extreme gesture space than the right, and the reverse is true in the common gesture space. We plot the perspective projection of all poses corresponding to the extreme and common gesture spaces in Fig. 6 for each speaker to visualise these differences. The top row reveals that the right arm of Deep3 does contribute to gesturing in the extreme gesture space, but the poses of the left arm are wider, 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.

Fig. 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 their left arm and raises the right and Shallow2 extends their 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 5.1. The percentage of poses within each gesture space as shown in Fig. 6 impacts the effect of mirroring. Given more symmetry being found in the common gesture space, if a speaker has a lower use of the extreme gesture space, the potential negative impact of mirroring is reduced.

5.3. Self-adaptor traits

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

Fig. 7 shows the poses of speaker Shallow25 who frequently touches their left hand to their right forearm. The reverse, right hand touching left forearm, is not present in any of the 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 has been found to significantly impact the perceived level of neuroticism of a speaker (Neff et al., 2011), and the effect of reversing the handedness of the behaviour is not well established.

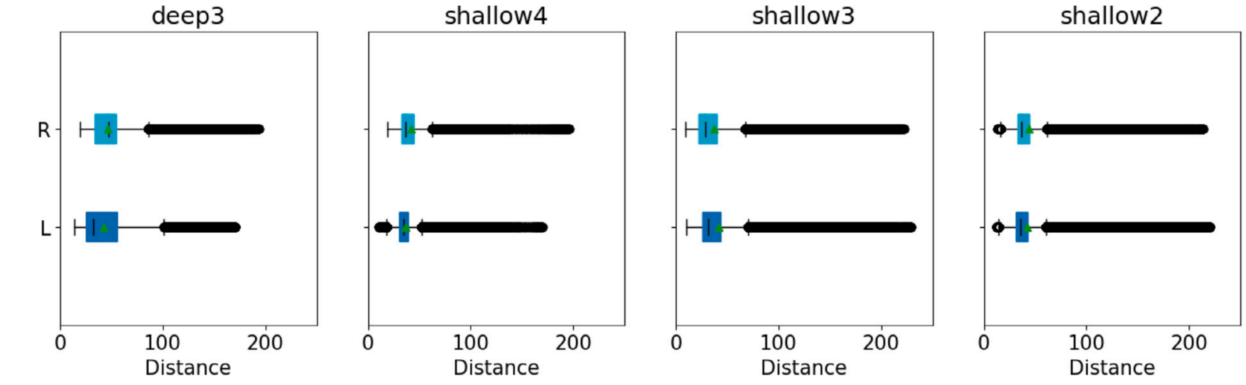

Fig. 3. Per-frame Euclidean distance from the mean of each arm for four speakers. L = Left arm, R = Right arm.

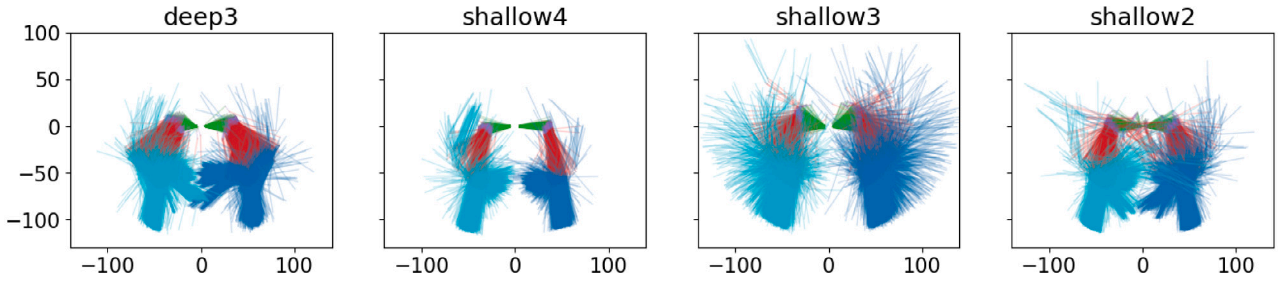

Fig. 4. A frontal perspective projection of all poses per speaker, taken at one-second intervals.

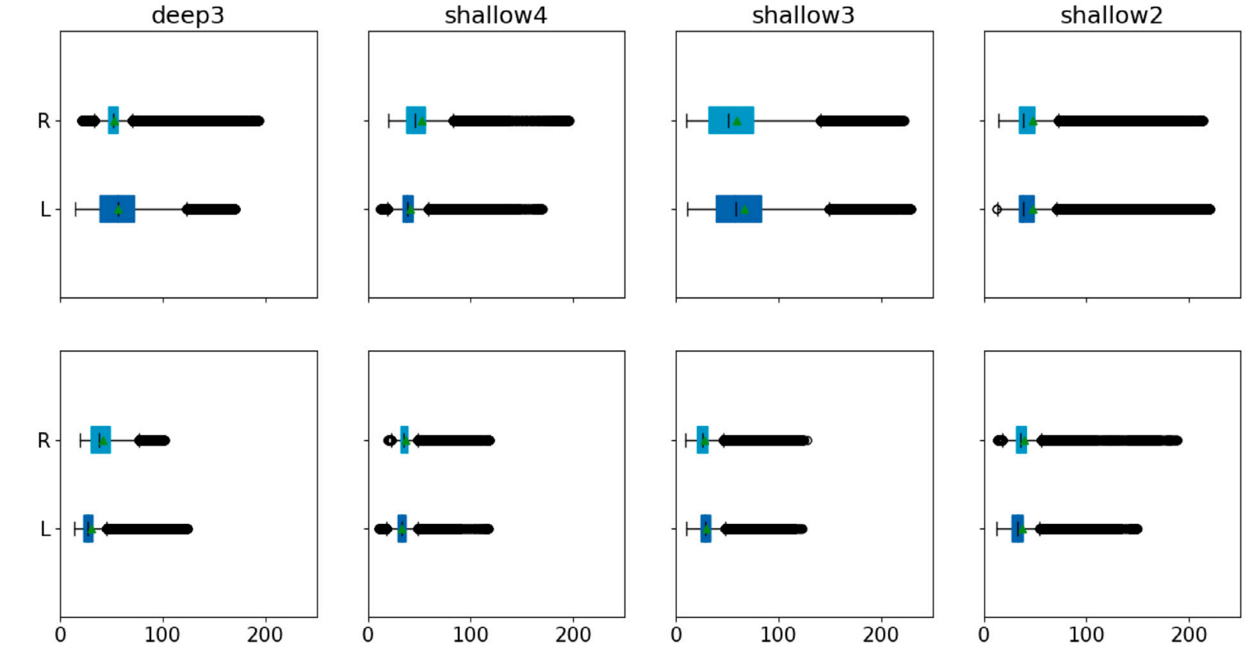

Fig. 5. Per-frame Euclidean distance from the mean of each arm, split into *Extreme Gesture Space* (Top) and *Common Gesture Space* (Bottom). L = Left Arm. R = Right Arm.

6. Symmetry in gesture types

When considering the impact of symmetry, the type of gesture being performed may be important. We reviewed a number of speech-motion pairs to determine what impact may occur from the gesture being mirrored. We cannot generalise from these few examples, but instead should be useful to consider specific aspects of gesture suitable when mirrored.

We observe that beat gestures are often performed by a single hand. Fig. 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 we do not know the dominant hand of this speaker, we observe some trends similar to those of Çatak et al. (2018) where one hand is performing the gesture.

Çatak et al. (2018) suggest that representational gestures are performed by a dominant hand for right-handed speakers but no dominant hand was found in left-handed speakers. While we cannot compare handedness in this work, we do consider that the context of the gesture can determine the symmetry of the gesture performed. Fig. 9 shows a

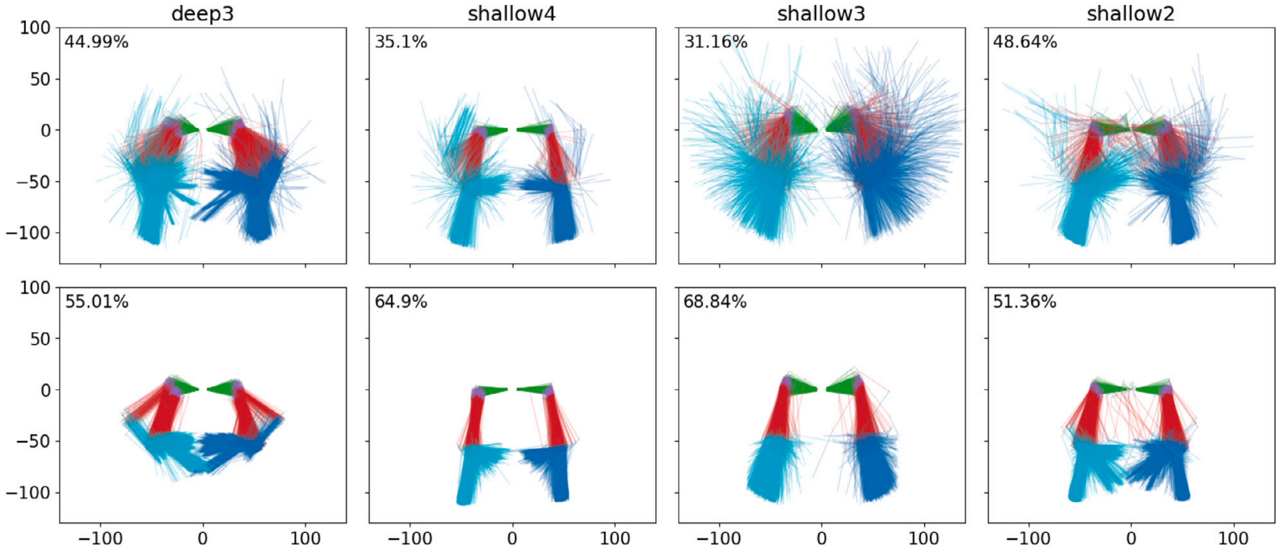

Fig. 6. Frontal projections of all poses from four speakers at one-second intervals, split into *Extreme Gesture Space* (Top) and *Common Gesture Space* (Bottom). Percentage in the corner denotes the percentage of poses belonging to the respective gesture space for the respective speaker.

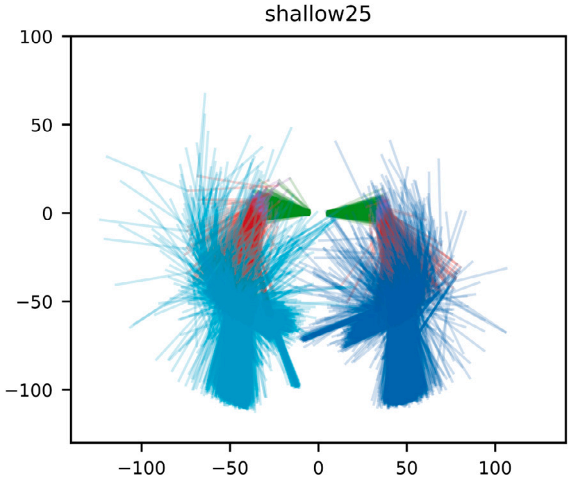

Fig. 7. Shallow25 poses taken at one-second intervals. This speaker exhibits self-adaptor movements whereby the left hand frequently touches the right forearm.

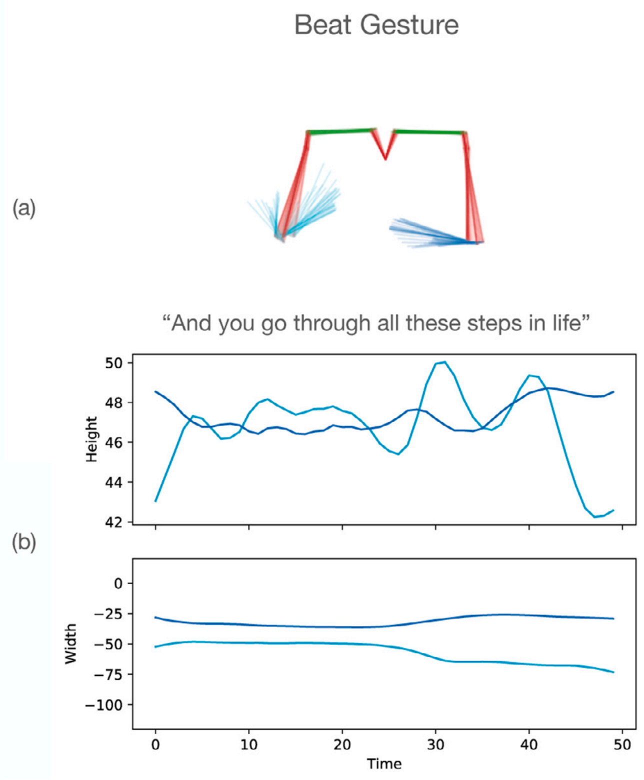

Fig. 8. A speaker performing a beat gesture. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left–right) and height (up-down) directions.

metaphoric gesture being performed, mimicking the use of an umbrella. It is typical for a person to only use a single hand while using an umbrella and 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. Fig. 10 is a gesture performed by another speaker, however, they are referring to moving a heavy object onto a table. Typically moving heavy objects in the manner outlined in the speech would require two hands and 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.

With regards to directional Deictic gestures, we observed that often the hand closest to the direction was used. Fig. 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 the end in which they are stood and a small movement of the left arm is used to refer to this. Fig. 11 time plot shows a clear spike as the right arm moves to the peak directional gesture, the left arm is lowering, suggesting asymmetry.

We describe some examples of symmetrical and asymmetrical poses and their associated gesture type. We find that in some cases a mirrored, symmetrical pose may well still portray the same meaning. A good example of this is when a metaphoric action requires the use of both hands to lift something. However, in the example Deictic gesture this would not continue to make sense when performed in the same location.

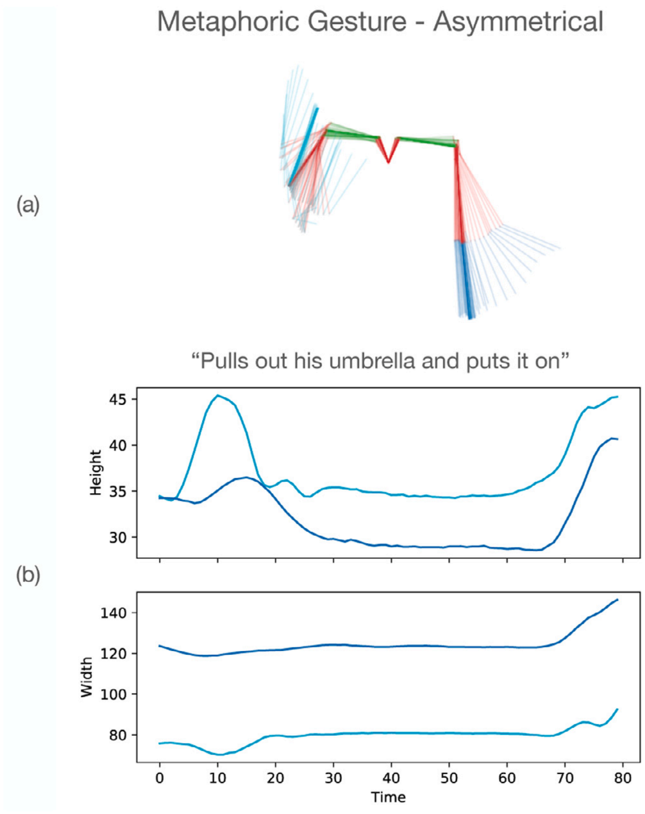

Fig. 9. A speaker performing a metaphoric gesture. In this case, the gesture is **asymmetric** due to context. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left–right) and height (up-down) directions.

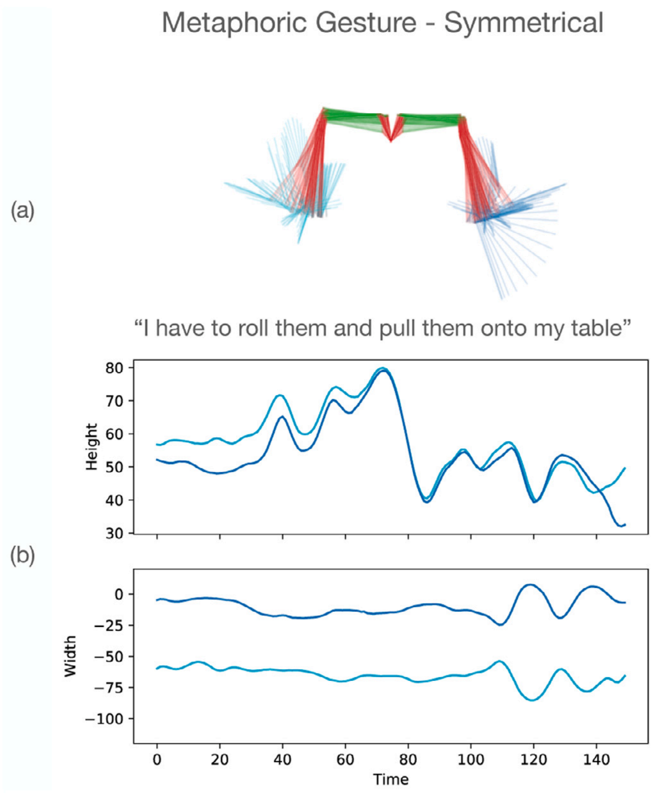

Fig. 10. A speaker performing a metaphoric gesture. In this case, the gesture is **symmetric** due to context. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left–right) and height (up-down) directions.

7. Mirrored pose validity

For some machine learning approaches, the goal of laterally mirroring body pose is 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. In this section we visualise and quantify mirrored pose validity using this definition.

We perform a nearest neighbour search of 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. We focus on the poses that fall within the extreme gesture space, defined as any pose outside of one standard deviation away from the mean pose (Section 5.2). We first present a visualisation of the nearest neighbours in Fig. 12. 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 Fig. 12 we observe 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 independently from the right arm. For both 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 not possible to match these in the original data.

Fig. 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 5), we observe that, in the extreme gesture space, they produce similar gestures with both left and right hands.

8. Temporal symmetry

Our analysis so far has considered only frame-wise statistics, which does 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, there may exist an asynchrony, or a temporal shift, between the motion of the two arms, particularly if the speaker gestures with a dominant hand. In this section we perform a cross-correlation analysis 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 (≈4.5 s), centred at frame t. For each windowed frame in the left hand data, $t = 0, \dots, T$, we slide the window over the right arm data from frames $t - 200$ to $t + 200$ and compute the correlation coefficient between the segments. A larger window size was not used since we observed that a lag longer than 2.2 s 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. We independently run the analysis on each directional axis and the Euclidean distance to the mean pose of each hand, and the results can be seen in Fig. 14.

Although Shallow2 has a relatively symmetrical gesture space (Fig. 4), Fig. 14 clearly shows a dominant hand in the temporal domain. This indicates that this speaker leads with their right hand with a mean

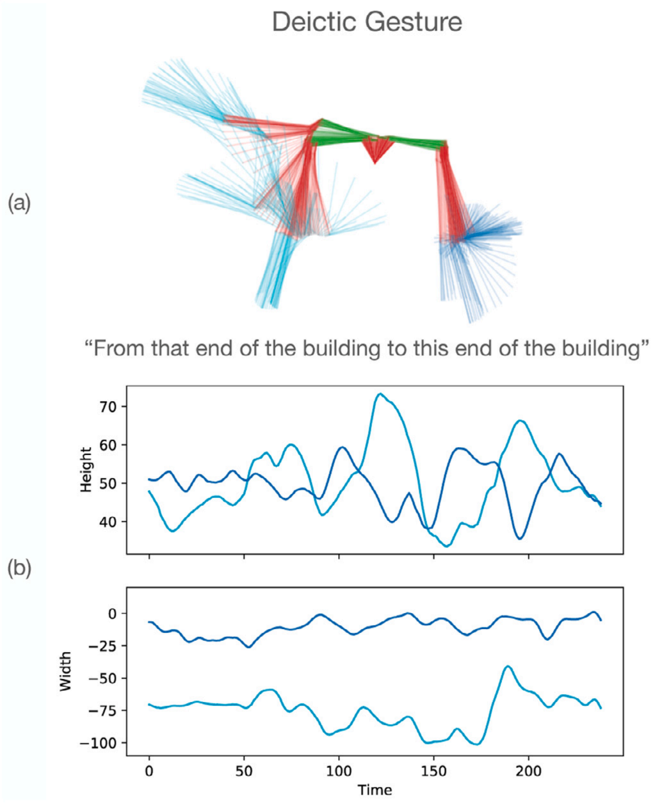

Fig. 11. A speaker performing a Deictic gesture. (a) shows each pose formed over the sequence with the sentence being said below. (b) shows the positions of each wrist in both lateral (left–right) and height (up-down) directions.

offset of 28 frames (≈0.31 s) when considering the distance from the mean pose. If we consider the individual axes, we observe that the right hand leads in all cases, and in the X and Y axes the offset is greater than 0.5 s. This suggests that, although a symmetrical pose is formed, there is a temporal offset between hands achieving this pose.

It is evident that other speakers' motions are more symmetrical and very small temporal offsets were found. Shallow3 in Fig. 14 is an example where the mean offset does not exceed a mean of 17 frames (0.19 s) in any axis.

9. Mutual information

In this Section we explore 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? We compute the mutual information between the original data and its mirrored counterpart to reveal the dependence between the two distributions.

We measure Normalised Mutual Information (NMI) (Strehl and Ghosh, 2002) on a per-speaker, per-axis basis at the wrist joint. NMI is computed using the following:

$$NMI(X,\tilde{X}) = \frac{I(X,\tilde{X})}{\sqrt{H(X)H(\tilde{X})}} \tag{1}$$

where $I(X,\tilde{X})$ is the mutual information between the original and mirrored data, 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 (Kozachenko and Leonenko, 1987).

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 Fig. 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, our 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 indicates 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 missing from the data. By removing or including potentially incorrect arm movement on one side, you are losing important information or introducing large amounts of uncharacteristic information.

10. Generative modelling

To further support our findings, we train a Long Short-Term Memory (LSTM) model on different splits of data and use various augmentation settings to map from speech to body pose. We aim 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.

10.1. Motion representation

Of the 36 speakers released, only 18 have both audio and motion capture available and therefore we use this subset. Mocap was down-sampled to 30fps to ensure realistic motion was maintained, 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. The global position for each speaker is inconsistent and 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 which are standardised by subtracting the mean pose and dividing by the standard deviation computed over all speakers across all training sequences.

10.2. Audio representation

Mel Spectrograms or Mel Frequency Cepstral Coefficients (MFCCs) are often used in speech-to-motion pipelines (Habibie et al., 2021; Alexanderson et al., 2020a; Taylor et al., 2021). We instead use a model trained using a multi-task learning framework that is comprised of 12 regression tasks. (PASE+) (Ravanelli et al., 2020) 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 44.1 kHz to 16 kHz.

10.3. Generative model

Using an LSTM-based model, we train using a single motion frame's worth of audio (33 ms) to predict a frame of motion. To ensure motion is speaker-specific, we condition the speech using a learned feature vector that encodes a speaker's identity. This learned feature vector should adequately associate the speaker and their gesturing style. With this learned feature vector, it should allow us to introduce a speaker's potentially uncharacteristic mirrored motion to the model, without affecting the gesturing style of the 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.

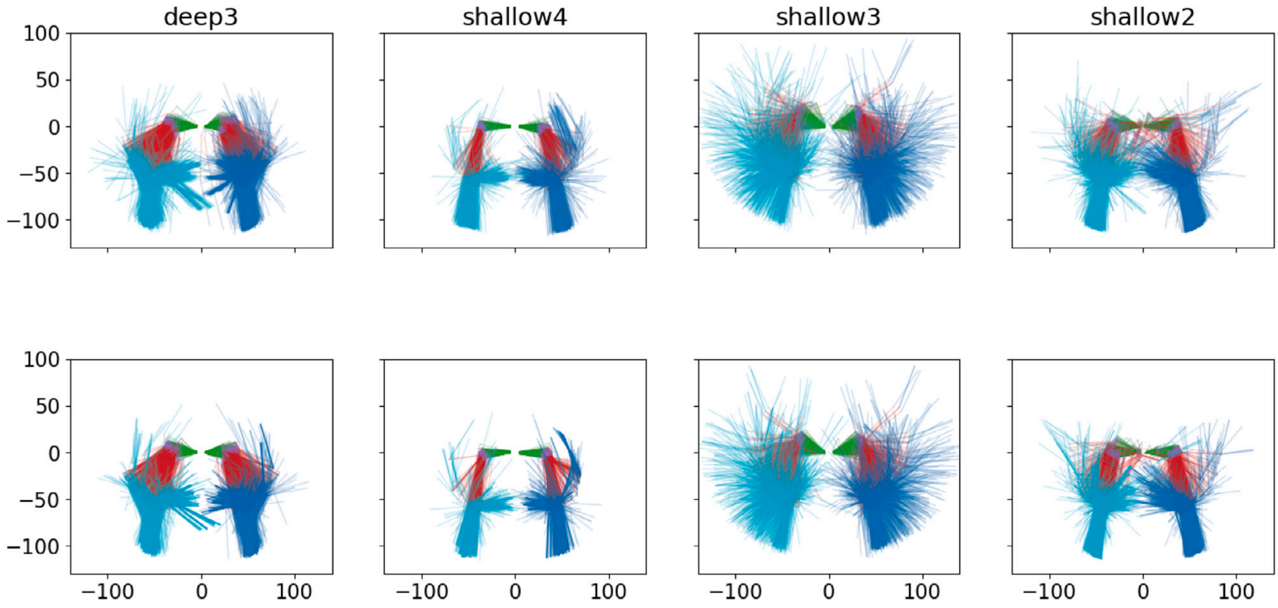

Fig. 12. The frontal 2D projections of mirrored poses that are at least 1 standard deviation away from their mean pose (top) and the closest respective mean poses from the original data (bottom).

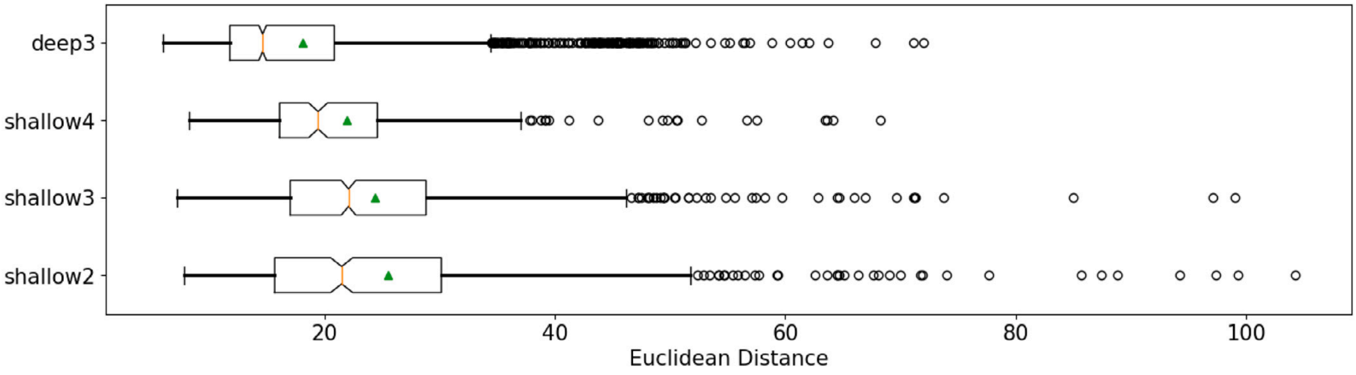

Fig. 13. Euclidean distance between mirrored arm position and the closest pose from the original data for poses in the extreme gesture space.

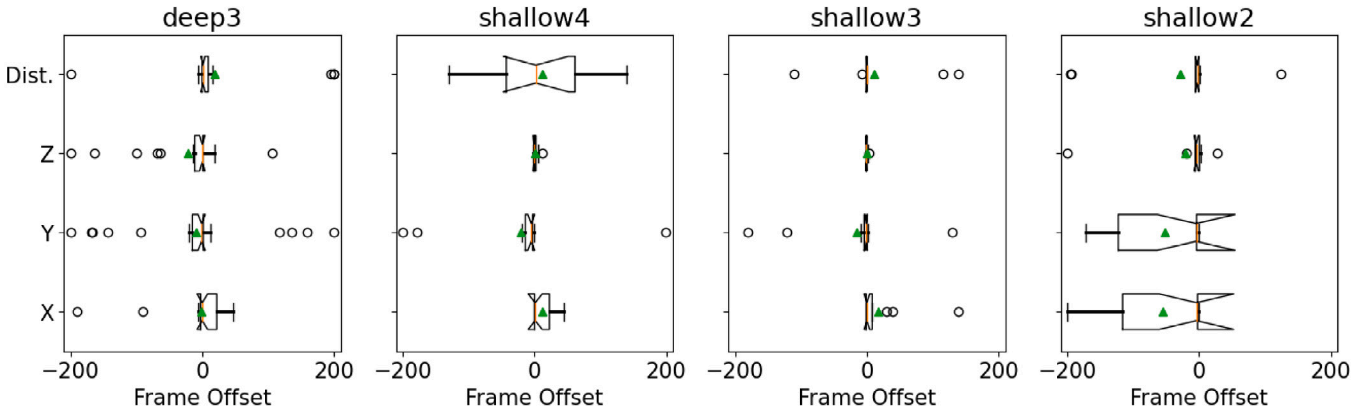

Fig. 14. Cross correlation analysis between left and right hand position for each directional axis and Euclidean distance from the mean.

10.4. Training procedure

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

We use a multi-term loss function. We minimise the position values as an L_2 loss on joint positions and also an L_2 loss on joint velocity and acceleration. Introducing the velocity and acceleration allows the

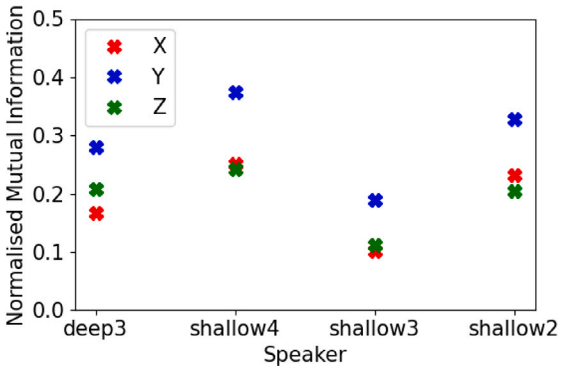

Fig. 15. Normalised Mutual Information per-speaker, per-axis measured between the original and mirrored wrist joints. Lower values represent a higher degree of independence.

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, we include an L_1 loss on bone length. The final loss L_c is computed as:

$$\begin{aligned}
&L_p = L_2(y, \hat{y})\\
&L_v = L_2(f'(y), f'(\hat{y}))\\
&L_a = L_2(f''(y), f''(\hat{y})) \qquad (2)\\
&L_b = L_1(y_{lengths}, \hat{y}_{lengths})\\
&L_c = L_p + L_v + L_a + L_b
\end{aligned}$$

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. The term L_p is representative of 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.

10.5. Experimental setup

We train the same model architecture on each of the settings defined as follows:

All Data. We form a baseline 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% We train a model using this reduced data to enable us to compare the effect of doubling the size of the training set by augmentation versus adding additional ground truth data.

Mirrored Same Identity. We augment 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 allows us to determine the impact of introducing uncharacteristic motion for a specific speaker.

Mirrored Virtual Identity. We augment the *Half Data* training set by laterally mirroring the pose at each frame. During training, we assign a **new** virtual identity label to the mirrored data. This allows us to determine if adding motion that could be considered characteristic for a different speaker aids or hinders performance.

All Data Mirrored Virtual Identity We additionally train our model on all available training data plus the laterally mirrored augmentation. As in the *Mirrored Virtual Identity* setting, the augmented sequences are assigned new virtual identity labels. This represents our optimal setting.

10.6. Results

We continue to use motion characteristics to evaluate performance. These include positional pose plots, distances from the mean pose and temporal handedness. Our analysis should provide an indication of how characteristic the predicted motion is and whether the introduction of motion has had an impact on performance. We follow the same procedure as in Section 3 and translate per frame so that the midpoint of the left and right shoulders and centred on the origin.

10.6.1. Using the same identity

We observe two key findings; the mirrored data produced far more muted and symmetrical motion than desired.

We found the movement generated to be positionally symmetrical over the whole pose but particularly with arm movements. Fig. 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, there are more asymmetrical poses which are closer to the characteristics performed in the ground truth.

Fig. 16b indicates the amount of 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 when compared to motion generated from half of the data and the ground truth. This is indicative of 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. Fig. 16 shows a strong correlation between the left and right wrists moving at a temporal lag offset of ±1 frame. When compared to the ground truth, this high temporal symmetry is very uncharacteristic of the speaker.

10.6.2. Augmenting with a virtual identity

With a detrimental effect of including mirrored data under the same identity, we examine the effects of including mirrored data under a virtual identity (*Mirrored Virtual Identity*). We identified improvements in generated motion quality varied between speakers, however, we did not find a negative impact on performance. Mirroring with a virtual identity was found to be competitive with a model trained with all of the available data, often improving positioning, adding some 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 Fig. 17. The distribution of distances from the mean pose shown in Fig. 17b decreases from half of the data and half mirrored as a virtual identity. We also note the poses in Fig. 17a appear closer to the predictions using all of the data and ground truth. By seemingly lowering the arms more often than the generated motion using half of the data, this supports the hypothesis that the addition of mirrored data as a virtual identity can be competitive with a model including all data.

11. Discussion

We discuss our findings on arm symmetry during dyadic conversation and its impact on lateral mirroring for body motion data augmentation. We present the potential issues that could arise, and when it would and would not be a suitable data augmentation approach.

If lateral mirroring is used for body data augmentation, caution should be taken if gesturing style and handedness of the speaker are to be preserved. From our analysis it is clear that mirroring can result in both valid poses and dynamics for certain 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, for these highly symmetrical speakers, the information gained from mirroring the arm motion might be minimal. In the majority of cases, the speakers did not move symmetrically, and the mirrored data would not reflect

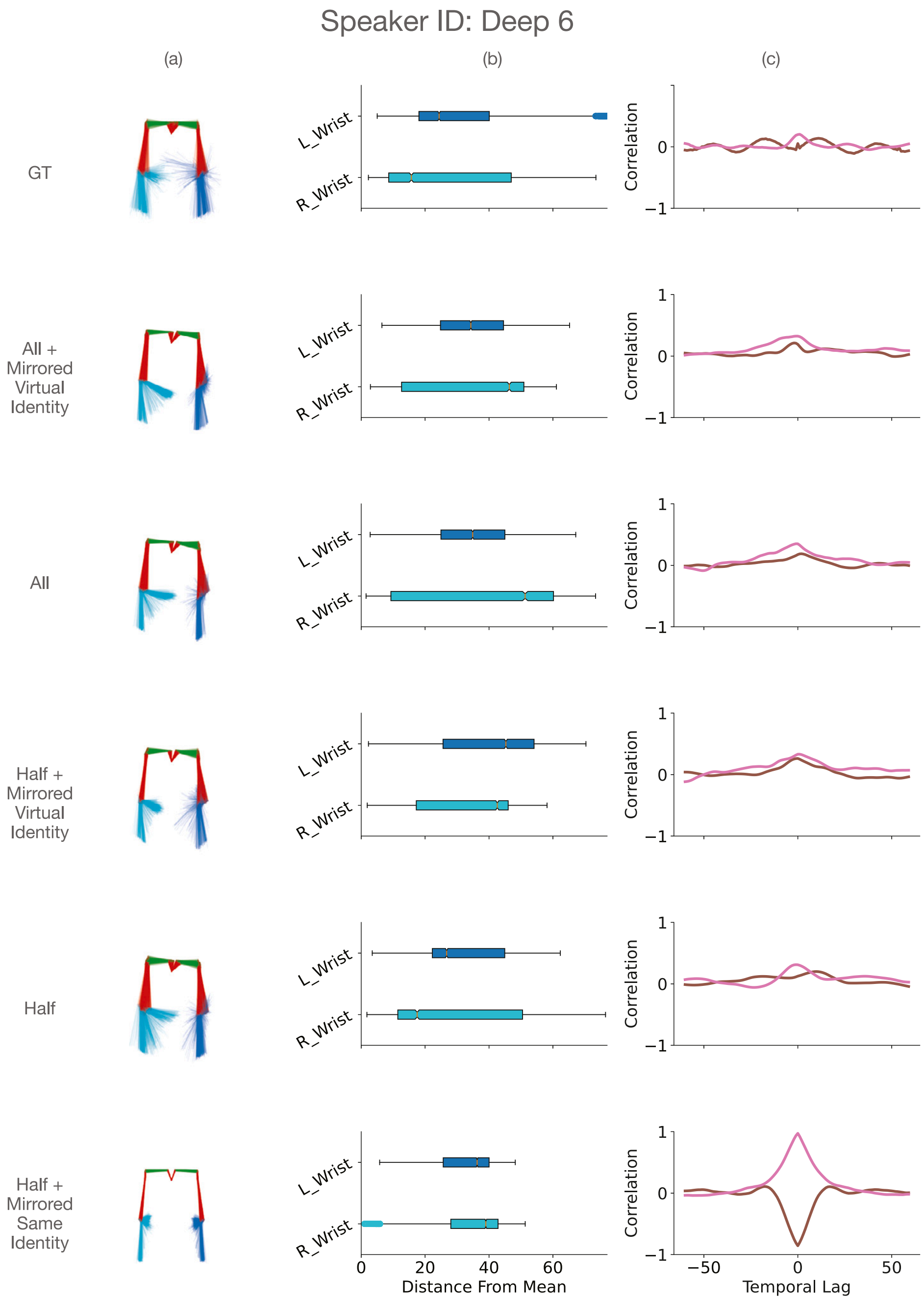

Fig. 16. A comparison for a single speaker's generated motion showing detrimental impact of including mirrored motion under the same identity. Each row corresponds to a different data split used. Column (a) contains the orthographic projection of a pose at every second in the sequence. Column (b) shows 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the true 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 our generative modelling, a naive mirroring implementation did not predict characteristic or plausible motion and was found to be detrimental to model performance. We instead suggest the use of a new *virtual identity* for the mirrored poses. We found that the amount of improvement was speaker-dependent. We speculate this may be due to the non-uniform distribution of data across the speakers. As the dataset used has *shallow* and *deep* speakers, 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 simply not present in the training data. We speculate the improvement may be due to an increase in generalised characteristics common across all speakers. If the aim is to preserve the gesturing style and handedness of the original speaker, 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 Fig. 7 is an example of an asymmetrical self-adaptor trait that is characteristic to 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 in the same manner. If this stylistic motion was to be maintained, simply mirroring the body pose would not suffice.

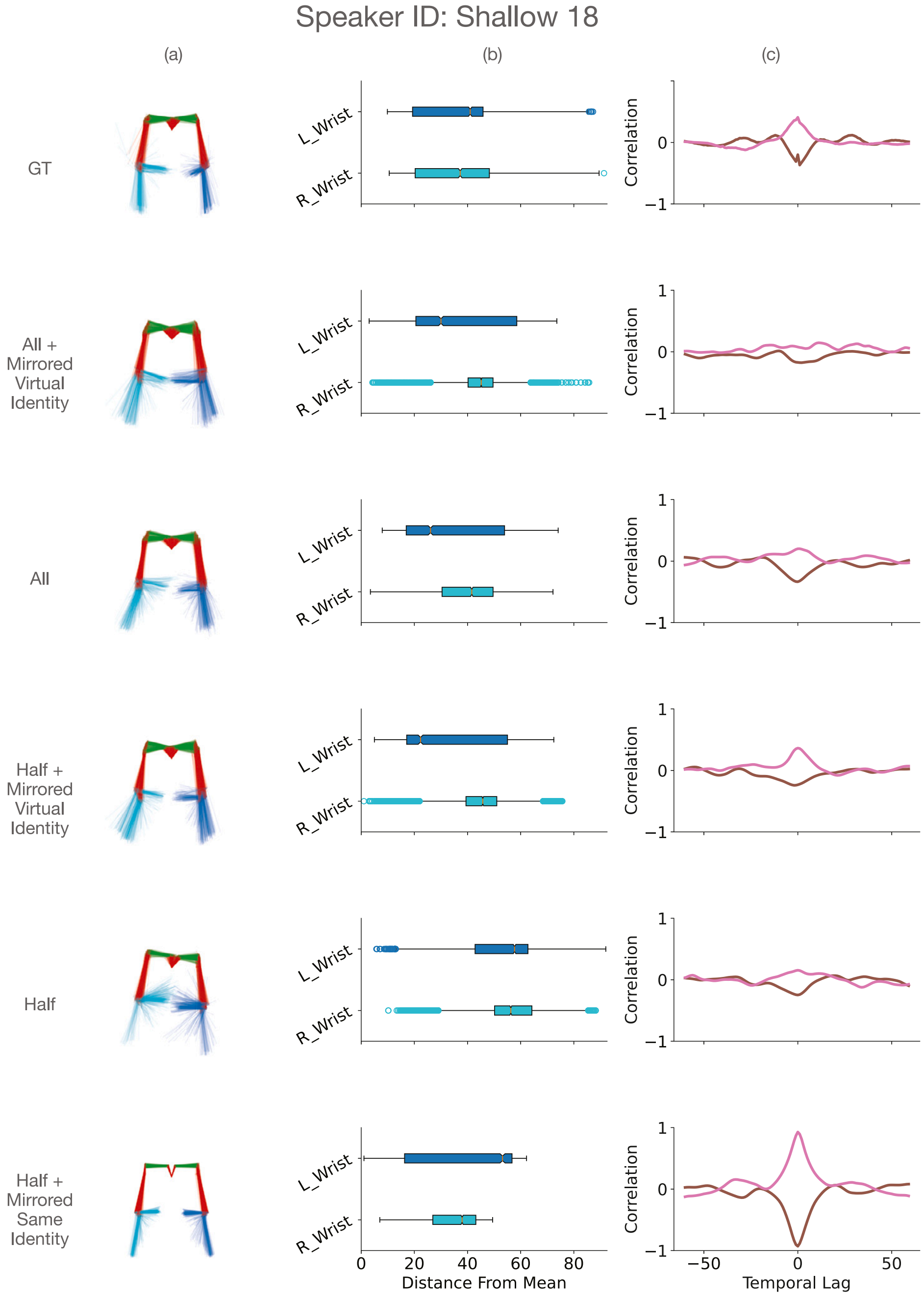

Fig. 17. A comparison for a single speaker's generated motion showing detrimental impact of including mirrored motion under the same identity. Each row corresponds to a different data split used. Column (a) contains the orthographic projection of a pose at every second in the sequence. Column (b) shows 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Mirroring the data has the potential to cancel out temporal offset characteristics. We have observed that certain speakers gesture with a leading hand. We found a generative model that has been trained on both the 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, we expect the inclusion of a symmetry statistic to aid in numerous tasks. We discuss the use of statistics for synthetic motion evaluation in Section 11.1, however, we also suggest considering the use of these statistics for identity classification. Motion symmetry could be important to the classification of speaker identity. We expect 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 original is speaker dependent, and we observe 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. (2018) has considered the impact of handedness on beat and representational gestures. They found that beat gestures had a preference for the dominant hand of the speaker, whereas representational gestures varied. In left-handed speakers, there was no preference, but in 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.

We analysed a few gesture types and their relationship to symmetry. While we cannot generalise from this small analysis, it would be sensible to consider when certain gesture types could be adequately mirrored. It is essential that handedness is 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 with no adaptation of the corresponding audio speech, this would lead to a disparity in the multi-modal context. When building gesture-generation systems, it would be beneficial to keep the handedness of gestures produced consistent.

Further study is required to determine the impact of modifying positional and temporal symmetry on realism and understanding. However, our 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.

11.1. Evaluating synthetic motion

A significant challenge in data-driven synthesis of embodied agents is how to evaluate the synthesised body animation. It is common to evaluate performance of generative models by means of a user study (Alexanderson et al., 2020a). 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, we would expect the animation to possess the same positional and temporal characteristics as the speaker's ground truth motion. We propose that statistical analyses based on the work presented in this paper would provide good indicators of these qualities. The per-speaker percentage of time spent in the extreme gesture space, degree of spatial symmetry and temporal lag of the animated result compared to the ground truth motion would be indicative of the similarities in both gesturing style and handedness.

12. Conclusion

We have studied four subjects from the *Talking with Hands* dataset to examine the symmetry in arm motion during dyadic conversation. We found that motion symmetry is highly speaker dependent. We derived a data-driven approach for defining a per-speaker gesture space, and found that the arms exhibited more lateral symmetry when in the common gesture space (closer to the mean pose) than when in the extreme space (further from the mean). We discovered that some speakers gesture with a leading hand, and others maintain left–right temporal alignment. We used information theory to find there is a large amount of information to be gained from both wrists. We employed a speech-to-motion model to support our findings.

Using these findings we have determined the efficacy of lateral mirroring for data augmentation and the considerations that should be made. If the goal is to maintain a speaker's gesturing style and handedness, mirroring for generating further examples of that speaker can only be used in certain cases, and is not suitable as a generic data augmentation approach. However, we suggest it is suitable for increasing the number of speakers in the training set by treating the mirrored data as a new *virtual identity*.

Finally, we propose our statistical analysis for evaluating the performance of speech-driven conversational agents to ensure that speaker characteristics have been retained in the synthesised motion.

CRediT authorship contribution statement

Jonathan Windle: Conceptualisation, Methodology, Software, Formal analysis, Writing – original draft. **Sarah Taylor:** Conceptualisation, Methodology, Writing – review & editing. **David Greenwood:** Conceptualisation, Methodology, Writing – review & editing. **Iain Matthews:** Conceptualisation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Iain Matthews reports a relationship with Epic Games that includes: employment.

Data availability

Data will be made available on request.

Acknowledgements

Sarah Taylor was supported by the Engineering and Physical Research Council (Grant number EP/S001816/1).

References

Alexanderson, Simon, Henter, Gustav Eje, Kucherenko, Taras, Beskow, Jonas, 2020a. Style-controllable speech-driven gesture synthesis using normalising flows. In: Computer Graphics Forum. 39, Wiley Online Library, pp. 487–496.

Alexanderson, Simon, Székely, Éva, Henter, Gustav Eje, Kucherenko, Taras, Beskow, Jonas, 2020b. Generating coherent spontaneous speech and gesture from text. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–3.

Bagautdinov, Timur, Wu, Chenglei, Simon, Tomas, Prada, Fabián, Shiratori, Takaaki, Wei, Shih-En, Xu, Weipeng, Sheikh, Yaser, Saragih, Jason, 2021. Driving-signal aware full-body avatars. ACM Trans. Graph. 40 (4).

Çatak, Esra Nur, Açık, Alper, Göksun, Tilbe, 2018. The relationship between handedness and valence: A gesture study. Q. J. Exp. Psychol. 71 (12), 2615–2626.

De Ruiter, Jan P., Bangerter, Adrian, Dings, Paula, 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. Top. Cognit. Sci. 4 (2), 232–248.

Drijvers, Linda, Özyürek, Asli, Jensen, Ole, 2018. Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. Human Brain Mapp. 39 (5), 2075–2087.

Ferstl, Ylva, McDonnell, Rachel, 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. pp. 93–98.

Ginosar, Shiry, Bar, Amir, Kohavi, Gefen, Chan, Caroline, Owens, Andrew, Malik, Jitendra, 2019. Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506.

Gong, Kehong, Zhang, Jianfeng, Feng, Jiashi, 2021. PoseAug: A differentiable pose augmentation framework for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8575–8584.

Habibie, Ikhsanul, Xu, Weipeng, Mehta, Dushyant, Liu, Lingjie, Seidel, Hans-Peter, Pons-Moll, Gerard, Elgharib, Mohamed, Theobalt, Christian, 2021. Learning speech-driven 3D conversational gestures from video. In: ACM International Conference on Intelligent Virtual Agents (IVA).

Henter, Gustav Eje, Alexanderson, Simon, Beskow, Jonas, 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Trans. Graph. 39 (6), 1–14.

Hostetter, Autumn B., 2011. When do gestures communicate? A meta-analysis. Psychol. Bull. 137 (2), 297.

Joo, Hanbyul, Liu, Hao, Tan, Lei, Gui, Lin, Nabbe, Bart, Matthews, Iain, Kanade, Takeo, Nobuhara, Shohei, Sheikh, Yaser, 2015. Panoptic studio: A massively multiview system for social motion capture. In: The IEEE International Conference on Computer Vision (ICCV).

Kendon, Adam, 1994. Do gestures communicate? A review. Res. Lang. Soc. Interact 27 (3), 175–200.

Kipp, Michael, Martin, Jean-Claude, 2009. Gesture and emotion: Can basic gestural form features discriminate emotions? In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, pp. 1–8.

Korzun, Vladislav, Dimov, Ilya, Zharkov, Andrey, 2020. The FineMotion entry to the GENEA challenge 2020.

Kozachenko, L.F., Leonenko, Nikolai N., 1987. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii 23 (2), 9–16.

Lee, Gilwoo, Deng, Zhiwei, Ma, Shugao, Shiratori, Takaaki, Srinivasa, Siddhartha S, Sheikh, Yaser, 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. pp. 763–772.

McNeill, David, 1985. So you think gestures are nonverbal? Psychol. Rev. 92 (3), 350.

McNeill, David, 2008. Gesture and Thought. University of Chicago Press.

McNeill, David, 2011. Hand and Mind. De Gruyter Mouton.

Neff, Michael, Toothman, Nicholas, Bowmani, Robeson, Tree, Jean E Fox, Walker, Marilyn A, 2011. Don't scratch! self-adaptors reflect emotional stability. In: International Workshop on Intelligent Virtual Agents. Springer, pp. 398–411.

Perez, Luis, Wang, Jason, 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Pouw, Wim, Harrison, Steven J, Esteve-Gibert, Núria, Dixon, James A, 2020. Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. J. Acoust. Soc. Am. 148 (3), 1231–1247.

Ravanelli, Mirco, Zhong, Jianyuan, Pascual, Santiago, Swietojanski, Pawel, Monteiro, Joao, Trmal, Jan, Bengio, Yoshua, 2020. Multi-task self-supervised learning for robust speech recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6989–6993.

Shorten, Connor, Khoshgoftaar, Taghi M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 1–48.

Strehl, Alexander, Ghosh, Joydeep, 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3 (Dec), 583–617.

Studdert-Kennedy, Michael, 1994. Hand and mind: What gestures reveal about thought. Lang. Speech 37 (2), 203–209.

Taylor, Sarah, Windle, Jonathan, Greenwood, David, Matthews, Iain, 2021. Speech-driven conversational agents using conditional flow-VAEs. In: European Conference on Visual Media Production. pp. 1–9.

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

Yoon, Youngwoo, Cha, Bok, Lee, Joo-Haeng, Jang, Minsu, Lee, Jaeyeon, Kim, Jaehong, Lee, Geehyuk, 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. Graph. 39 (6), 1–16.