Pose Augmentation: Mirror The Right Way

Jonathan Windle
University of East Anglia
United Kingdom

Sarah Taylor
University of East Anglia
United Kingdom

David Greenwood
University of East Anglia
United Kingdom

Iain Matthews
University of East Anglia
United Kingdom

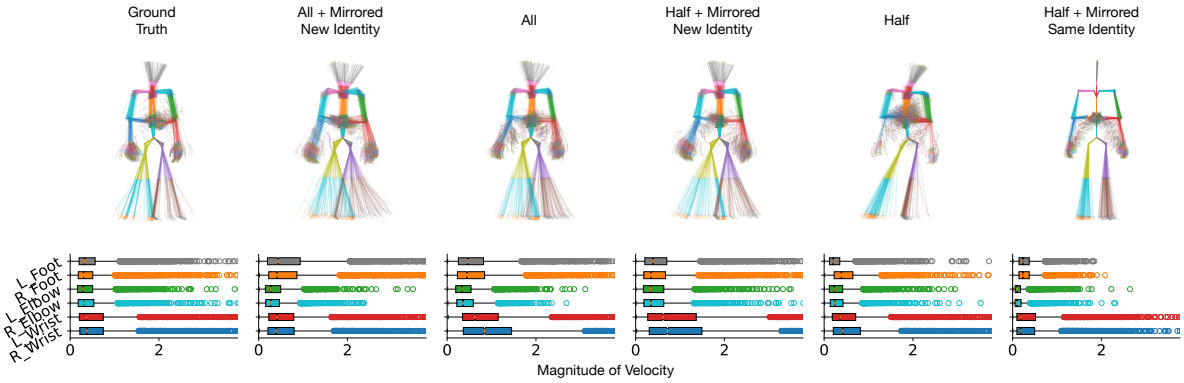

Figure 1: A comparison for a single speaker's generated motion. Including mirrored motion under the same identity produces undesirable muted and symmetrical poses. Performance increases between half and half mirrored with a virtual identity performing competitively to all of the data available. When including the most optimal setting of all of the data mirrored as a virtual identity, performance of magnitude of change and pose positions improves towards the ground truth. Each column corresponds to a different data split used. Top row is a plot containing the orthographic projection of a pose at every second in the sequence. Bottom row shows the distribution of velocity magnitude, this is how far a joint moves between two frames.

ABSTRACT

We demonstrate an effective method of augmenting speech animation data, and show comparable performance to *double* the quantity of real data. We investigate the effect of lateral mirroring as a means of data augmentation for 3D poses in multi-speaker, speech-to-motion modelling. Our approach uses a bi-directional LSTM to generate 3D joint positions from audio features extracted using problem-agnostic speech encoder (PASE+) [7]. We demonstrate that naive mirroring for augmentation has a detrimental effect on model performance. We show our method of providing a virtual speaker identity embedding improved performance over no augmentation and was competitive with a model trained on an equal number of samples of real data.

CCS CONCEPTS

• **Computing methodologies** → Intelligent agents; *Animation.*

KEYWORDS

Audio-driven gesture generation, 3D Pose, Data augmentation

ACM Reference Format:
Jonathan Windle, Sarah Taylor, David Greenwood, and Iain Matthews. 2022. Pose Augmentation: Mirror The Right Way. In *ACM International Conference on Intelligent Virtual Agents (IVA '22), September 6–9, 2022, Faro, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3514197.3549677

1 INTRODUCTION

Speech-driven generation of human body motion is an active research area [1–5, 8, 9]. A common challenge faced during speech-to-motion generation is acquiring sufficient data. Motion capturing with synced high-quality audio is time-consuming and financially and computationally expensive. Data augmentation is a logical solution. The ability to effectively double the amount of data using a lateral mirroring technique is attractive. When mirroring data in a multi-speaker model, there are two speaker labelling options: a) original identity, or b) a new virtual identity. We aim to clearly define how this data augmentation method should be used for multi-speaker speech-to-motion generation.

This study uses the Talking With Hands dataset [6]. The dataset consists of dyadic conversations across 50 different speakers. We train a Long Short-Term Memory (LSTM) model to map from speech to body pose using different data augmentation settings. Specifically, we set out to answer the questions: 1) Does data augmentation

IVA '22, September 6–9, 2022, Faro, Portugal

ACM ISBN 978-1-4503-9248-8/22/09.
https://doi.org/10.1145/3514197.3549677

from lateral mirroring result in a better prediction? 2) Does this performance match a model trained using double the amount of ground truth data? 3) How should we label the identity of the mirrored data? We confirm that more data *is* better than augmented data for motion prediction. We show that naive pose mirroring is detrimental to model performance. We introduce an identity embedding and show laterally mirrored motion as a new identity is comparable to the same quantity of real data.

2 MODEL

We train a simple but effective LSTM-based model to compare augmentation settings. The model is trained using a single motion frame's worth of audio (33ms) encoded using the pre-trained problem-agnostic speech encoder (PASE+) [7] to predict a frame of motion Each speaker in the dataset is assigned a unique ID that is passed to an embedding layer that consists of trainable weights. The speakers that move similarly should be closer in this embedding space. The speaker embedding is concatenated with the PASE+ features to create an input size of 776. 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.

The loss function contains multiple terms. An L_2 loss on ground truth and predicted pose positions. We also observed that including an L_2 loss on the joint velocity and acceleration helped to produce smoother motion with more realistic transitions. We include an L_1 loss term on bone length to address any bone stretching and squeezing artefacts. Models are trained using the Adam optimiser with a learning rate of 0.0001 and batch size of 256. Hand motion is absent from some sequences. Where this is the case we compute the loss against all joints in the body except the hands. Each sequence of motion and corresponding audio is split into 30 frame chunks with a 25 frame overlap. This ensures a balance of number of samples and samples not being too similar to each other.

3 EXPERIMENTAL SETUP

To fully explore the effect of data augmentation, we train a model on different splits of data and using various augmentation settings defined as follows:

All Data. We use all available training data with no additional augmentation. This forms our baseline model.

Half Data. We randomly subsample the training data to reduce 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 Virtual Identity. We augment the *Half Data* training set by laterally mirroring the pose at each frame. This effectively doubles the amount of data available for training. During training, we assign a **new** virtual identity label to the mirrored data.

Mirrored Same Identity. We augment the *Half Data* training set by laterally mirroring the pose at each frame. Here, we assign the mirrored data the **same** identity label as the original speaker data. This enables us to determine whether it is better to have more identities or more samples per identity.

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.

4 RESULTS

We evaluate the realism of the predicted motion from different models and discuss its relationship to the speaker identity.

We first compare the generated motion using half of the data (*Half Data*) against using the same half of data mirrored under the same identity (*Mirrored Same Identity*). We observe two key findings; the mirrored data produced far more muted and symmetrical motion than desired. Figure 1 (top row) shows each of the arms consistently raising simultaneously when using mirrored data as the same identity.

We found that mirroring half of the data as a virtual identity (*Mirrored Virtual Identity*) was competitive with a model trained with all available data. We noticed an improvement in the position of poses formed and the magnitude of velocity change. The speaker shown in Figure 1 benefited from augmenting with a virtual identity. The positions formed in Figure 1 (top row) show that when only half of the data is used, the poses are asymmetrical and often default to the same leaning position. When we laterally mirror this half and include the data as a virtual identity, this improves the positioning, adding some more symmetry and movement that closely resembles the ground truth and generated motion from all of the data.

The distribution of the magnitude of velocity shown in Figure 1 (bottom row) increases from half of the data and half mirrored as a virtual identity. When comparing these velocities against the ground truth it appears to be losing performance. However, we found that it is very similar to the velocities shown in motion generated with all 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.

We found further improvements when augmenting all of the data with a virtual identity. Figure 1 shows that while the poses formed are similar to that of the model using all of the data, the distribution of velocity magnitude used to form the poses are much closer to the ground truth. This is a good example of how the same poses can be formed but the speed and range of motion used to achieve them can be drastically different.

5 CONCLUSION

We conclude that more data is always preferred and lateral mirroring as a virtual identity should be considered as a data augmentation technique. The inclusion of mirrored data is introducing a bias to motion that a human did not perform. The bias introduced from mirrored data under the same identity reduces the model's ability to fit the data and causes muted, highly symmetrical gesturing. Given the scarce and expensive nature of speech-to-motion datasets, we present a more useful way of expanding a dataset through the use of mirrored data under a virtual identity. We found using this technique generated motion that improves training and generative performance to be competitive with a model trained on equal amounts of real data.

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

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

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

[4] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. In *ACM International Conference on Intelligent Virtual Agents (IVA)*. arXiv:Todo

[5] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[6] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 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*. 763–772.

[7] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 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, 6989–6993.

[8] Sarah Taylor, Jonathan Windle, David Greenwood, and Iain Matthews. 2021. Speech-Driven Conversational Agents using Conditional Flow-VAEs. In *European Conference on Visual Media Production*. 1–9.

[9] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.