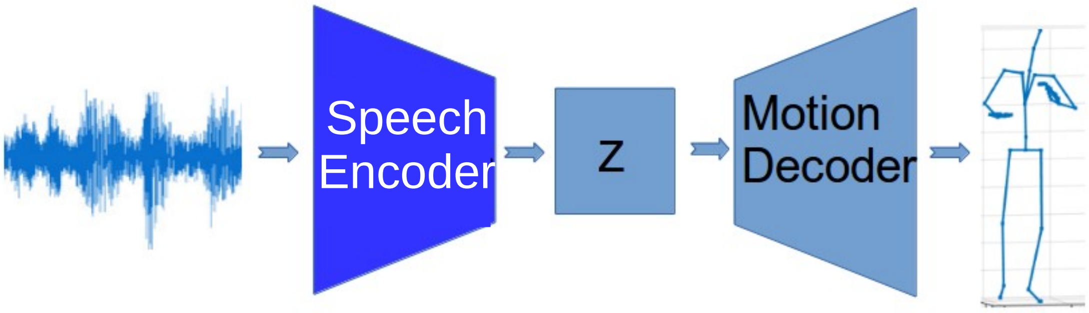
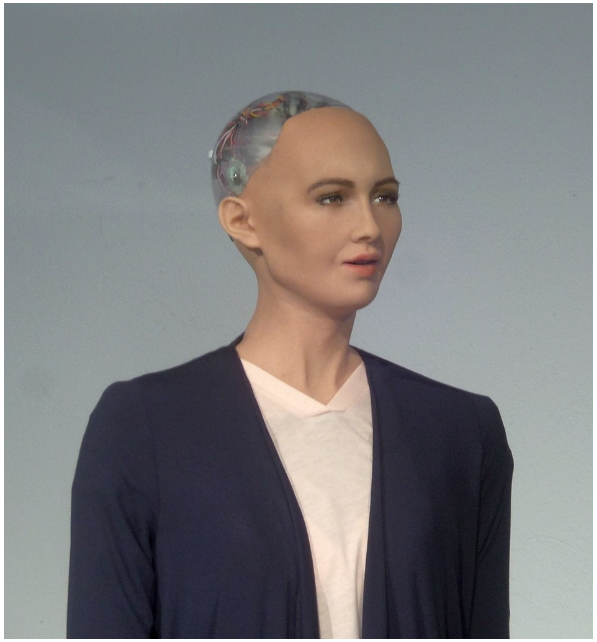
KTH ROYAL INSTITUTE OF TECHNOLOGY

Analyzing Input and Output Representations for Speech-Driven Gesture Generation



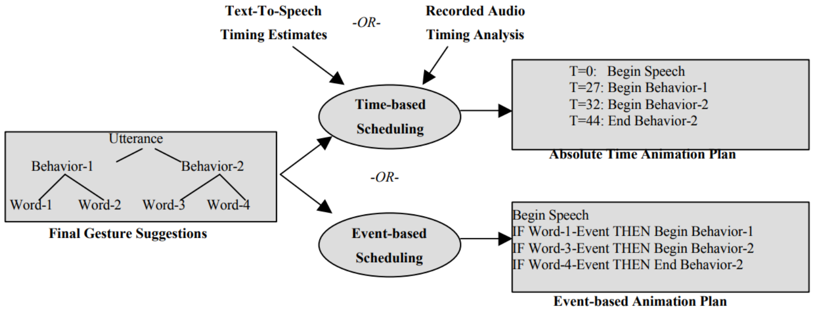
Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, Hedvig Kjellström

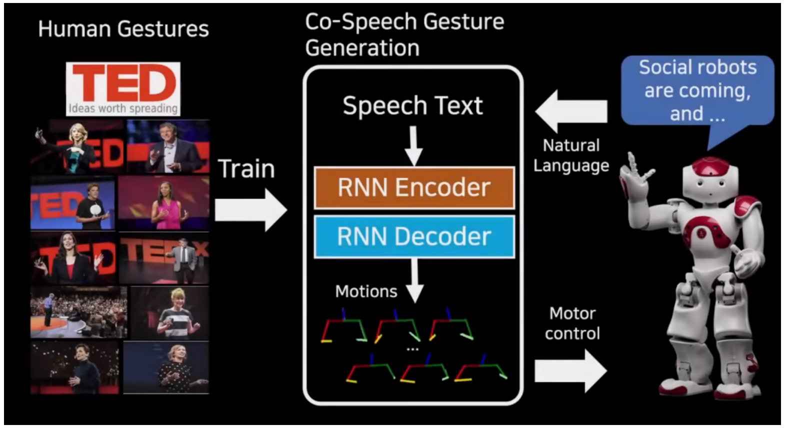
Importance of body language



≠

Why data-driven?



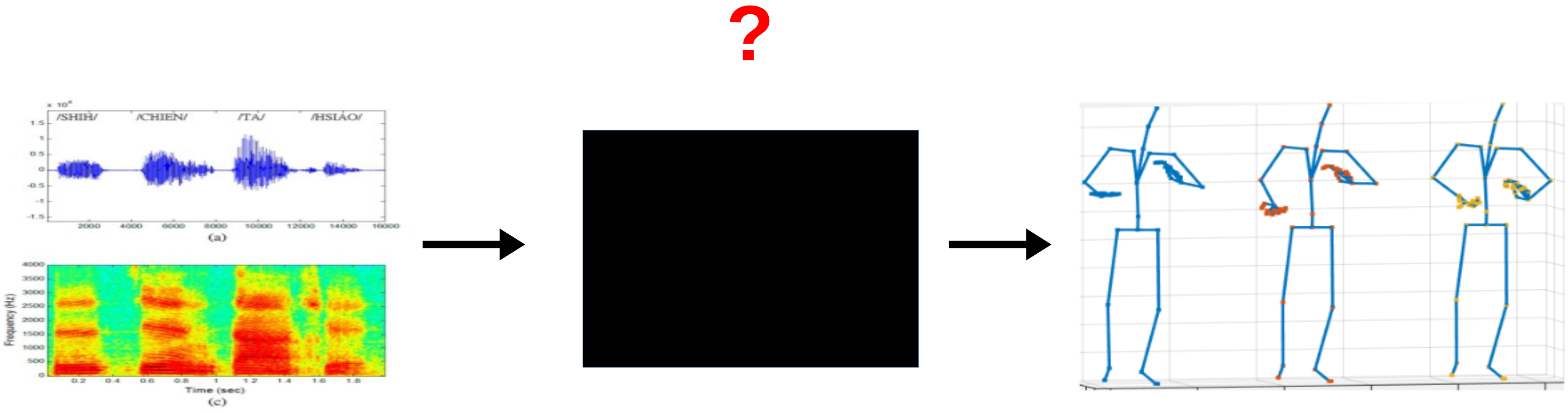


Cassell et al. "BEAT: the Behavior Expression Animation Toolkit" In SIGGRAPH, 2001.

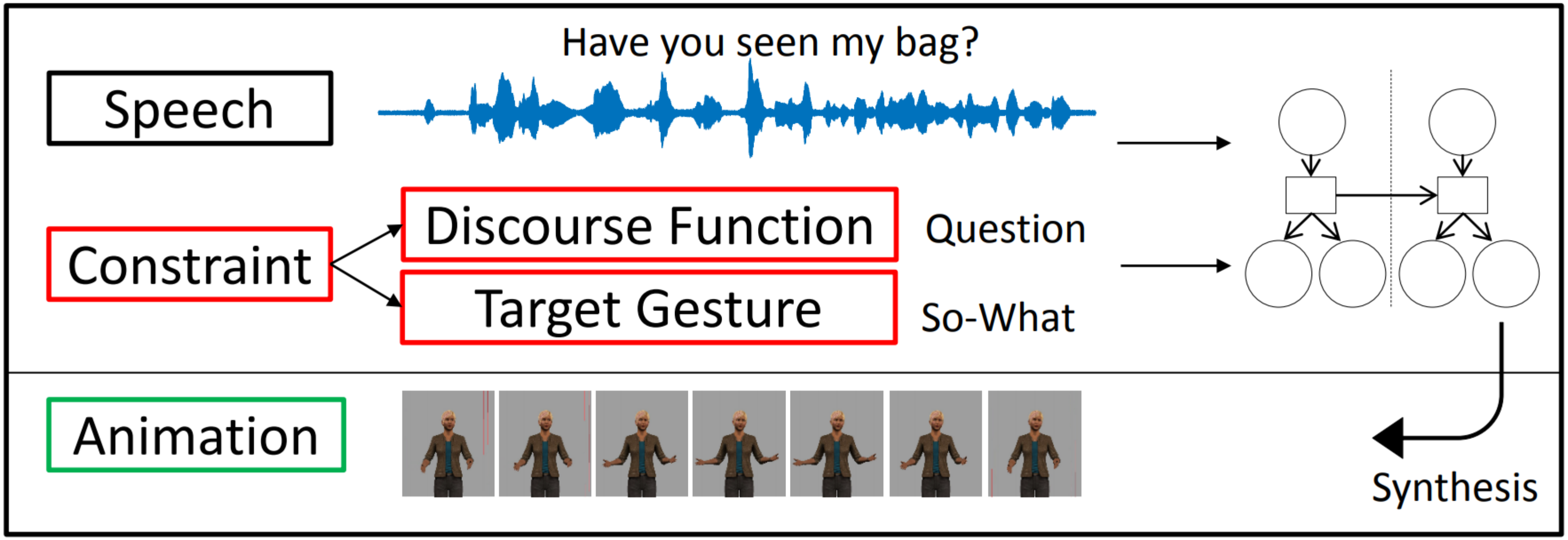
Yoon et al. "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots." In ICRA. 2019

- ✔ Scalability
- ✔ Adaptability
- ✔ Variability

Speech-driven gesture generation



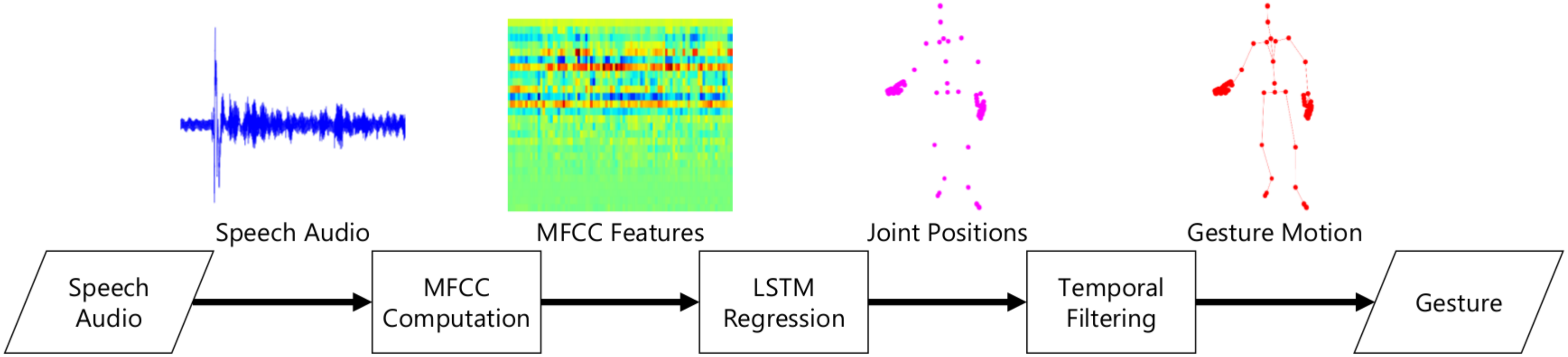
Related work



- Hybrid between data-driven and rule-based approaches
- Based on PGM with an additional hidden node for a constraint
- Evaluate 3 hand gestures and 2 head motions.
- Do smoothing afterwards

Sadoughi et al. "Speech-driven animation with meaningful behaviors."
Speech Communication 110. 2019

Related work



- From speech to 3D motion
- Deep-learning based approach
- Applied a lot of smoothing as post-processing

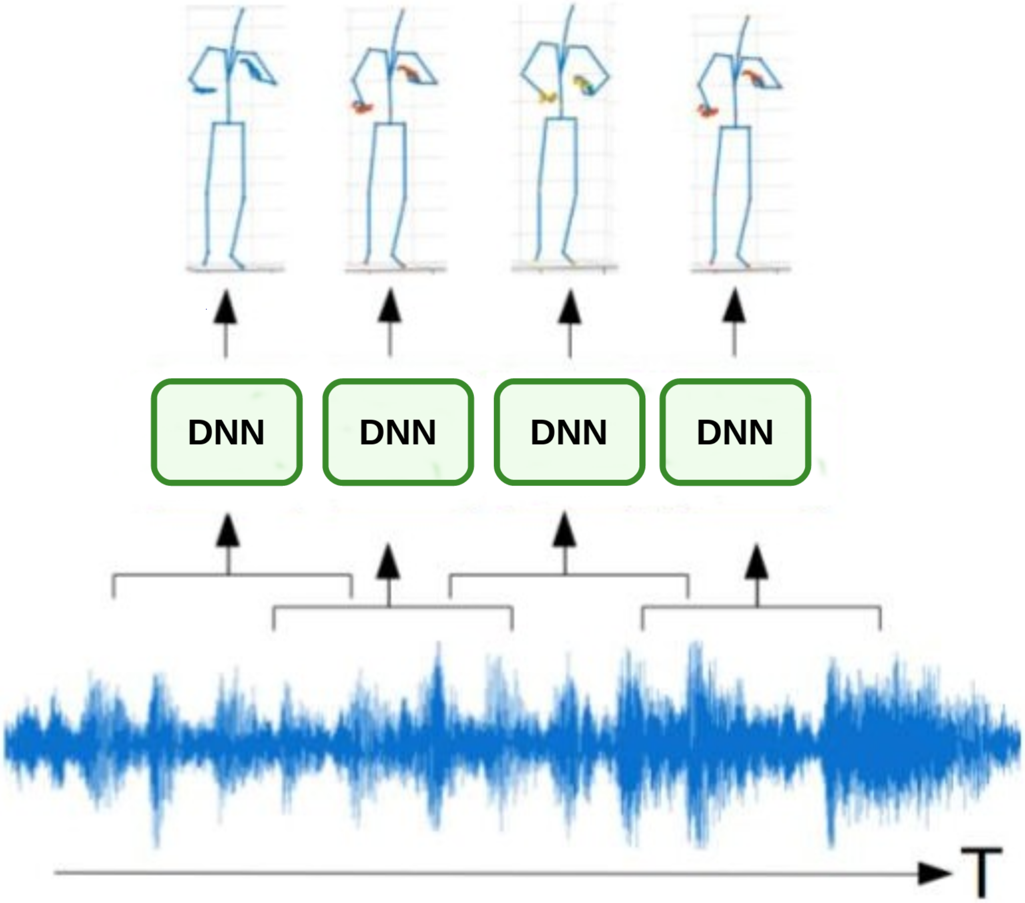
Hasegawa et al. "Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network."
In IVA'18. ACM. 2018.

Contributions

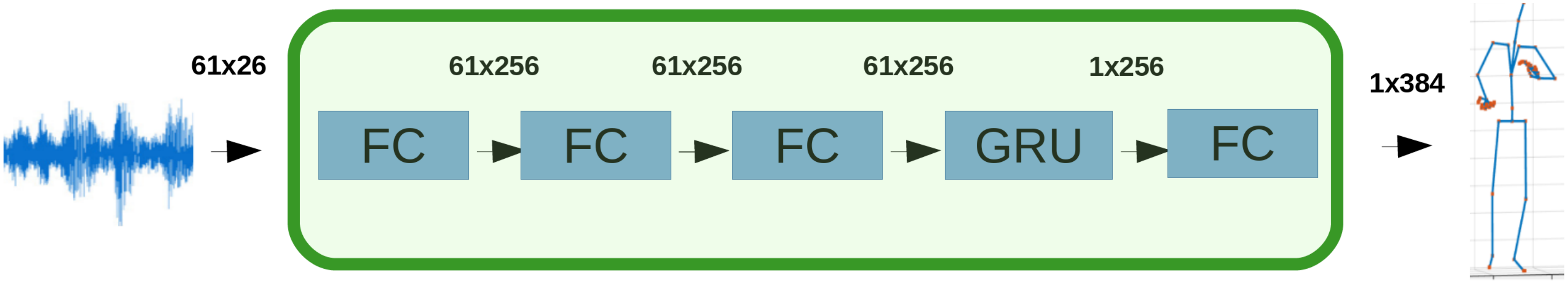
1. A novel speech-driven method for non-verbal behavior generation that can be applied to any embodiment.

2. Evaluation of the importance of representation both for the motion and for the speech

General framework

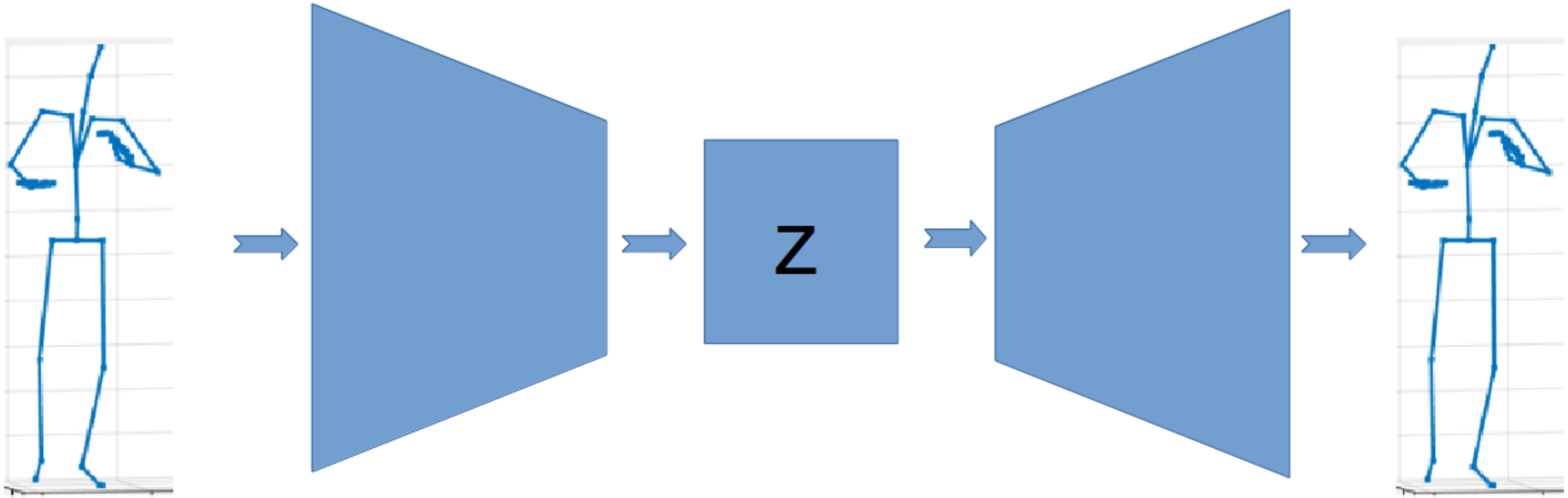


Our baseline model



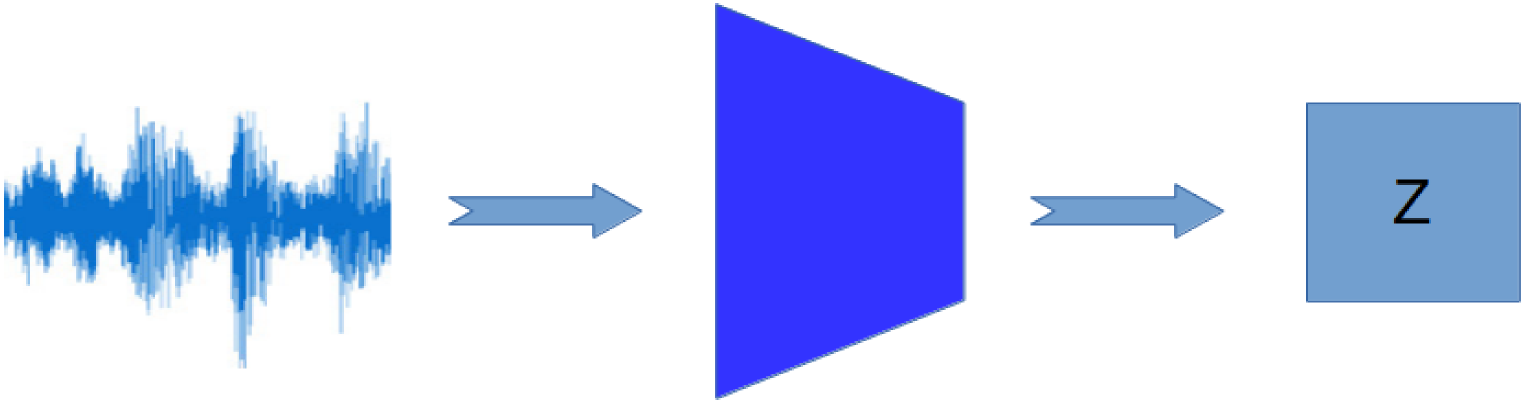
Hasegawa, Dai, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi.
"Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network."
In Proceedings of the 18th International Conference on Intelligent Virtual Agents. ACM, pp. 79-86. 2018.

Proposed method



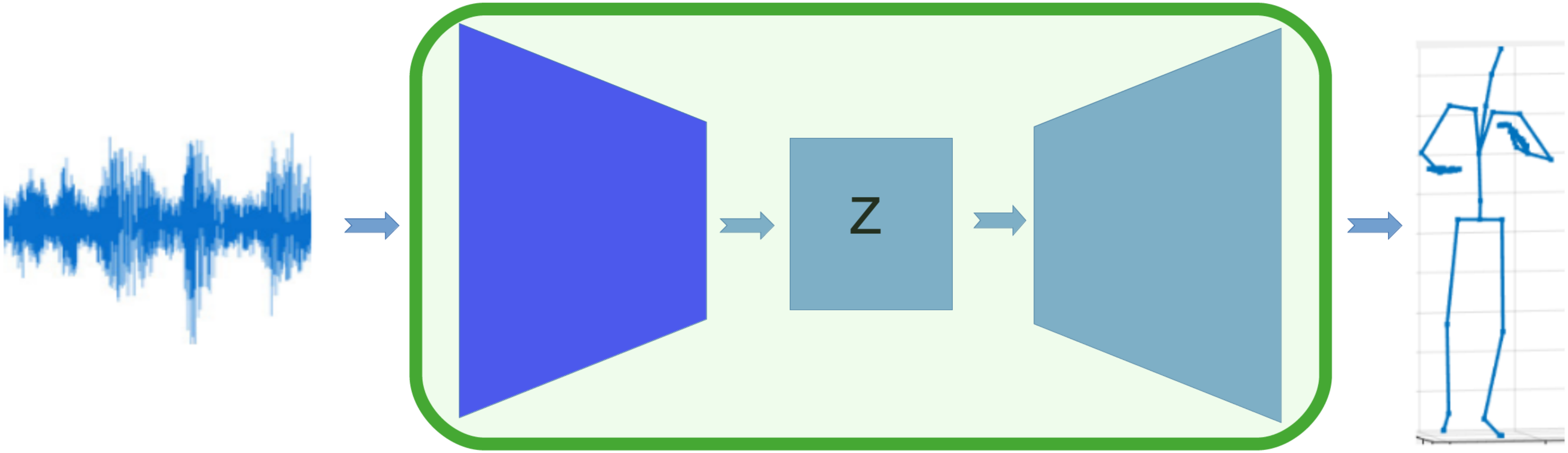
Step 1

Proposed method



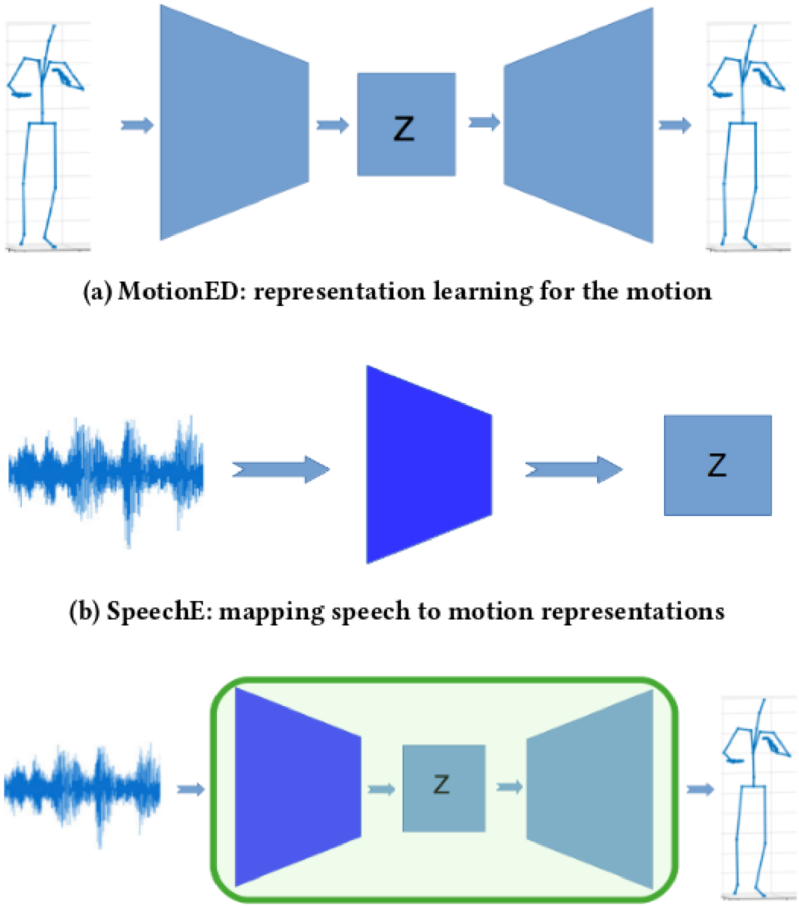
Step 2

Proposed method



Step 3

Proposed method



(a) MotionED: representation learning for the motion

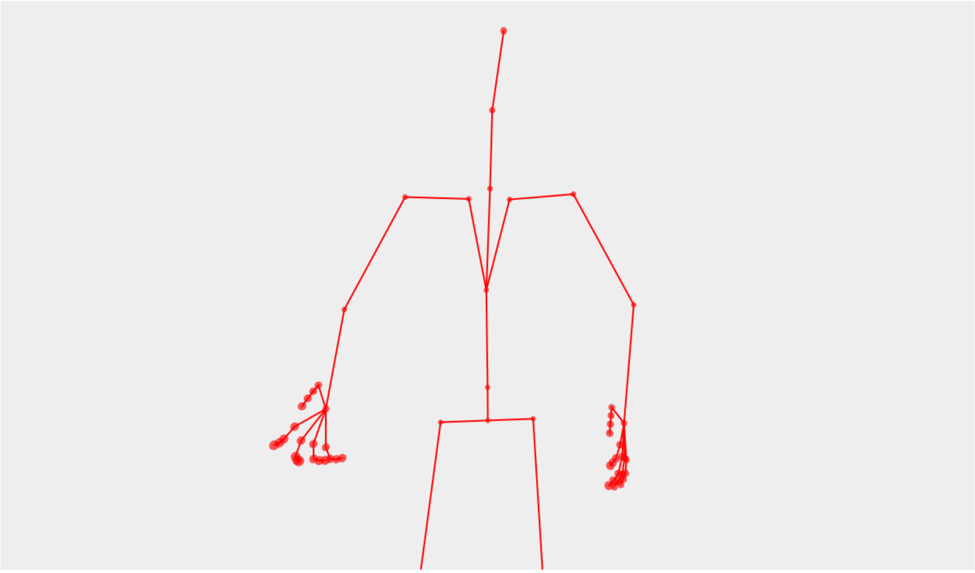
(b) SpeechE: mapping speech to motion representations

(c) Combining the learned components: SpeechE and MotionD

Experimental results

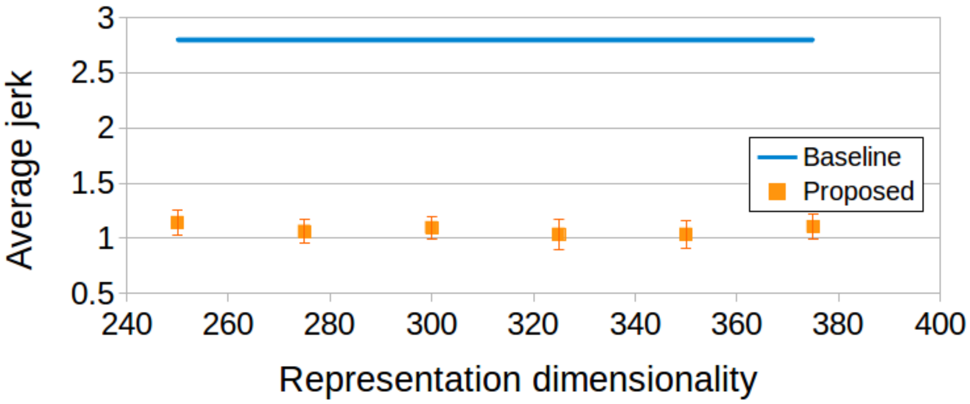
Dataset used

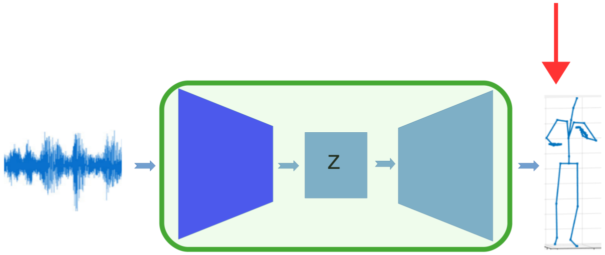
- **Japanese** language
- 171 min of speech and 3D motion
- Speech in mp3 format
- Motion in bvh format



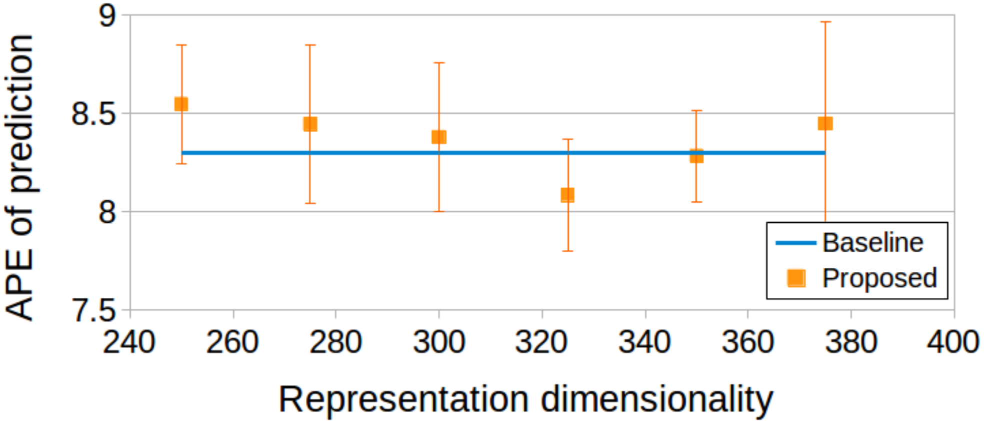
Takeuchi et al. "Creating a gesture-speech dataset for speech-based automatic gesture generation."
In HCII. 2017.

Dimensionality choice

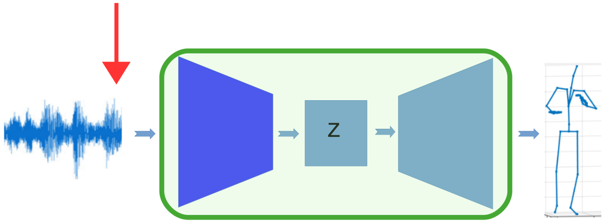




Original dim. was 384

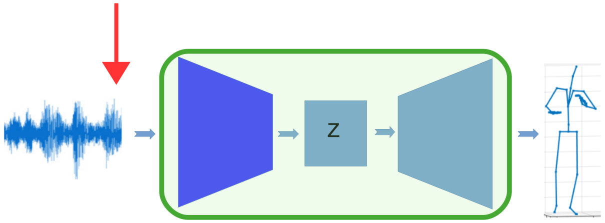


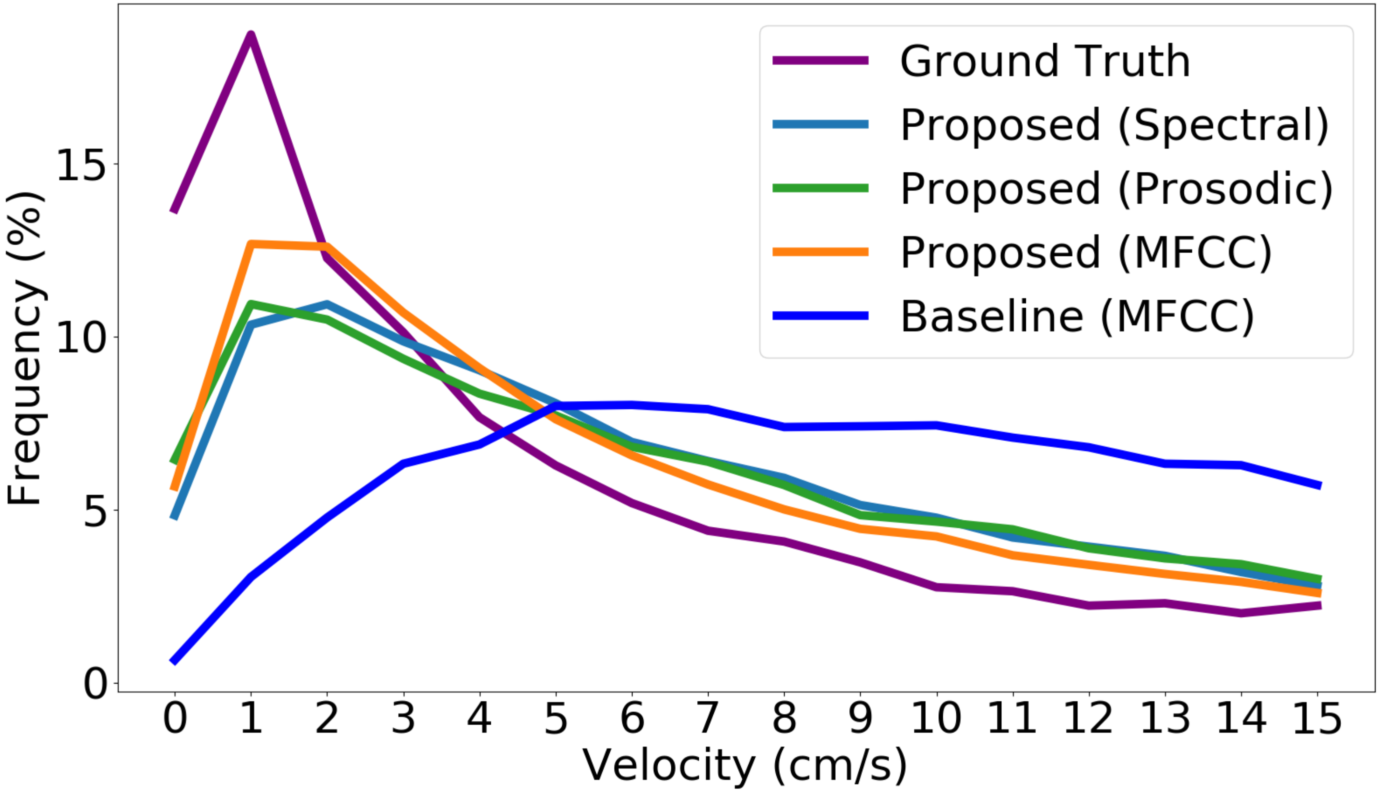
Input feature analysis



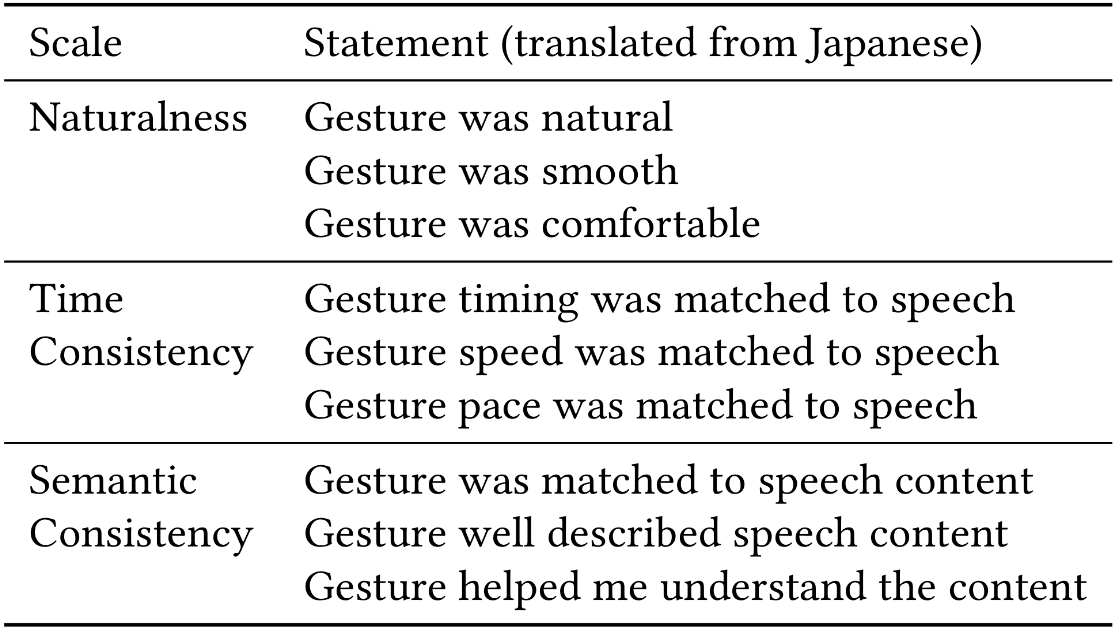
Model	Features	APE	Acceleration	Jerk
Static mean pose		8.95	0	0
Proposed	Prosodic	8.56±0.2	0.90±0.03	1.52±0.07
Proposed	Spectral	8.27±0.4	**0.51**±0.07	**0.85**±0.12
Proposed	Spec. + Pros.	8.11±0.3	0.57±0.08	0.95±0.12
Proposed	MFCC	**7.66**±0.2	0.53±0.03	0.91±0.05
Proposed	MFCC + Pros.	**7.65**±0.2	0.58±0.06	0.97±0.11
Baseline	MFCC	8.07±0.1	1.50±0.03	2.62±0.05
Ground truth		0	0.38	0.54

Histogram for wrists joints





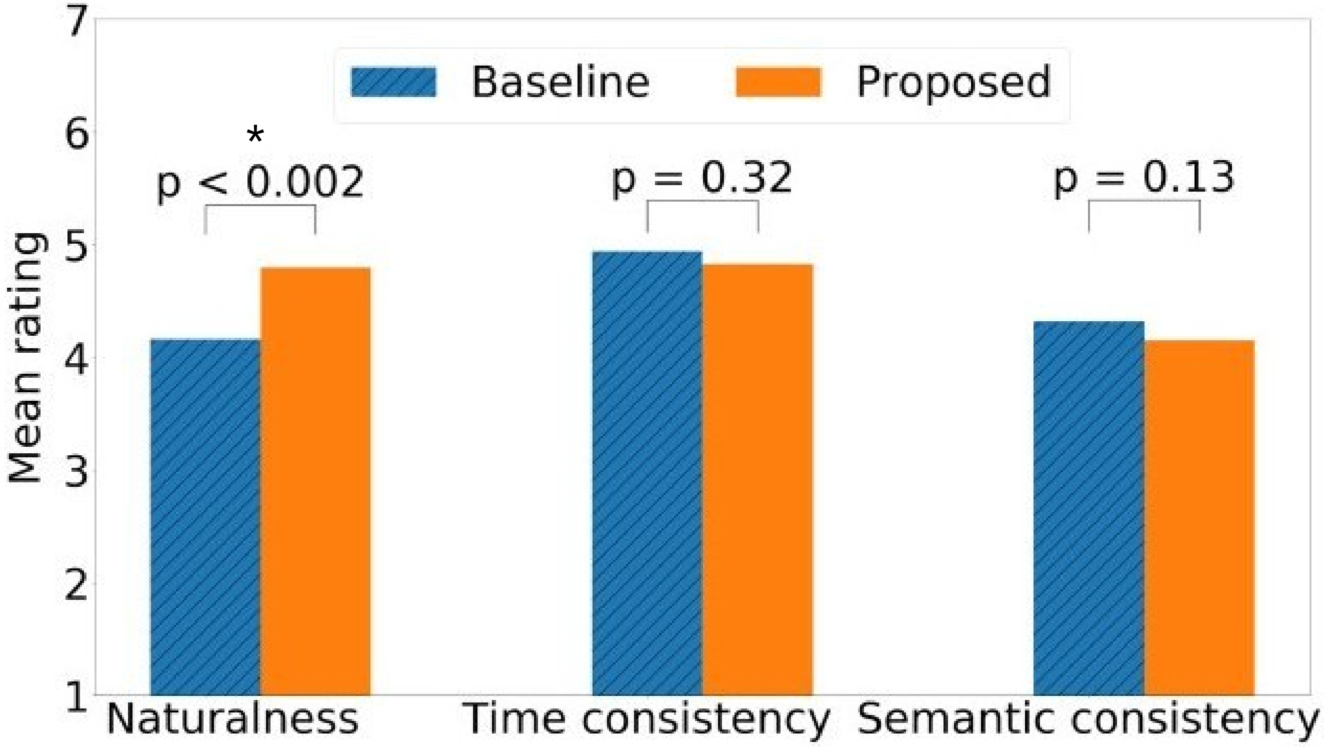
User study measures



Scale	Statement (translated from Japanese)
Naturalness	Gesture was natural Gesture was smooth Gesture was comfortable
Time Consistency	Gesture timing was matched to speech Gesture speed was matched to speech Gesture pace was matched to speech
Semantic Consistency	Gesture was matched to speech content Gesture well described speech content Gesture helped me understand the content

All were evaluated in the Likert scale from 1 to 7

User study results



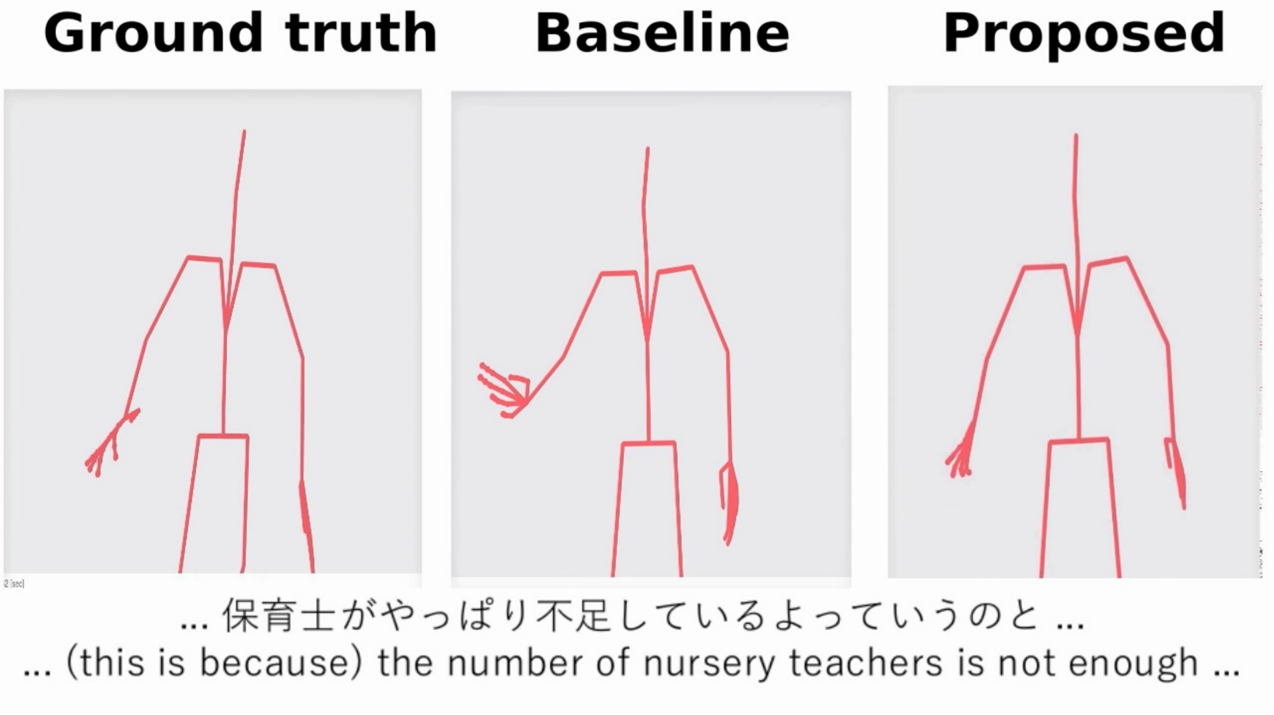
19 participants with
10 videos x 9 questions x 2 conditions = 180 ratings each

Visual comparison

Baseline model

No smoothing was applied

Visual comparison



… 保育士がやっぱり不足しているよっていうのと …

… (this is because) the number of nursery teachers is not enough …

No smoothing was applied

Conclusion

Deep-**learning** based speech-driven **gesture generation** becomes more natural using **representation learning**

The team

Taras Kucherenko

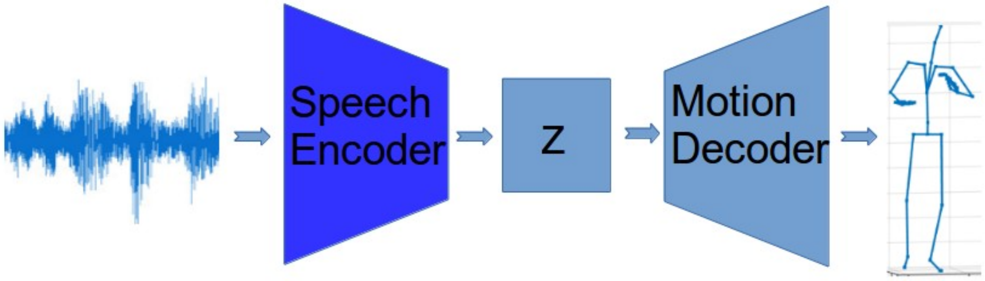
Dai Hasegawa

Gustav Eje Henter

Naoshi Kaneko

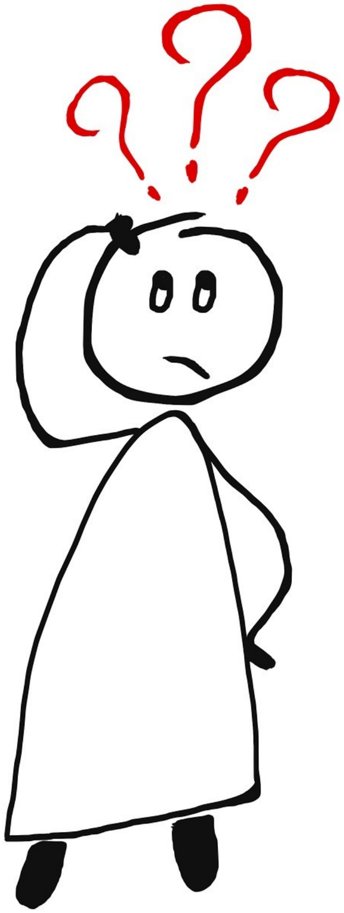
Hedvig Kjellström

Analyzing Input and Output Representations for Speech-Driven Gesture Generation



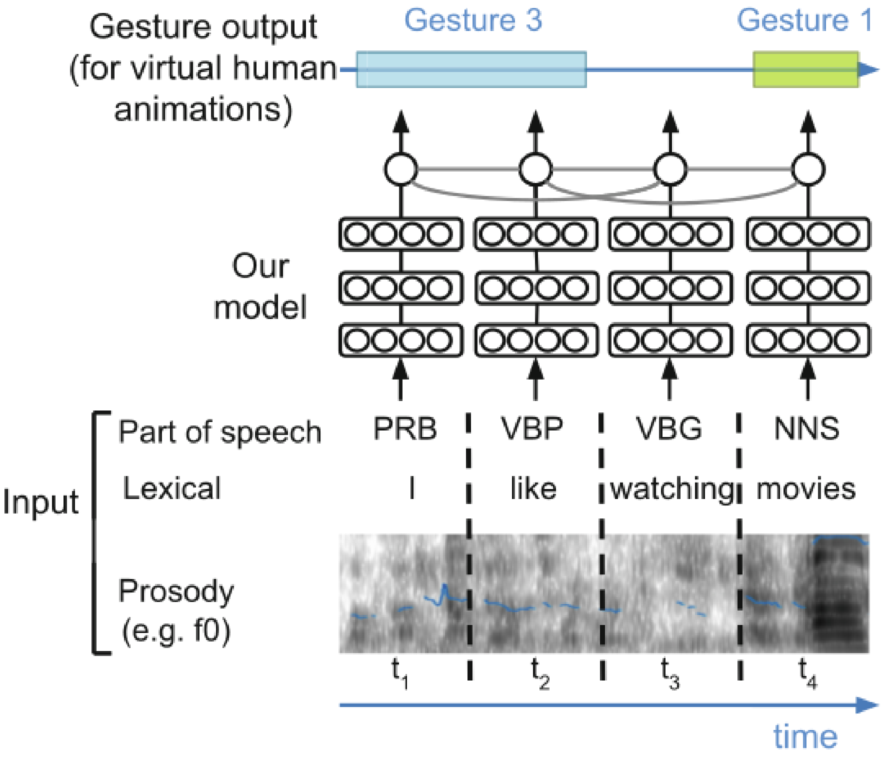
Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, Hedvig Kjellström

Questions?



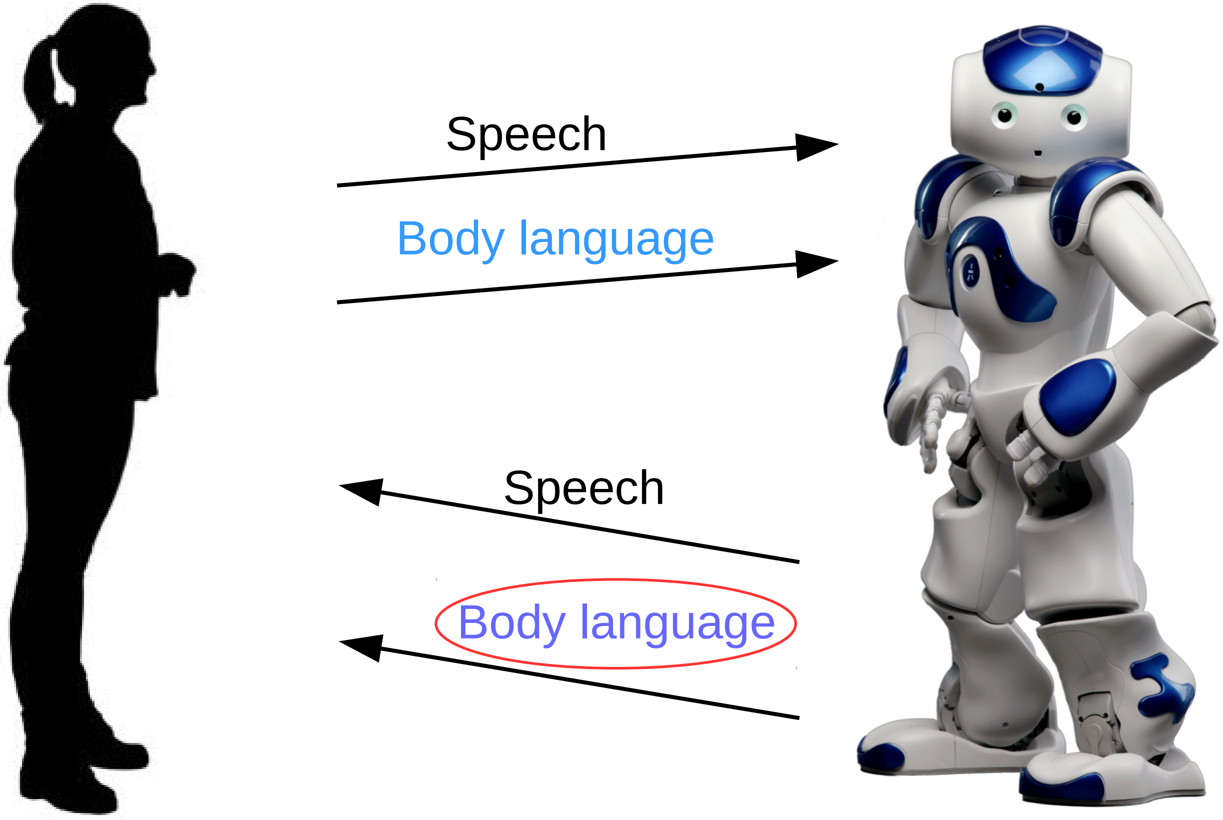
Related work

- DNN + CRF = DCNF
- Virtual character
- Discrete set of motions



Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella.
Predicting co-verbal gestures: a deep and temporal modeling approach.
International Conference on Intelligent Virtual Agents. Springer, Cham, 2015.

Human-robot communication



Speech

Body language

Speech

Body language

https://www.ald.softbankrobotics.com