KTH ROYAL INSTITUTE OF TECHNOLOGY

# Move over, MSE!
# New probabilistic models of motion

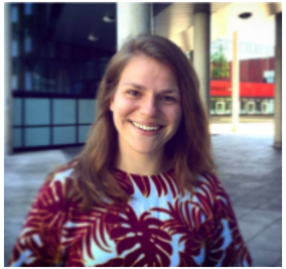
Gustav Eje Henter
Division of Speech, Music and Hearing (TMH), KTH Royal Institute of Technology, Stockholm, Sweden

# Motion and speech co-authors


Simon
Alexanderson


Taras
Kucherenko


Patrik
Jonell


Sanne
van Waveren


Éva
Székely


Jonas
Beskow


Joakim
Gustafson

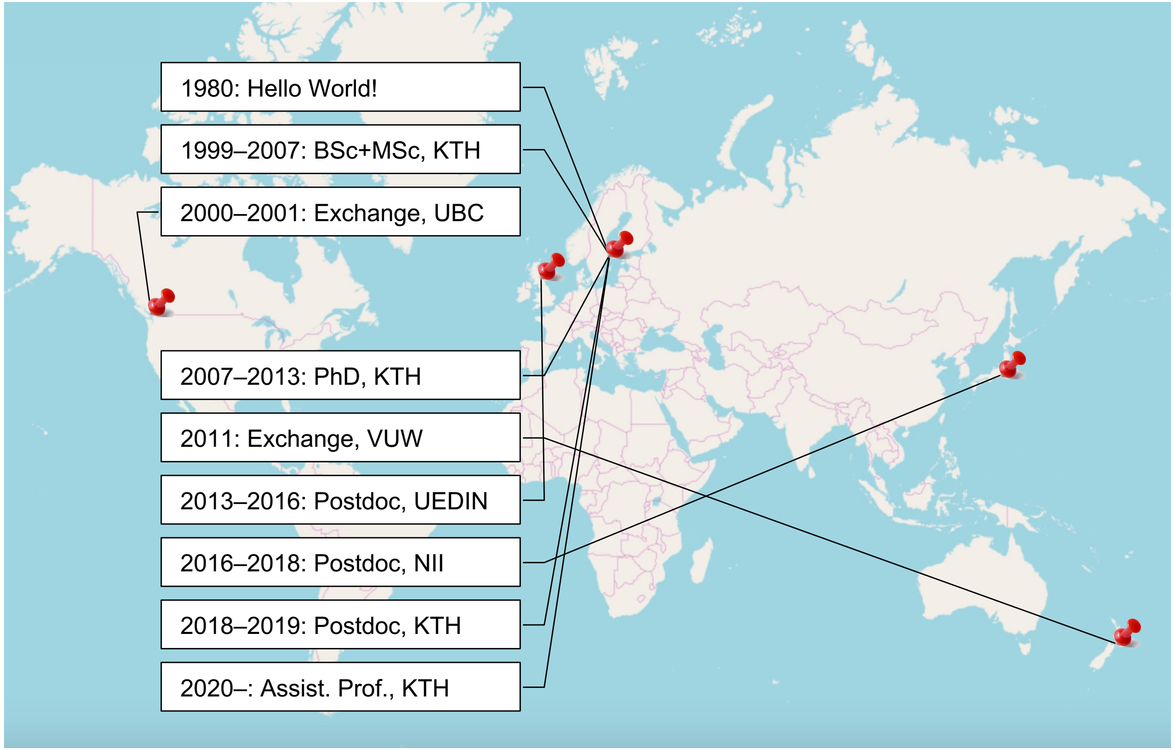

Dai
Hasegawa


Naoshi
Kaneko

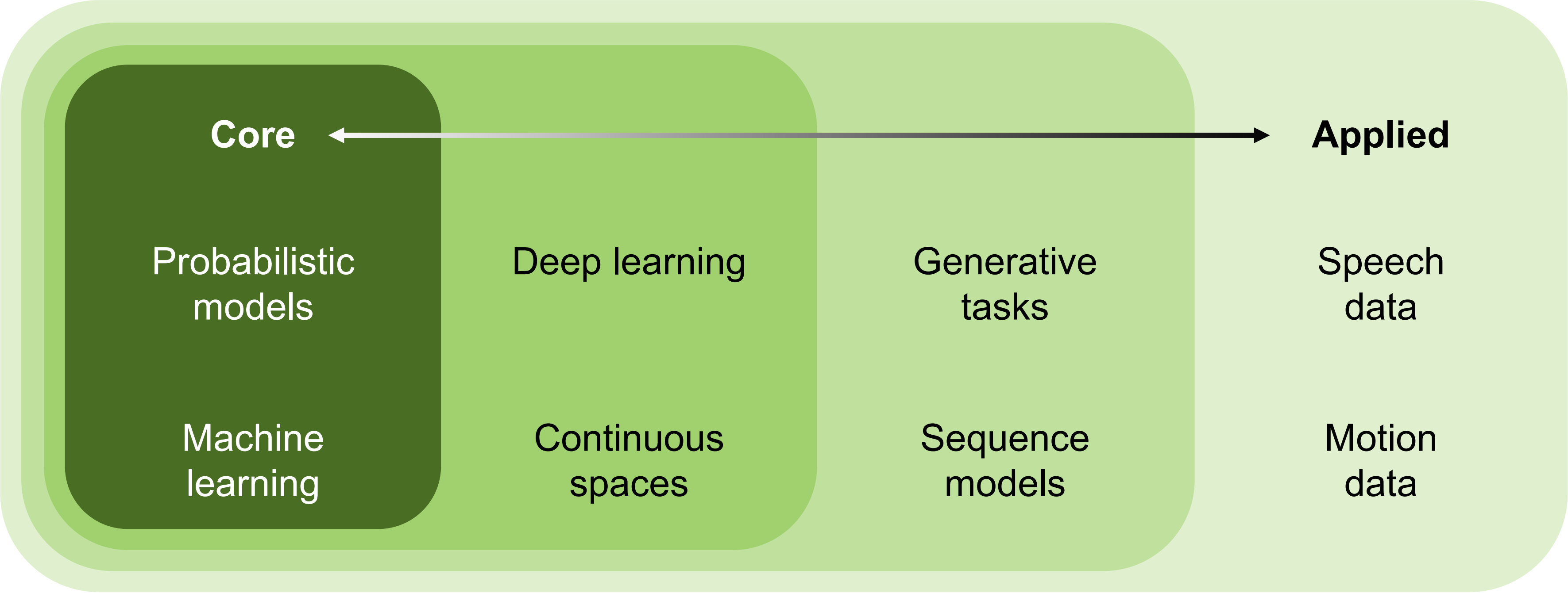

Iolanda
Leite


Hedvig
Kjellström

# Personal background



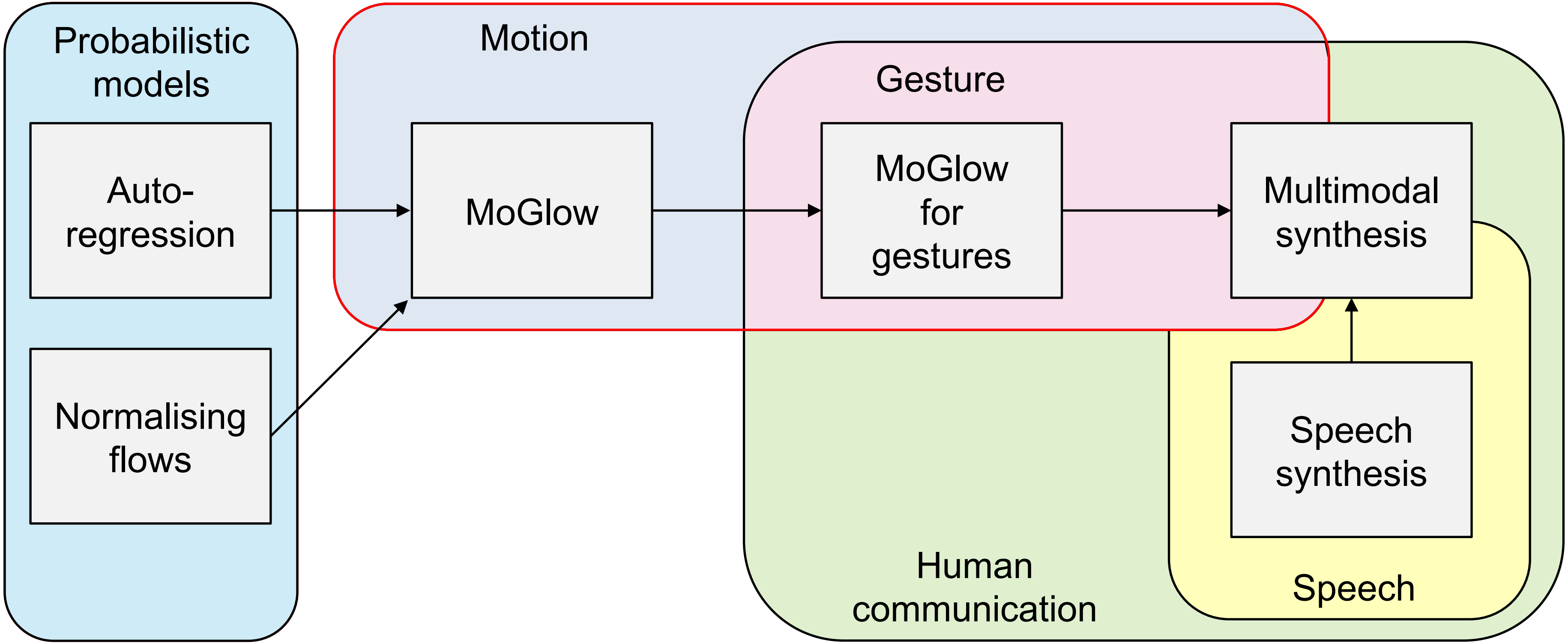
# Research interests



# This talk in a nutshell

- Automated character animation is a challenging and interesting problem
- The world is probabilistic; our motion models should be, too
- MoGlow is a new probabilistic model for motion
  - Autoregressive sequence model with normalising flows
  - Reaches the state of the art in a range of different applications
- Text-to-speech → text-to-behaviour

# Graphical overview of this talk

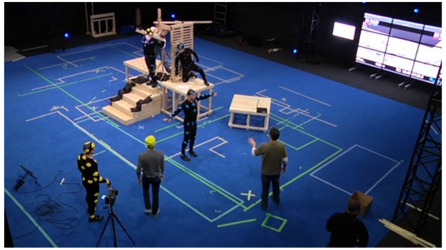


# Where do we need character animation?

- Computer and video games
- Film and SFX
- Architectural visualisations
- Virtual avatars
- Social robots

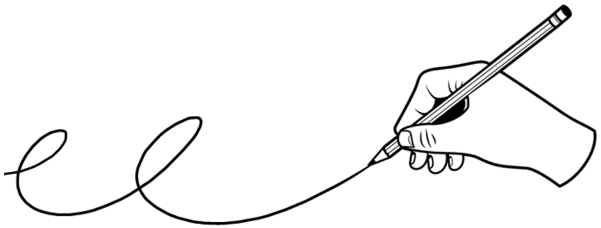
# Animation is a complex process

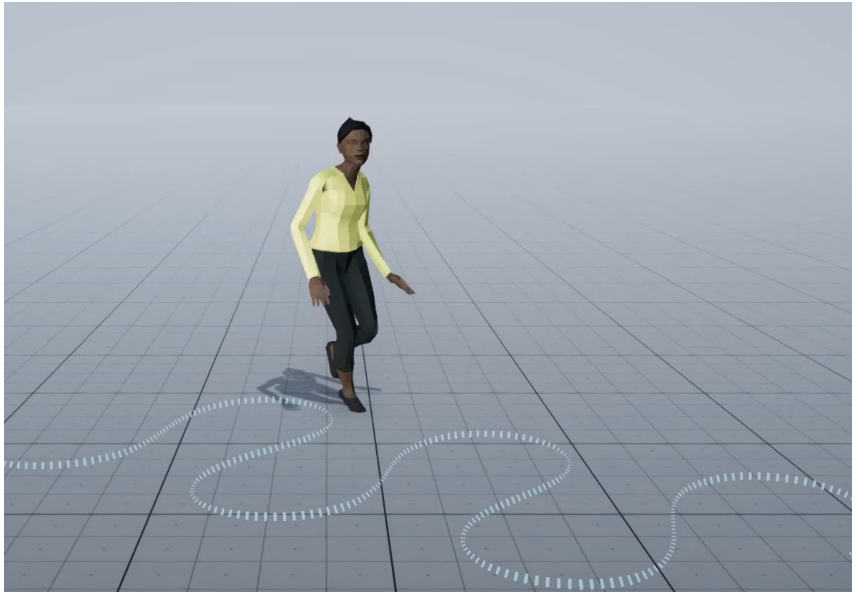
- 3D character animation requires several steps
  1. Planning motion
     > *Storyboarding, previsualisation*
  2. Creating motion
     > *Motion capture (mocap) or keyframing + inbetweening*
  3. Editing motion
     > *Cleanup, retargeting, etc.*
- Issues
  - Time-consuming
  - Expensive
    > *Requires coordination among many different experts*
    > *Director, mocap actors, technicians, animators…*
  - Rigid process



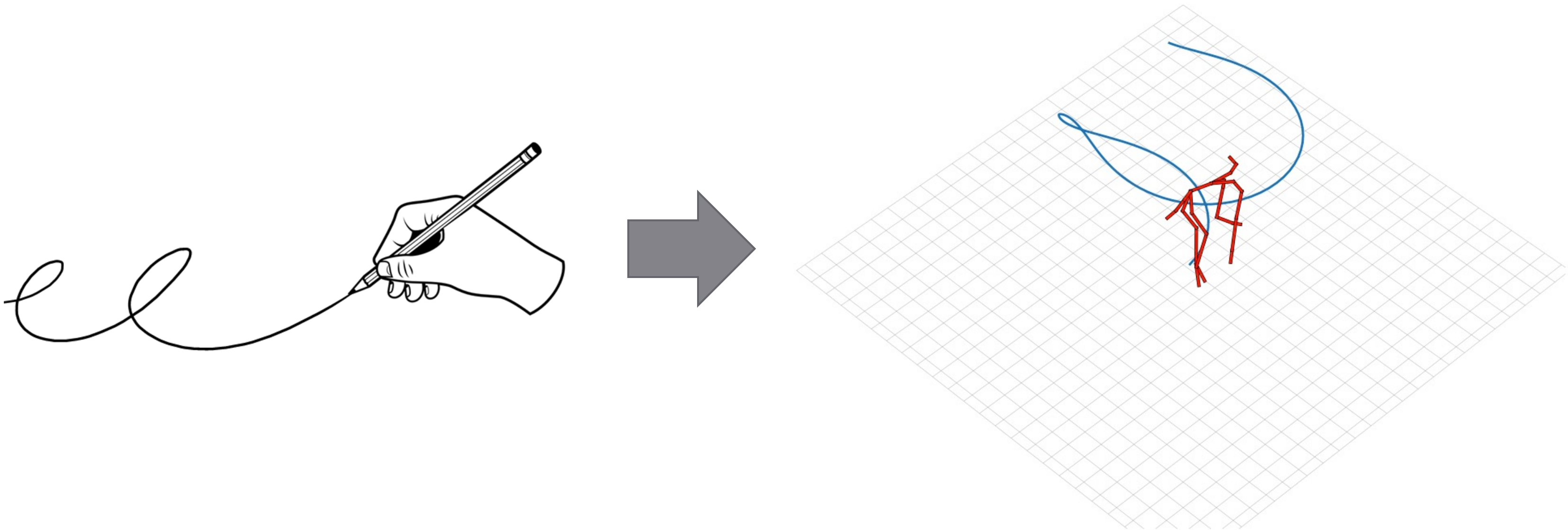
# Character animation example

# Sketch to locomotion

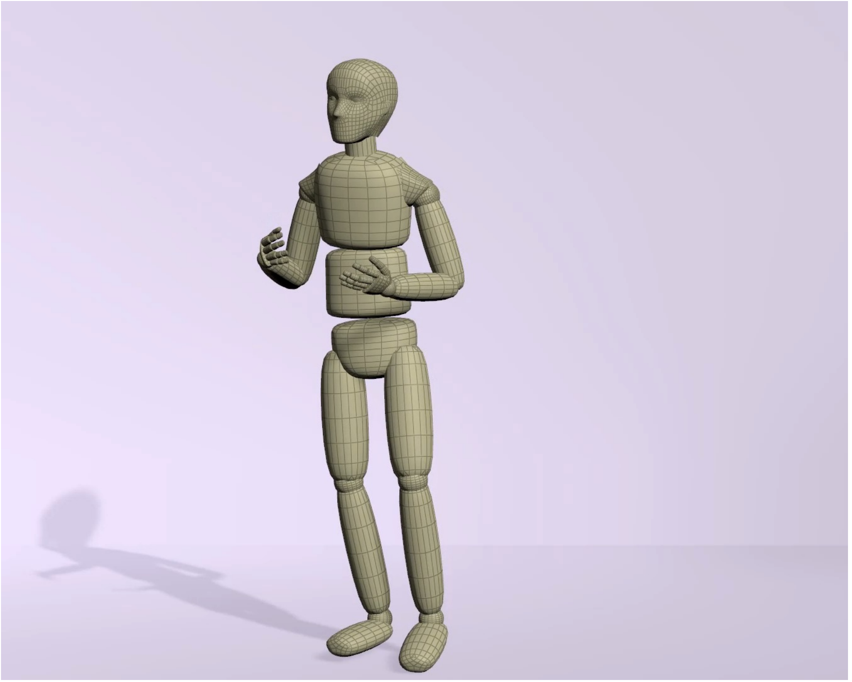




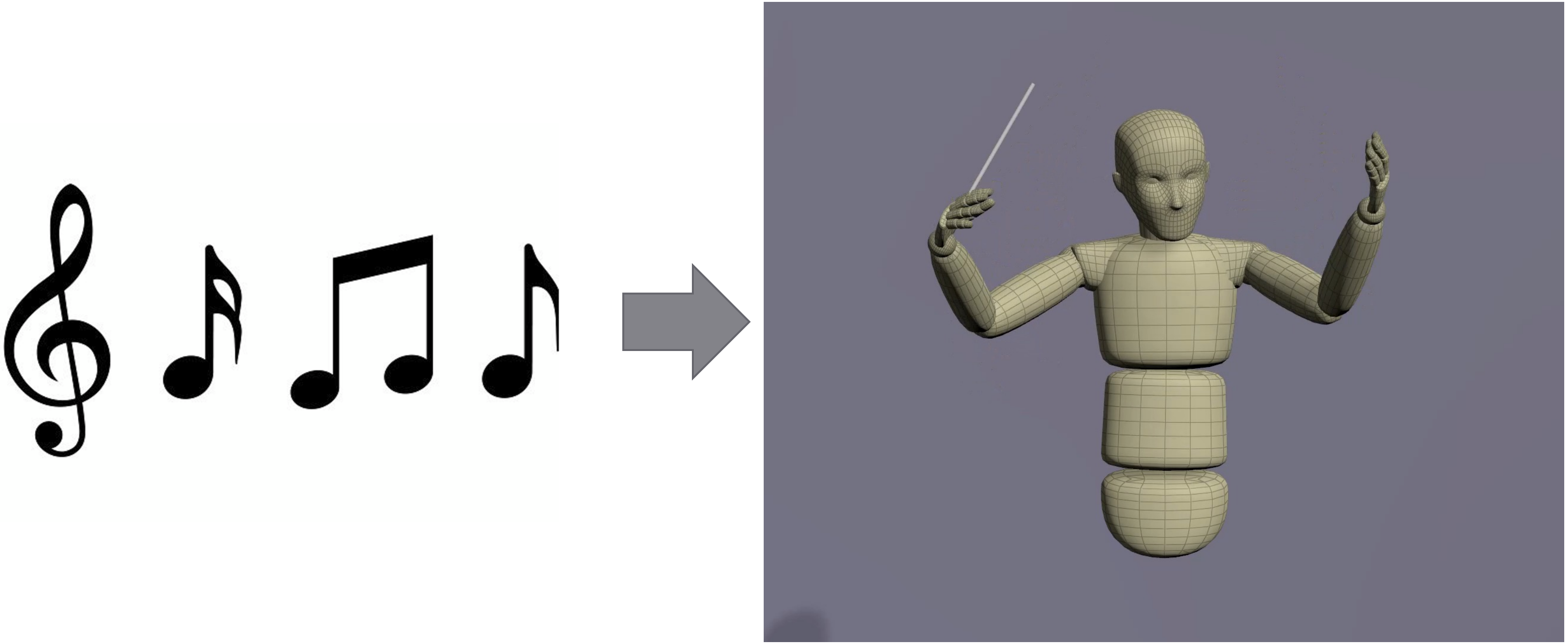
# Sketch to locomotion



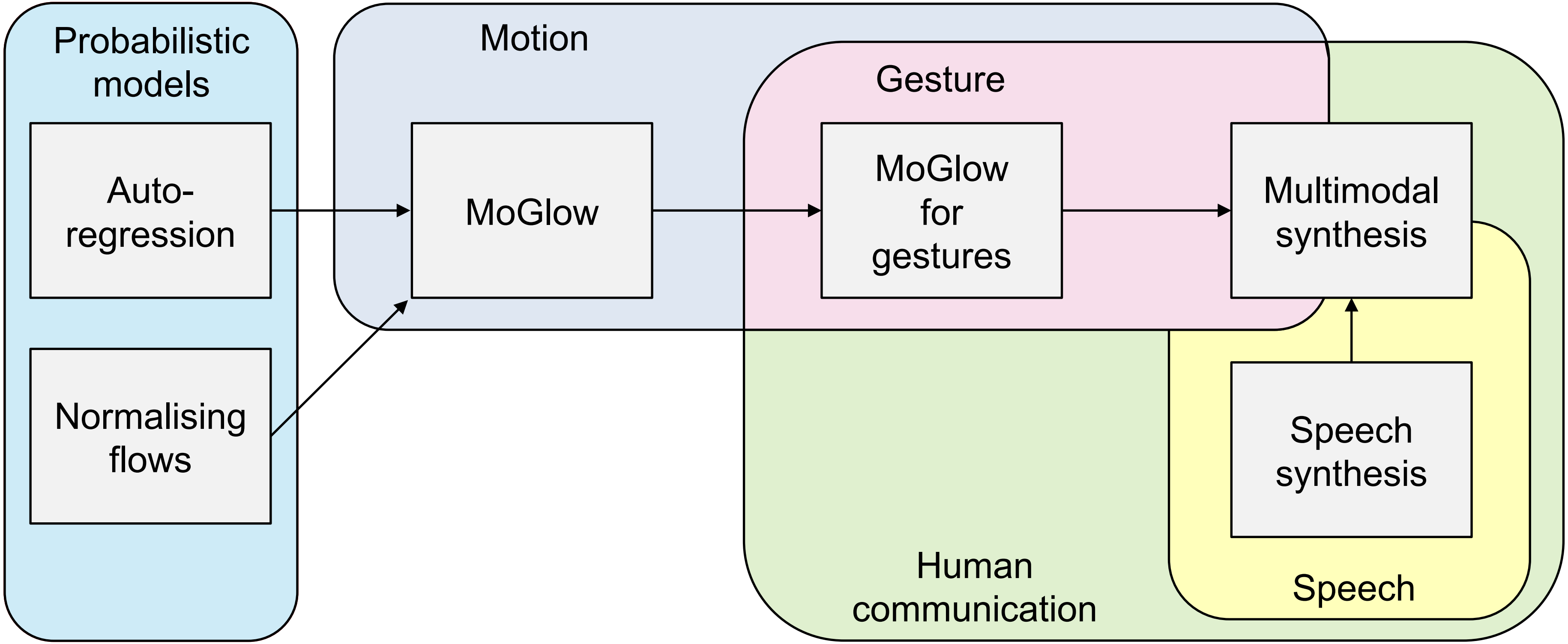
# Speech to gesture



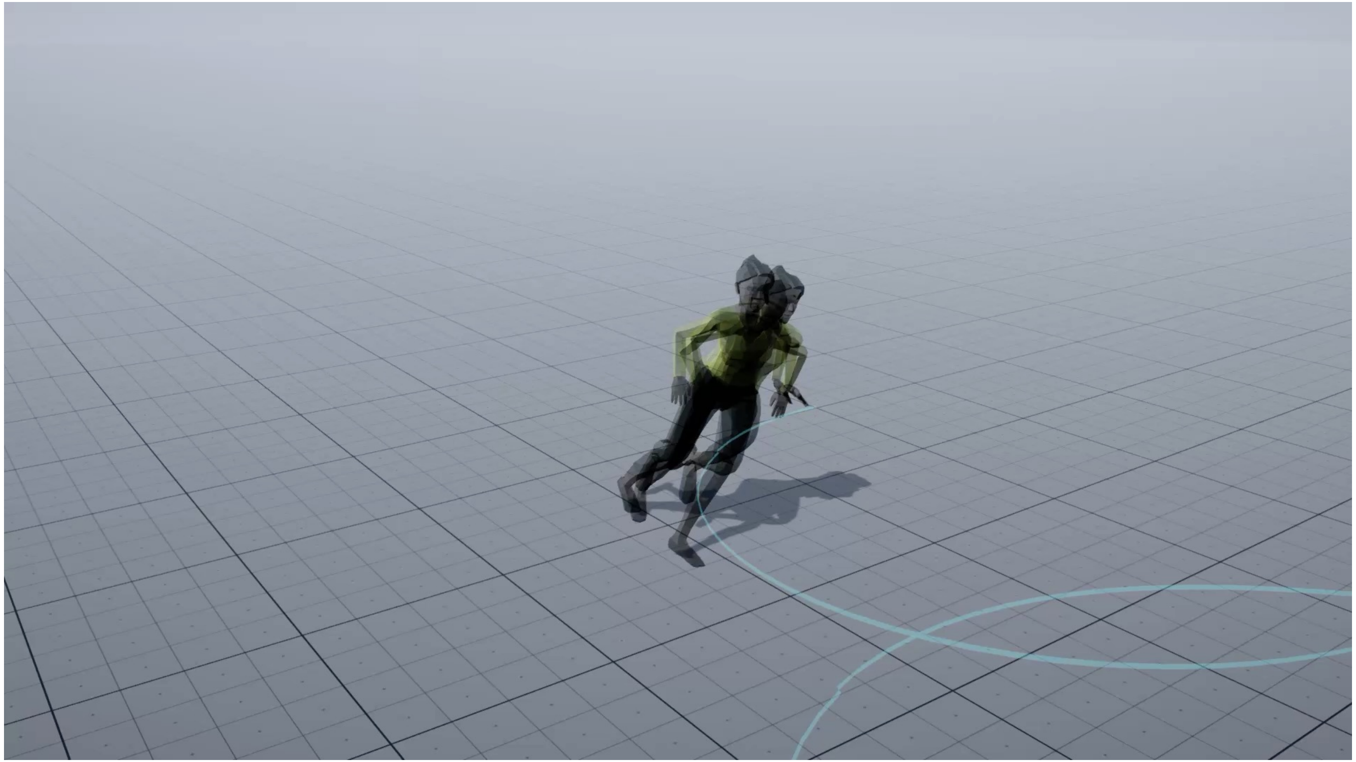
# Music to motion



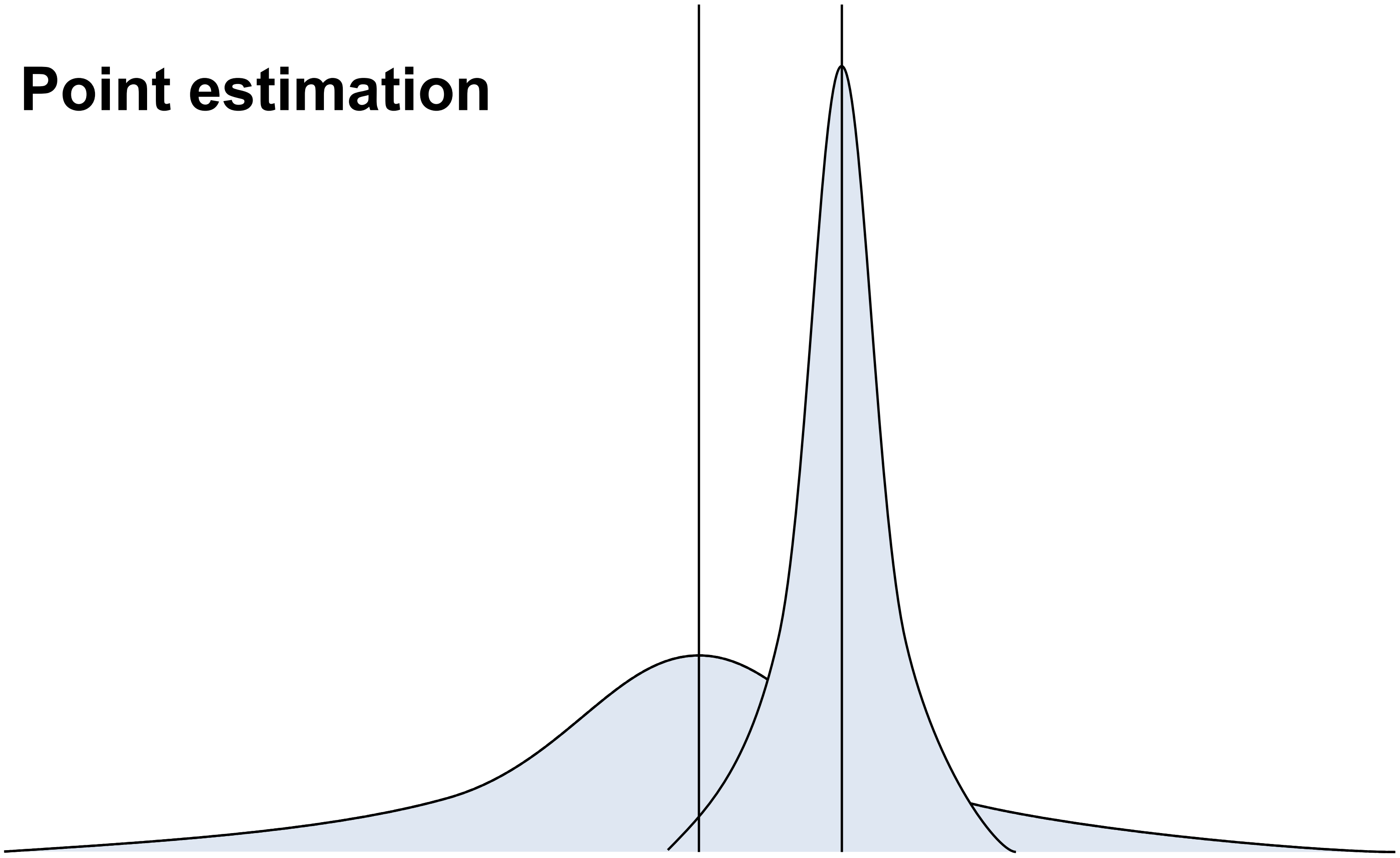
# Graphical overview of this talk



# Why be probabilistic?



# Point estimation



# Understanding minimum mean squared error

- In the limit of infinite data, the expected mean squared error (MSE) is minimised by the conditional mean

$$\widehat{y}\left(x\right)=\min_{y}\mathbb{E}\left[\left(Y-y\right)^{2}\,\middle|\,X=x\right]=\mathbb{E}\left[Y\,|\,X=x\right]$$

- Models trained to minimise the MSE loss converge on “average motion”
  - “Mean collapse” or “regression to the mean”
  - This does not necessarily look natural

# Motivating example 1: D6

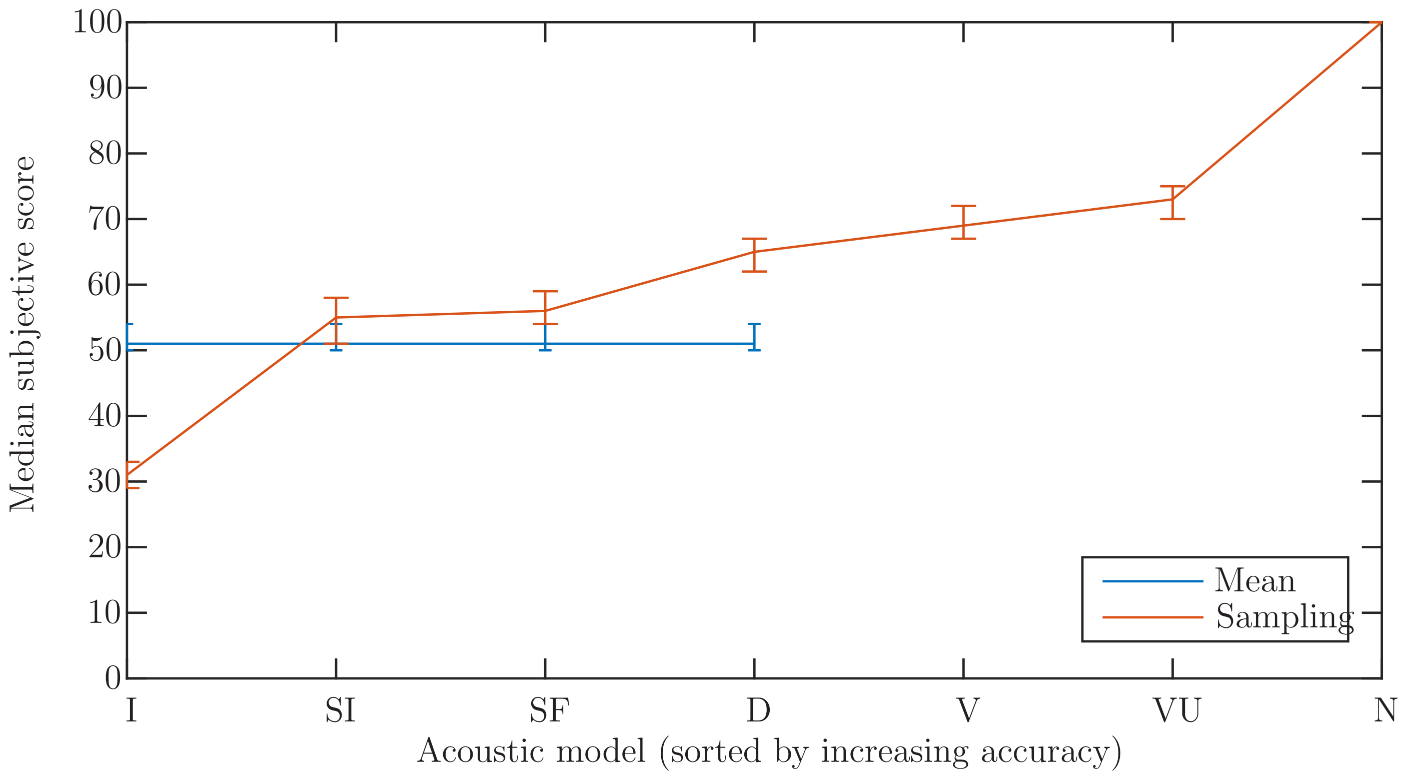
Mean outcome
= 3.5

Where's the side with 3.5 pips?



Image source: https://commons.wikimedia.org/wiki/File:Dice_2005.jpg

# Motivating example 2: Speech

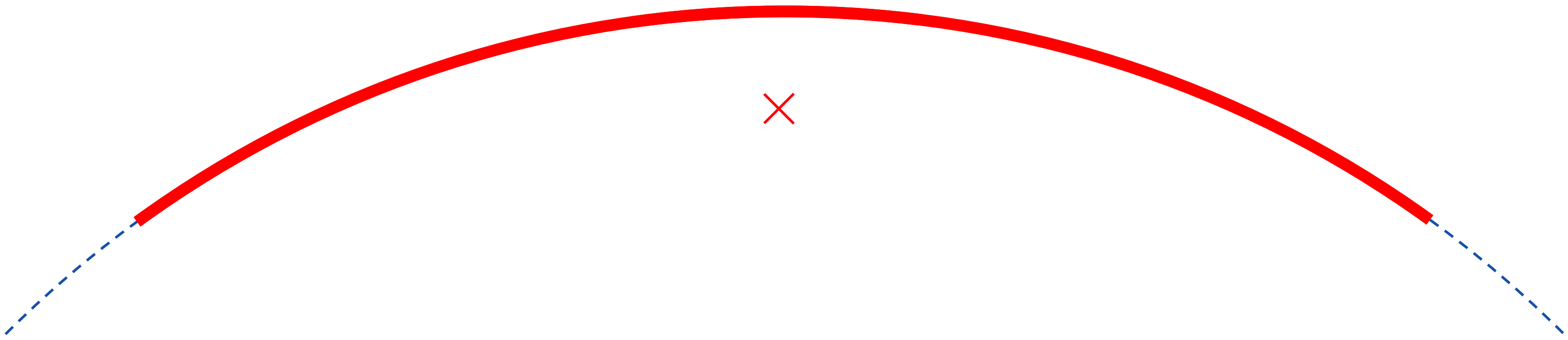




G. E. Henter et al., "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech", Proc. Interspeech, 2014

# Averaging in higher dimensions

- In higher dimensions, the data typically sits on a low-dimensional manifold
  - Additional information (context, control) helps narrow down the range of possible motion
- The (conditional) mean is the centre of gravity of the probability mass
  - The greater the degree of averaging, the more noticeably unnatural the result



# Control of different motion types

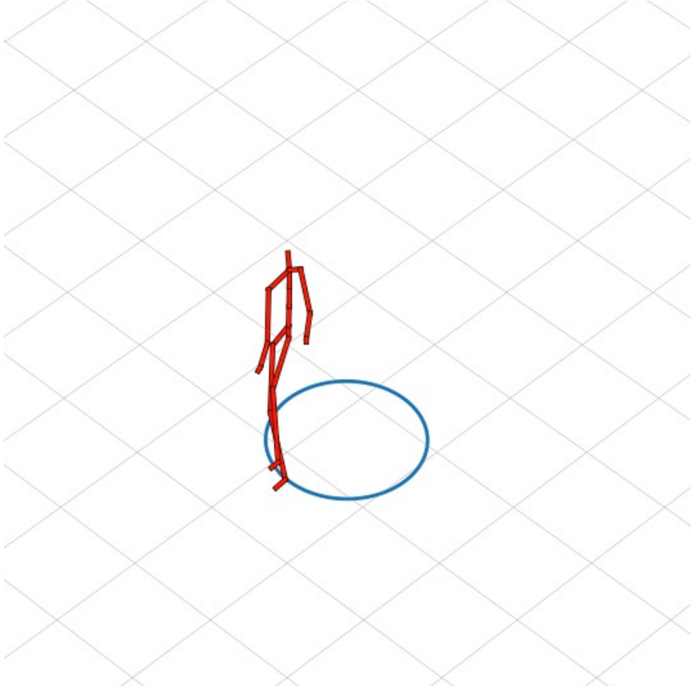
- Lip motion
  - Highly predictable from speech audio
    - We call this a “strong control signal”
- Locomotion
  - Not highly predictable from the path
    - A “weak control signal”
  - Highly predictable from path and, e.g., phase (cyclic locomotion) or foot contacts
- Head motion, hand gestures, stance in conversation, etc.
  - Not well-determined by co-occurring speech
  - No strong control signals are available



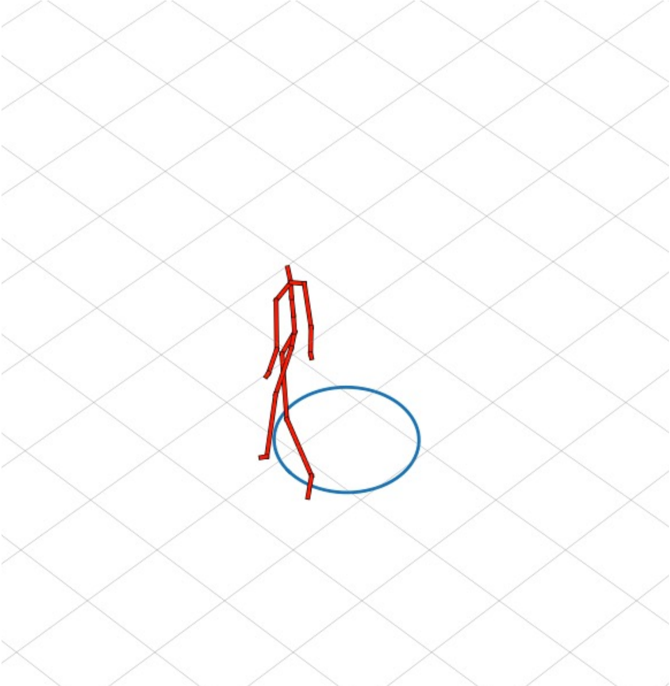
D. Holden et al., “Phase-functioned neural networks for character control” & D. Holden et al., “A deep learning framework for character motion synthesis and editing”, ACM Trans. Graph. 2018, 2016

# Motion synthesis from weak control signals

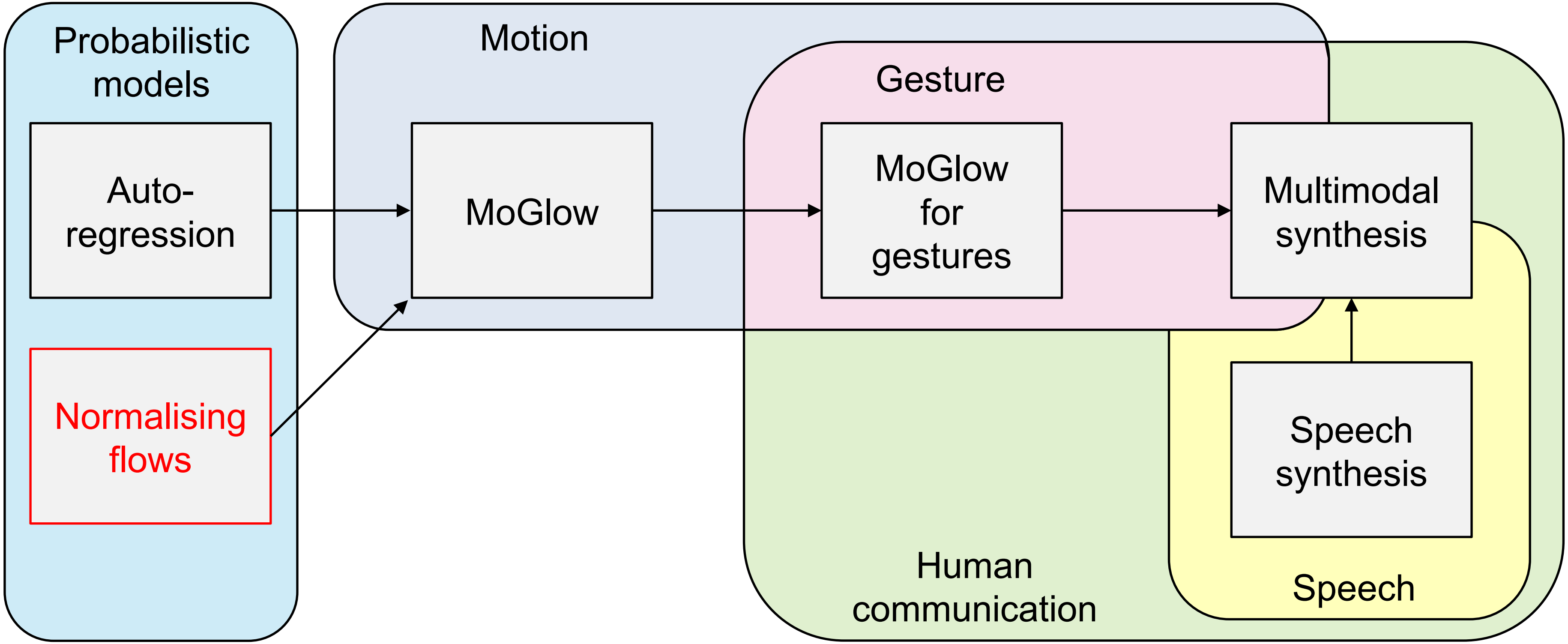
Deterministic (MMSE)



Probabilistic (MoGlow)



# Graphical overview



# Desiderata

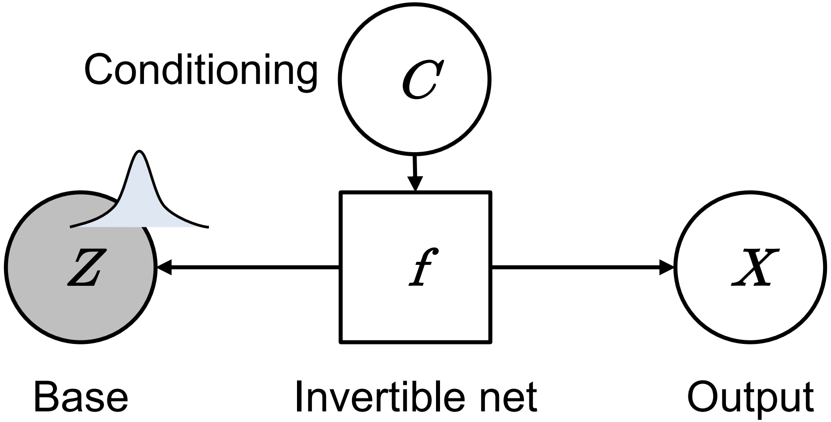
- Tractable statistical inference
  - It should be easy to compute the exact probability of an observation
  - This enables efficient maximum-likelihood training
- Straightforward output generation
  - Drawing random samples from the learned distribution should be fast and easy
- Flexibility
  - Mathematically, all the probability distributions our model can represent constitute a *parametric family*

$$\mathcal{F} = \{p(x_{1:T}; \theta)\}_{\theta \in \Theta}$$

    - Example: Multivariate Gaussians with diagonal covariance matrices
  - We want this parametric family to be sufficiently rich to well fit the true distribution

# Different probabilistic approaches

| | Gaussian (MMSE) | MDN | VAE | GAN | Normalising flow |
|---|---|---|---|---|---|
| Training | ✓ | ✓ | ✗ | ✗ | ✓ |
| Sampling | ✓ | ✓ | ✓ | ✓ | ✓ |
| Flexibility | ✗ | ✗ | ✗ | ✓ | ✓ |



# Key idea of normalising flows

- Change a simple distribution into a more complex distribution using an *invertible* transformation $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{Z})$
  - The change-of-variables formula gives the log-likelihood after transformation

$$\ln p_{\boldsymbol{X}}(\boldsymbol{x}) = \ln p_{\boldsymbol{Z}}\left(\boldsymbol{f}^{-1}(\boldsymbol{x})\right) + \ln\left|\det \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{f}^{-1}(\boldsymbol{x})\right|$$

- These invertible, nonlinear transformations can be chained together

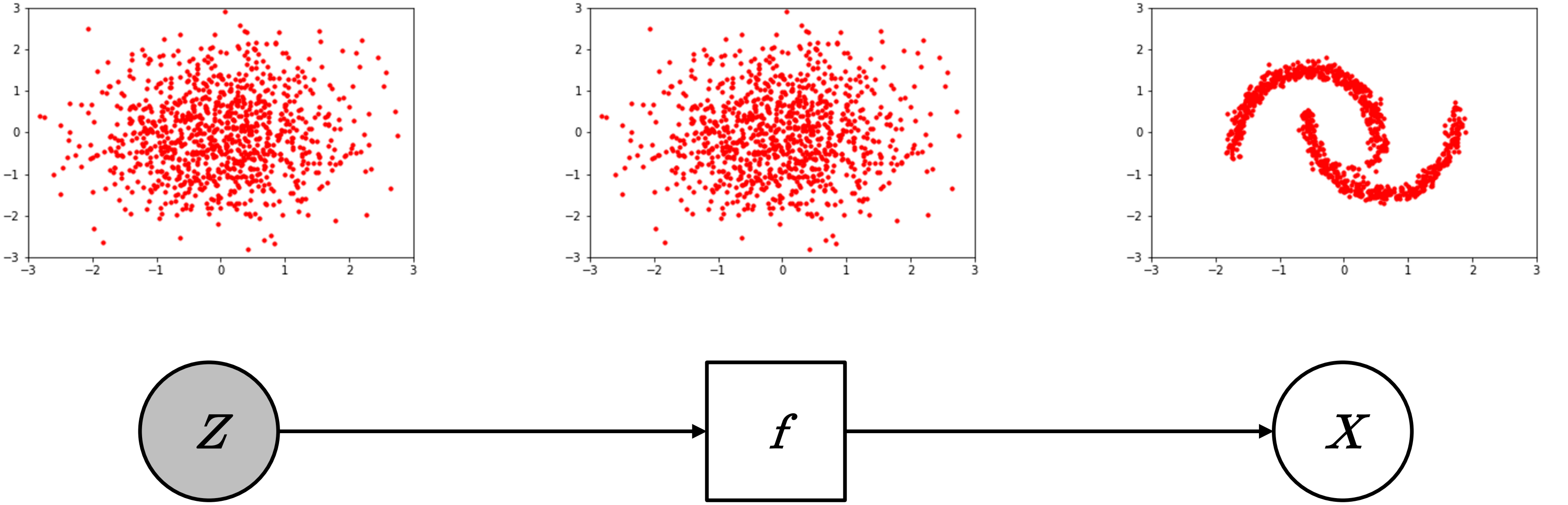
$$\boldsymbol{z} = \boldsymbol{z}_N \stackrel{\boldsymbol{f}_N}{\to} \boldsymbol{z}_{N-1} \stackrel{\boldsymbol{f}_{N-1}}{\to} \ldots \stackrel{\boldsymbol{f}_2}{\to} \boldsymbol{z}_1 \stackrel{\boldsymbol{f}_1}{\to} \boldsymbol{z}_0 = \boldsymbol{x}$$

  - This is the same idea that gives power to neural networks and GANs
  - The *base*, *latent*, or *source distribution* $\boldsymbol{Z}$ can be a standard Gaussian

# An analogy to baking

# An analogy to baking

# A 2D toy example




Simple base distribution

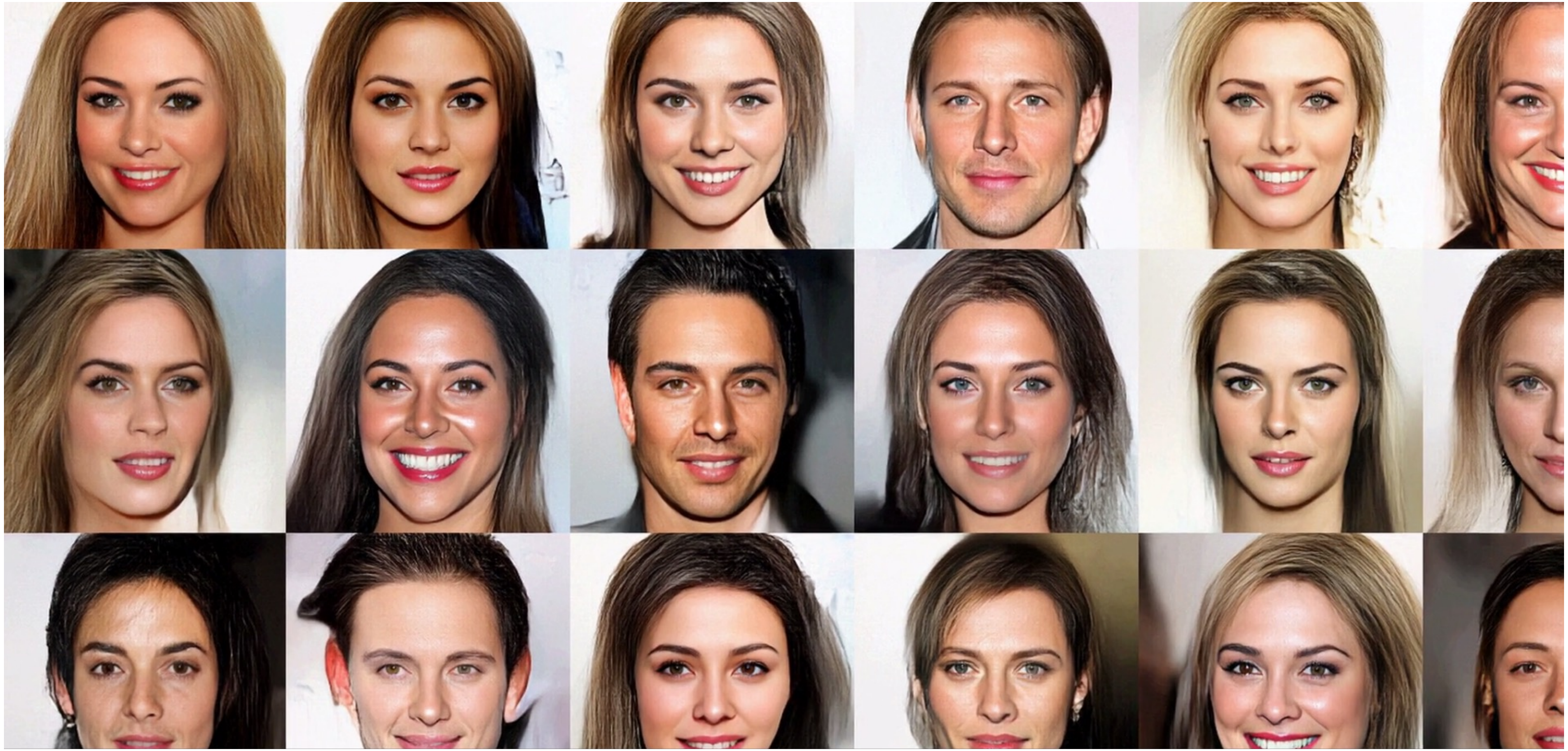
Invertible transformation(s)

Natural output distribution



Image source Eric jang: https://blog.evjang.com/2018/01/nf2.html

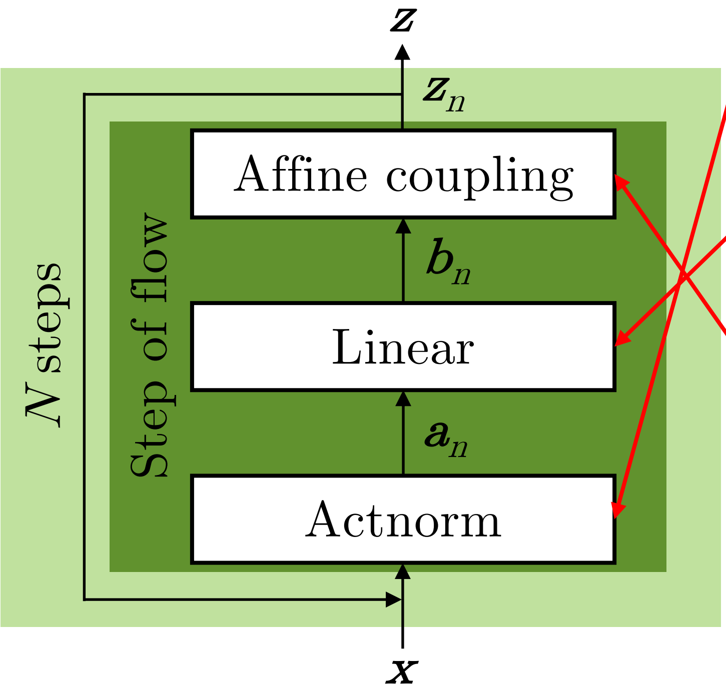
# Face generation using Glow





D. P. Kingma & P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions”, Proc. NeurIPS, 2018

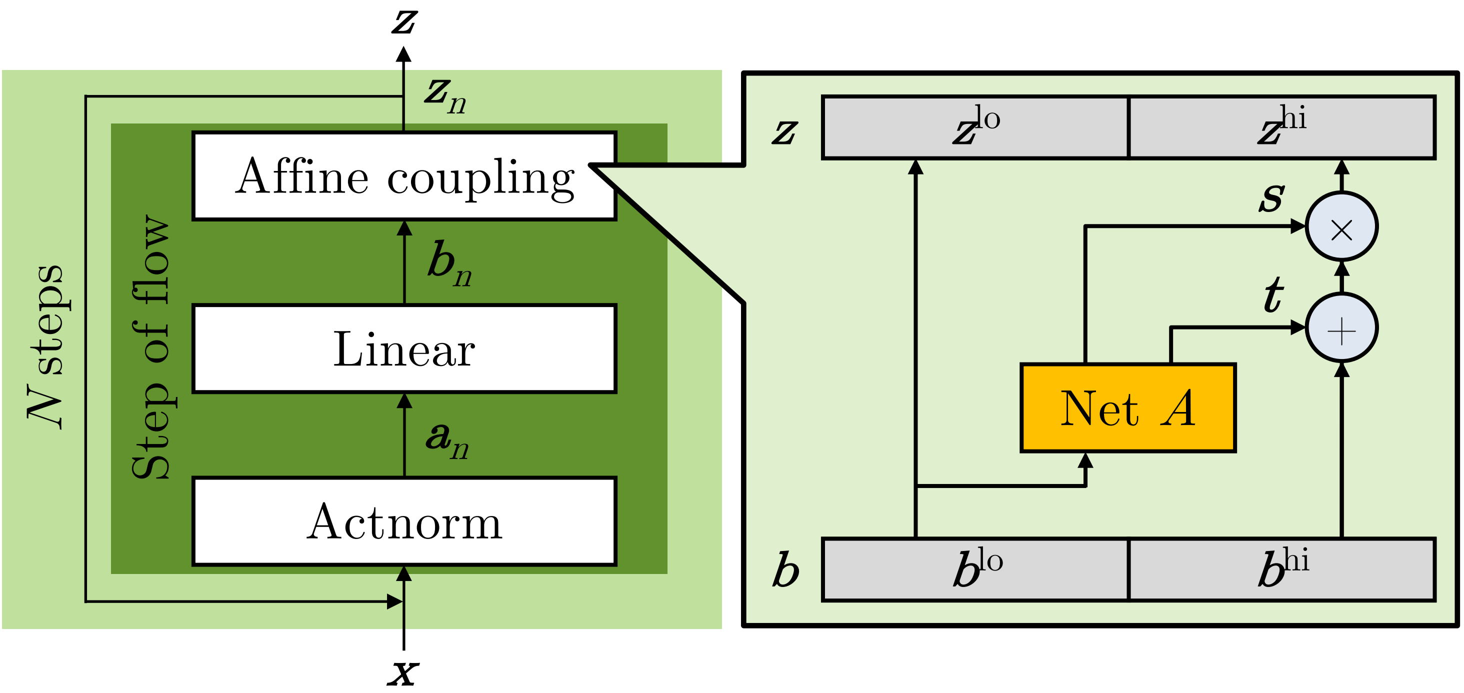
# The Glow invertible neural network



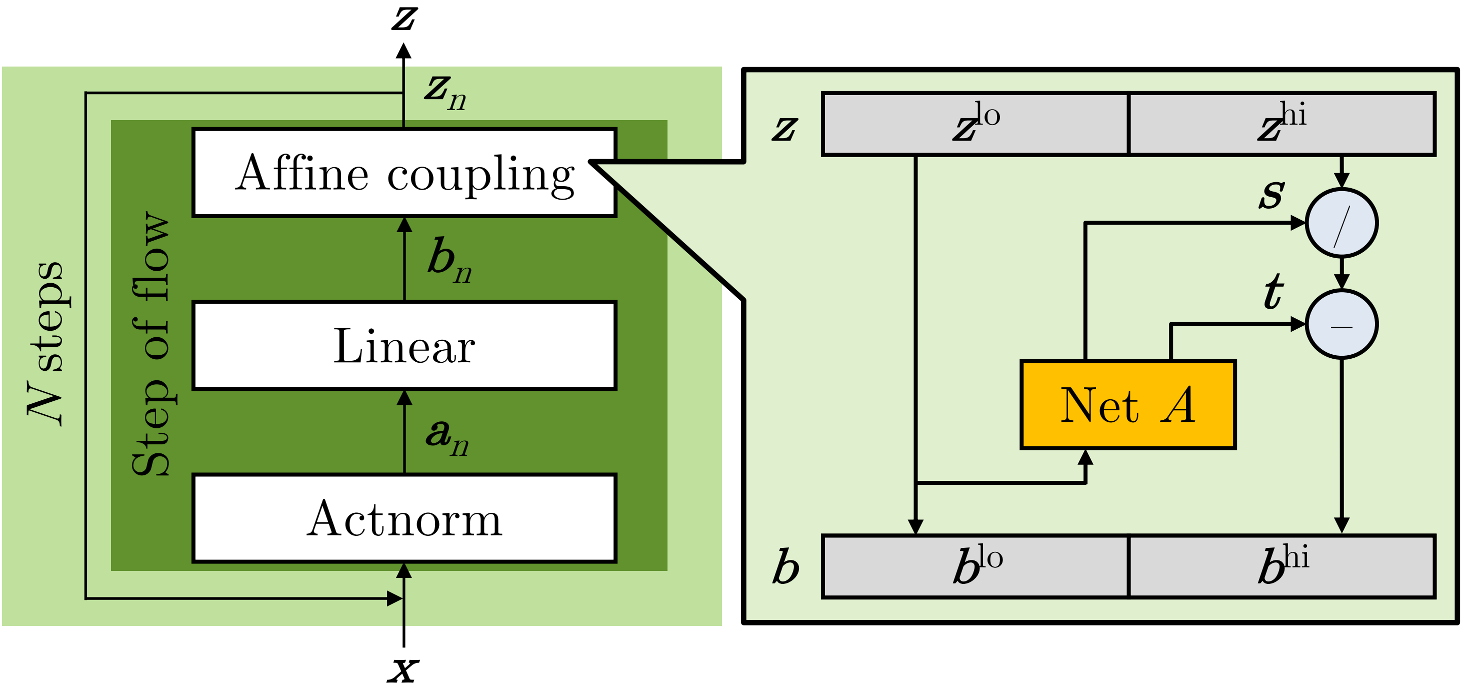

- Activation normalisation
  - Elementwise affine transformation
  - Special initialisation to stand in for batch norm
- Linear transformation
  - Also called “1x1 convolution”
  - Extension of a learned permutation
- Coupling layer
  - Nonlinearly transforms one half of the input variables, based on the remaining half
  - Transformation is invertible
  - A neural network computes the transformation parameters


D. P. Kingma & P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions”, Proc. NeurIPS, 2018

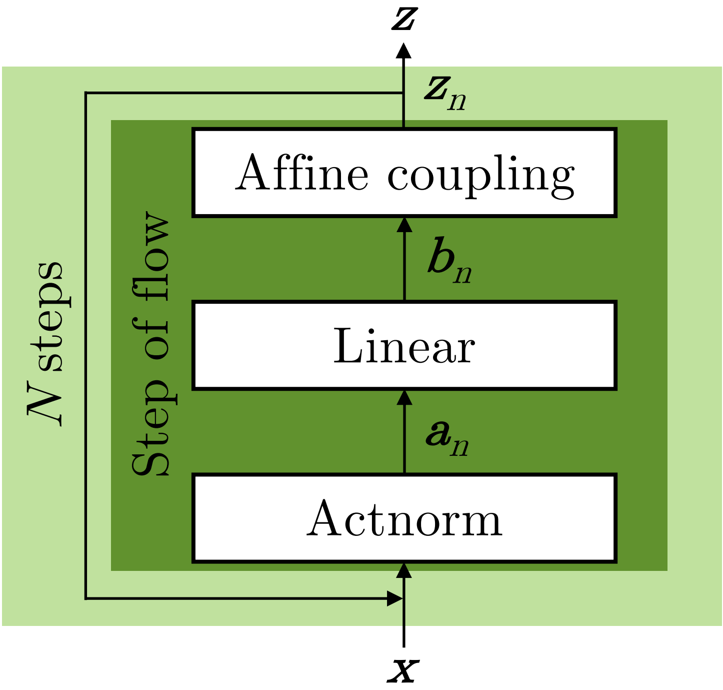
# The affine coupling layer



# Inverting the affine coupling layer



# The effect of the different substeps



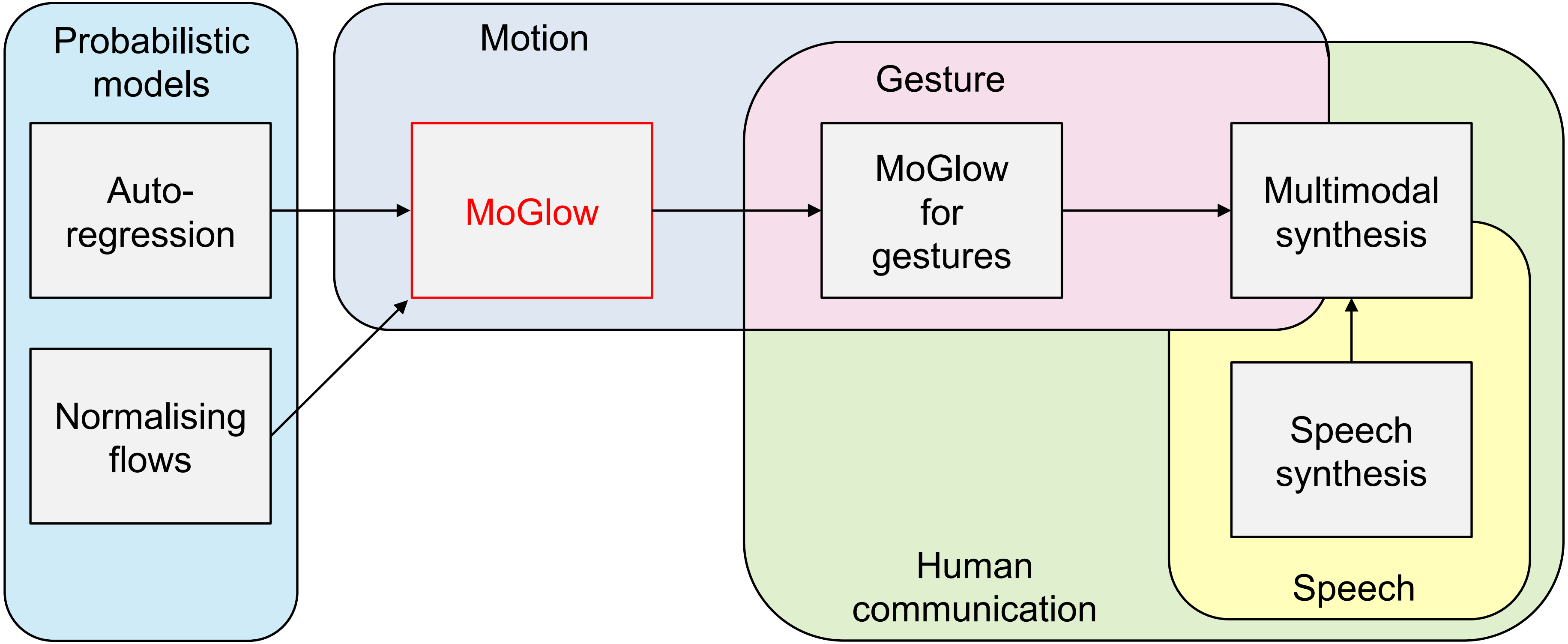

- Without activation normalisation
  - The network may never learn to perform well due to poor initialisation
- Without the linear transformation
  - Half of the output elements will follow a simple Gaussian distribution
    - Since the elements have never been reordered, these elements have never been nonlinearly transformed
- Without the coupling layer
  - The remaining network layers collapse to a simple affine transformation of the input distribution

# Pros and cons of Glow

- Advantages
  - Exact inference
  - Likelihood can be optimised using gradient-based methods
  - Equally fast to compute the forward and backward transformations
- Disadvantages
  - More computations than GANs since $\dim Z = \dim X$
    - Hierarchical structure can reduce computation
  - More layers needed since the transformations are weak
    - Thus more parameters
- My view: “It’s easier to make a good model fast than it is to make a fast model good”


L. Dinh et al., “Density estimation using Real NVP”, Proc. ICLR, 2017

# Graphical overview



# MoGlow: Probabilistic and controllable motion synthesis using normalising flows

Gustav Eje Henter

Simon Alexanderson

Jonas Beskow

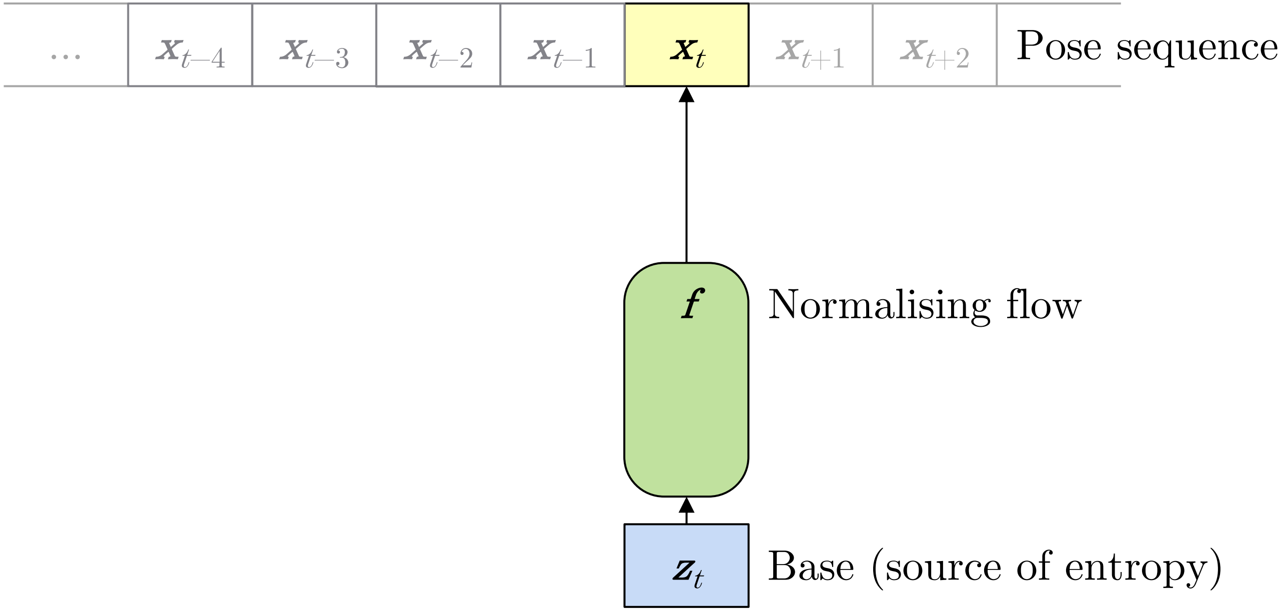
SIGGRAPH Asia 2020

# Sequence modelling

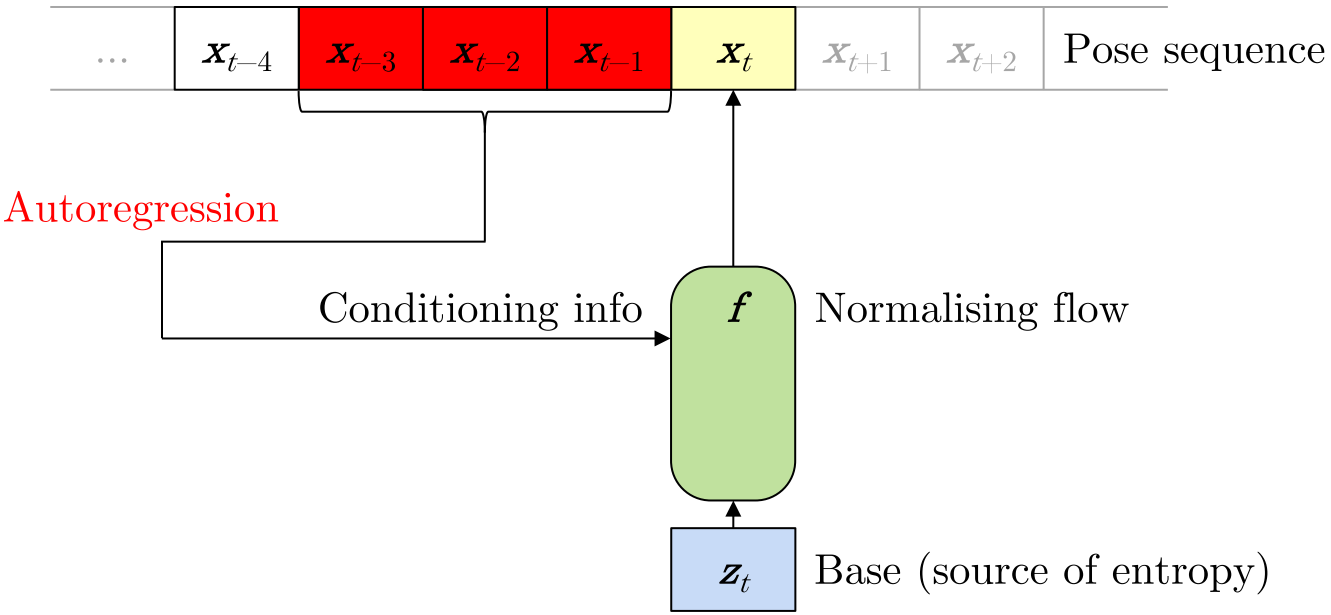
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\cdots$ | $\boldsymbol{x}_{t-4}$ | $\boldsymbol{x}_{t-3}$ | $\boldsymbol{x}_{t-2}$ | $\boldsymbol{x}_{t-1}$ | $\boldsymbol{x}_{t}$ | $\boldsymbol{x}_{t+1}$ | $\boldsymbol{x}_{t+2}$ |

Pose sequence

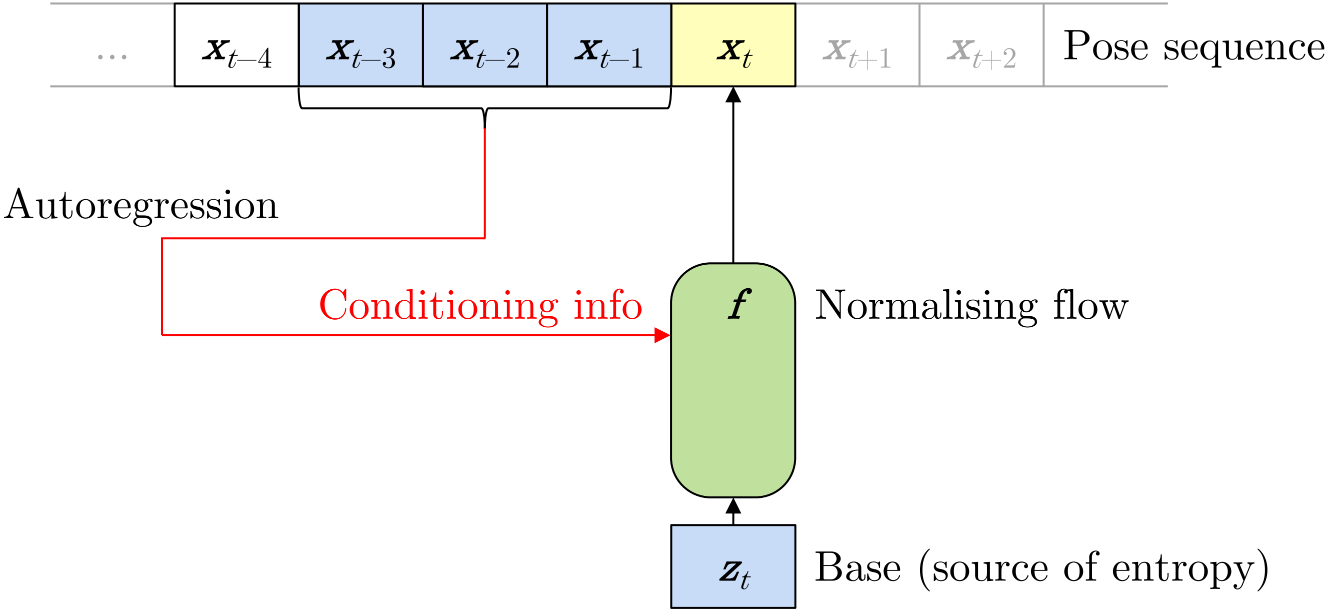
# Sequence modelling



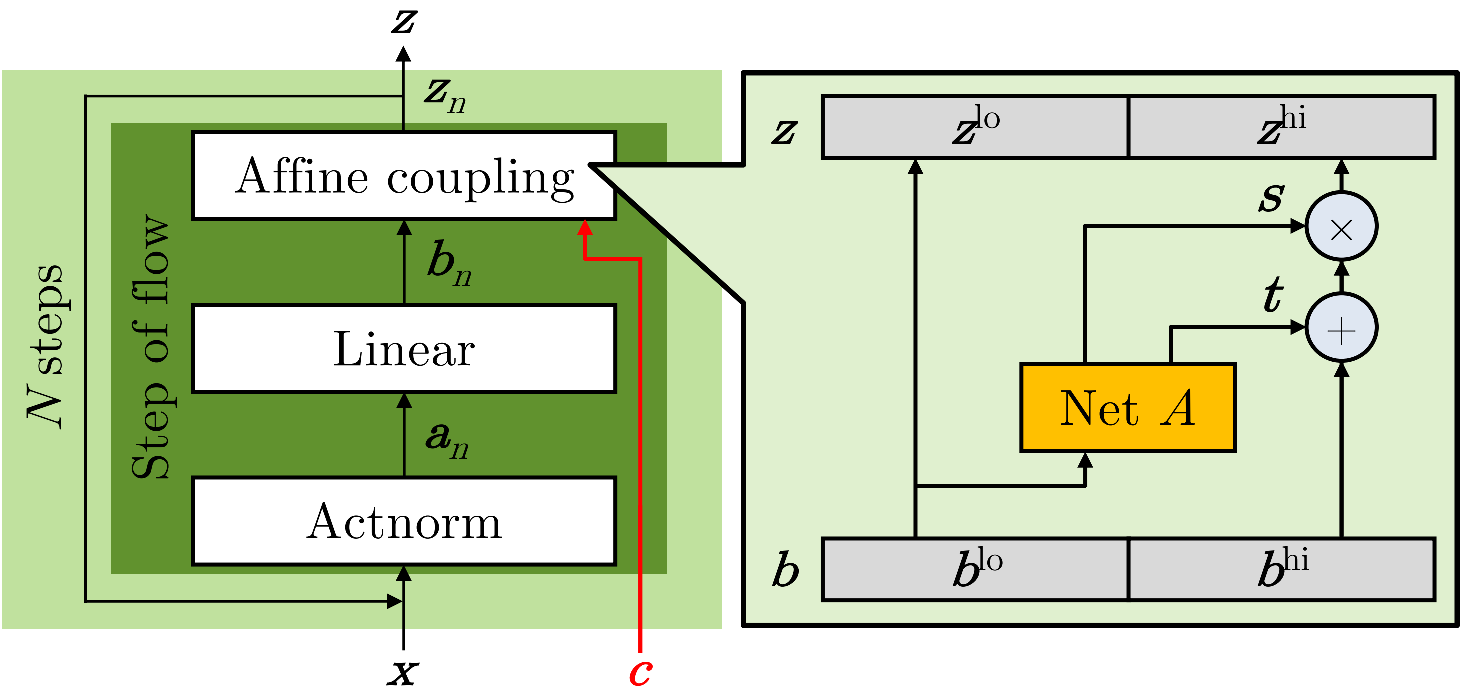
# Autoregression



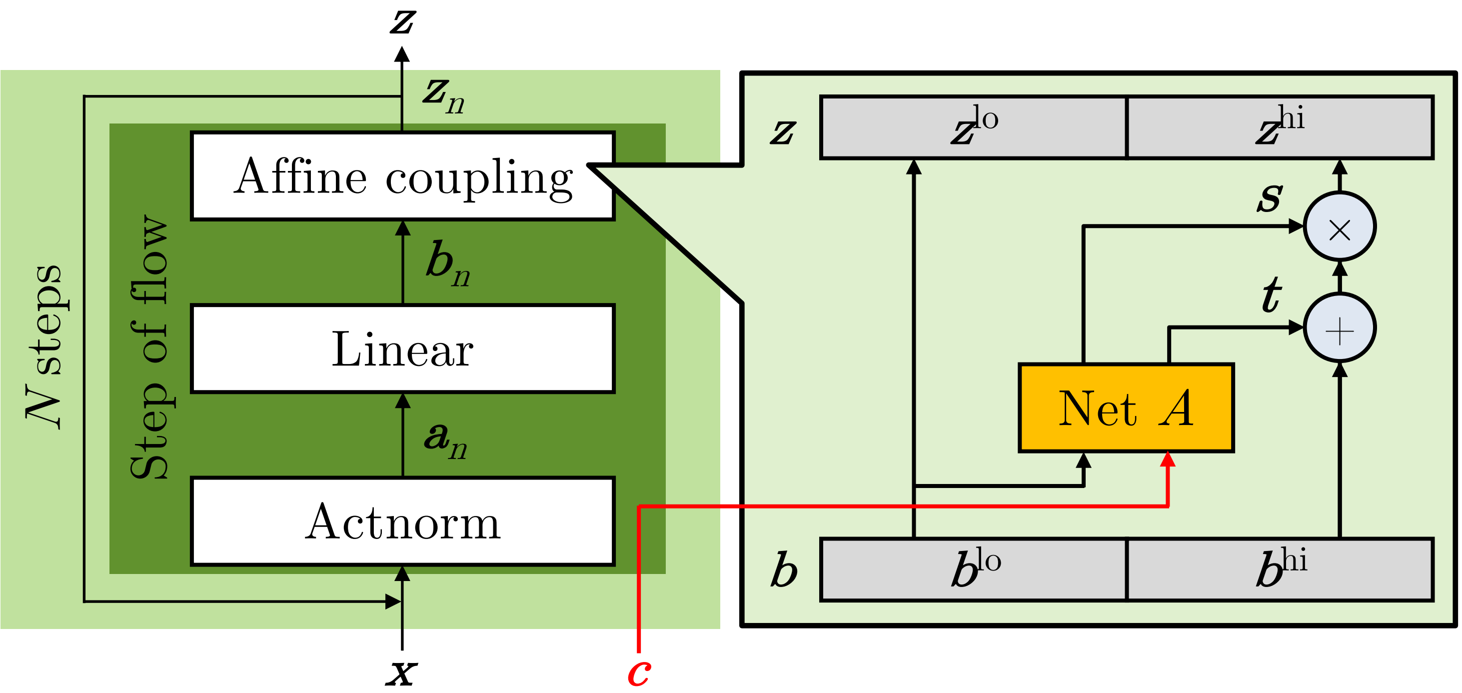
# Conditional Glow



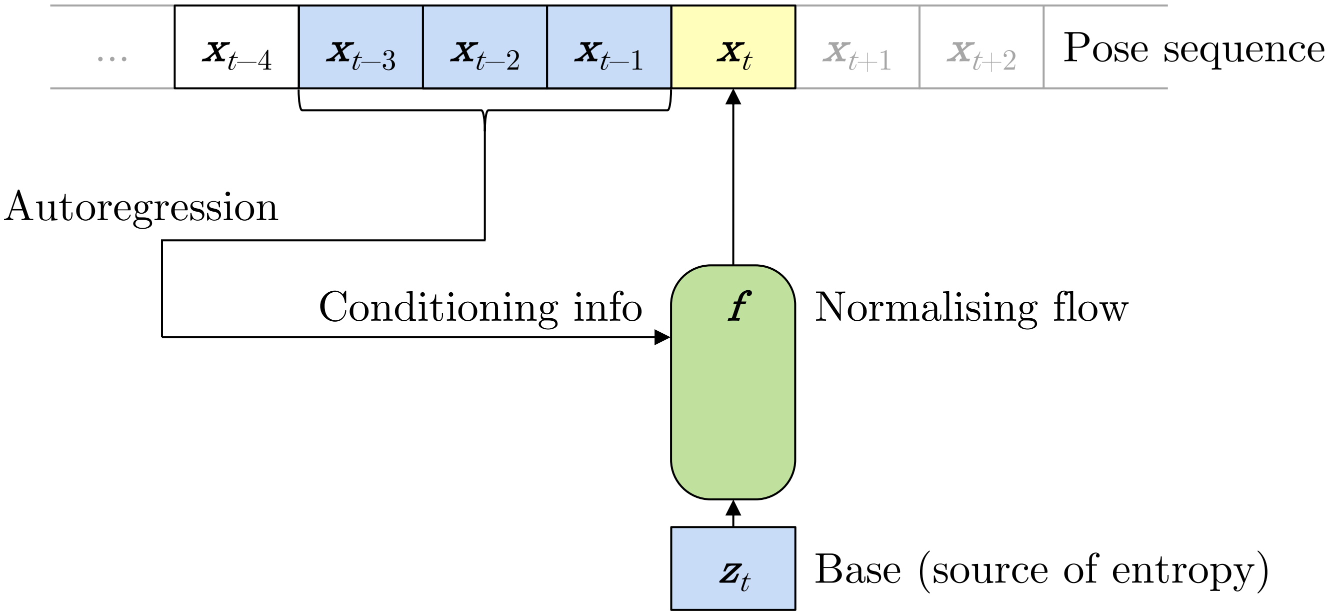
# Conditional Glow



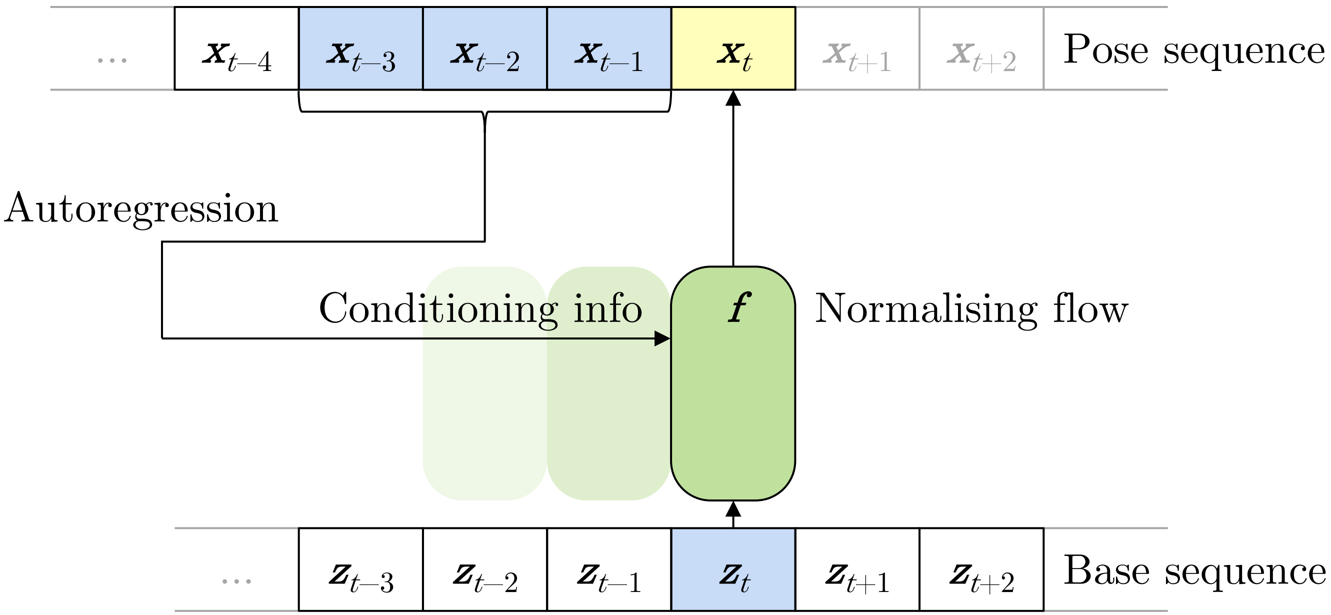
# Conditional Glow



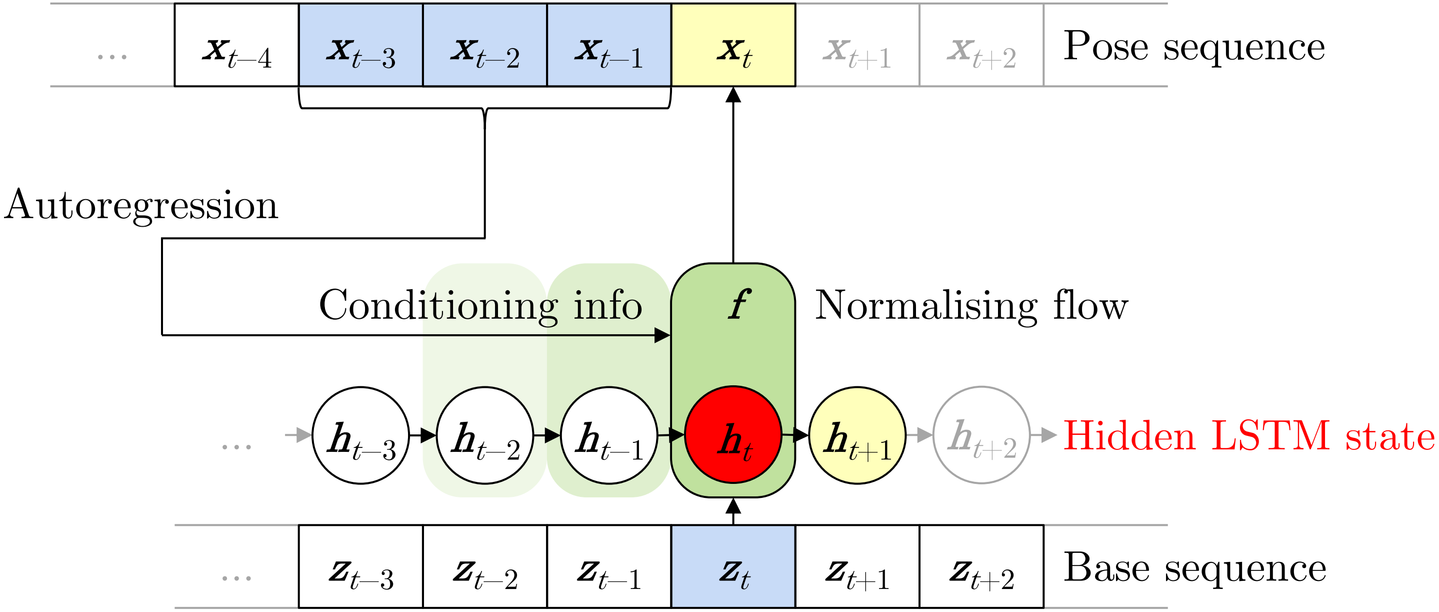
# Autoregression



# Autoregression

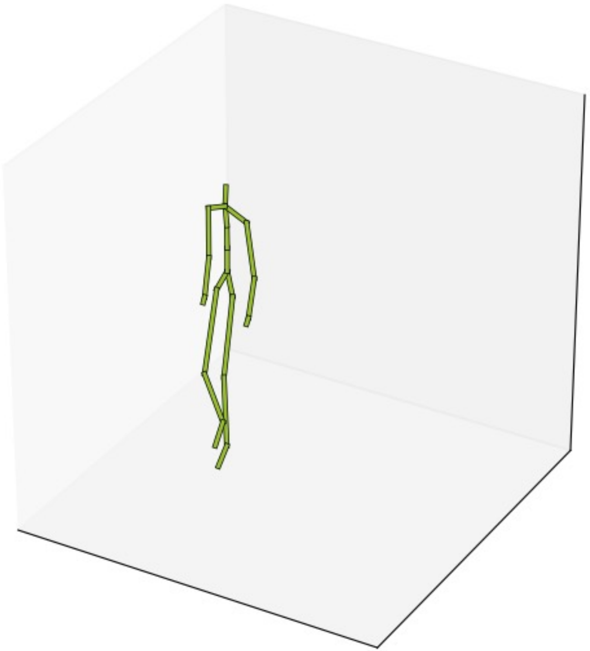


# Long-term memory

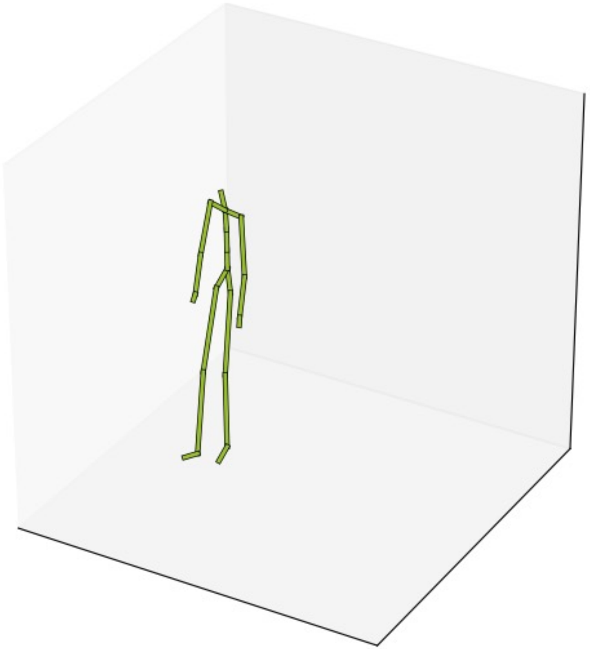


# Effect of long memory on stability

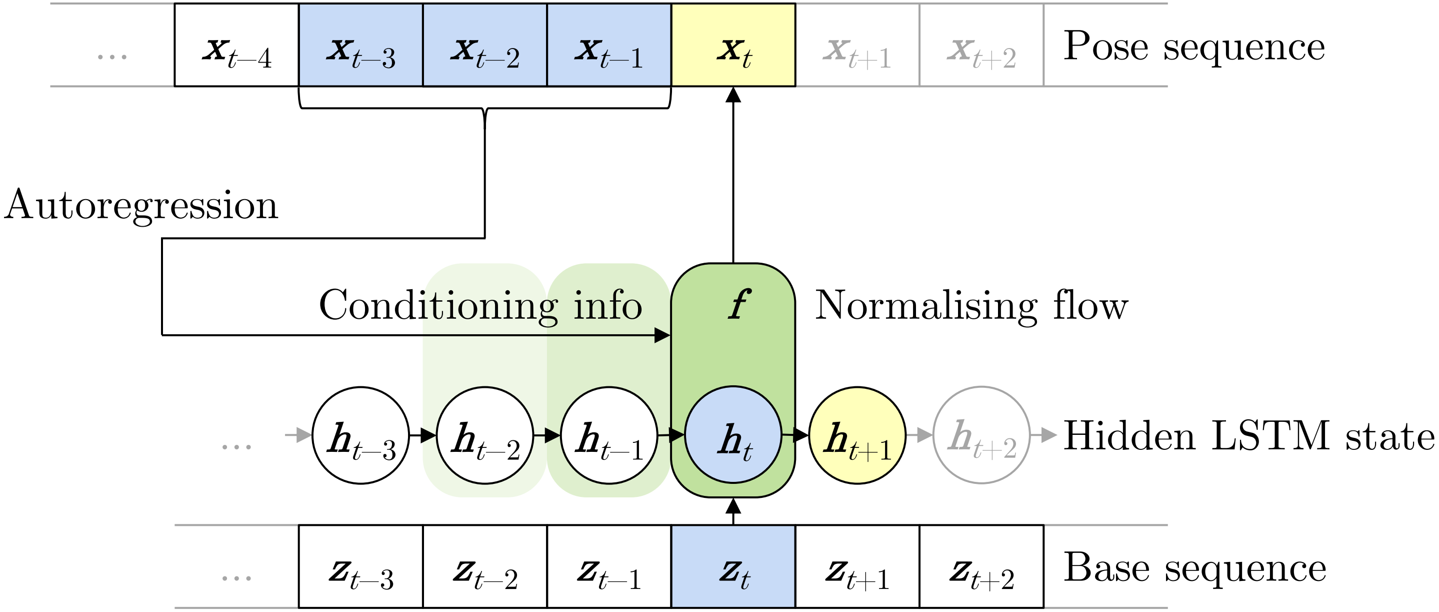
Without LSTMs



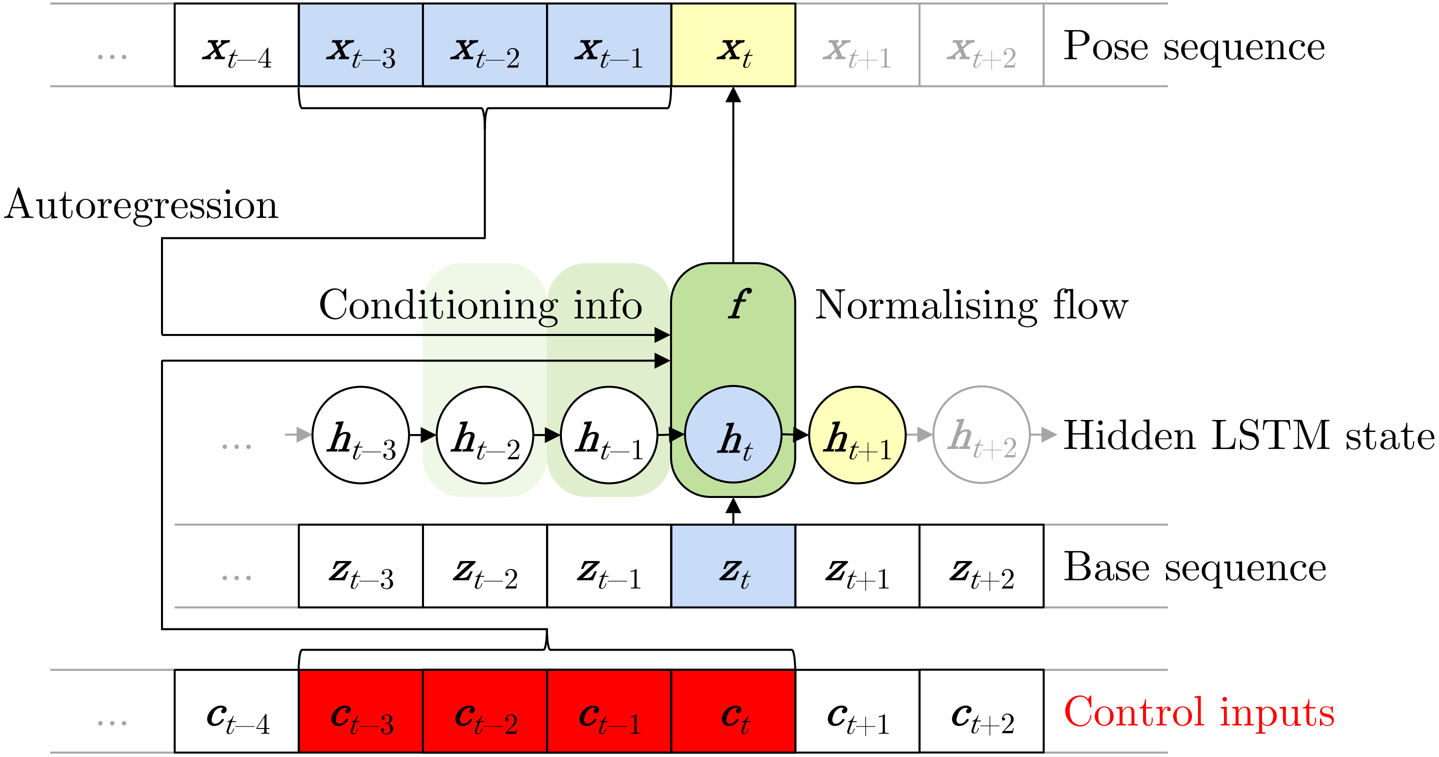
With LSTMs



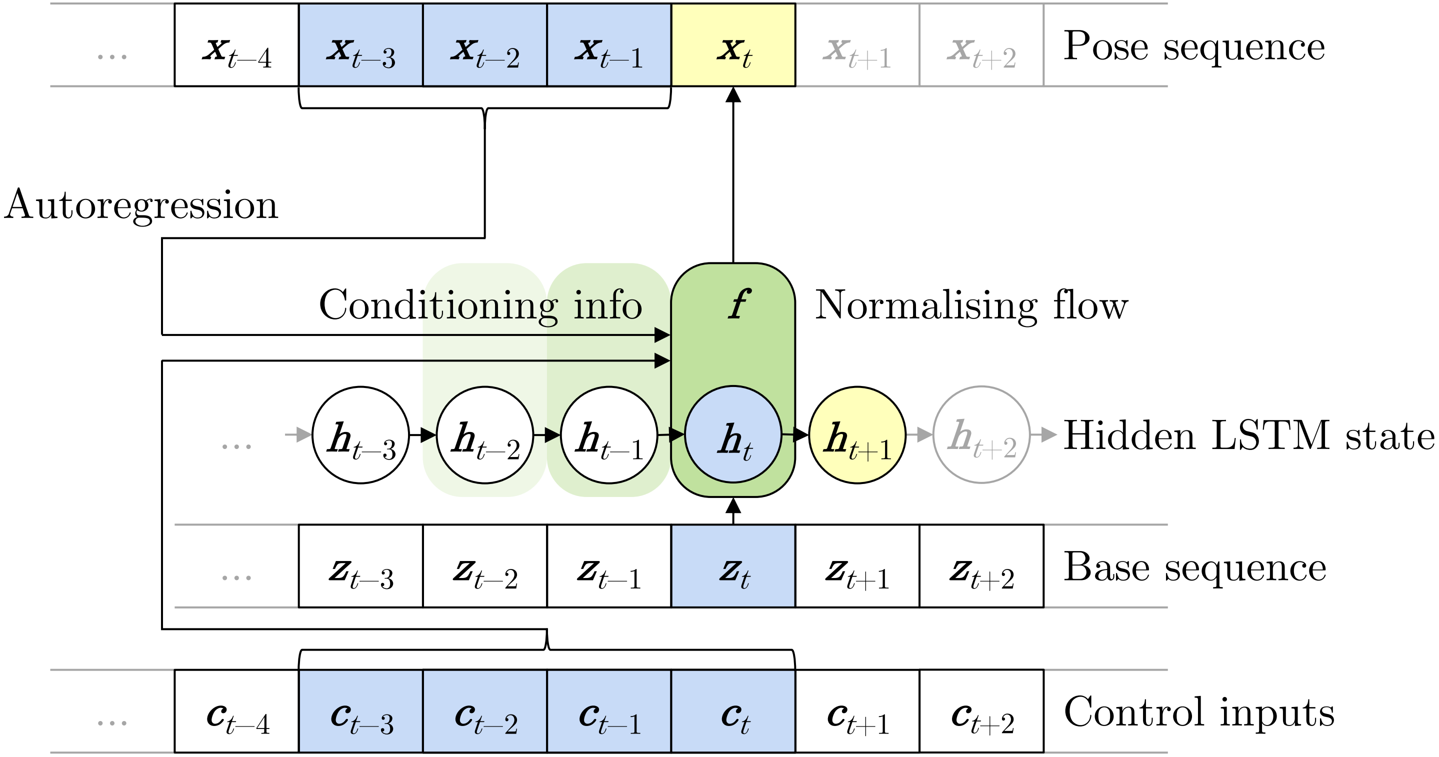
# Achieving control



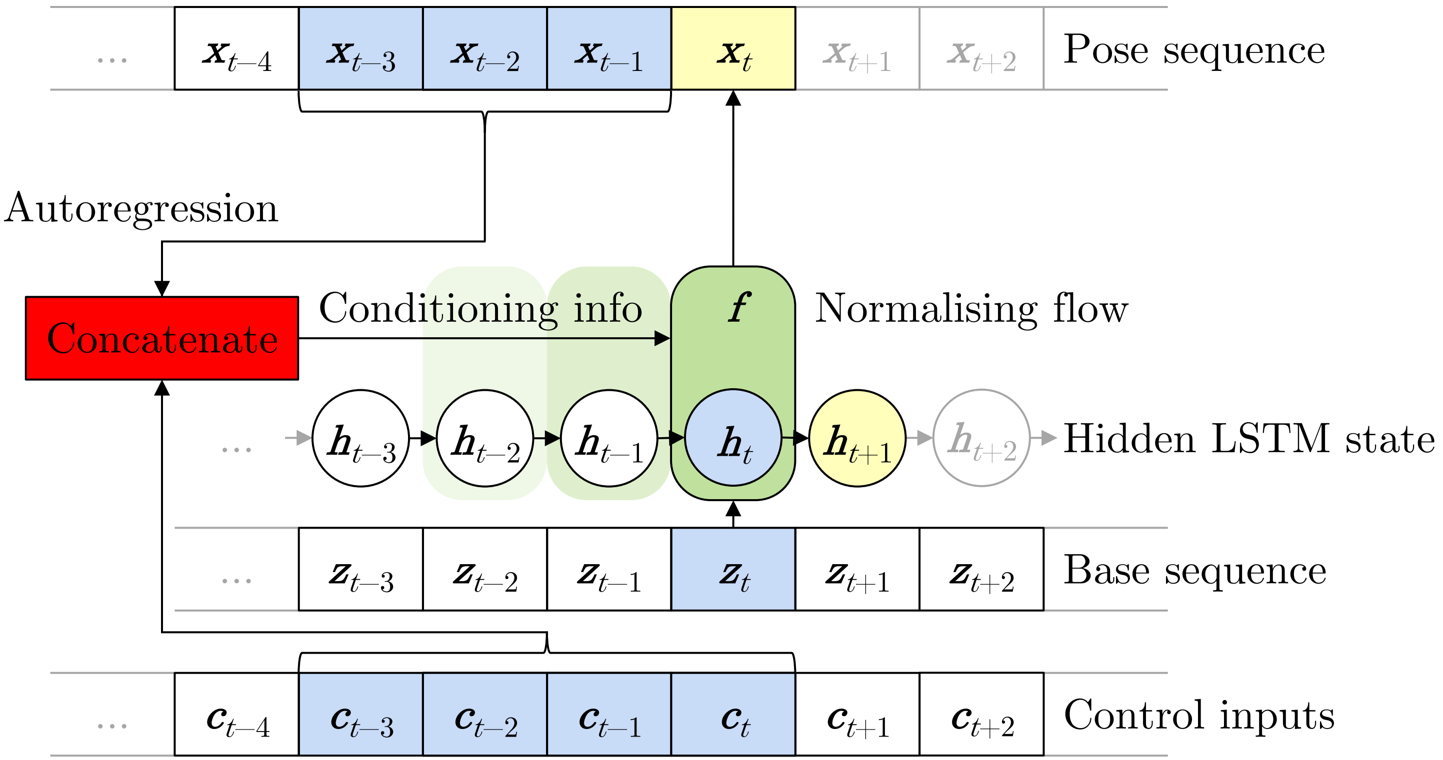
# Achieving control



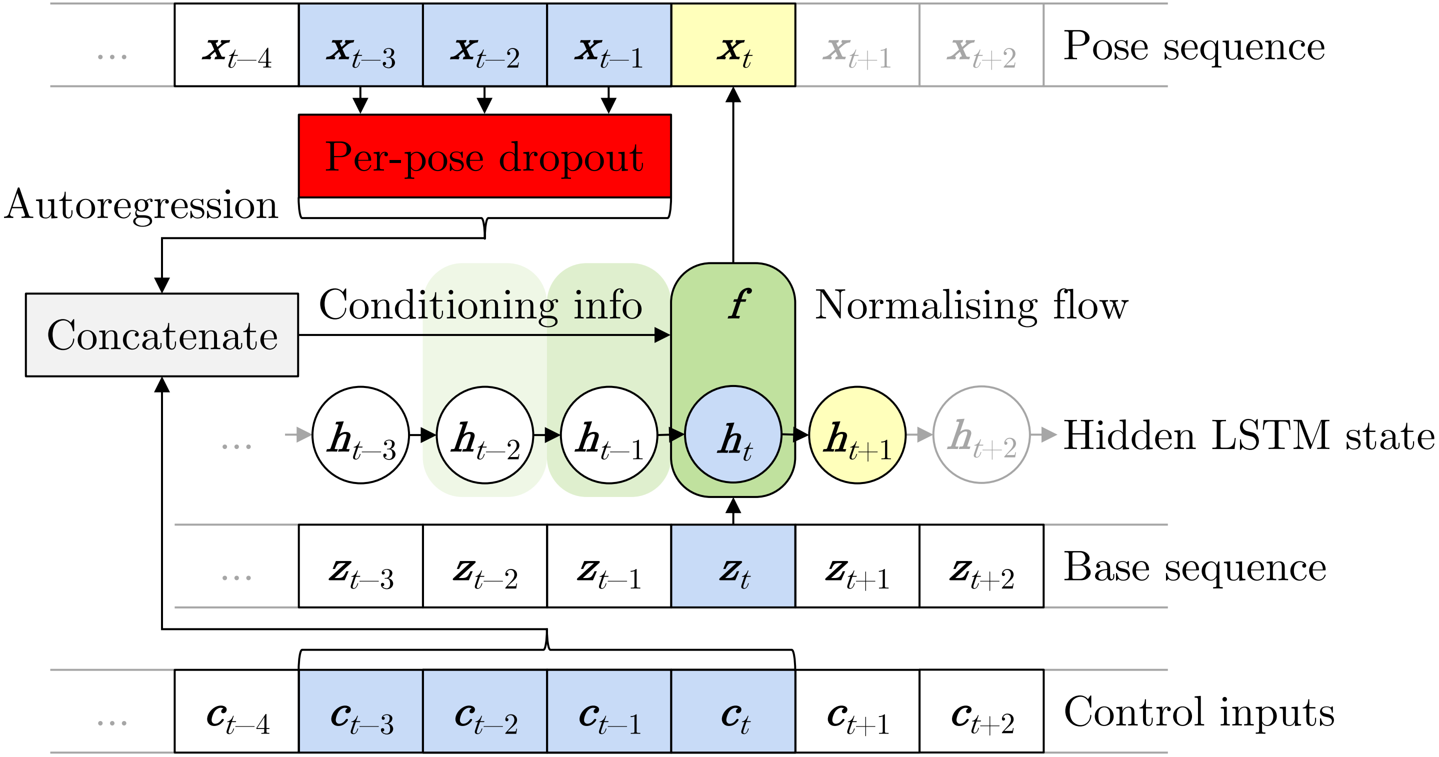
# Achieving control



# Achieving control



# Achieving control

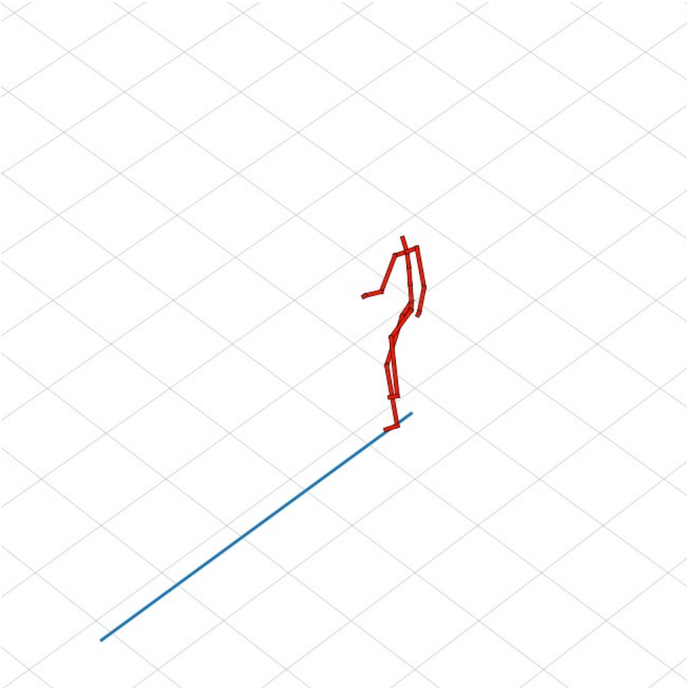




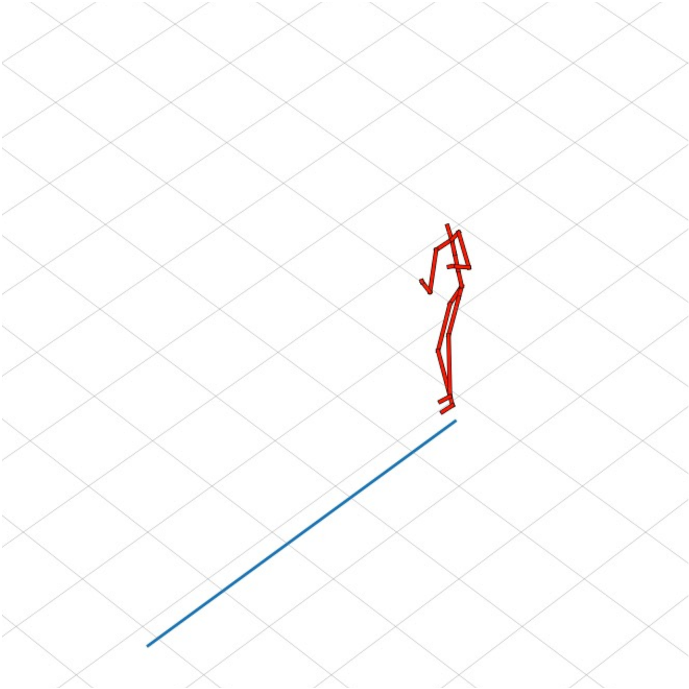

X. Wang et al., "Autoregressive neural f0 model for statistical parametric speech synthesis," IEEE/ACM T. Audio Speech, 2018.

# Effect of data dropout

No dropout

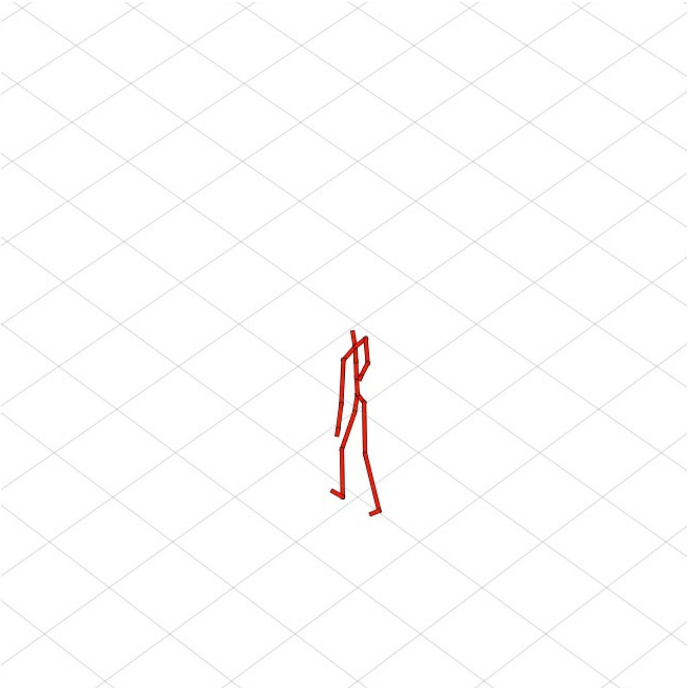


Pose dropout rate = 0.95

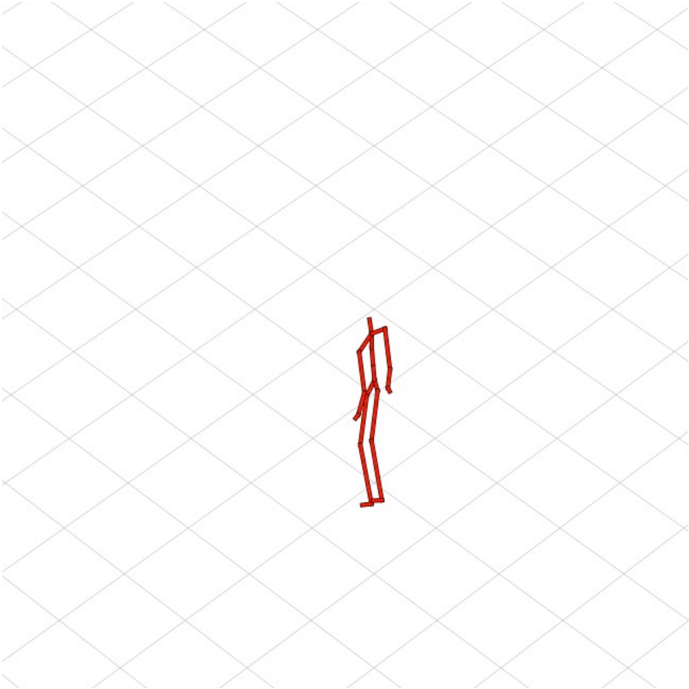


# Effect of data dropout

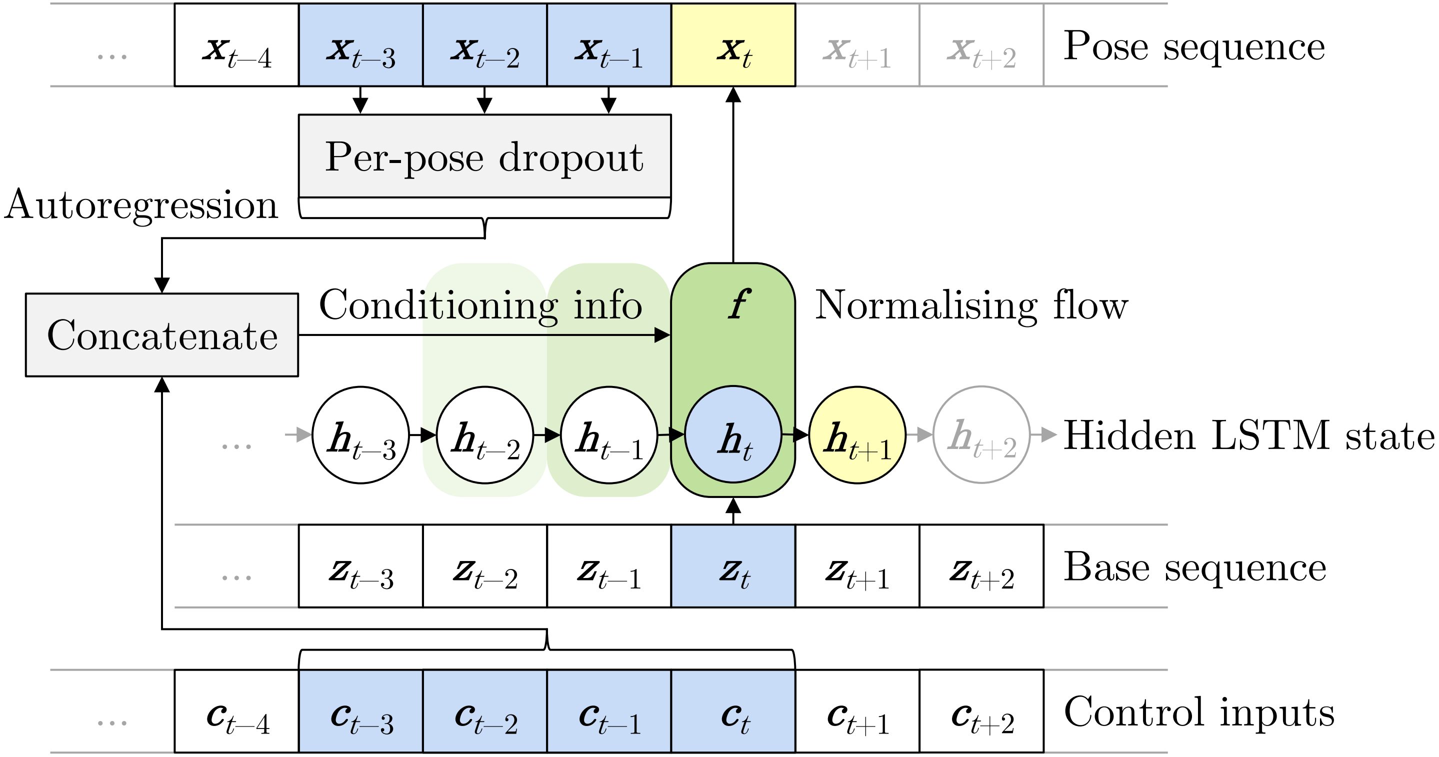
No dropout



Pose dropout rate = 0.95



# Complete MoGlow architecture



# MoGlow advantages

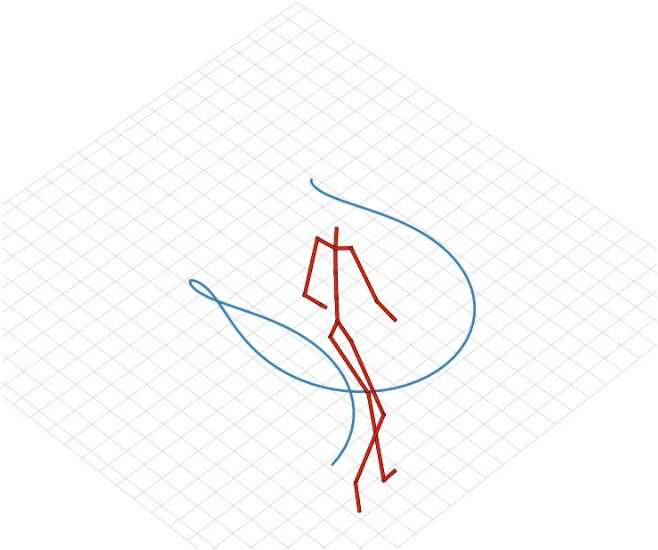
- Probabilistic model
  - Describes *all* possible outcomes, not just a single one
- Implicit generator structure
  - Flexible and fast to sample from, like GANs
- Tractable statistical inference
  - Can be trained to maximise likelihood
- General
  - No assumptions about the nature of the motion (or even that the data is motion at all!)
- Interactively controllable
  - No algorithmic latency
- Gives high-quality results

# Experiments

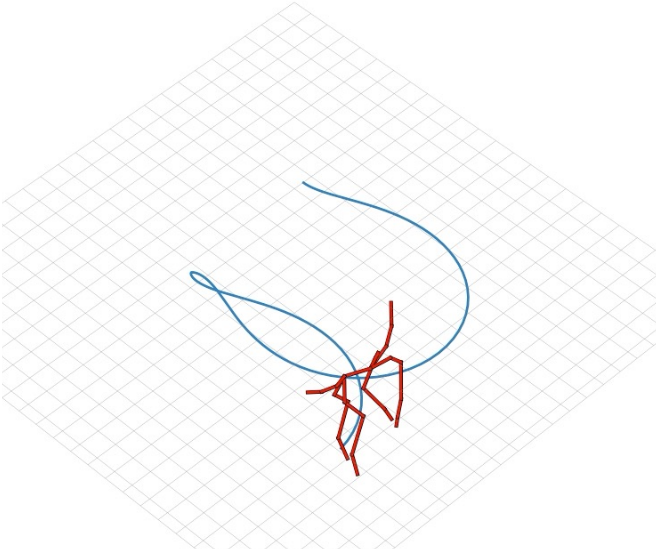
- Initial application: Locomotion synthesis with path control
- Studying locomotion has several advantages:
  - It is easy to spot artefacts and poor adherence to the control
  - Foot-sliding can be quantified objectively
- Control signal: Forward, sideways, and angular velocity of the root node
  - Result: The root node exactly follows a given path through space; the model has to generate a consistent series of poses along the way
  - The path dictated by the control signal is visualised as a blue curve projected onto the ground plane in videos

# Locomotion synthesis tasks

Bipedal locomotion



Quadrupedal locomotion

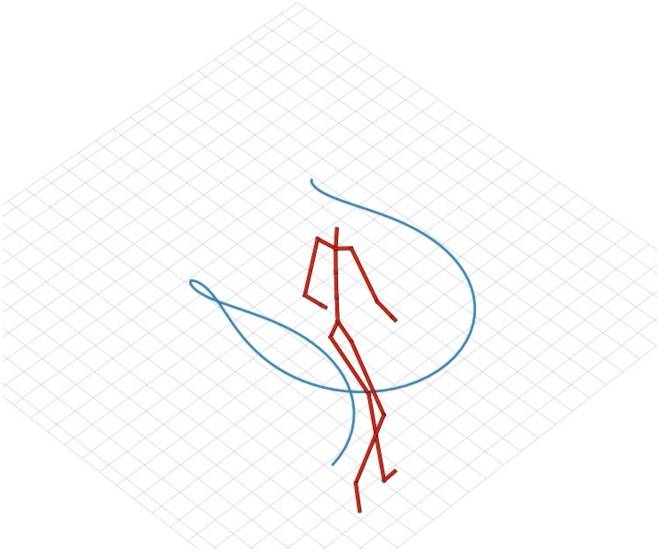




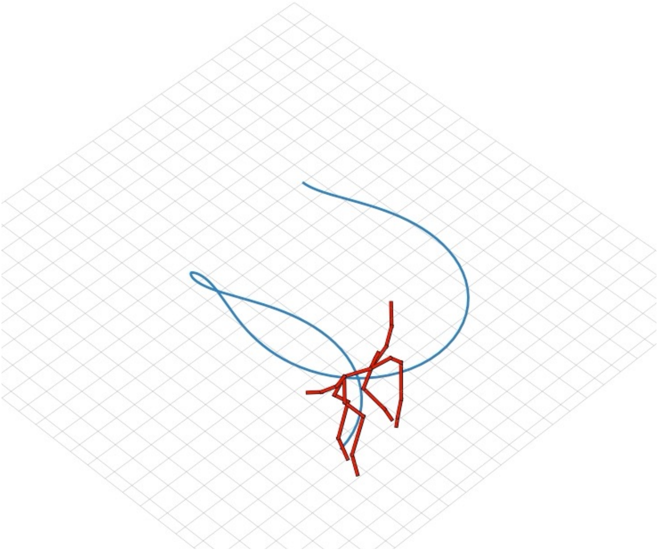
Data from I. Habibie et al., “A recurrent variational autoencoder for human motion synthesis”, Proc. BMVC, 2017 & CMU Graphics MoCap Database & HDM05

# Locomotion synthesis tasks

Bipedal locomotion



Quadrupedal locomotion

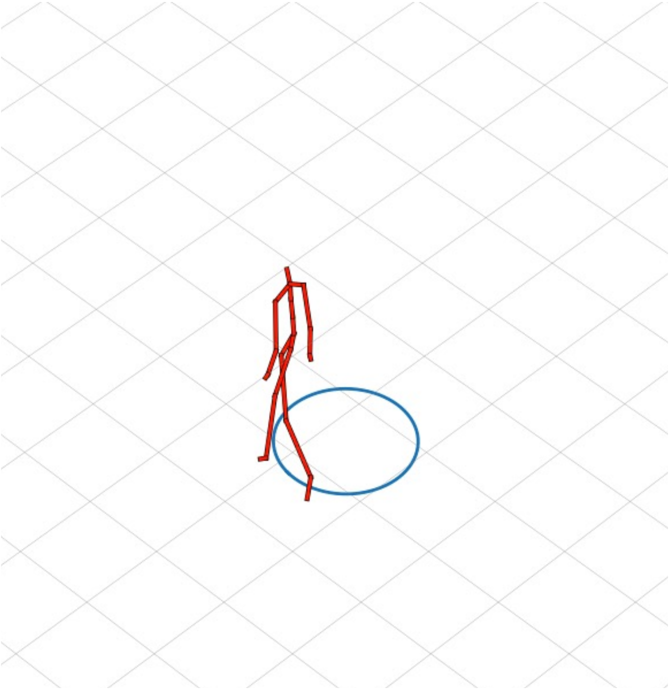




Data from H. Zhang et al., “Mode-adaptive neural networks for quadruped motion control”, ACM T. Graphic., 2018

# Pose representations

Joint positions



Joint angles/rotations

# Systems trained

| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Num. params. Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |

# Systems trained

| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |



D. Greenwood et al., “Predicting head pose in dyadic conversation”, Proc. IVA., 2017.

# Systems trained

| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Num. params. Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |


D. Pavllo et al., “QuaterNet: A quaternion-based recurrent model for human motion,” Proc. BMVC, 2018.

# Systems trained

| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Num. params. Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |



H. Zhang et al. “Mode-adaptive neural networks for quadruped motion control,” ACM Trans. Graph, 2018.

# Systems trained

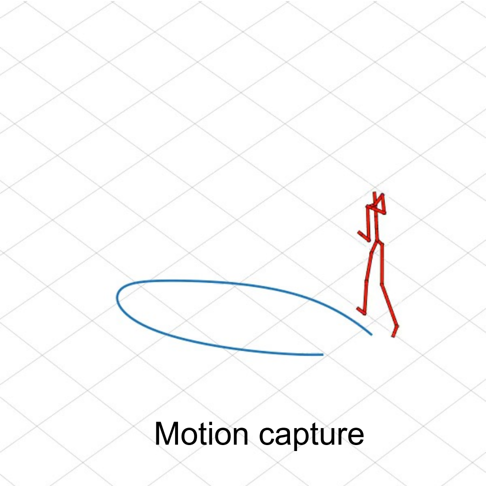
| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |

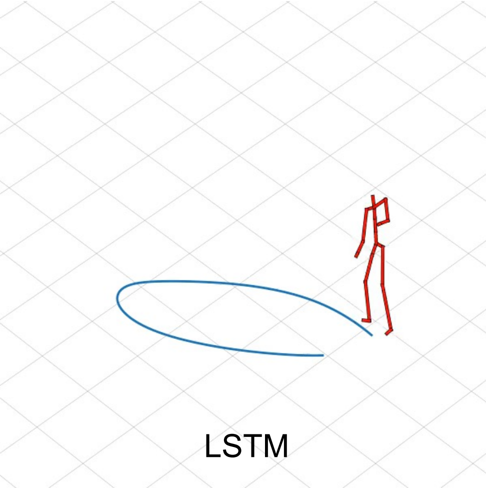
# Systems trained

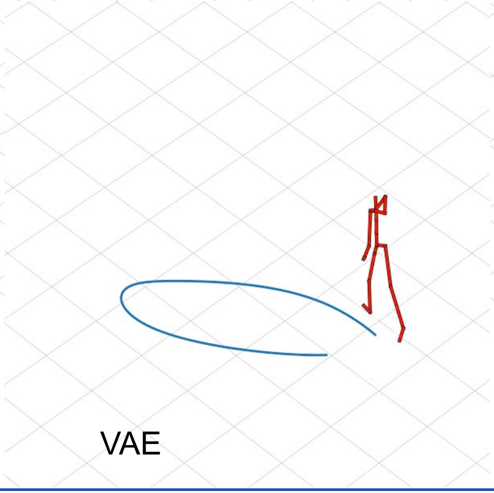
| | Configuration | ID | Proba-bilistic? | Task-agnostic? | Algo. latency | Context frames | Hidden state | Pose dropout | Num. params. Man | Dog |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Plain LSTM | RNN | ✗ | ✓ | None | - | LSTM | - | 1M | 1M |
| | Greenwood et al. [2017a] | VAE | Partially | ✓ | Full seq. | - | BLSTM | - | 4M | 4M |
| | Pavllo et al. [2018] | QN | ✗ | ✗ | 1 sec. | - | GRU | - | 10M | - |
| | Zhang et al. [2018] | MA | ✗ | ✗ | 1 sec. | 12 | - | - | - | 5M |
| | MoGlow | MG | ✓ | ✓ | None | 10 | LSTM | 95% | 74M | 80M |
| Ablats. | No pose dropout | MG-D | " | " | " | 10 | " | 0% | 74M | - |
| | No pose context | MG-A | " | " | " | 10 | " | 100% | 74M | - |
| | Minimal history | MG-H | " | " | " | 1 | " | 95% | 54M | - |

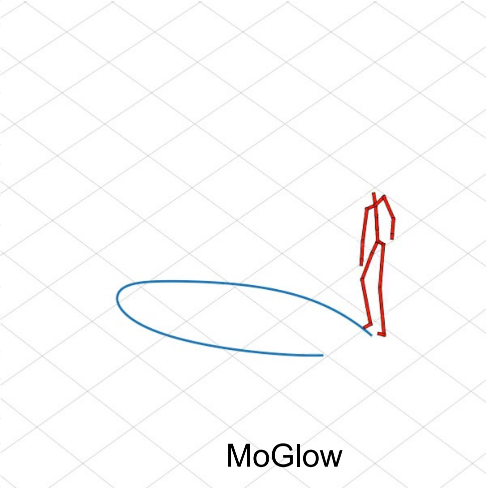
# Comparisons

Held-out control signal



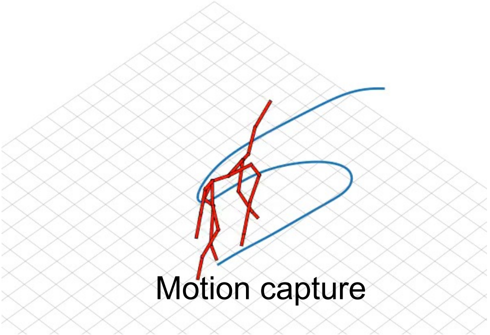



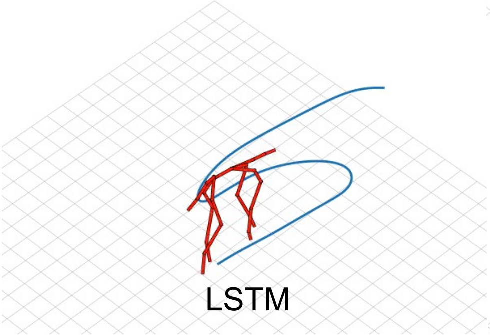



# Comparisons

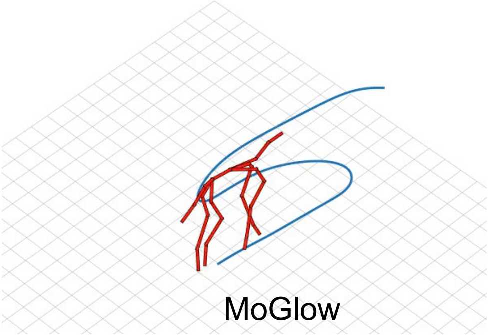
Held-out control signal



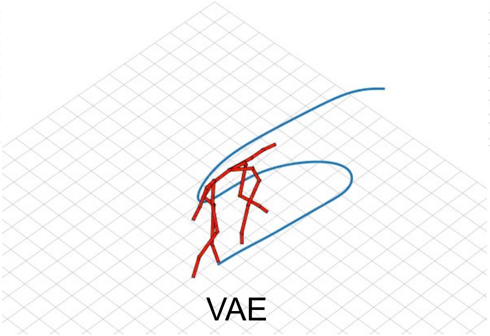
Motion capture



LSTM



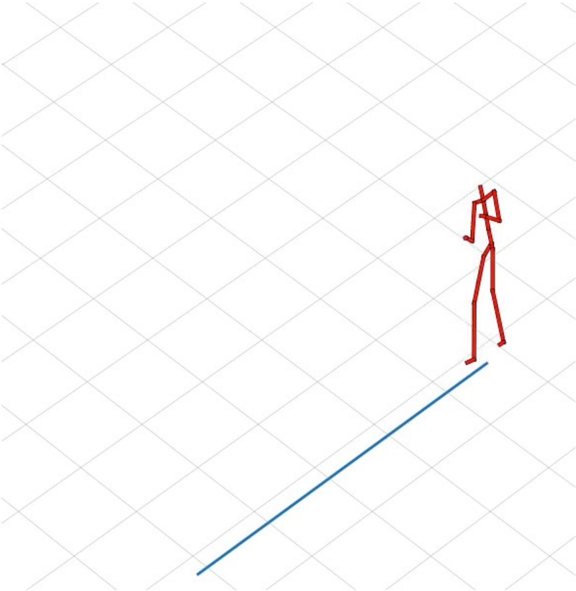
MoGlow



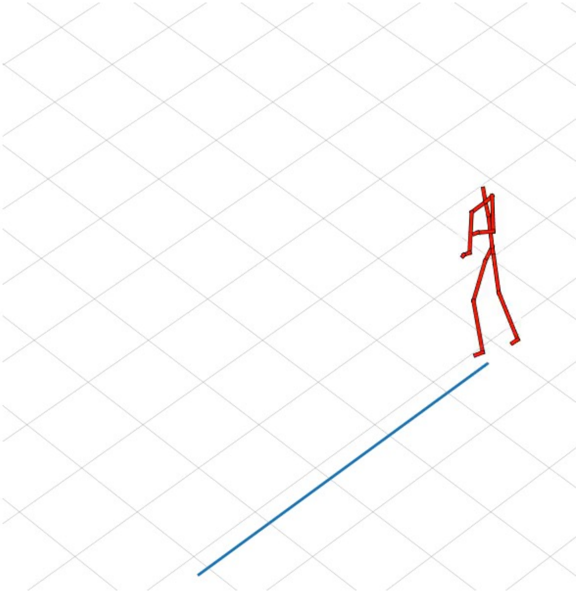
VAE

# Comparisons

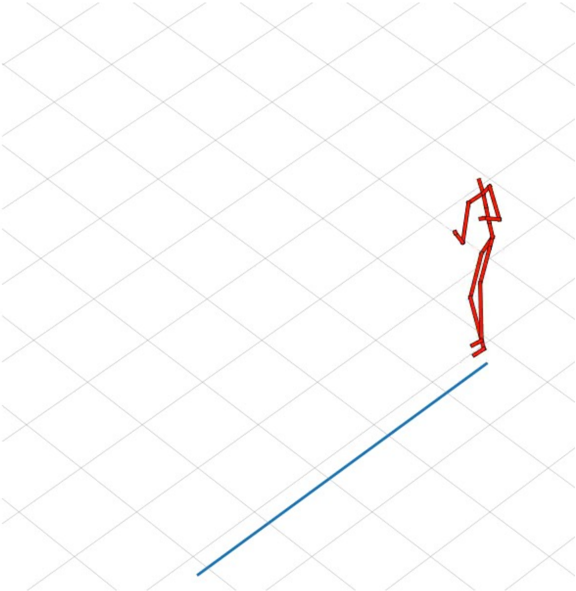
Synthetic control signal



LSTM



VAE



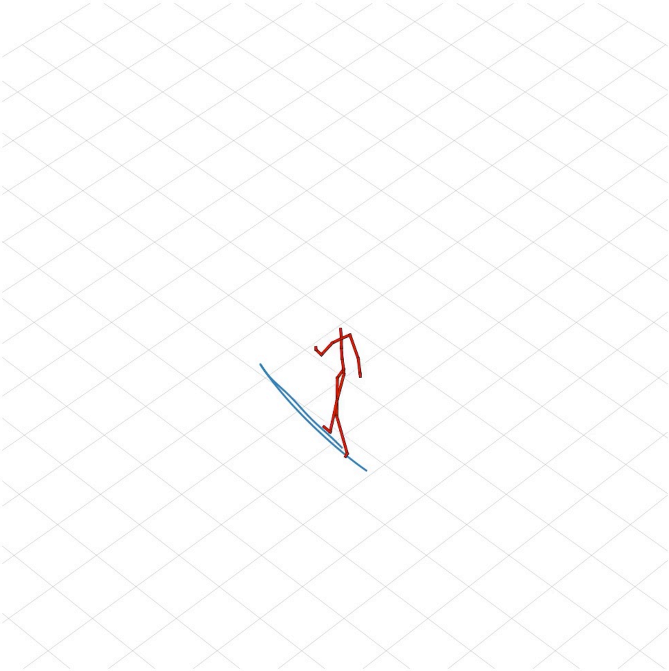
MoGlow

# Comparisons

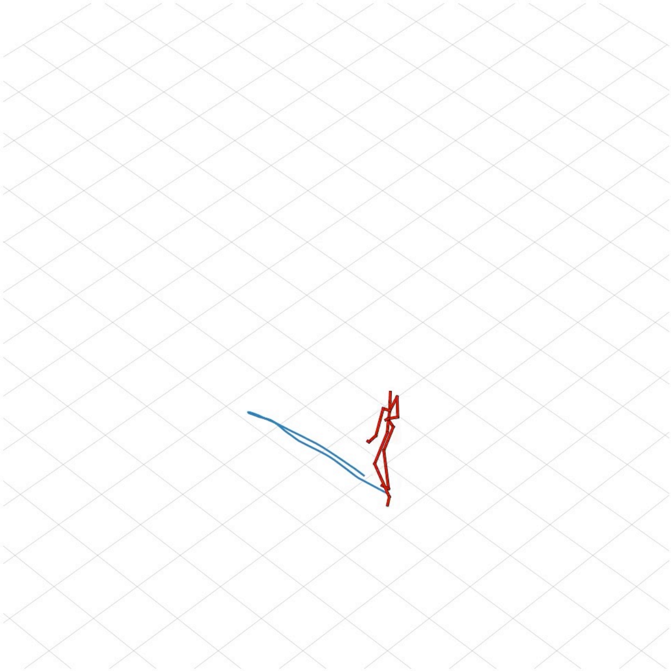
Synthetic control signal

# QuaterNet on held-out control signals

NAT

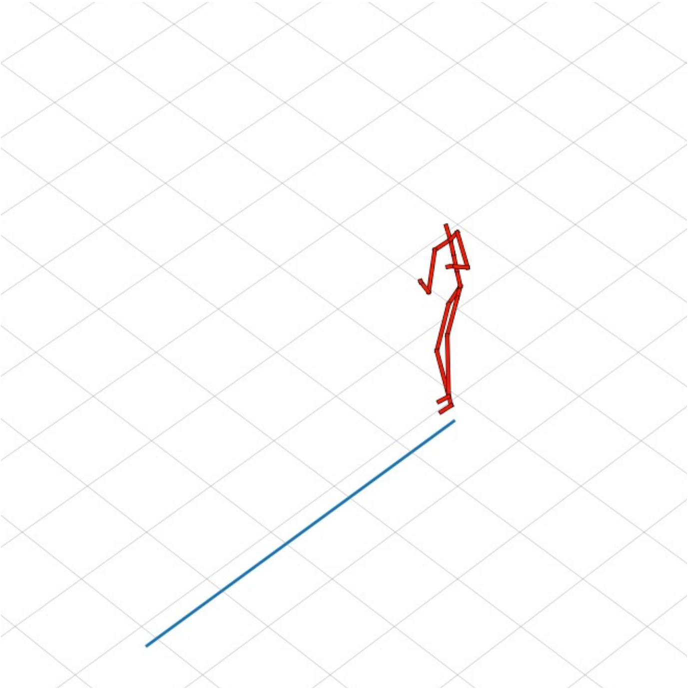


QN

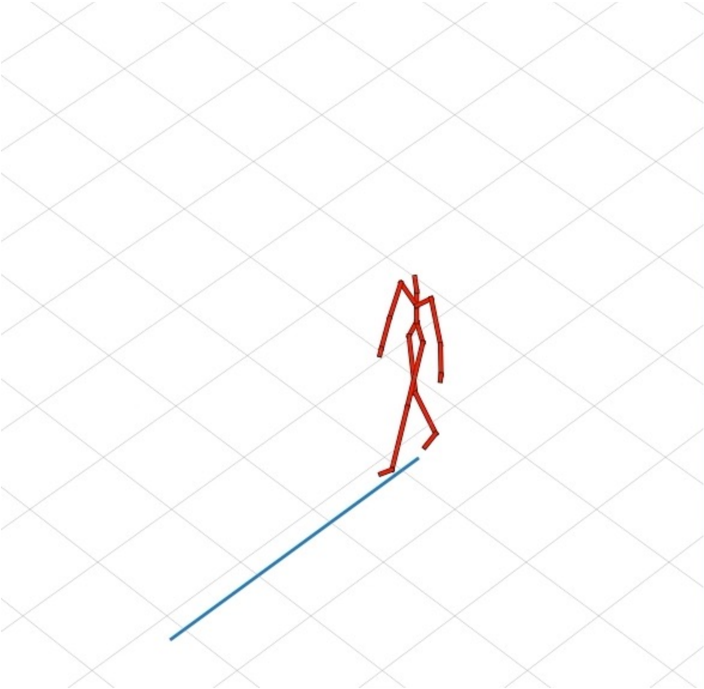


# QuaterNet on synthetic control signals

MG



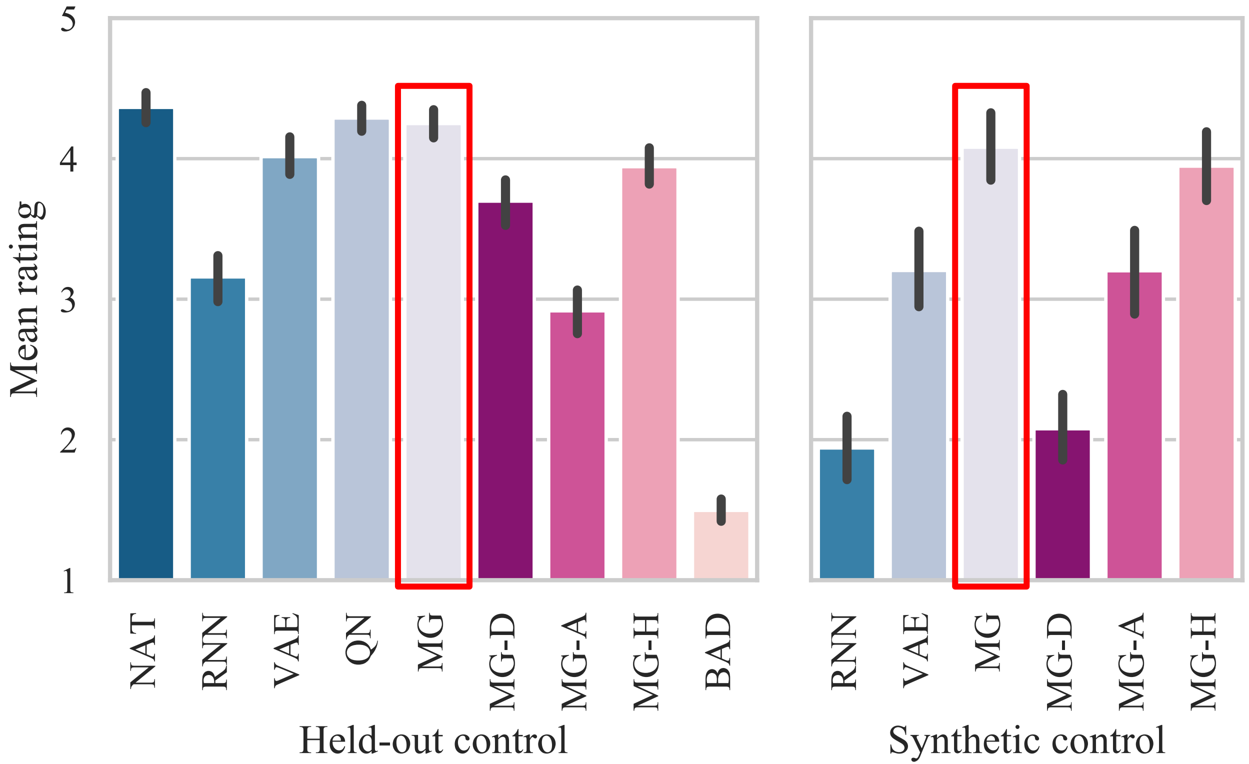
QN



# Evaluations

- Footstep analysis
  - In locomotion generation, the most noticeable artefacts are foot sliding, which is easy to quantify objectively
  - Please see the paper for the results
- Crowdsourced subjective evaluation
  - Figure Eight platform
  - “Grade the perceived naturalness of the animation from 1 to 5”
  - Held-out and synthetic control input
  - Bad clips and too rapid responses were used to filter out unreliable raters
  - 3,550/4,289 ratings analysed (human/dog)
- No foot stabilisation or other post-processing used

# Results of user study on the human data



# Average subjective ratings

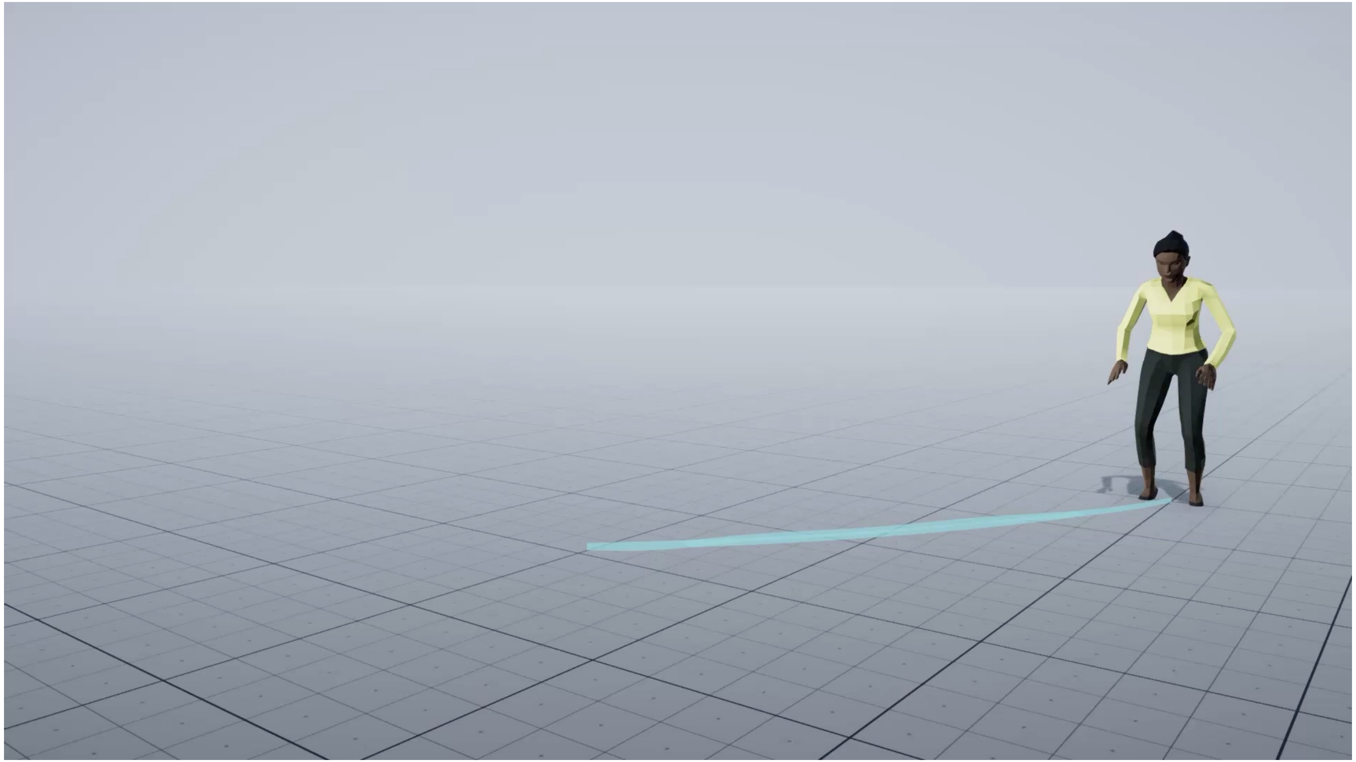
| ID | Human | | Quadruped | |
|---|---|---|---|---|
| | Held-out $\boldsymbol{c}$ | Synthetic $\boldsymbol{c}$ | Held-out $\boldsymbol{c}$ | Synthetic $\boldsymbol{c}$ |
| NAT | 4.27±0.11 | - | $4.25 \pm 0.06^{**}$ | - |
| RNN | $3.10 \pm 0.15^{**}$ | $1.9 \pm 0.2^{**}$ | $2.81 \pm 0.10^{**}$ | $1.14 \pm 0.04^{**}$ |
| VAE | 3.95±0.13 | $3.1 \pm 0.3^{**}$ | 3.55 ± 0.08 | $2.14 \pm 0.20^{**}$ |
| QN | 4.21±0.10 | - | - | - |
| MA | - | - | - | 3.78 ± 0.10 |
| MG | 4.17±0.11 | 4.0±0.2 | 3.71 ± 0.18 | 3.57 ± 0.20 |
| MG-D | $3.66 \pm 0.16^{**}$ | $2.1 \pm 0.2^{**}$ | - | - |
| MG-A | $2.86 \pm 0.16^{**}$ | $3.2 \pm 0.3^{**}$ | - | - |
| MG-H | $3.87 \pm 0.13^{*}$ | 3.9±0.3 | - | - |

# Validating the probabilistic aspects

- Can we get meaningfully different output for the same control input?
- Will more diverse data enable more diverse output motion?
- Also demonstrated on skinned characters
  - Trained on two different motion-capture datasets for video games applications
    - *LaFAN1 dataset from Ubisoft*
    - *Kinematica Demo dataset from Unity*
  - Joint angles (represented using the exponential map)
  - 60% dropout rate gave smoother motion


LaFAN1 data from F. G. Harvey et al., “Robust motion in-betweening”, ACM Trans. Graph., 2020.

# Random samples with the same control input



# Complicated and unusual motion

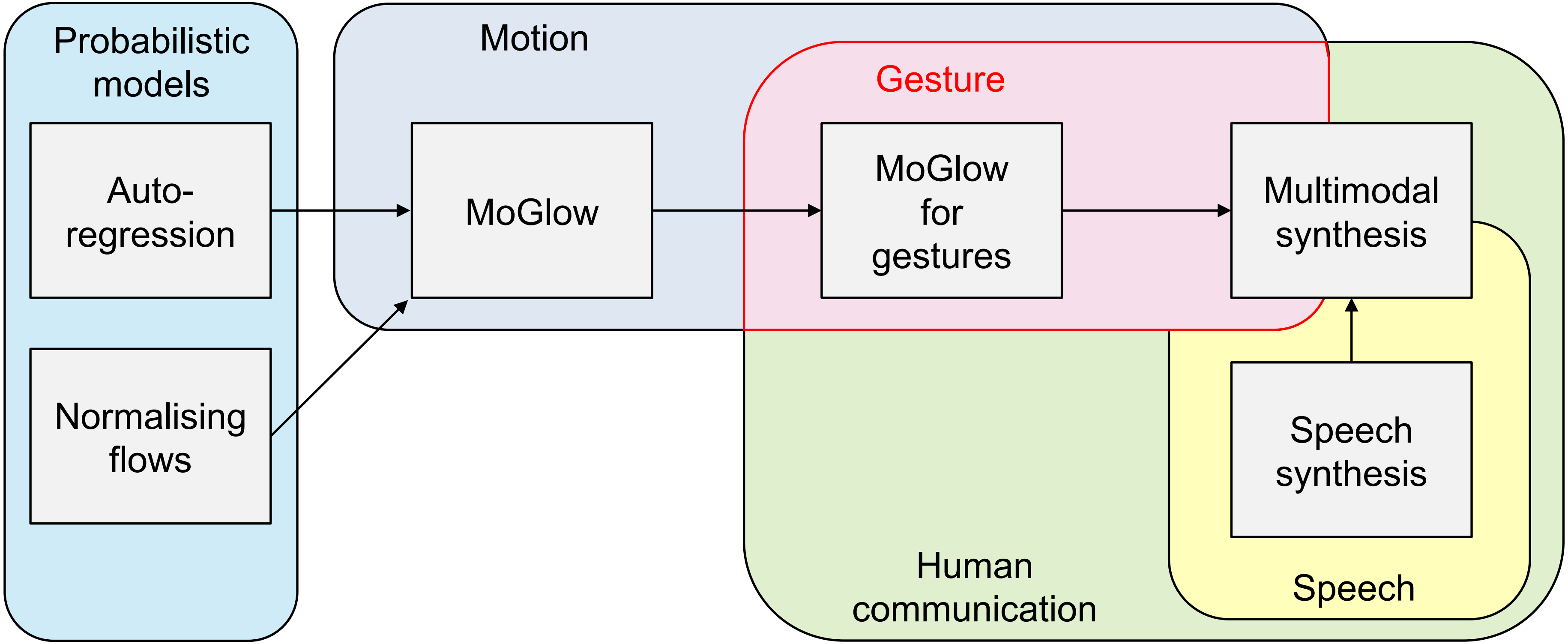
# What we learned

- Normalising flows deliver on their promise
  - Easy to train, fast to generate from, and flexible enough to describe believable motion
- Probabilistic motion modelling *works!*
  - We can describe many different outcomes in one model
- Interactive motion control without algorithmic latency is possible
- Results score close to natural motion
  - The same approach works for two different morphologies
  - And different pose representations
  - The model generalises well to synthetic motion trajectories

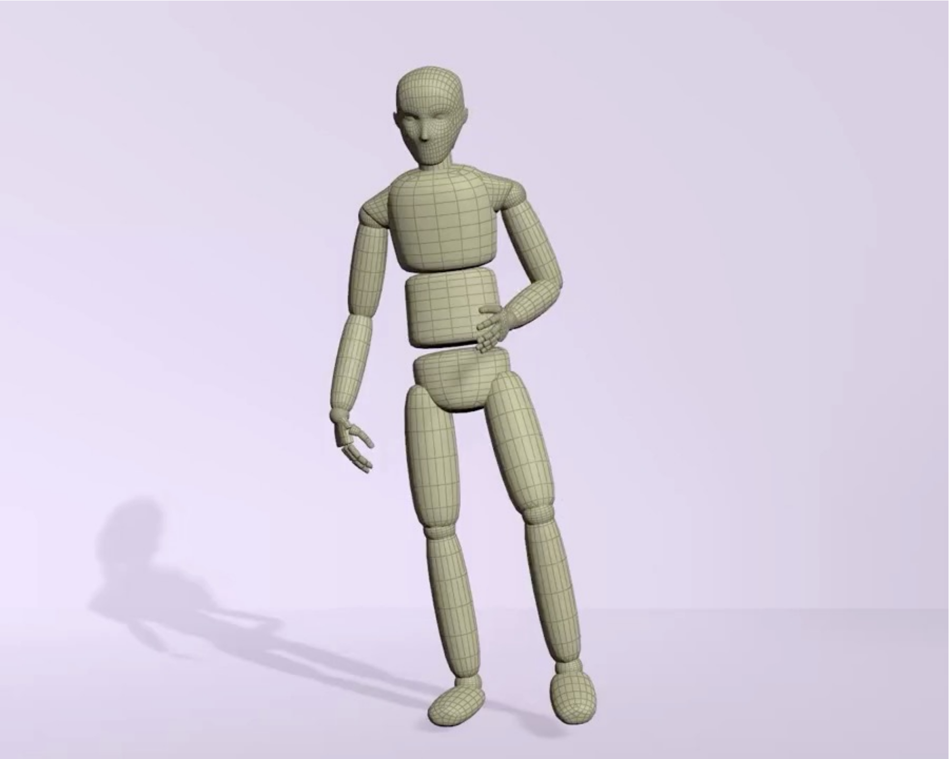
# What we learned

- Data dropout is a simple and effective trick to make autoregressive models respect the control
- Adding a recurrent hidden state (the LSTMs) stabilised synthesis
- There was no need to “reduce the temperature” to improve visual quality when drawing samples
  - Unlike Glow, BigGAN, GPT-3, etc.
- Data augmentation helps
  - Reversing the data in time taught the models to walk backwards
- Standing still was the most challenging control input for leading motion-synthesis methods

# Graphical overview



# Co-speech gesture example





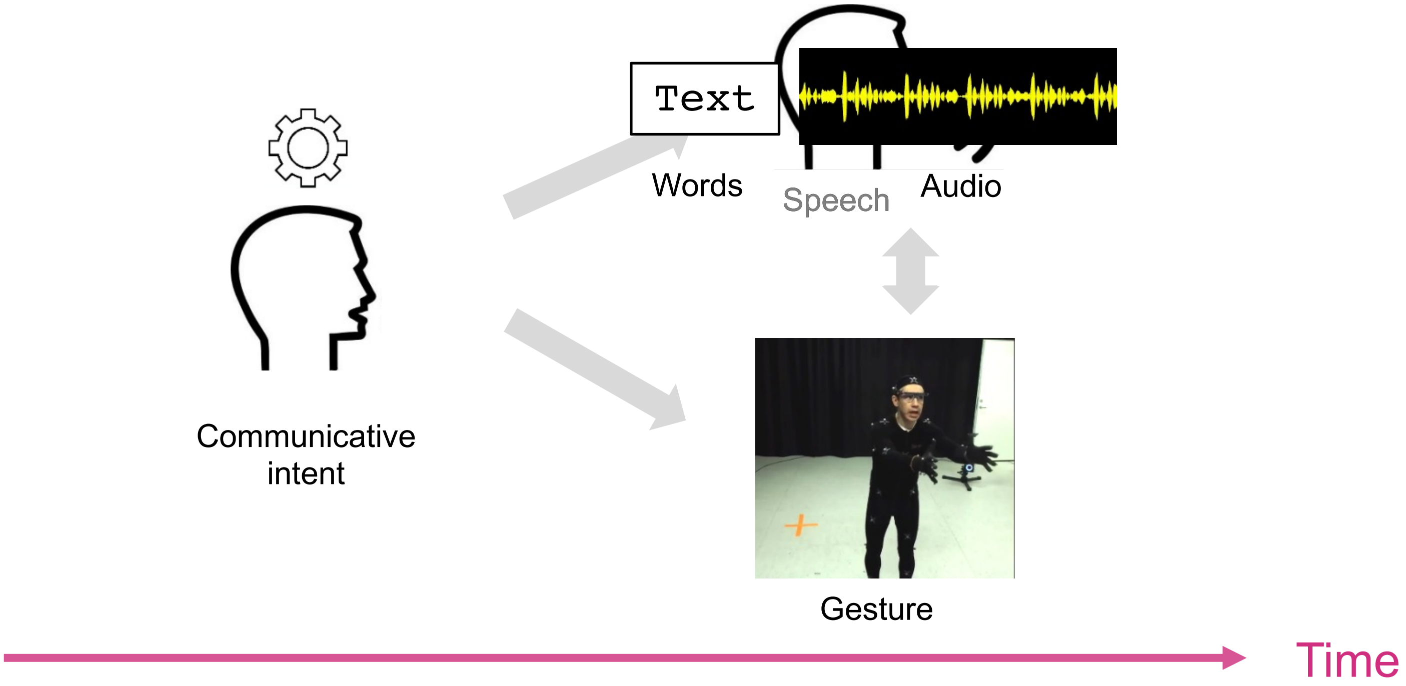
Data from Y. Ferstl & R. McDonnell, “Investigating the use of recurrent motion modelling for speech gesture generation”, Proc. IVA, 2018

# Synthetic gesture applications




Image sources: SoftBank Robotics (pepper) & Polygon.com (The Witcher 3: Wild Hunt by CD Projekt)

# Speech and gesture in communication





A. Kendon, “Gesticulation and speech: Two aspects of the process of utterance”, in M. R. Key (ed.), The Relationship of Verbal and Nonverbal Communication, Mouton, 1980

# Hand-gesture categories

- Deictic gestures
  - Pointing gestures and similar references to the space of the interaction
- Iconic gestures
  - Illustrate physical properties and actions
- Metaphorical gestures
  - Illustrate abstract meaning
- Beat gestures
  - Follow speech prosody
    - *Rhythm, emphasis, etc.*
- Beats primarily correlate with speech acoustics; the other categories primarily correlate with speech semantics



D. McNeill, Hand and Mind: What Gestures Reveal about Thought, University of Chicago Press, 1992

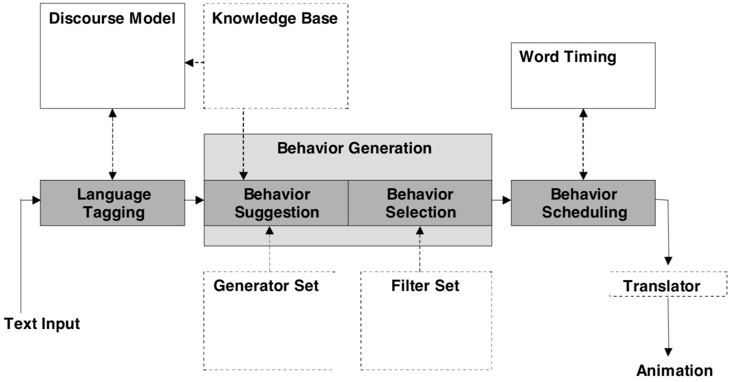
# Gesture-generation paradigms

Classic human-designed approach

- Hand-animated behaviour
- Triggered by rule

Emerging data-driven approach

- More adaptable and generalisable
- More diverse output
- Less interpretable
- Requires more data



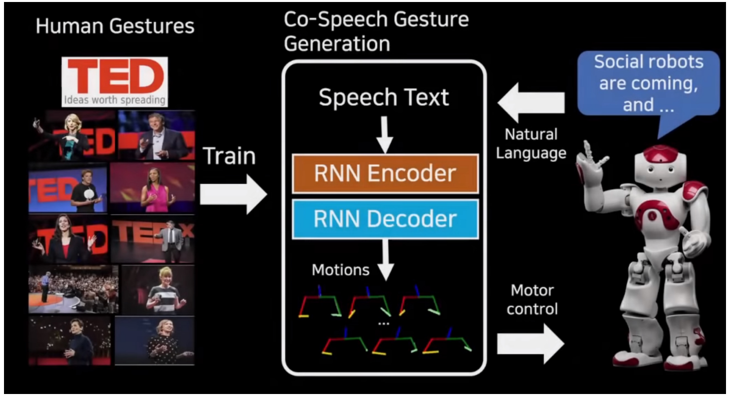
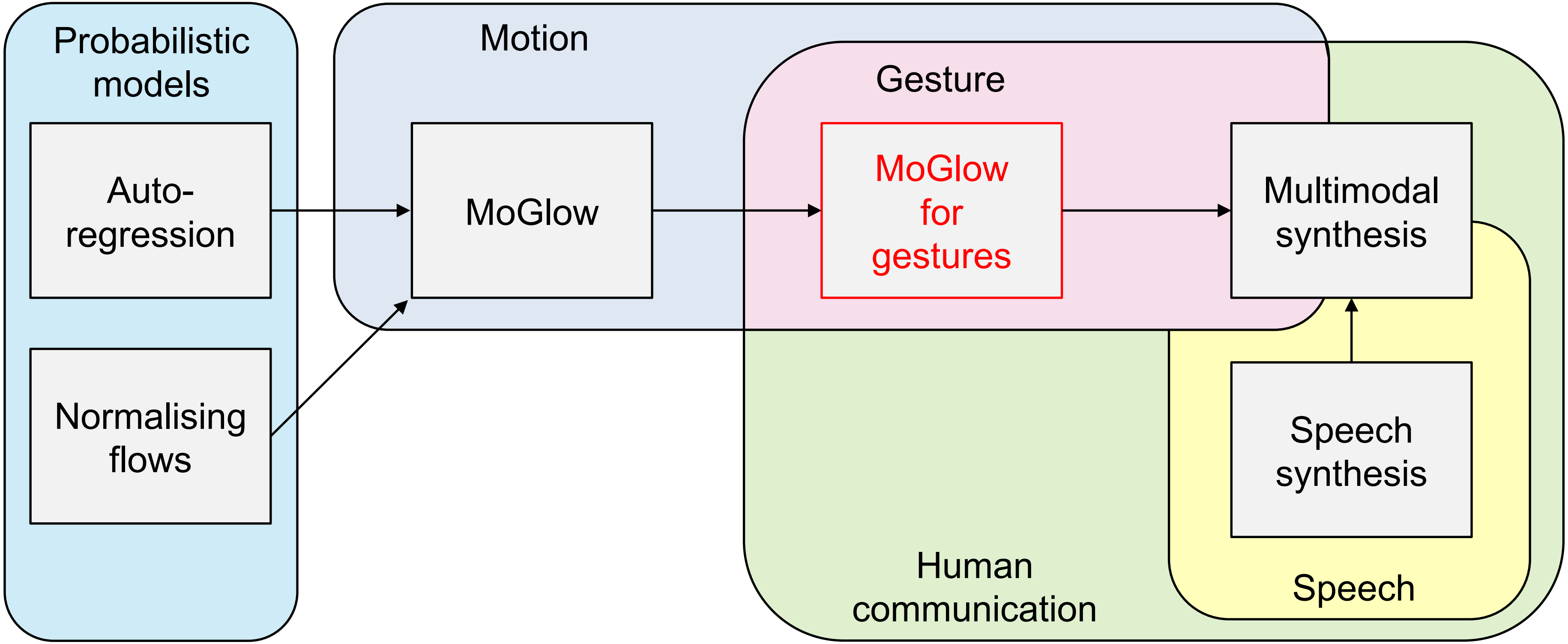






Image sources: J. Cassell et al. "BEAT: The behavior expression animation toolkit", Proc. SIGGRAPH, 2001 & Y. Yoon et al. "Robots learn social skills", Proc. ICRA, 2019

# Graphical overview



# Style-controllable speech-driven gesture synthesis using normalising flows

Simon Alexanderson

Gustav Eje Henter

Taras Kucherenko

Jonas Beskow

Honourable mention at EUROGRAPHICS 2020

# Problems with existing gesture synthesis

- Gesture synthesis is challenging due to massive variation
  - Rule-based methods cannot express this well
  - Deterministic methods also fail to capture variation and are prone to artefacts
- Synthesisers provide limited control over output
  - People gesture differently according to, e.g., personality and mood
- It is common to focus on hands and upper body only
  - But we use our entire body to express ourselves!

# Desired system

Speech


Random sampling

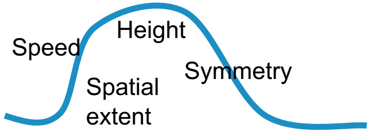
# Desired system

Speech

Try again!

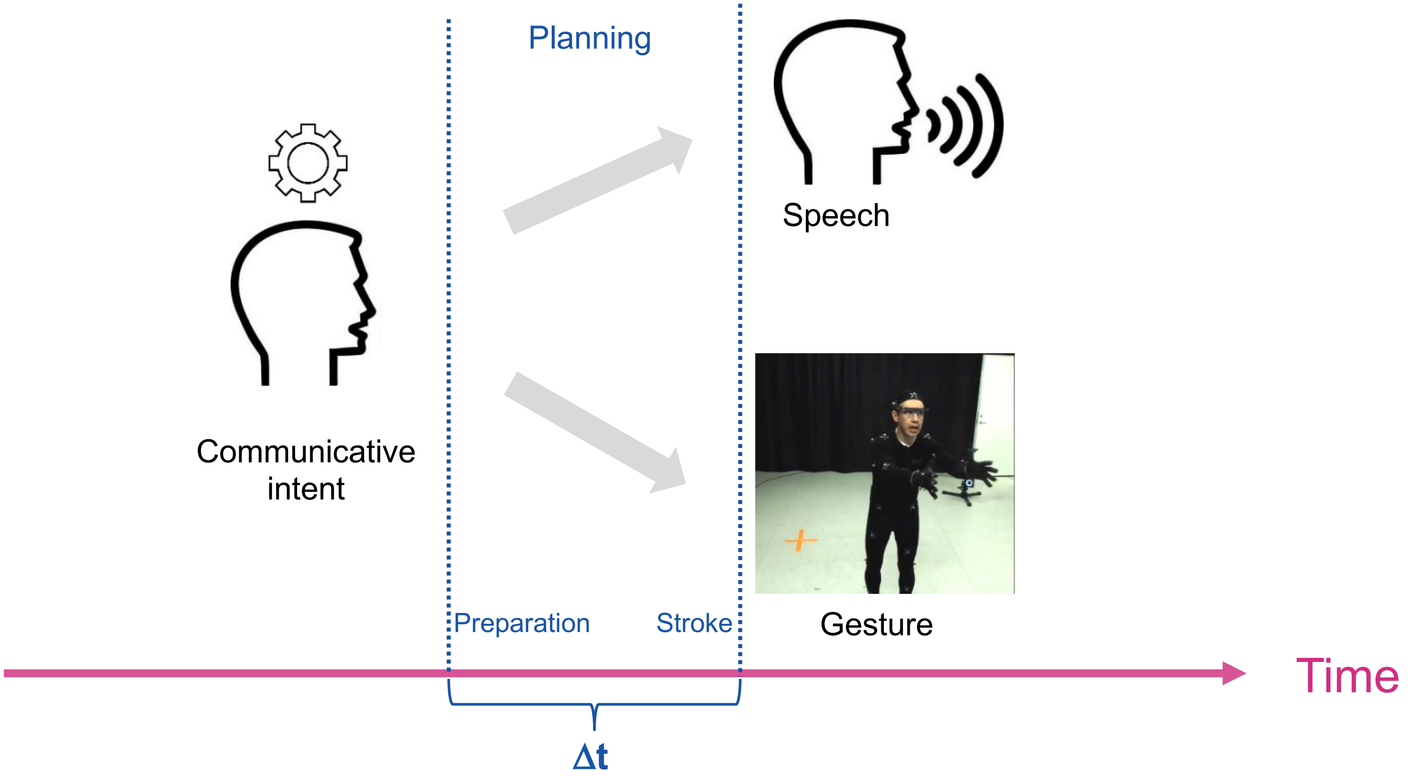
# Desired system

Speech

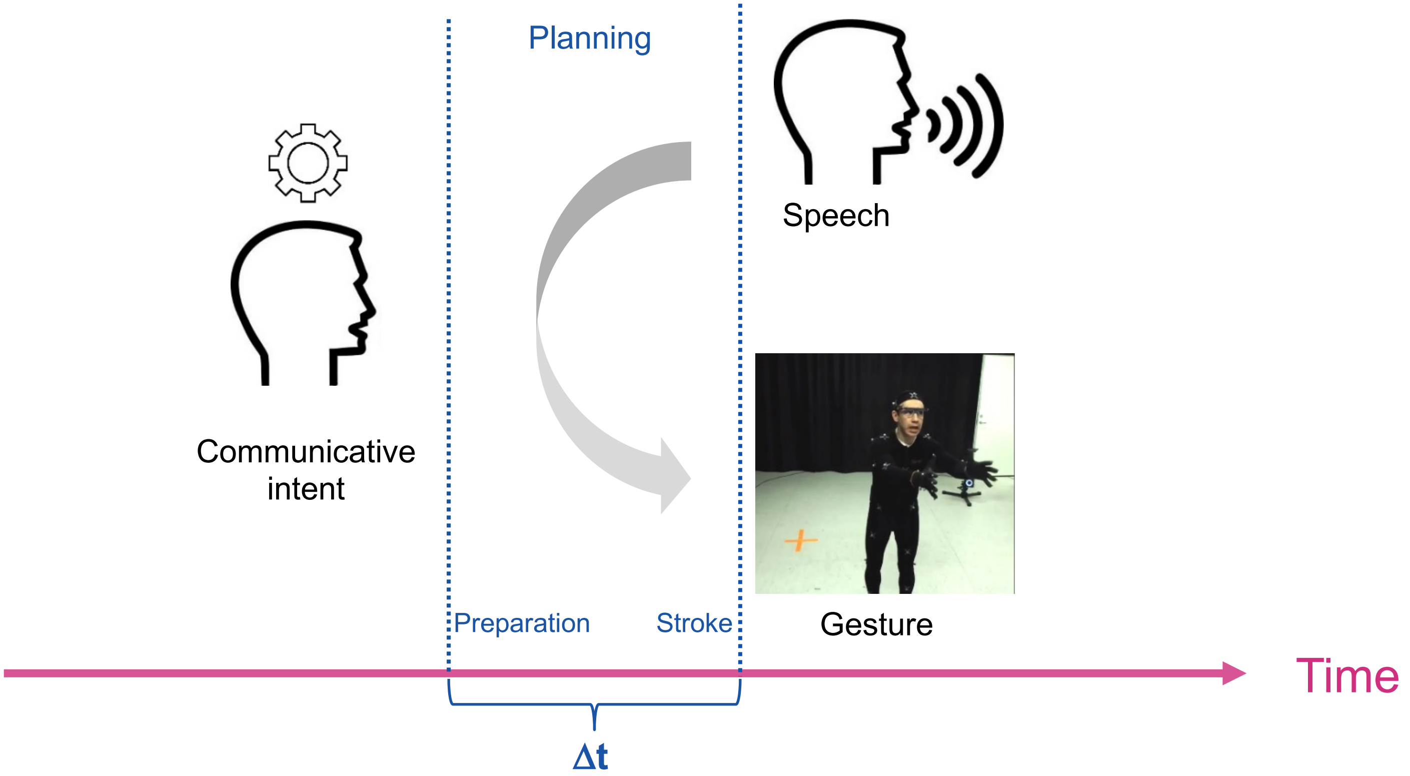



High-level style

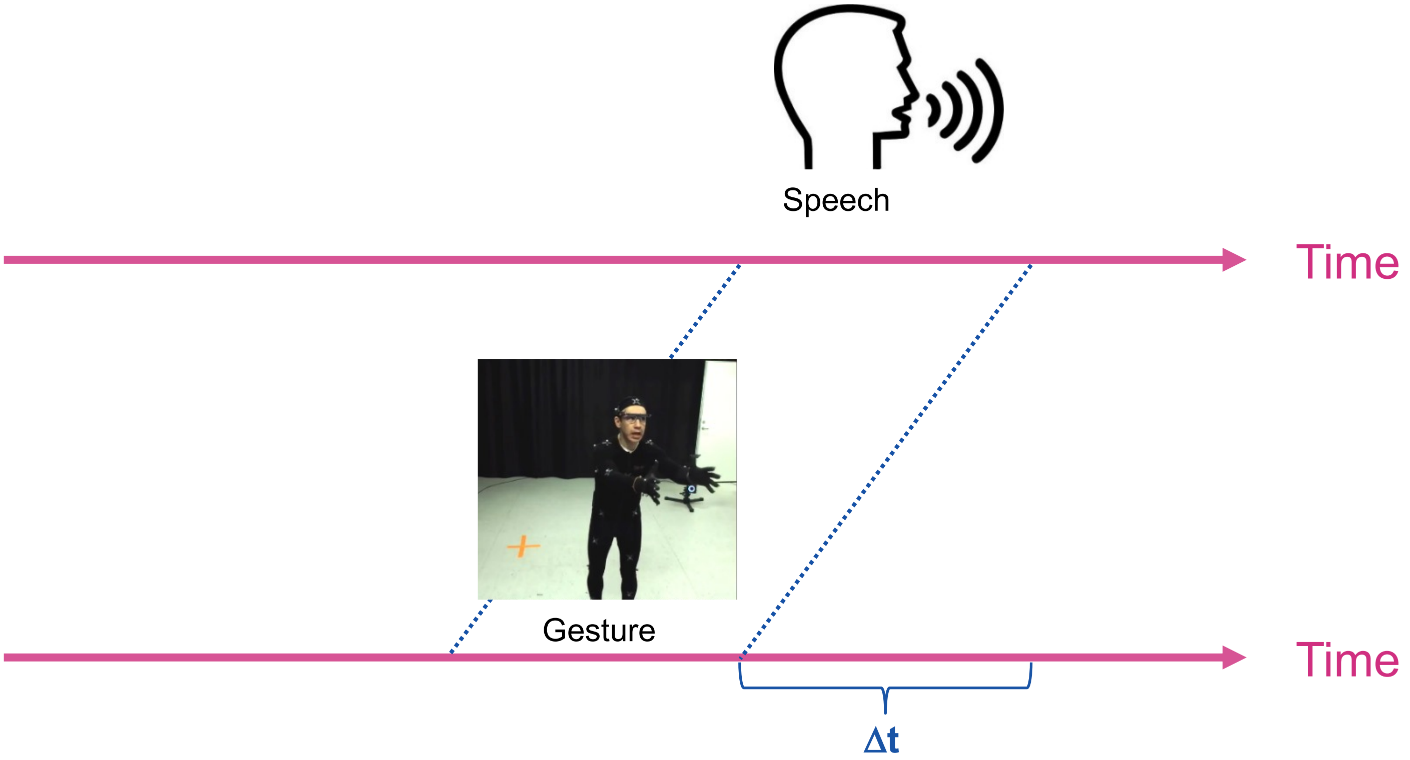
# Gesture production



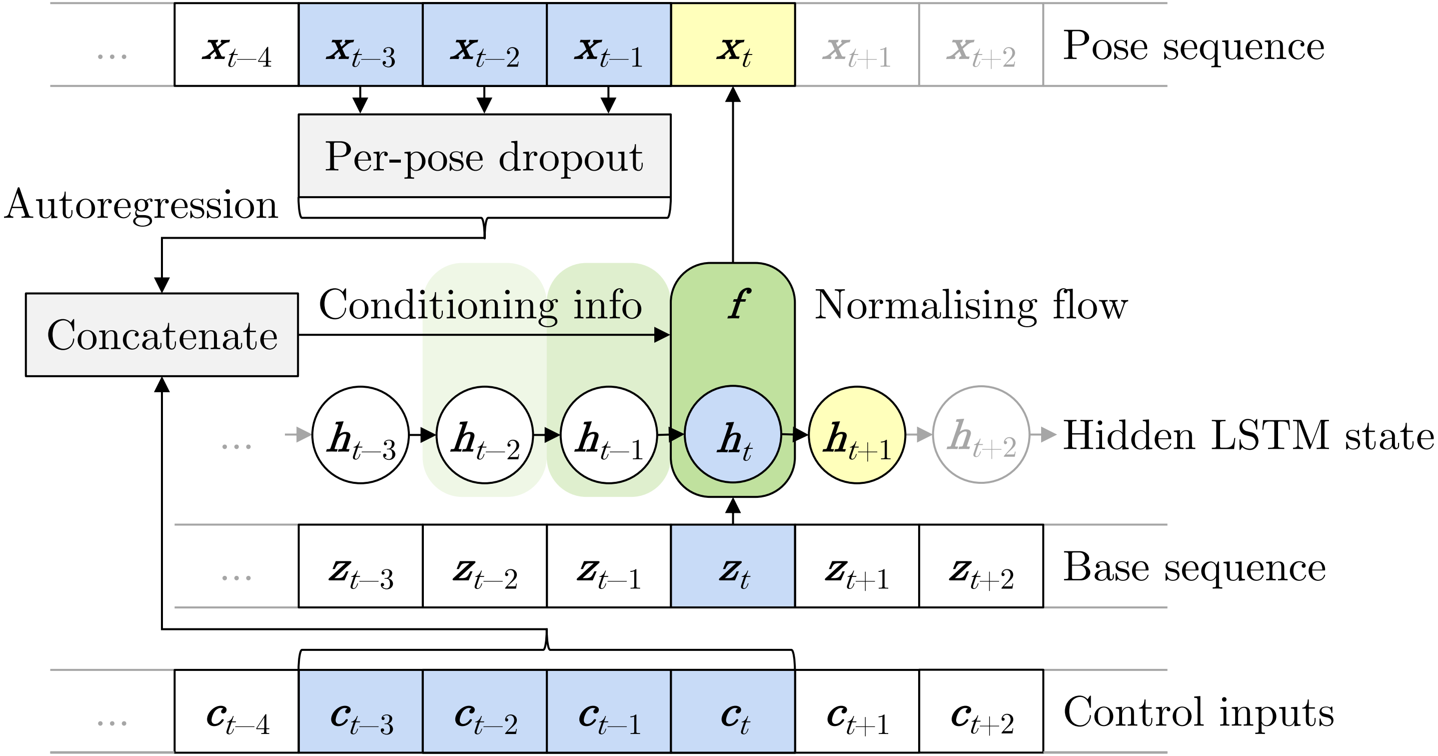
# Gesture production



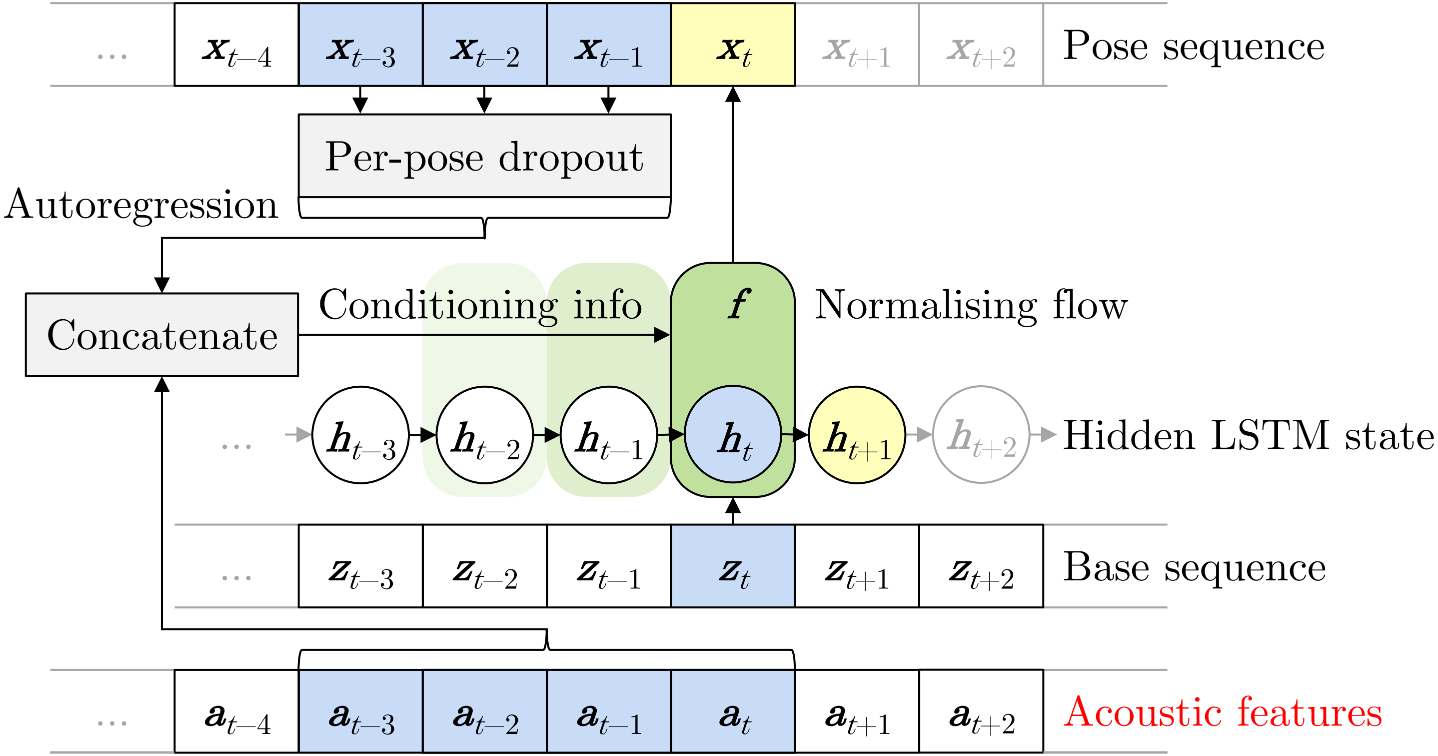
# Speech-driven gestures



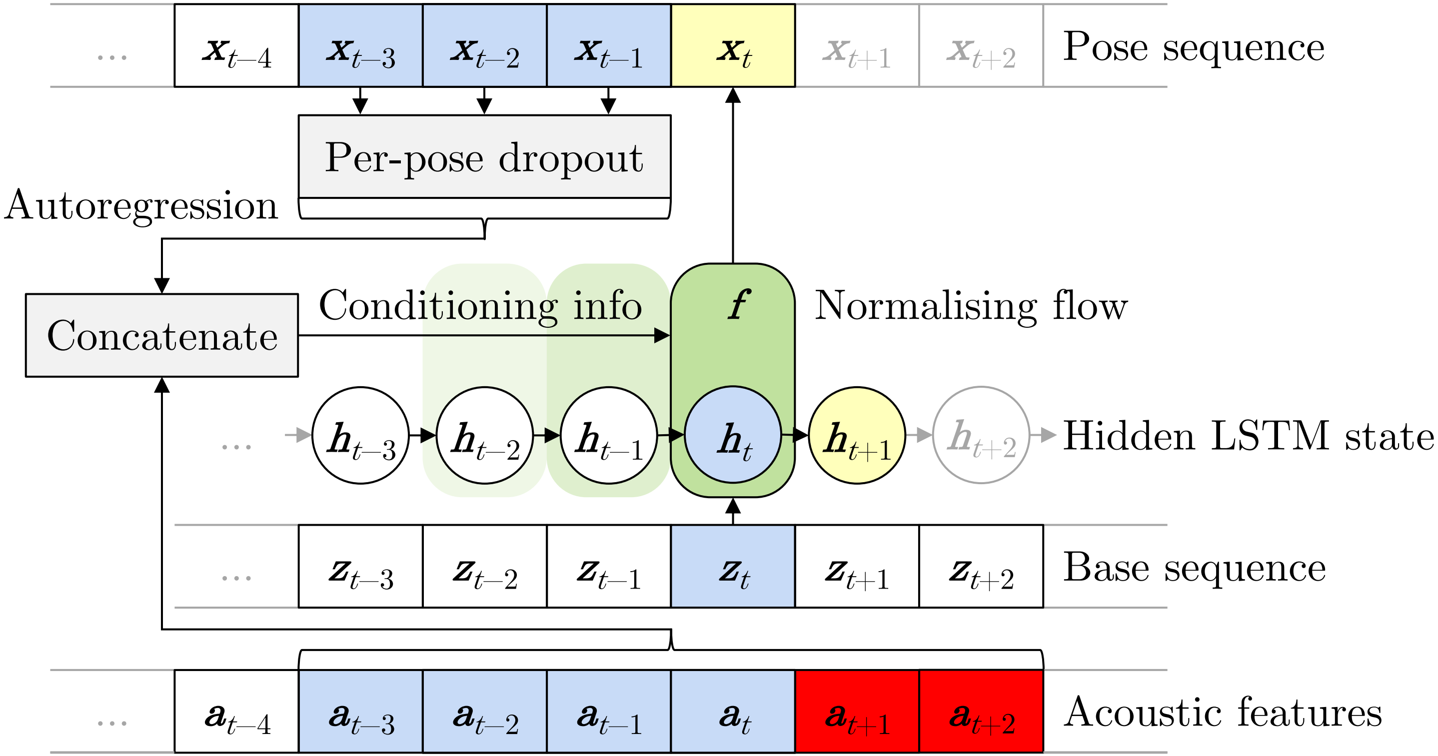
# Adapting MoGlow to gesture synthesis



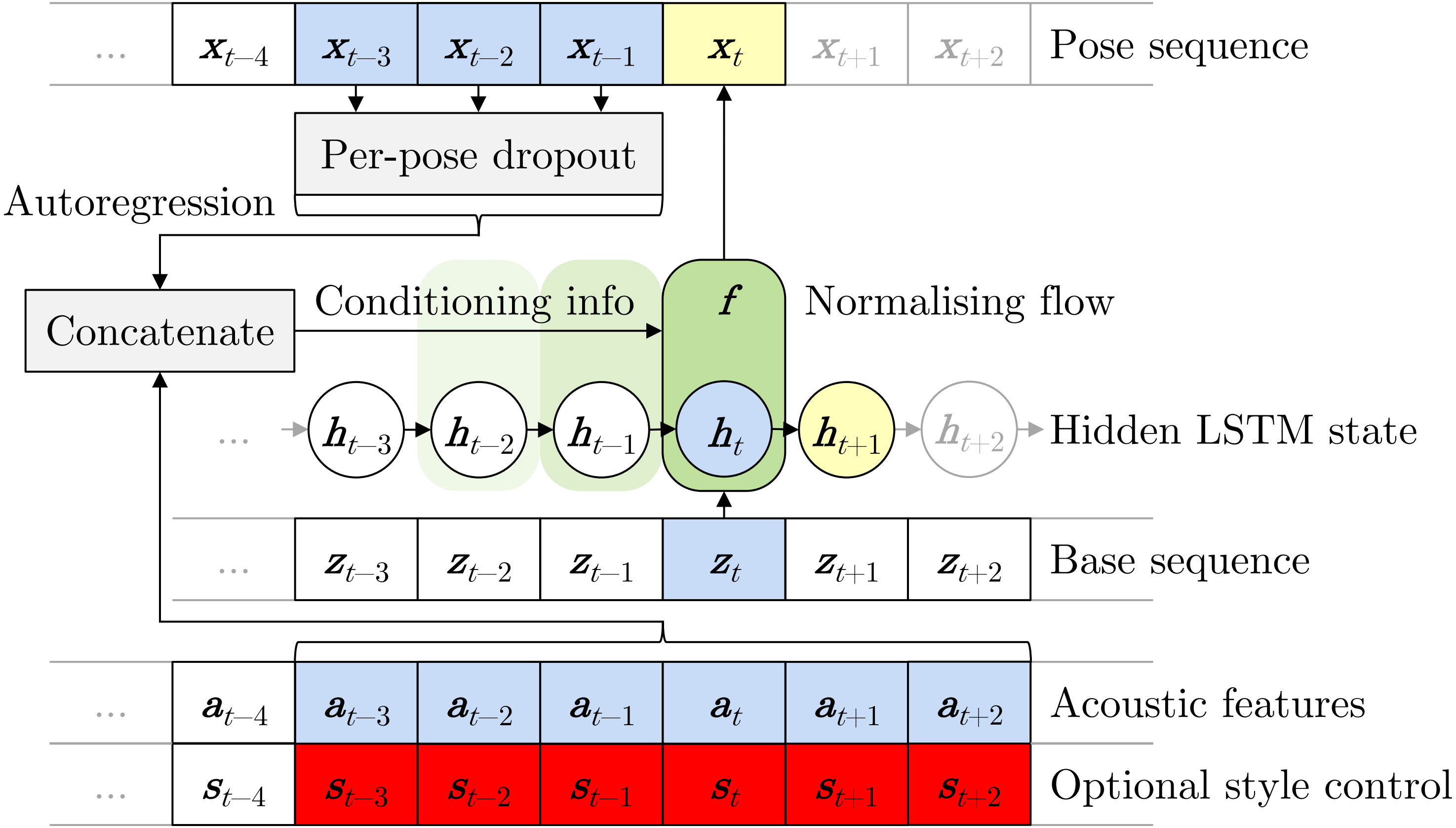
# Adapting MoGlow to gesture synthesis



# Adapting MoGlow to gesture synthesis

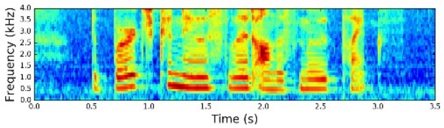


# Adapting MoGlow to gesture synthesis

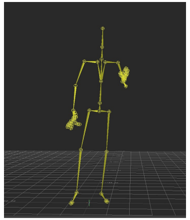


# Co-speech gesture data

- Trinity speech-gesture dataset
  - One male actor speaking spontaneously
  - 244 minutes of parallel audio and 3D motion capture
  - Post-processed to correct synchronisation issues
    - > *Corrected data available in the original repository*
    - > *Hands use a fixed pose due to low finger-capture quality*




27 log-magnitude
mel-spectrogram features



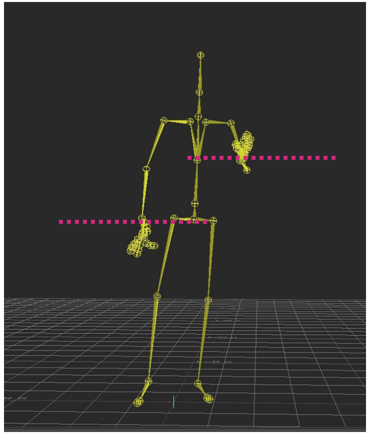
45/65 joint
rotations



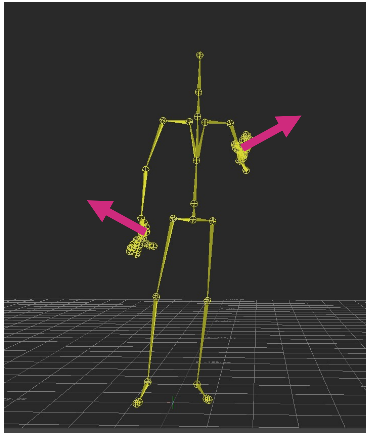
Y. Ferstl & R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation", Proc. IVA, 2018.

# Style inputs

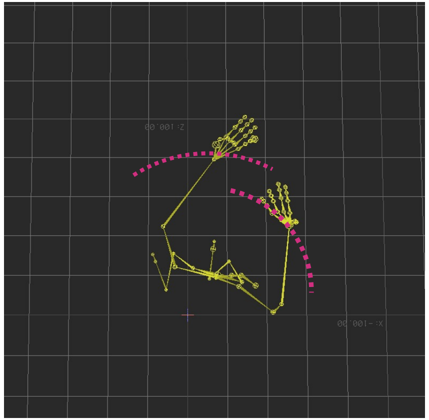
- The database does not come with stylistic annotations
  - For demonstration purposes, we used automatically-extracted hand-motion statistics
    - > *E.g.: “The hand speed should be X m/s on average over in a 4 s time window”*



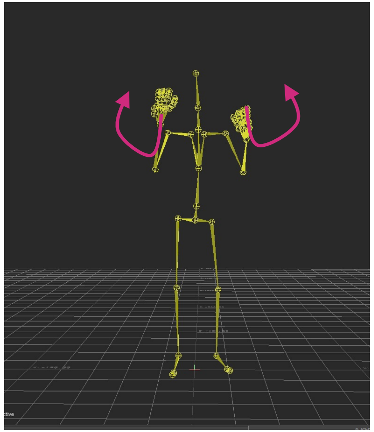
Height



Speed

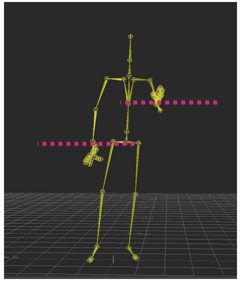


Radius

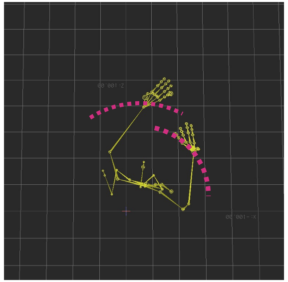


Symmetry

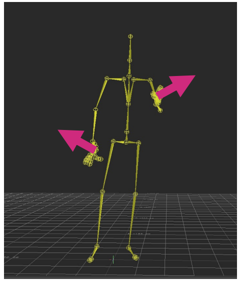
# Style inputs



Height

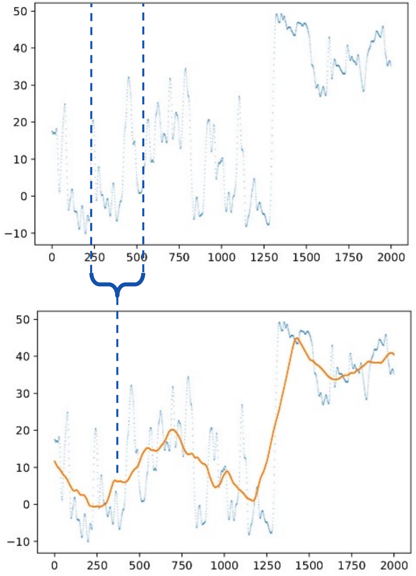


Radius

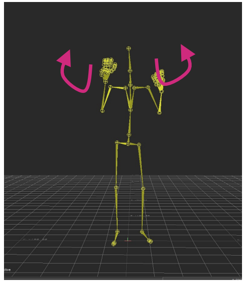


Speed

Moving average (4 s)

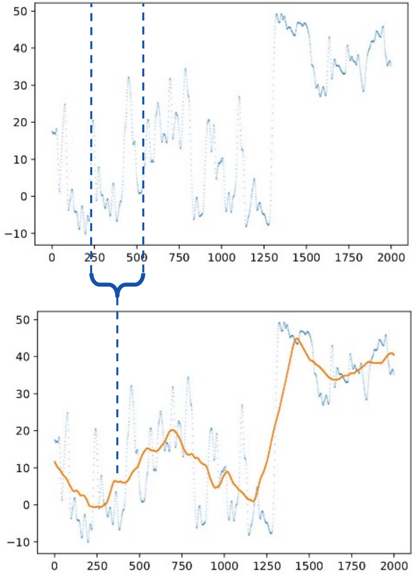


# Style inputs

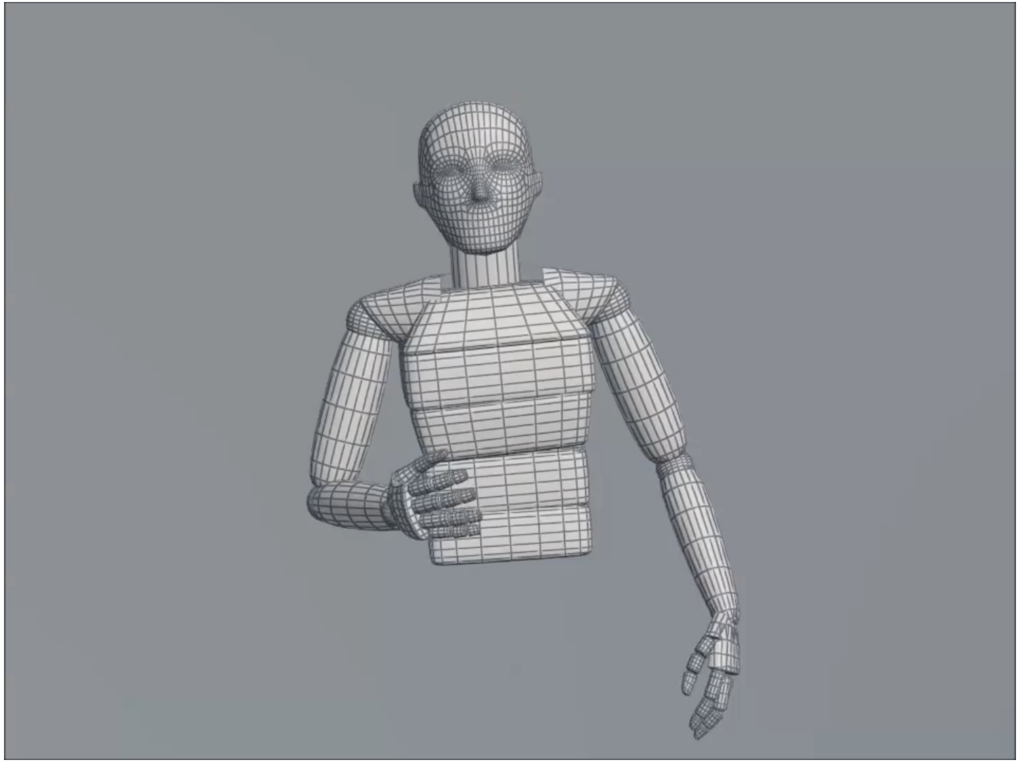


Symmetry

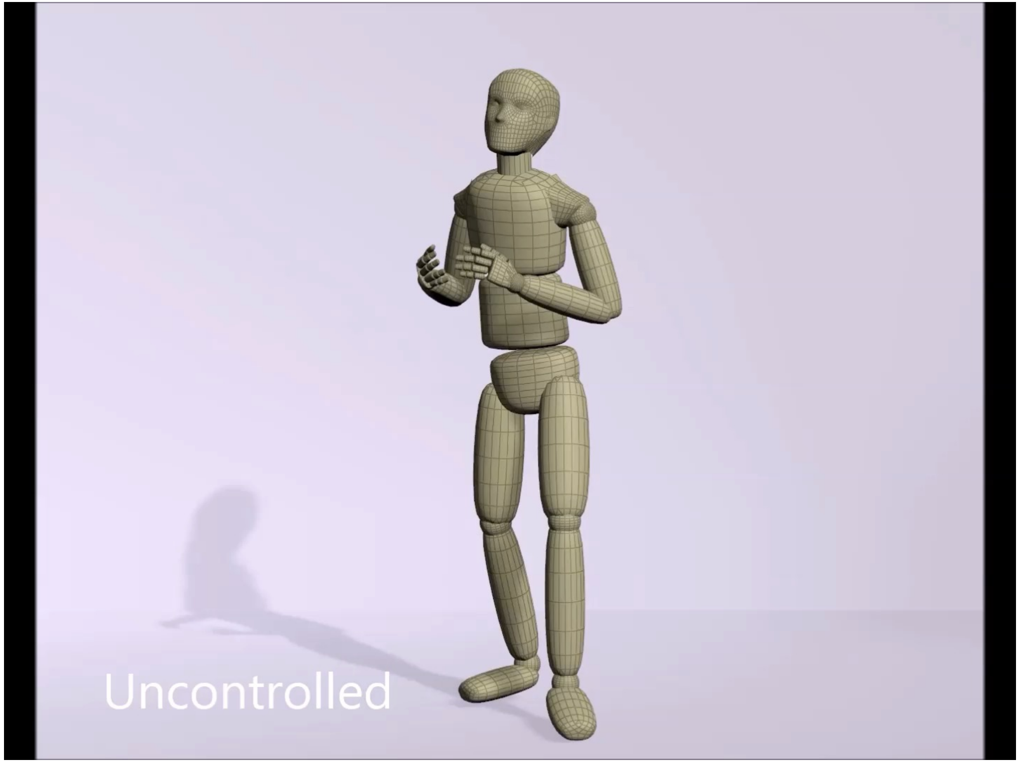
Correlation (4 s)



# Constant-input style-controlled gestures



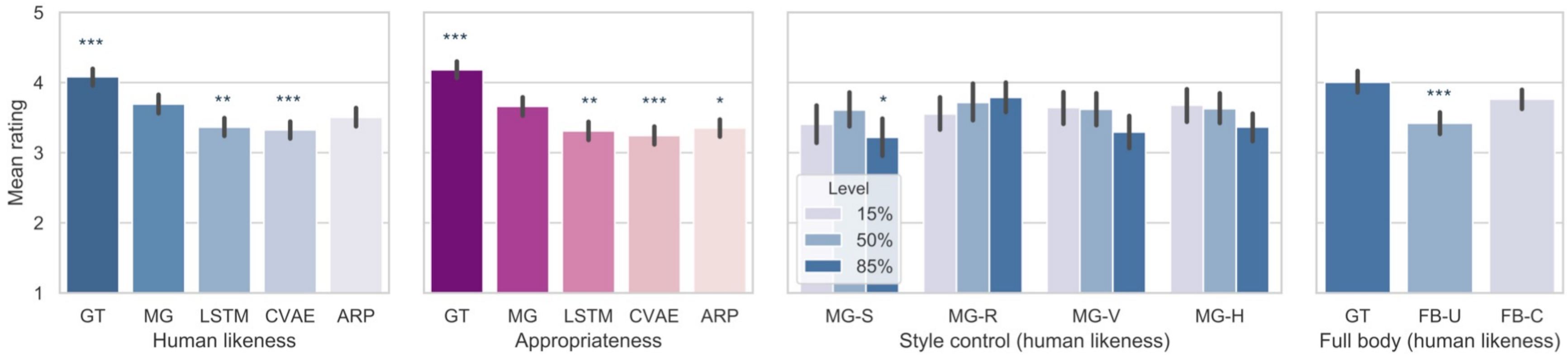
# Full-body gestures



# Subjective evaluations

- Crowdsourced subjective evaluation
  - Same Figure Eight platform and 1-to-5 MOS scale as before
  - Bad clips and too rapid/slow responses were used to filter out unreliable raters
  - 40 independent crowdworkers took part
- Two different aspects were rated
  - Human-likeness
    - > *"To what extent does the motion of the character look like the motion of a real human being?"*
  - Appropriateness
    - > *"To what extent does the motion match the audio?"*
    - > *Evaluating appropriateness is not a solved problem*

# Subjective evaluation



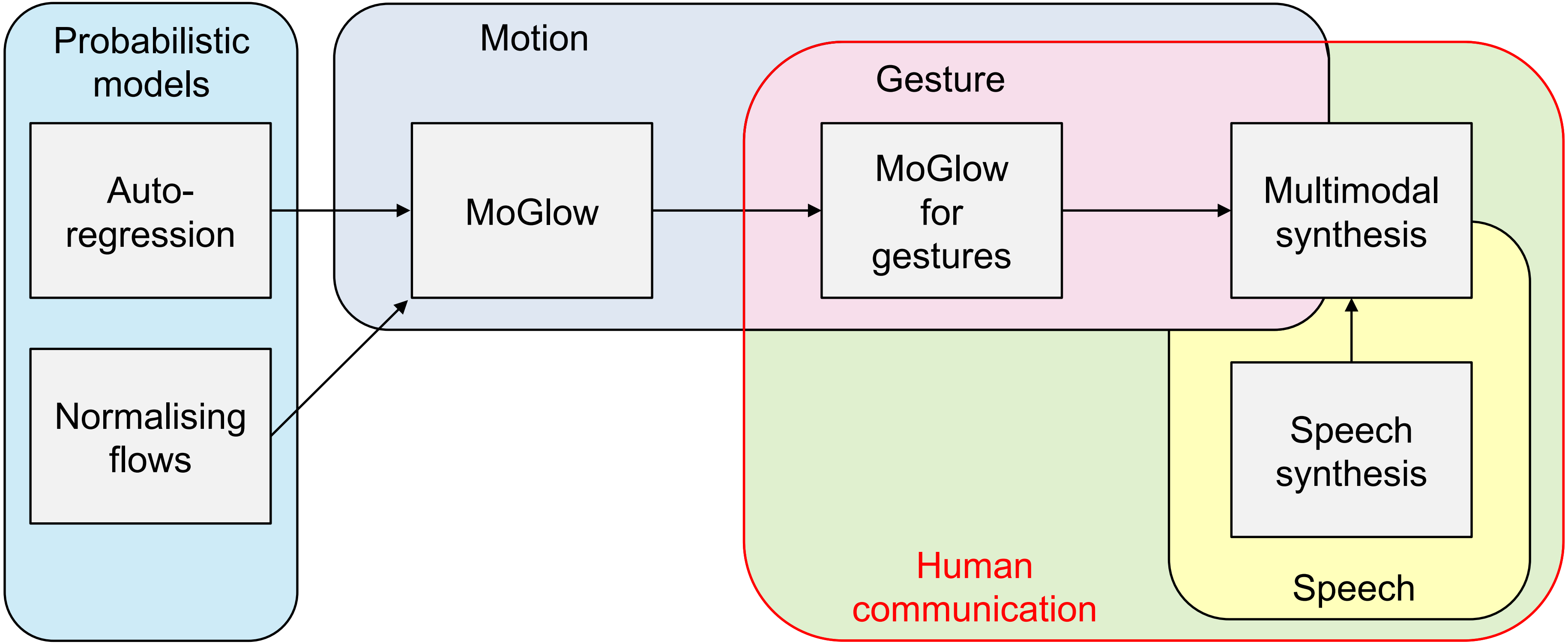


The ARP system is T. Kucherenko et al., “Analyzing input and output representations for speech-driven gesture generation”, Proc. IVA, 2019.

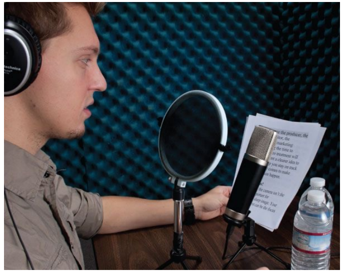
# What we learned

- MoGlow works for gesture generation
  - Appears to be a new state of the art in continuous gesture generation human-likeness
  - Strong showing in the 2020 GENEA Challenge
    - > *Like the Blizzard Challenge, but for gesture generation*
- Gesture style control is possible without degrading motion quality
- Tuning gesture-generation models is tricky
  - The output exhibits a large amount of random variation
  - The only useful objective measure we found was the training set log-likelihood
- Validation-set likelihood does not reflect the visual quality of the motion
  - Overfitted models look best
  - Possibly due to data dropout
  - Mismatch can be reduced by methods from robust statistics (see our INNF+ publication)

# Graphical overview



# Speech synthesis vs. gesture synthesis



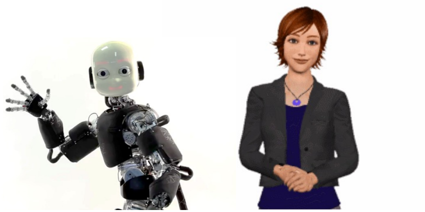
Read

Spontaneous

TTS

Gesture

Embodied agent



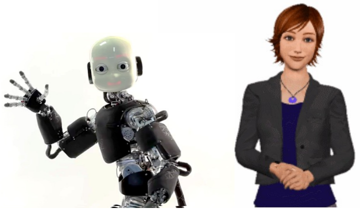
Incoherent!

# Objective

- Merge speech and gesture synthesis
  - Enable multimodal communication from text input

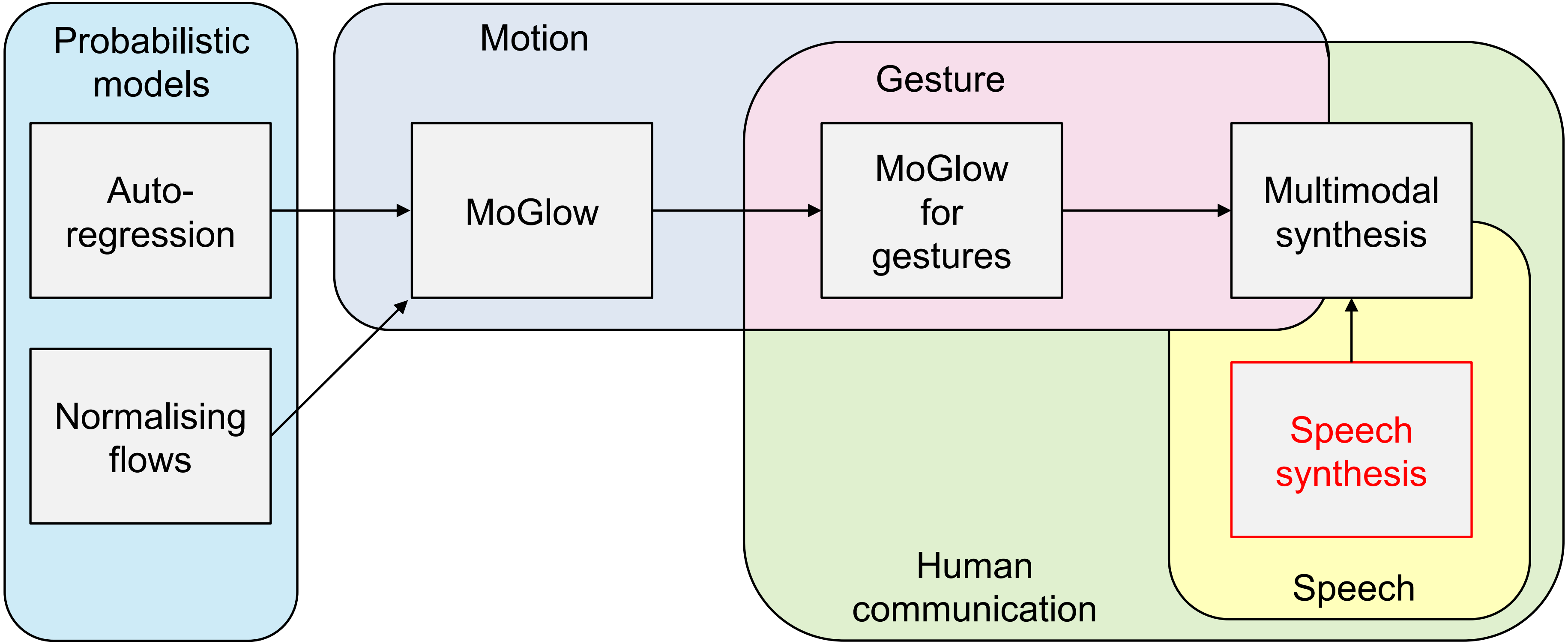
Speech + motion

TTS
+
gesture



Coherent!

# Graphical overview



# Spontaneous speech synthesis team

Éva
Székely

Gustav Eje
Henter

Jonas
Beskow

Joakim
Gustafson

# Recent publications

Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector

Spontaneous conversational speech synthesis from found data

Off the cuff: Exploring extemporaneous speech delivery with TTS

How to train your fillers: uh and um in spontaneous speech synthesis

Breathing and speech planning in spontaneous speech synthesis

Published at ICASSP 2019, 2020, Interspeech 2019, and SSW 2019

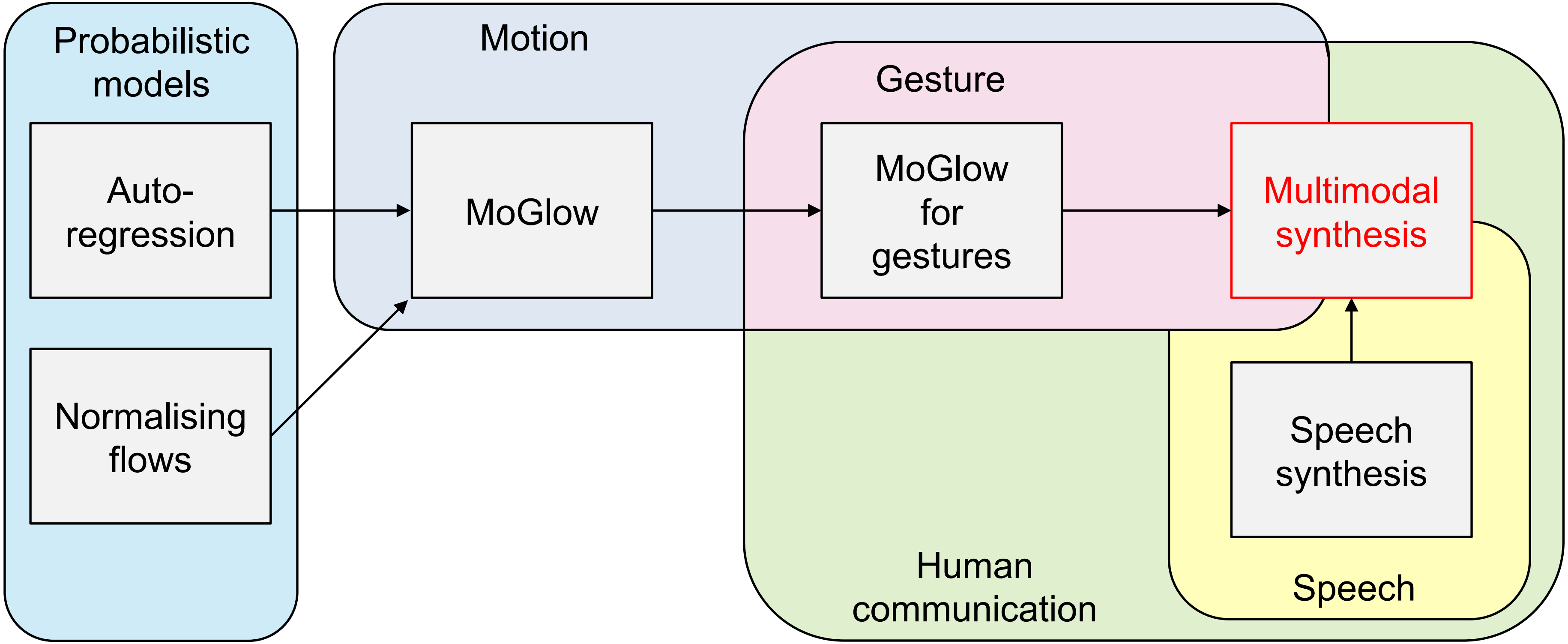
# Spontaneous TTS from found data

| Training data<br>Text prompt source | Read speech (24 h found audiobooks) | Spontaneous (9 h found podcast) | Spontaneous (1.5 h studio-recorded) |
|---|---|---|---|
| Books | | | |
| Public speaking | | | |
| Casual conversation | | | |



É. Székely et al., "Spontaneous conversational speech synthesis from found data", Proc. Interspeech, 2019 https://www.speech.kth.se/tts-demos/is19/

# Graphical overview



# Generating coherent spontaneous speech and gesture from text



Simon Alexanderson


Éva Székely


Gustav Eje Henter


Taras Kucherenko

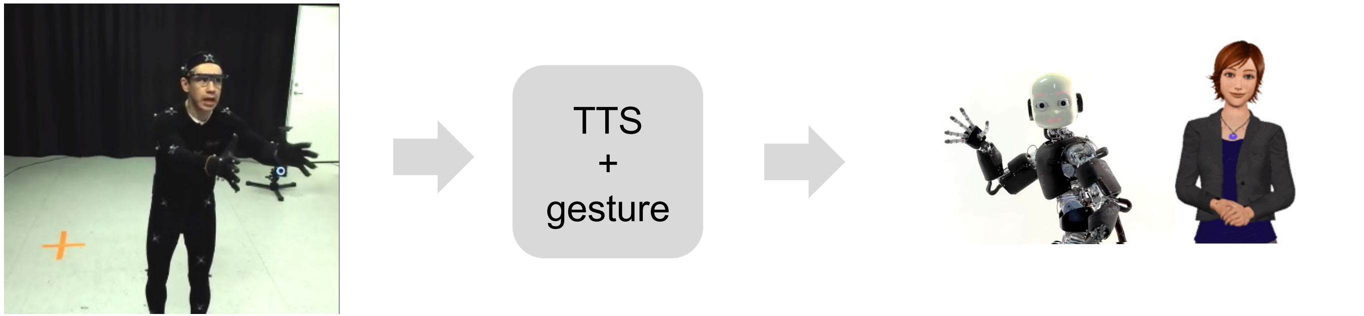

Jonas Beskow


IVA 2020

# Objective

- Merge speech and gesture synthesis
  - Enable multimodal communication from text input
- First step:
  1. Text to spontaneous speech
  2. Spontaneous speech to gesture

  Using the same multimodal recordings




Speech + motion

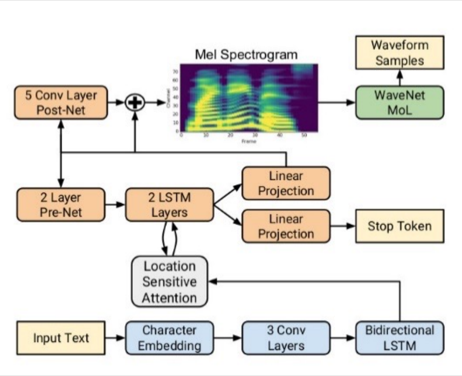
# Approach

Spontaneous synthetic speech

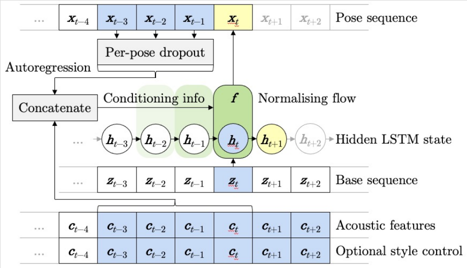
TEXT

Gesture

## Tacotron 2 + Griffin-Lim



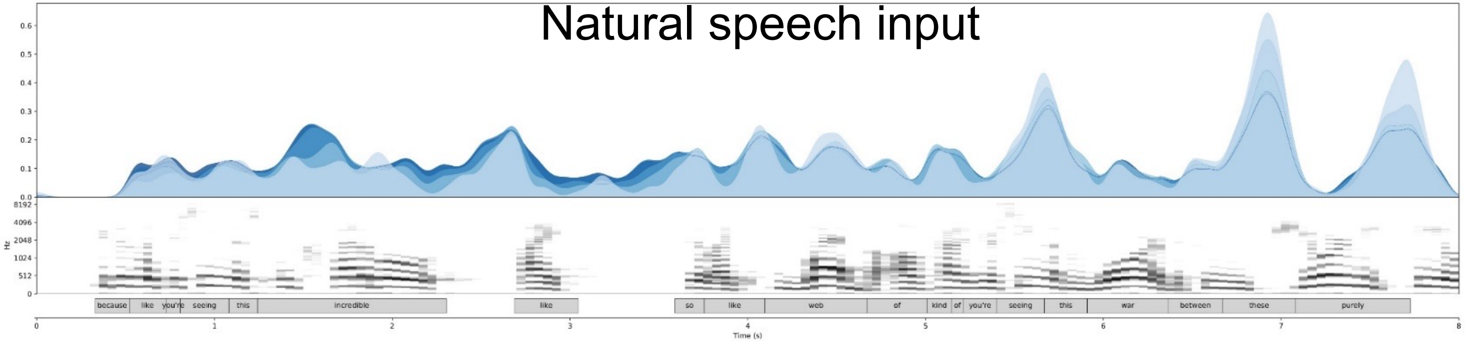

## Gesture MoGlow

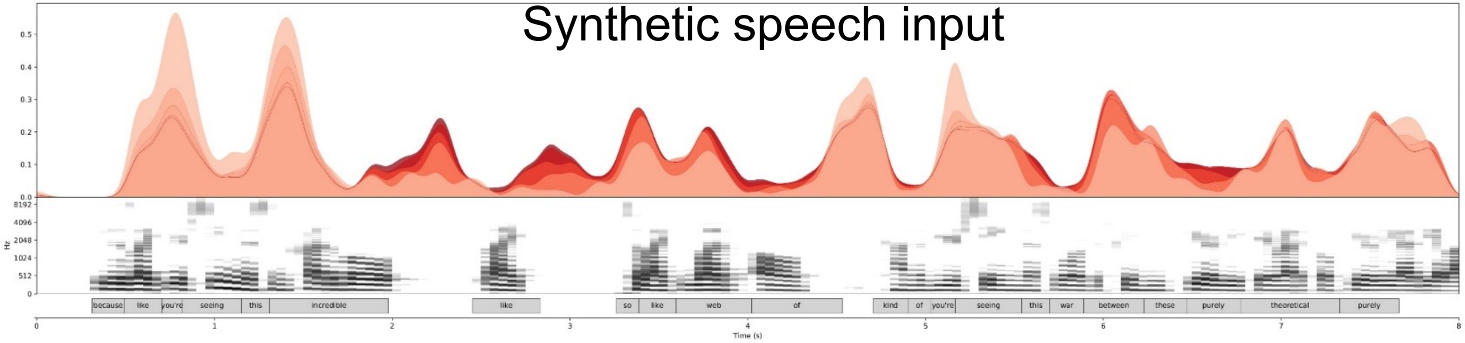


# Result

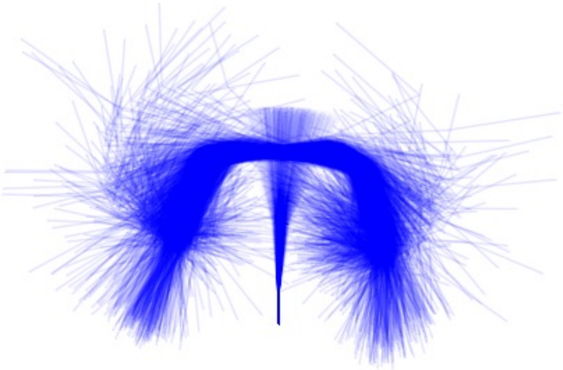
The following clip is generated from TEXT only

# Hand peak velocities across 300 samples

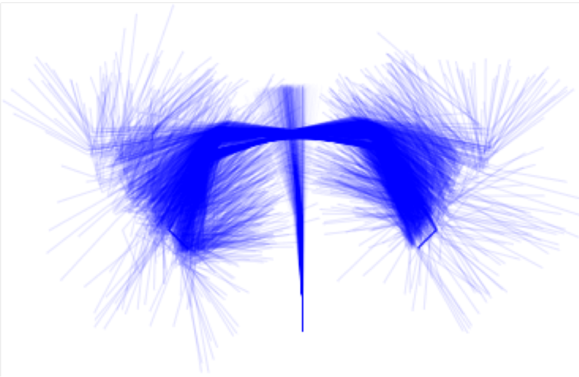


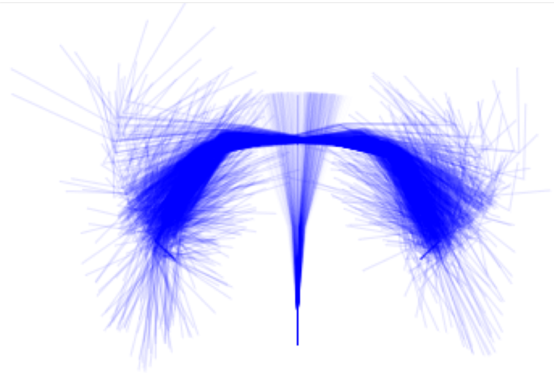
# Gesture-space visualisation



In the
training data



Generated from
natural speech input



Generated from
synthetic speech input

# Conclusion

- Automated character animation is a challenging and interesting problem
- The world is probabilistic; our motion models should be, too
- MoGlow is a new probabilistic model for motion
  - Task-agnostic
  - Meaningfully probabilistic
  - No (or adjustable) algorithmic latency
- MoGlow reaches or surpasses the state of the art in a wide variety of applications
- Text-to-speech → text-to-behaviour

# Project homepages

https://simonalexanderson.github.io/MoGlow

https://github.com/simonalexanderson/StyleGestures

https://simonalexanderson.github.io/IVA2020/

# Additional gesture publications


- T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, & G. E. Henter, “A Large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020”, *Proc. IUI*, 2021.
- T. Kucherenko, D. Hasegawa, N. Kaneko, G. E. Henter, & H. Kjellström, “Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation”, *Int. J. Human Comput. Interact.*, 2021.
- P. Jonell, T. Kucherenko, G. E. Henter, & J. Beskow, “Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings”, *Proc. IVA*, 2020.
- T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexanderson, I. Leite, & H. Kjellström, “Gesticulator: A framework for semantically-aware speech-driven gesture generation”, *Proc. ICMI*, 2020.
- S. Alexanderson & G. E. Henter, “Robust model training and generalisation with Studentising flows”, *Proc. INNF+*, 2020.
- T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, & H. Kjellström, “Analyzing input and output representations for speech-driven gesture generation”, *Proc. IVA*, 2020.
- T. Kucherenko, D. Hasegawa, N. Kaneko, G. E. Henter, & H. Kjellström, “On the importance of representations for speech-driven gesture generation”, *Proc. AAMAS*, 2019.

# Thank you for listening!

# Thank you for listening

Any questions?

# Backup slides

# Probabilistic approaches compared

| | Gauss. (MSE) | MDN | HMM / SLDS | Kalman filter | GP-LVM / GPDM | VAE | GAN | Norm. flow |
|---|---|---|---|---|---|---|---|---|
| Rand. $X$ | Gauss. | $\mathbb{R}$ | $\mathbb{R}$ | Gauss. | Gauss. | $\mathbb{R}$ | - | - |
| Map $f$ | Deep | Deep | Deep | Linear | Non-linear | Deep | Deep | Invertible deep |
| Rand. $Z$ | - | Discrete | Discrete | Gauss. | $\mathbb{R}$ | $\mathbb{R}$ | $\mathbb{R}$ | $\mathbb{R}$ |
| Map $g$ | - | Deep | Deep | Linear | Non-linear | Deep | - | - |
| Inference | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |
| Sampling | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ |
| Flexibility | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |

# Mathematical model

- Probability of a sequence of vector-valued, continuous observations
  - Assume limited memory – a *Markov model*

$$p\left(\boldsymbol{x}_{1:T}\right) = p\left(\boldsymbol{x}_{1:p}\right) \prod_{t=p+1}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}\right)$$

$$\approx p\left(\boldsymbol{x}_{1:p}\right) \prod_{t=p+1}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{1-p:t-1}\right)$$

  - The initial pose distribution is not modelled
- The *next-step distribution* $p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{1-p:t-1}\right)$ also depends on a parameter $\boldsymbol{\theta}$
  - Here, the parameters are the matrices and network weights inside Glow

$$p\left(\boldsymbol{x}_{1:T};\, \boldsymbol{\theta}\right) = p\left(\boldsymbol{x}_{1:p}\right) \prod_{t=p+1}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{1-p:t-1};\, \boldsymbol{\theta}\right)$$

# Long-term memory

- Introduce a hidden state $\boldsymbol{h}_t$ to the model that also influences the next-step distribution
  - HMMs, Kalman filters, and LSTMs all do this

$$p\left(\boldsymbol{x}_{1:T};\, \boldsymbol{\theta}\right) = p\left(\boldsymbol{x}_{1:p}\right) \prod_{t=p+1}^{T} p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{1-p:t-1},\, \boldsymbol{h}_{t-1};\, \boldsymbol{\theta}\right)$$

$$\boldsymbol{h}_{p-1} = \boldsymbol{0}$$

$$\boldsymbol{h}_t = \boldsymbol{g}\left(\boldsymbol{x}_{1-p:t-1},\, \boldsymbol{h}_{t-1};\, \boldsymbol{\theta}\right)$$

  - Concretely, this is done by using LSTMs in the coupling layer neural network
  - "Long memory" since $p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{1:t-1}\right) = p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{1-p:t-1};\, \boldsymbol{\theta}\right) \neq p\left(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{1-p:t-1}\right)$
- The main advantage appears to be avoid unstable autoregressive models
  - Crucial to get the approach to work in practice

# Achieving control

- The next-step distribution now also depends on a per-frame control input
  - No future control information is used

$$p\left(\boldsymbol{x}_{1:T};\,\boldsymbol{\theta}\right)=p\left(\boldsymbol{x}_{1:p}\right)\prod_{t=p+1}^{T}p\left(\boldsymbol{x}_{t}\mid\boldsymbol{x}_{1-p:t-1},\,\boldsymbol{c}_{1-p:t},\,\boldsymbol{h}_{t-1};\,\boldsymbol{\theta}\right)$$

$$\boldsymbol{h}_{p-1}=\boldsymbol{0}$$

$$\boldsymbol{h}_{t}=\boldsymbol{g}\left(\boldsymbol{x}_{1-p:t-1},\,\boldsymbol{c}_{1-p:t},\,\boldsymbol{h}_{t-1};\,\boldsymbol{\theta}\right)$$

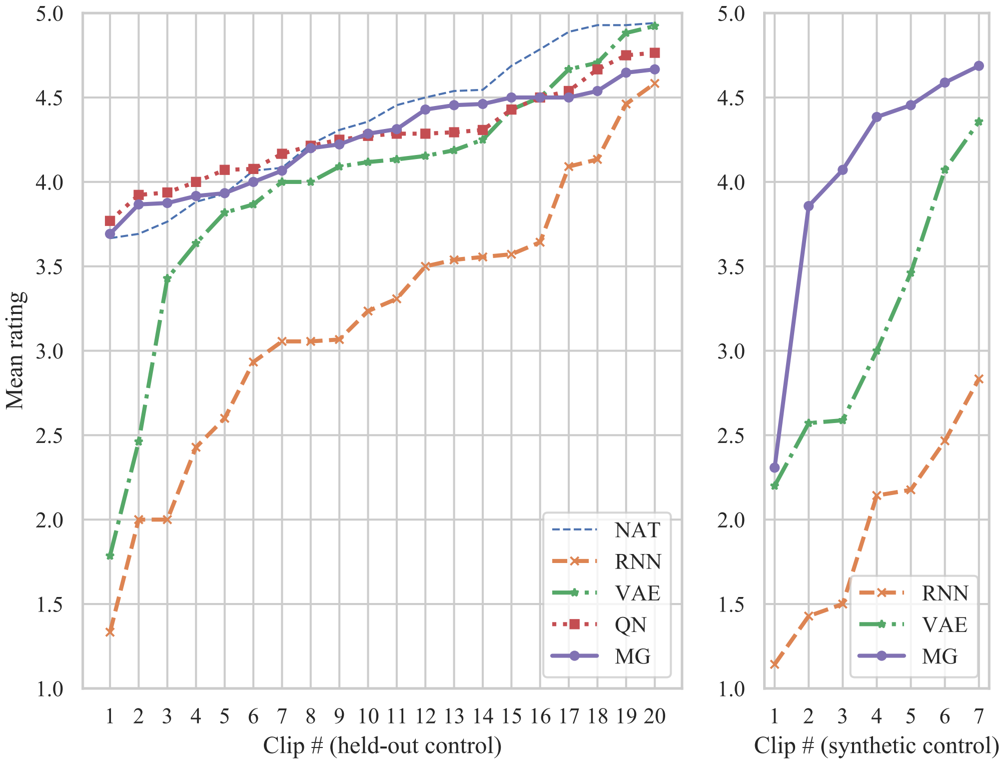
# Likelihood function

- For completeness, the likelihood of a single sequence in MoGlow is

$$\ln p_X(x;\theta) = \text{const.} + \frac{1}{2} z_N^{\mathrm{T}}(x) z_N(x) + \sum_{n=1}^{N} \sum_{d=1}^{D} (\ln s'_{n\,d} + \ln u_{n\,dd} + \ln s_{n\,d}(x))$$

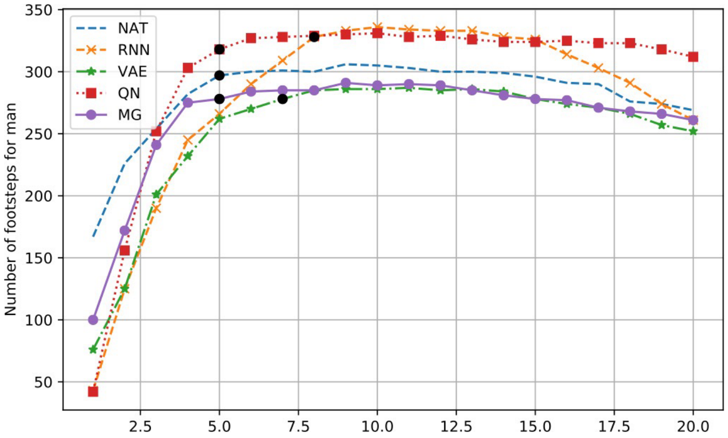
  - The constant is just the normalisation constant of a $D$-dimensional standard normal
  - There's one term each for the actnorm layer, the linear layer, and the coupling layer
  - The contribution from the linear layer is fast to compute by parametrising the transformation (matrix multiplication) using an LU-decomposition
  - Only the coupling term depends on $x$; the other terms are global and fixed

# Results per motion clip (human locomotion)

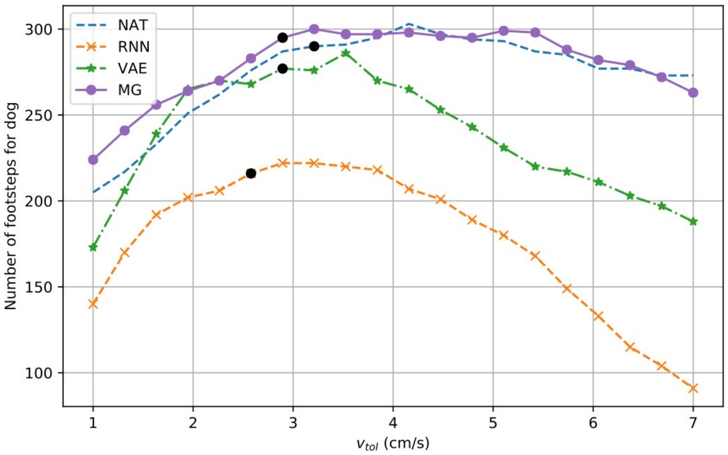


# Footstep analysis

Human




Dog

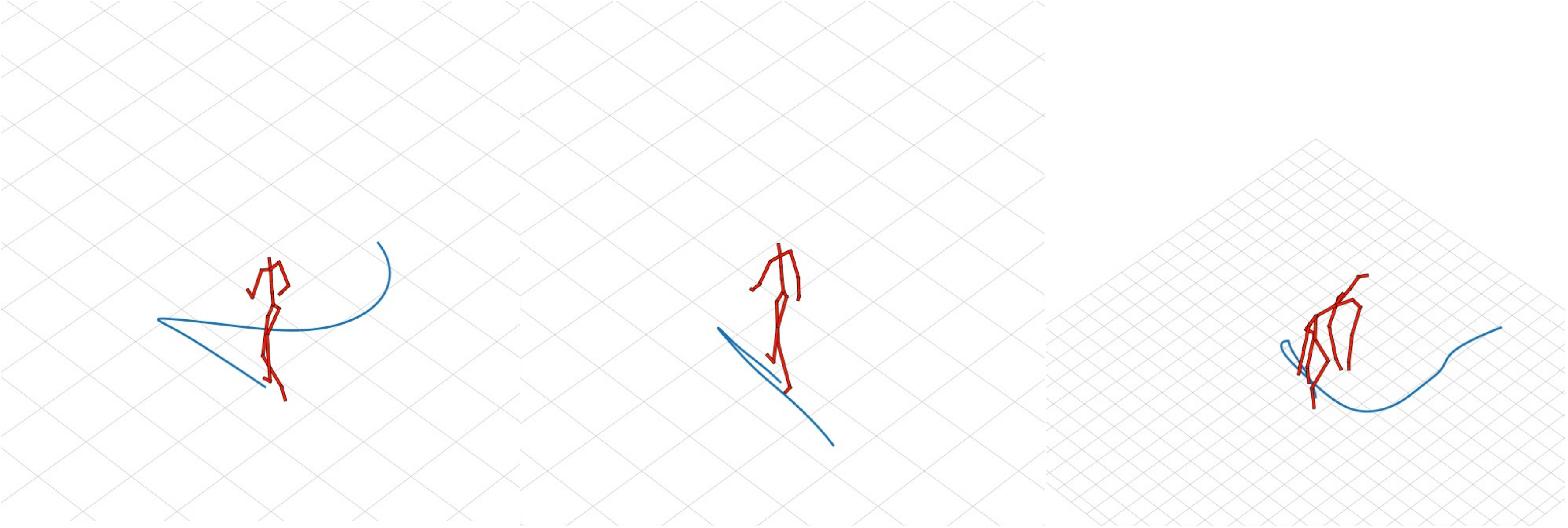


# Footstep analysis

| ID | Human | | | | | Quadruped | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $f_{\text{est}}$ | $v_{\text{tol}}^{(95)}$ | $\mu$ | $\sigma$ | RMSE | $f_{\text{est}}$ | $v_{\text{tol}}^{(95)}$ | $\mu$ | $\sigma$ | RMSE |
| NAT | 297 | 5.0 | 0.31 | 0.26 | - | 290 | 3.2 | 0.61 | 0.71 | - |
| RNN | 328 | 8.0 | 0.39 | 0.39 | 1.7 | 216 | 2.6 | 0.72 | 1.05 | 2.3 |
| VAE | **278** | 7.0 | 0.35 | 0.30 | 1.7 | 277 | **2.9** | **0.61** | 0.90 | 2.0 |
| QN | 318 | **5.0** | 0.23 | 0.19 | **0.07** | - | - | - | - | - |
| MG | **278** | **5.0** | **0.32** | **0.23** | 0.50 | **295** | **2.9** | 0.57 | **0.75** | **0.51** |

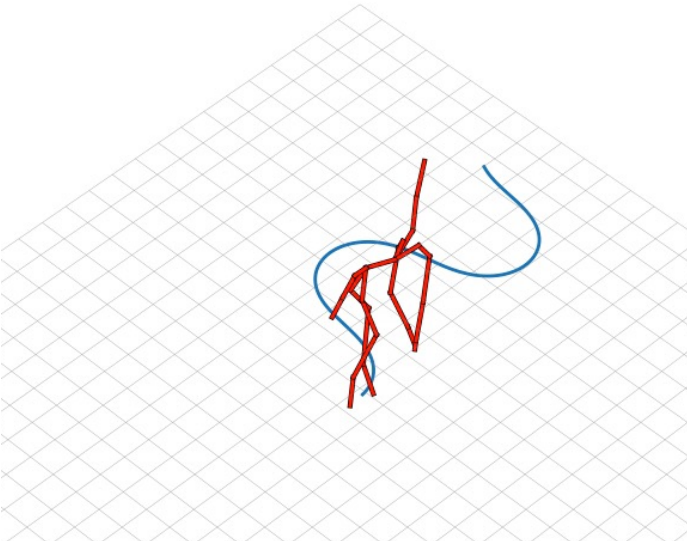
# Additional examples

Held-out control signal

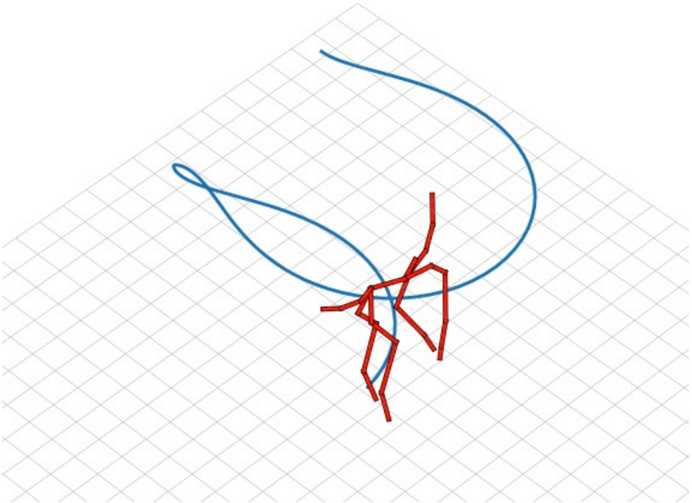


# Additional examples

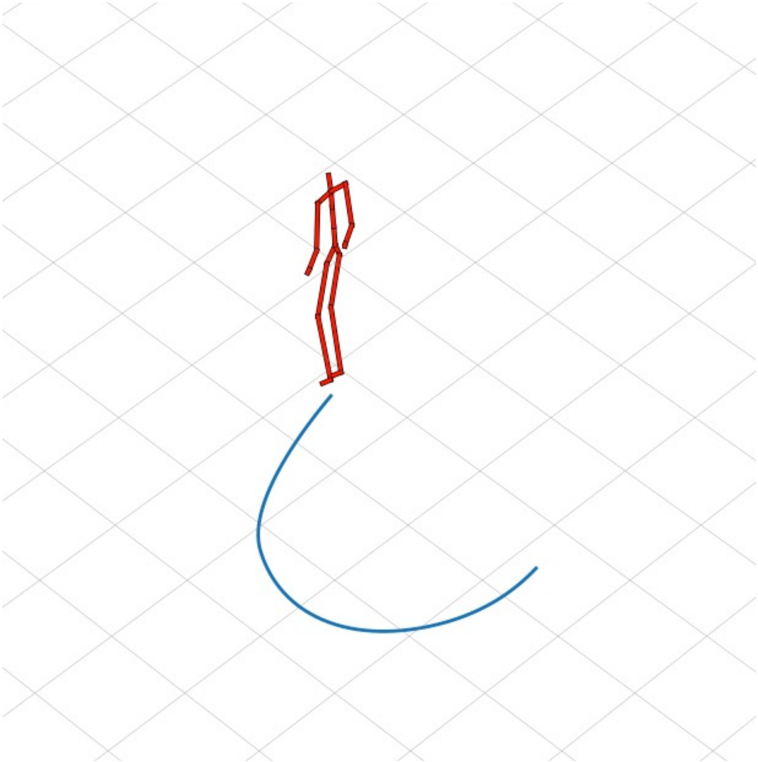
Sinusoidal heading, constant speed



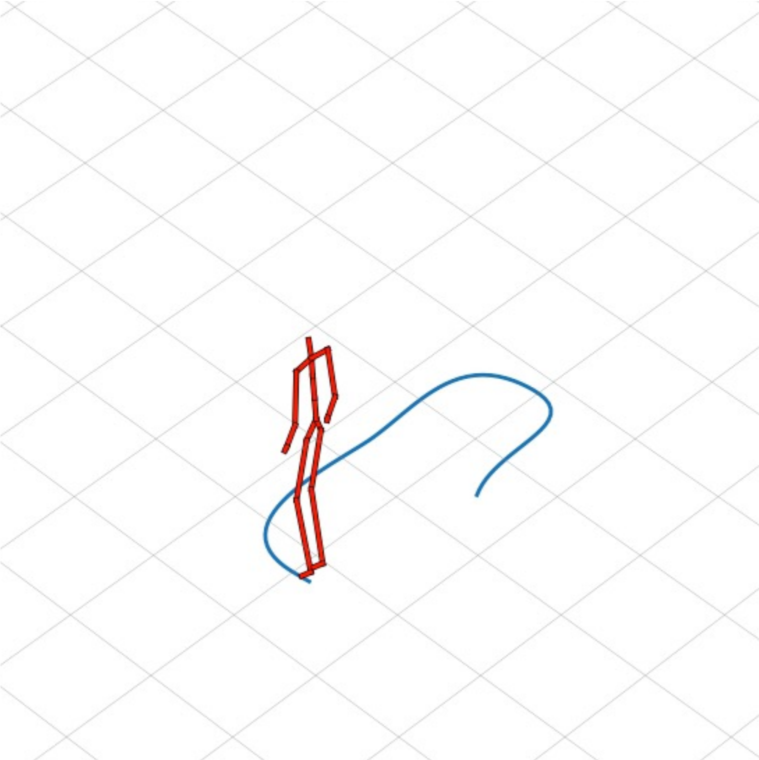
Sinusoidal heading and speed



# Stability and recovery

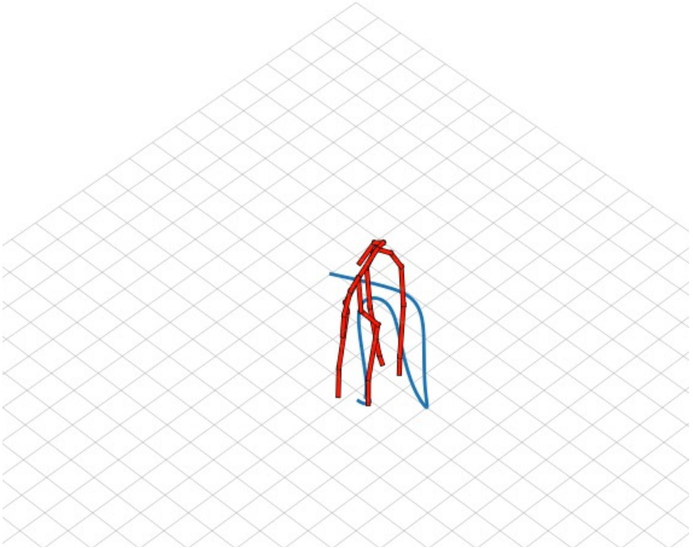


# Stability and recovery

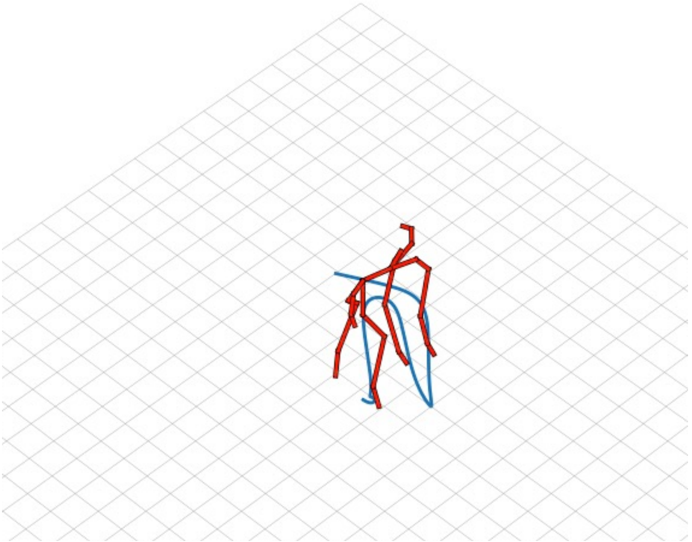


# Random samples with the same control input

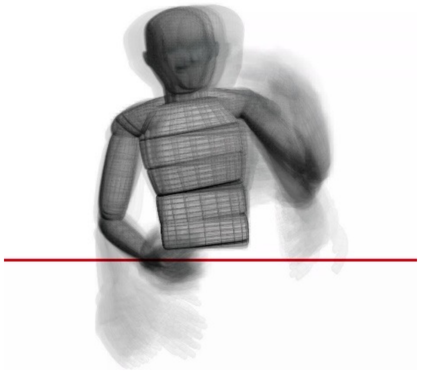
Random sample 1



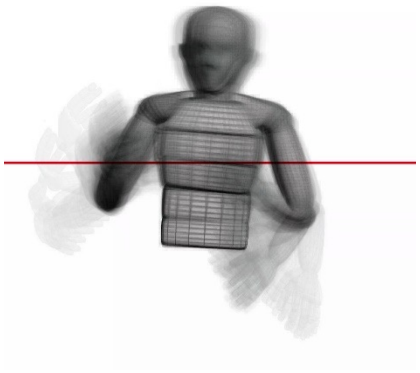
Random sample 2



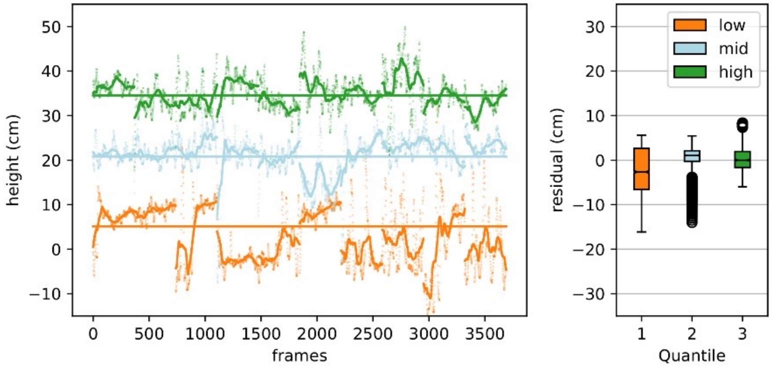
# Objective evaluation



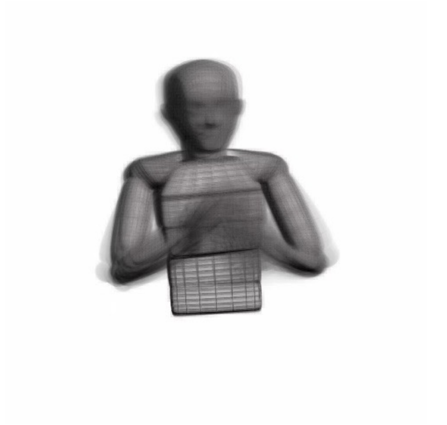
Low right hand



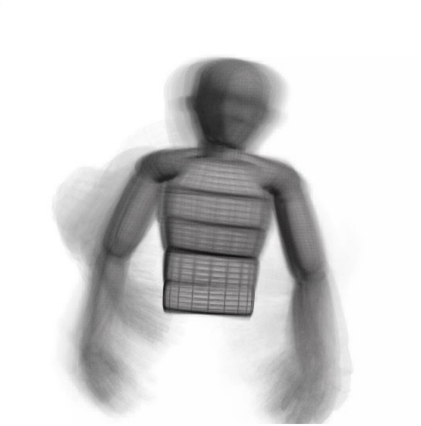
High right hand



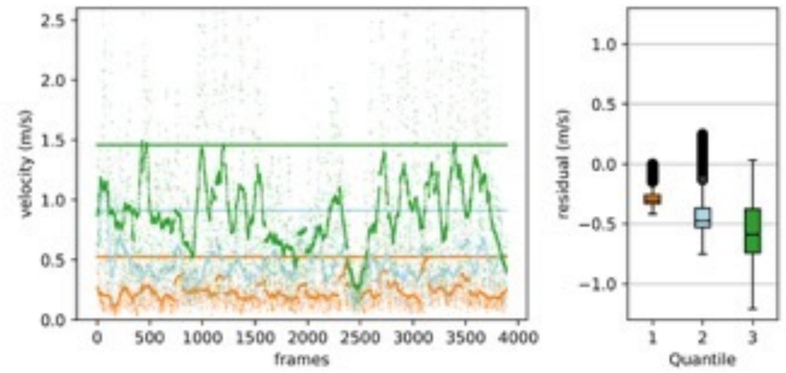
# Objective evaluation



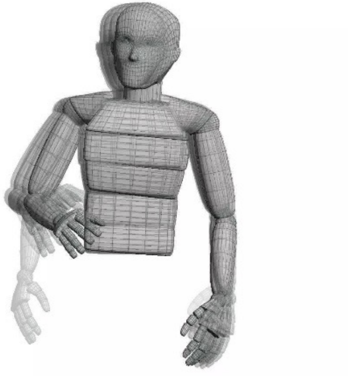
Low speed



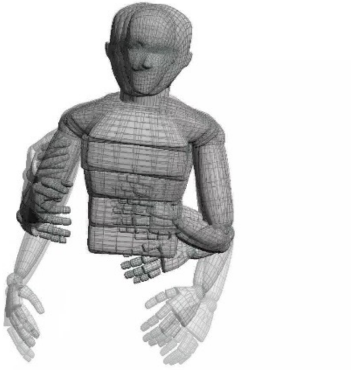
High speed



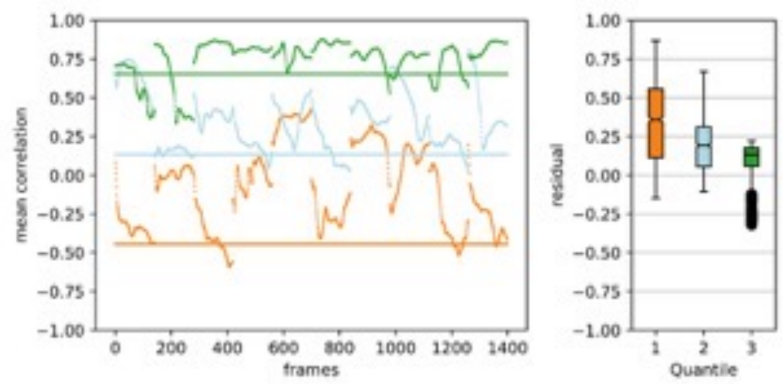
# Objective evaluation



Low symmetry

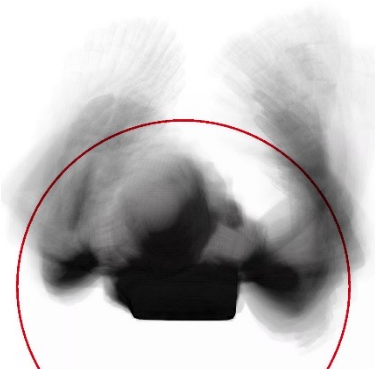


High symmetry



# Objective evaluation

Low radius



High radius