KTH ROYAL INSTITUTE
OF TECHNOLOGY

Speech2Properties2Gestures: Gesture-Property Prediction as a Tool for Generating Representational Gestures from Speech

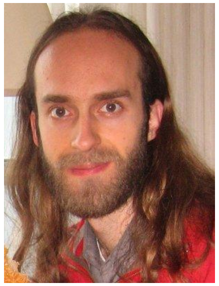
Taras Kucherenko

Rajmund Nagy

Patrik Jonell

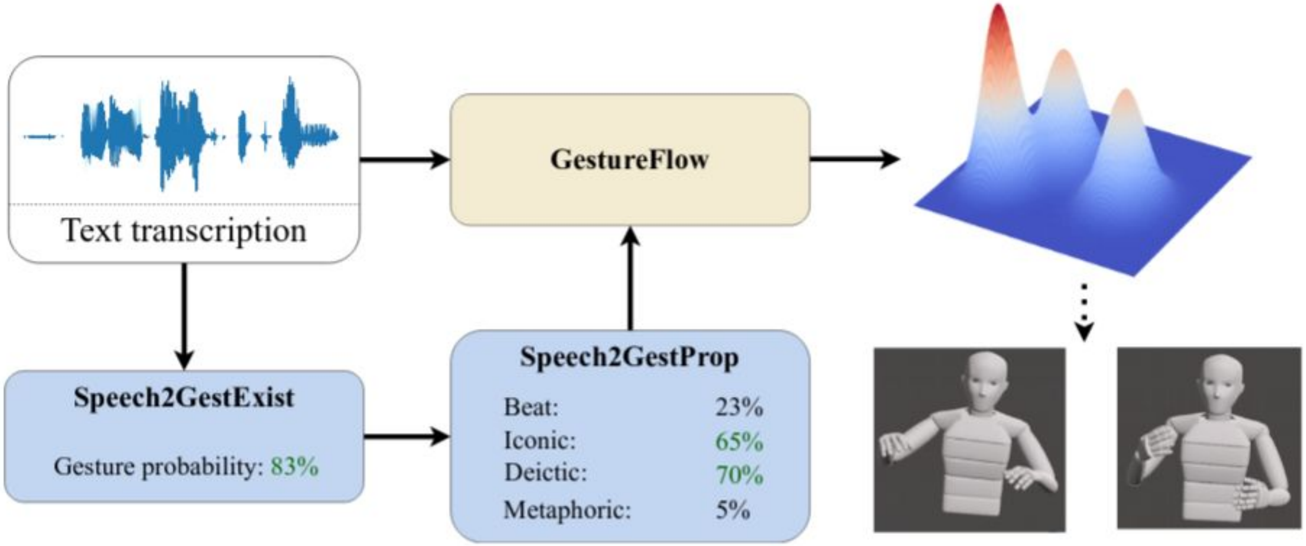
Michael Neff

Hedvig Kjellström


Gustav Eje Henter

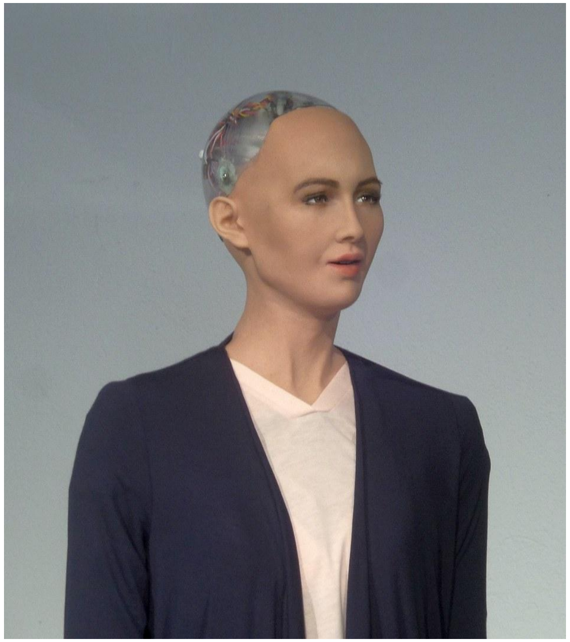
International Conference on Intelligent Virtual Agents (IVA’21). 2021

Takeaway / TL;DR



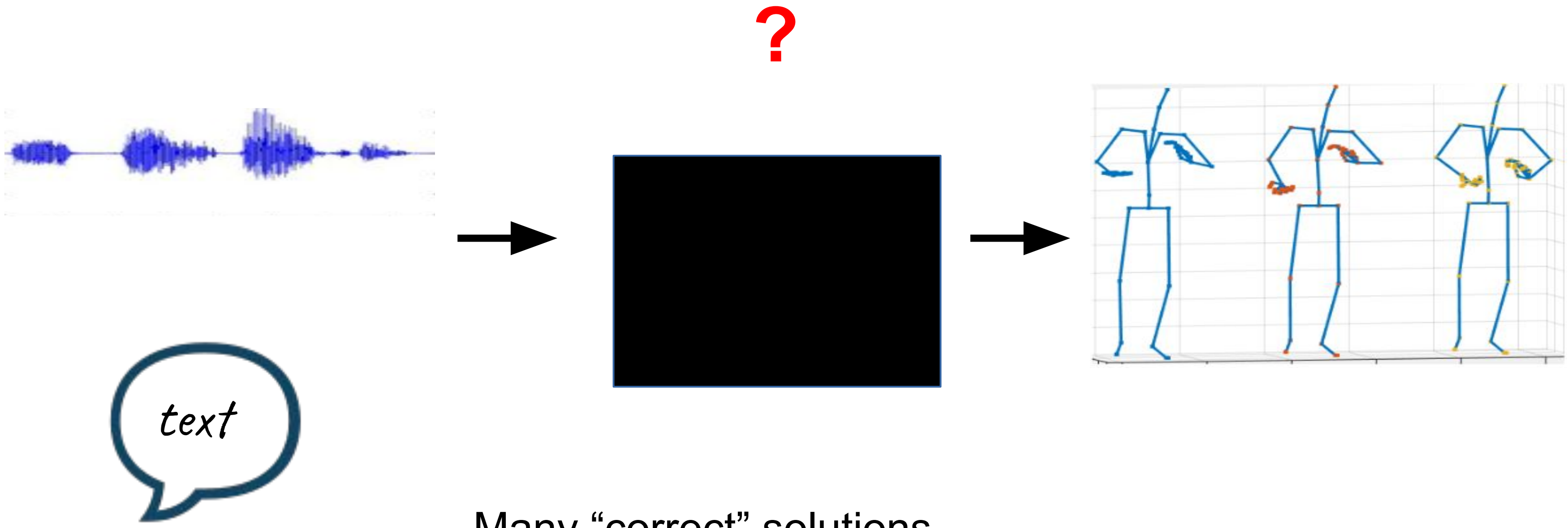
- Intent-driven methods and direct-synthesis can be married …
- We propose a system that predicts gesture properties and uses them to condition the generation
- Early results are promising

Importance of body language



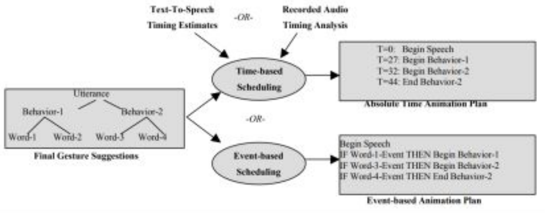
≠

Speech-driven gesture generation

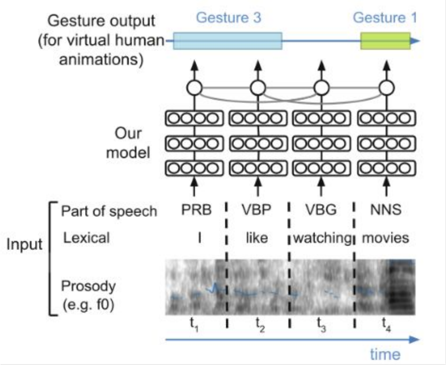


- Many “correct” solutions
- Little data available
- Depends on culture, context and mental state

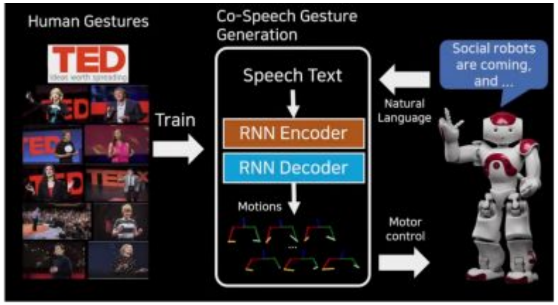
Previous work



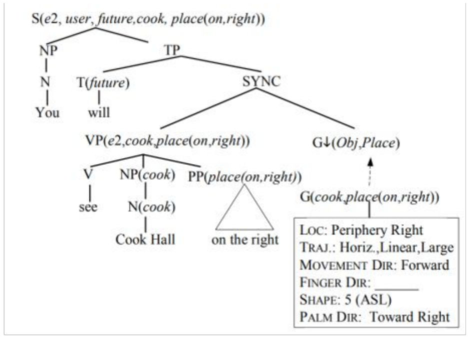
Cassell et al. "BEAT: the Behavior Expression Animation Toolkit" In SIGGRAPH, 2001.



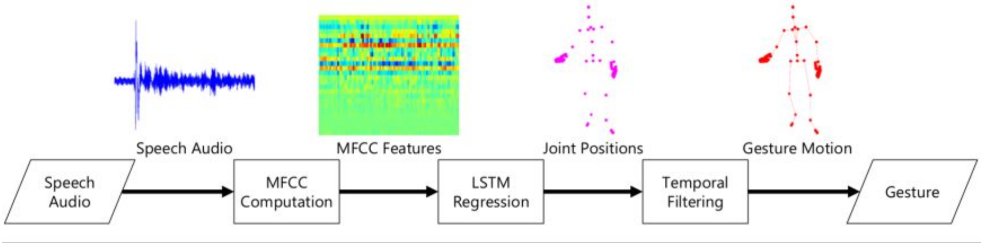
Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella.
Predicting co-verbal gestures: a deep and temporal modeling approach.
International Conference on Intelligent Virtual Agents. 2015.



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

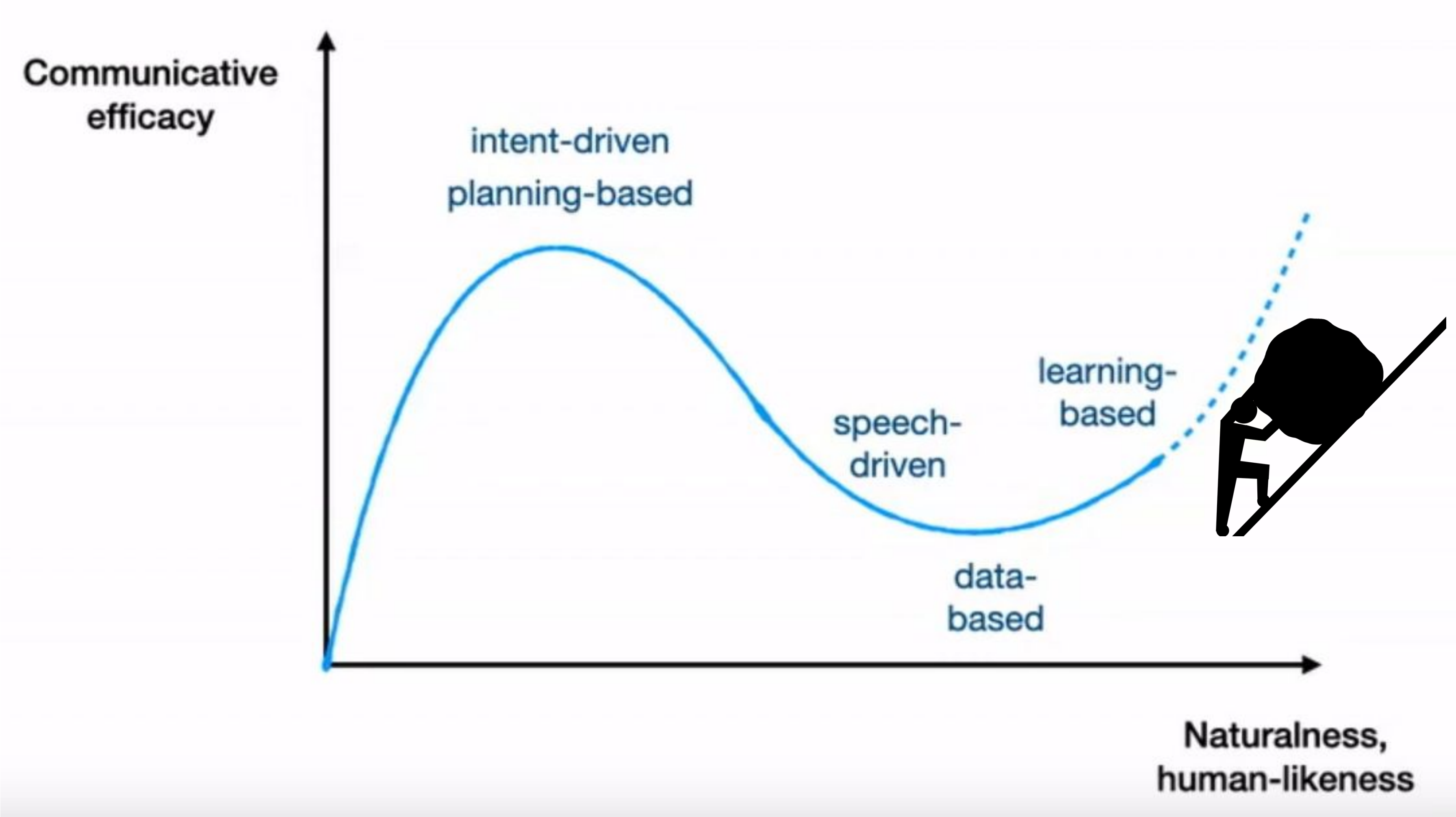


Stefan Kopp, Paul Tepper, and Justine Cassell. 2004.
Towards integrated microplanning of language and iconic gesture for multimodal output.
In Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04).



Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi
"Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network."
International Conference on Intelligent Virtual Agents. 2018.

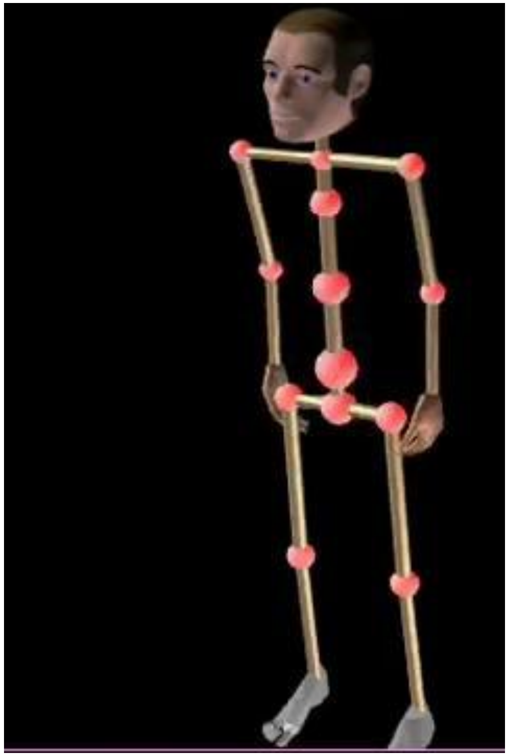
Field development



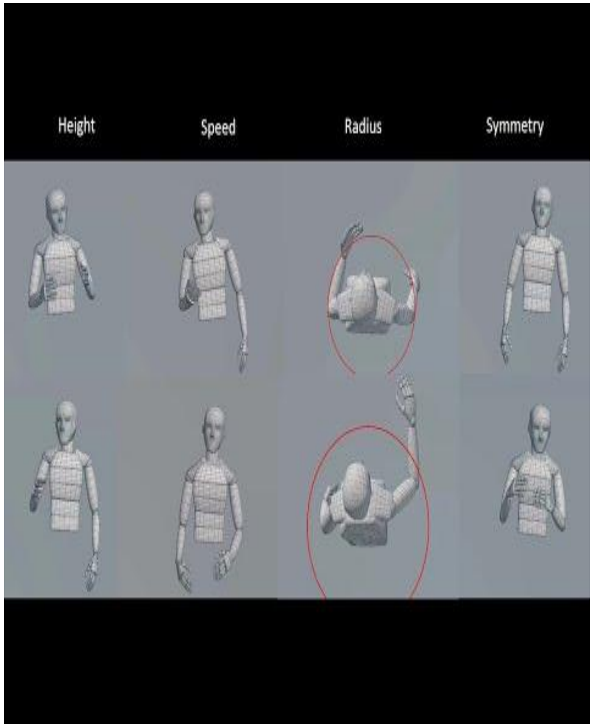
Slide credit - Stefan Kopp

https://www.maxpixel.net/Struggle-Push-Effort-Strength-Uphill-Man-Battle-1271657

Data-driven vs Rule-based

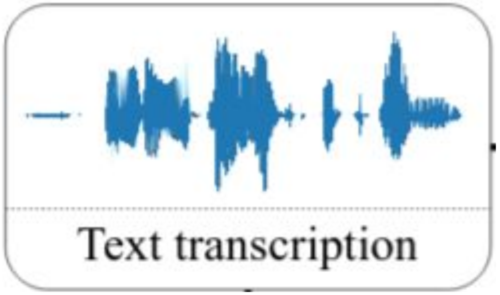


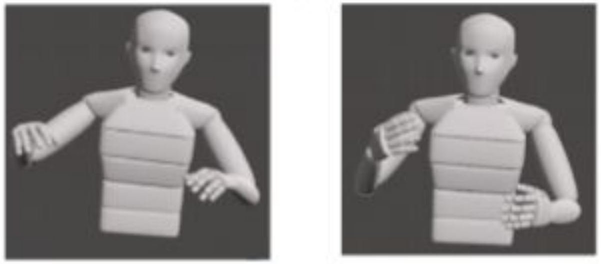
S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth, “Max – a multimodal assistant in virtual reality construction,” KI – Künstliche Intelligenz, vol. 17, no. 4, pp. 11–17, 2003



Alexanderson, S., Henter, G. E., Kucherenko, T., & Beskow, J. (2020, May). Style‐Controllable Speech‐Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum* (pp. 487-496).

Speech2Properties2Gestures



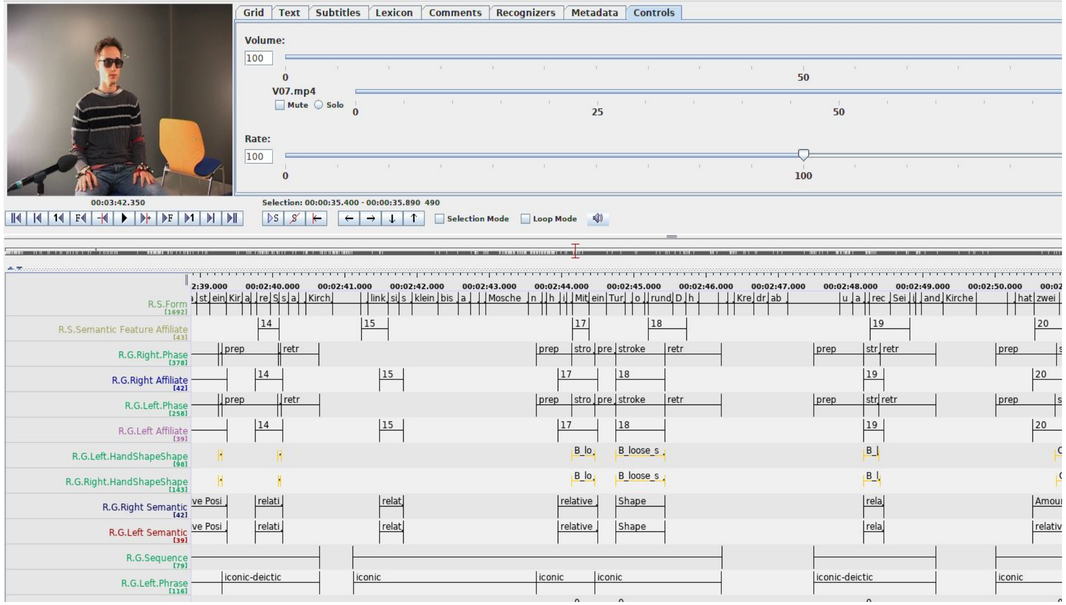


Dataset used

- German
- 240 min
- Speech and motion format
- Transcribed
- Annotated for various gesture properties including gesture category, gesture phase and semantic content

Lücking, Andy, et al. “Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications.” *Journal on Multimodal User Interfaces* 7.1 (2013): 5-18.

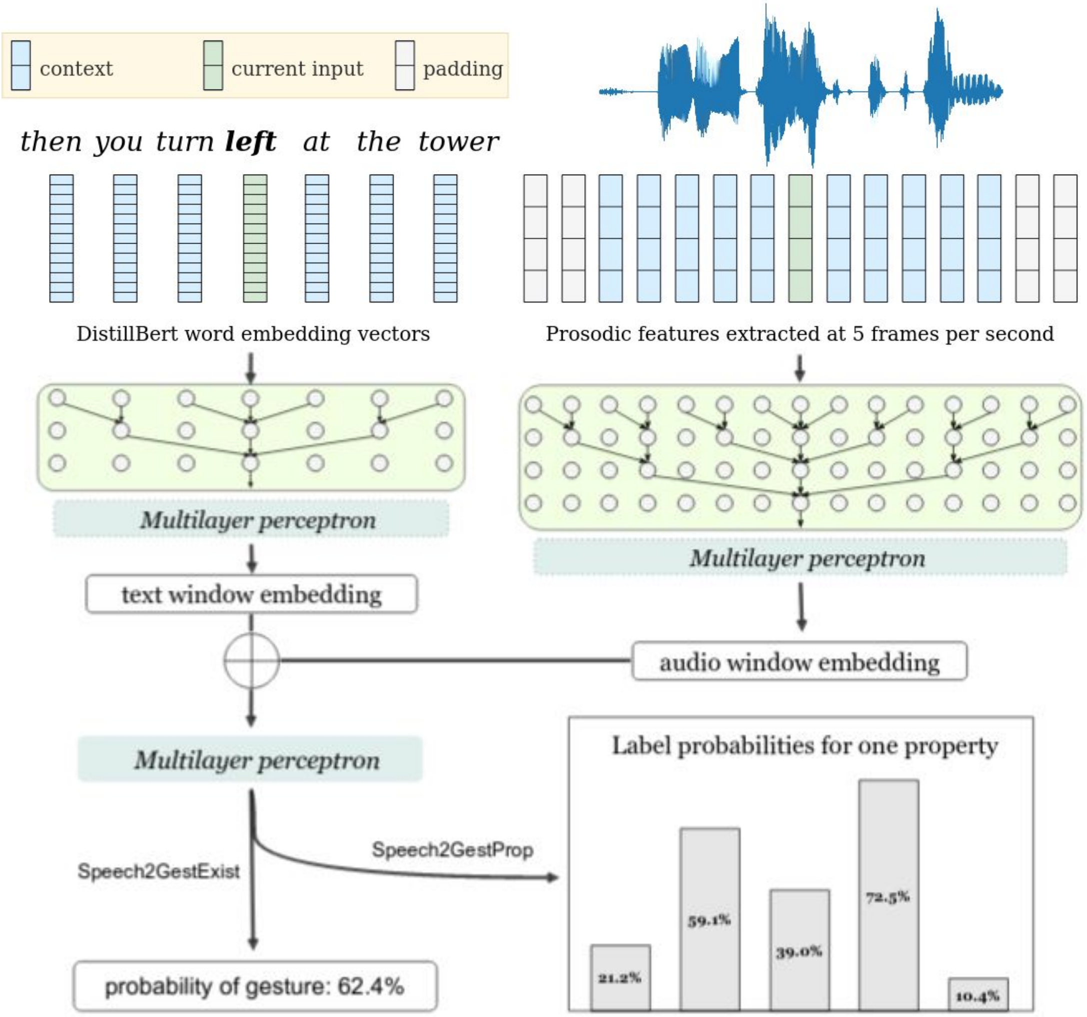
Gesture properties



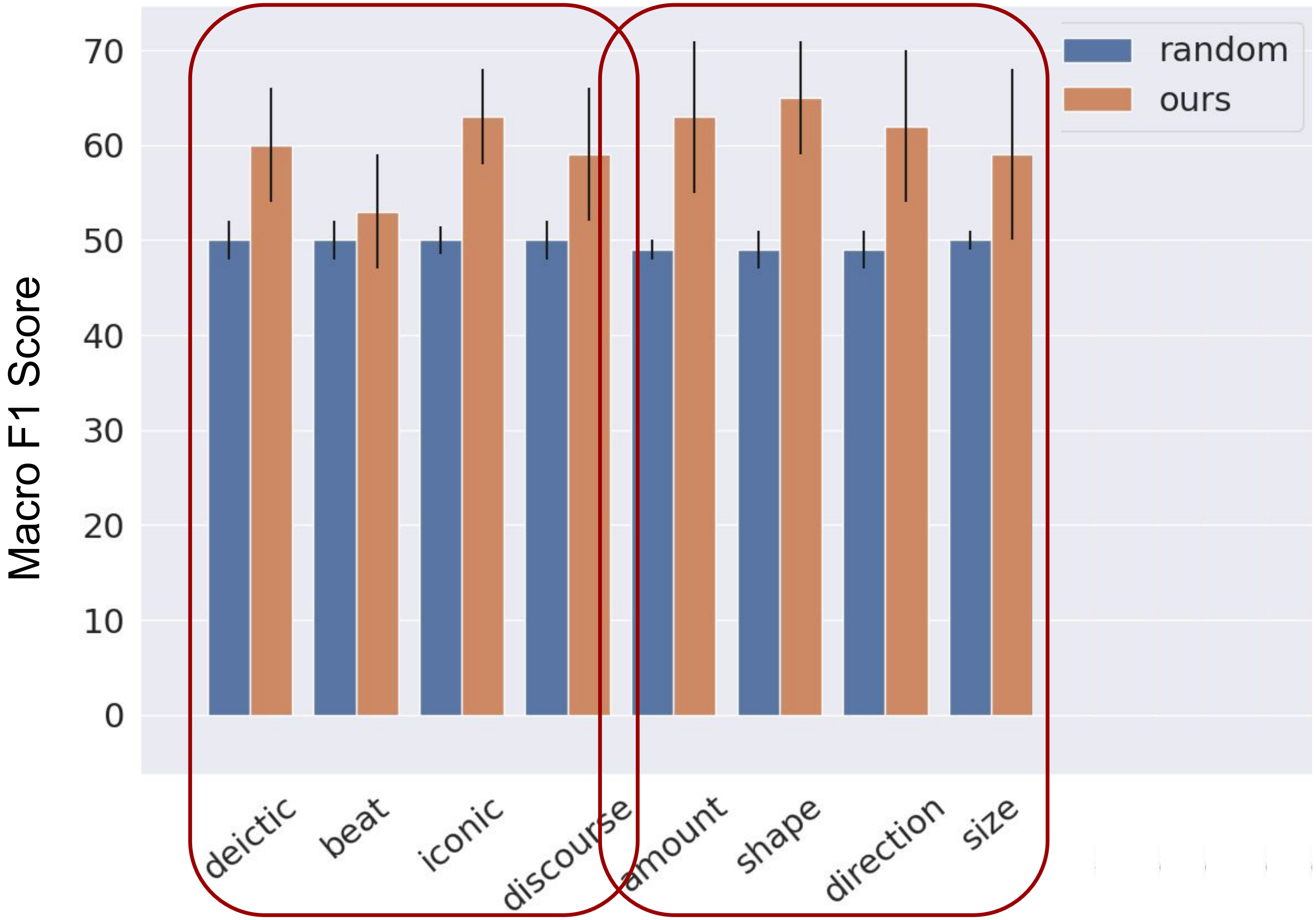
	gesture category [Macro F_1]				gesture semantics [Macro F_1]				gesture phase [F_1]				
label	deictic	beat	iconic	discourse	amount	shape	direction	size	pre-hold	post-hold	stroke	retr	prep
relative frequency	29.05%	14.47%	72.03%	12.78%	4.7%	13.1%	13.7%	1.9%	0.6%	12.2%	40.9%	14.8%	30.8%

KTH ROYAL INSTITUTE
OF TECHNOLOGY

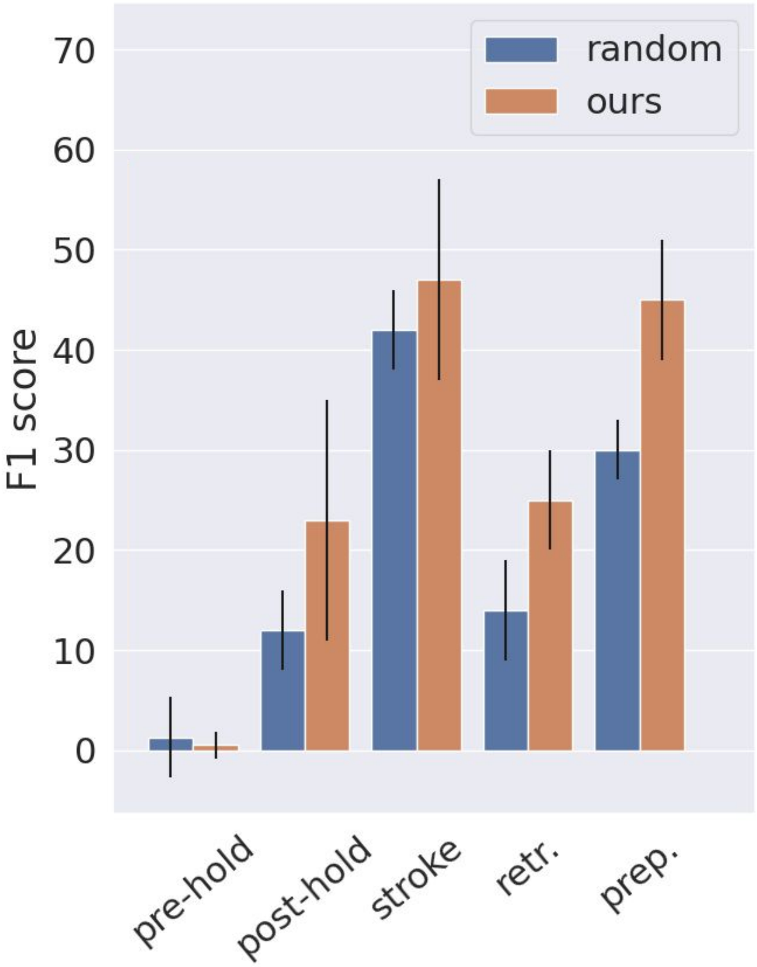
Speech2Properties



Speech2Prop results



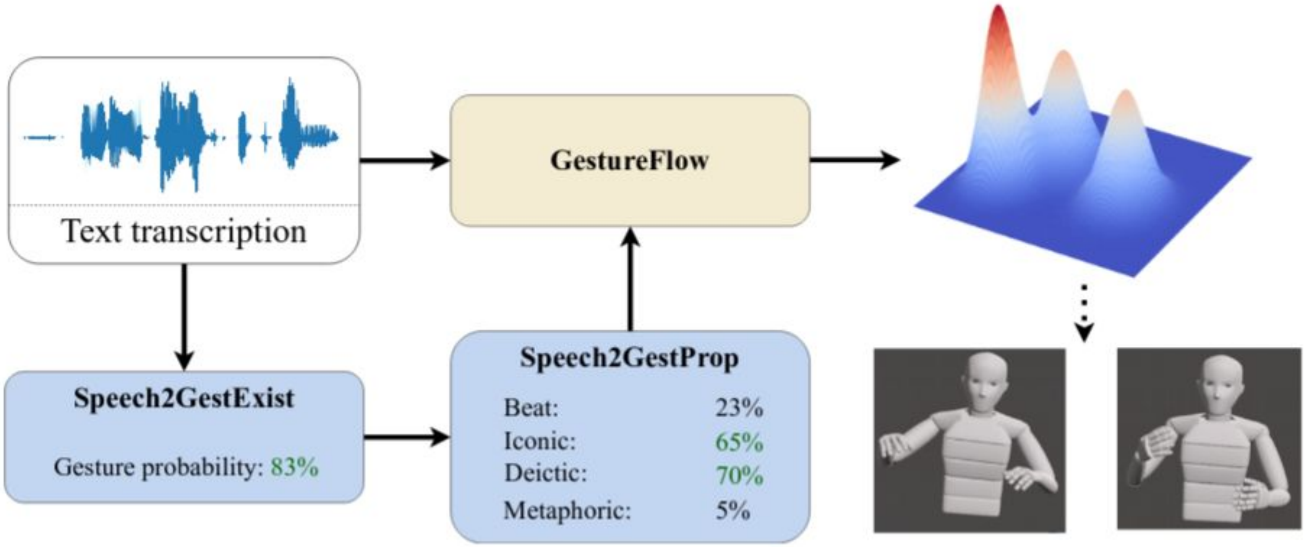
Speech2Prop results



Summary

- We presented a novel gesture-generation framework to create gestures that are communicative and natural at the same time.
- Our method first predicts if a gesture is needed and what kind of gesture is needed. Once this prediction is made, it is used to condition the gesture-generation model.
- Our gesture-property prediction results are promising and indicate that the proposed approach is feasible.

Conclusions



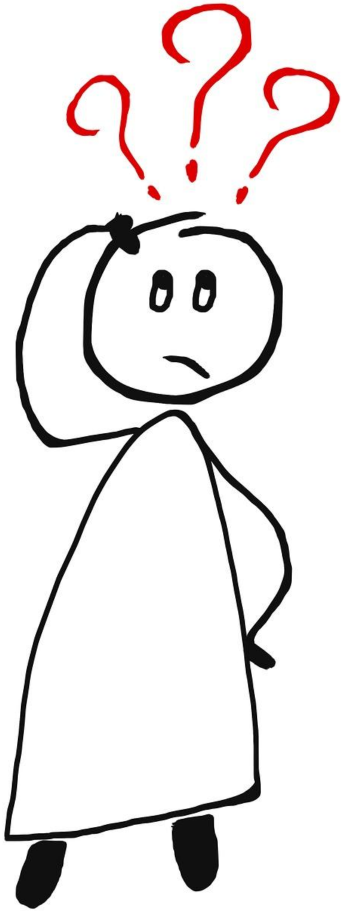
When learning something new we should not forget the old

Speech2Properties2Gestures: Gesture-Property Prediction as a Tool for Generating Representational Gestures from Speech

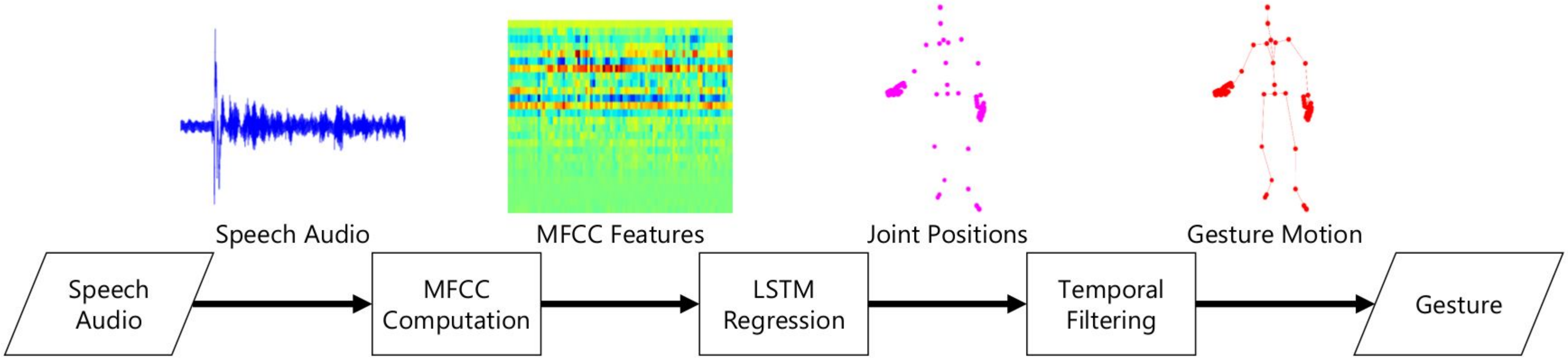
Taras Kucherenko Rajmund Nagy Patrik Jonell
Michael Neff Hedvig Kjellström Gustav Eje Henter

Our project page
with follow up work

Questions?



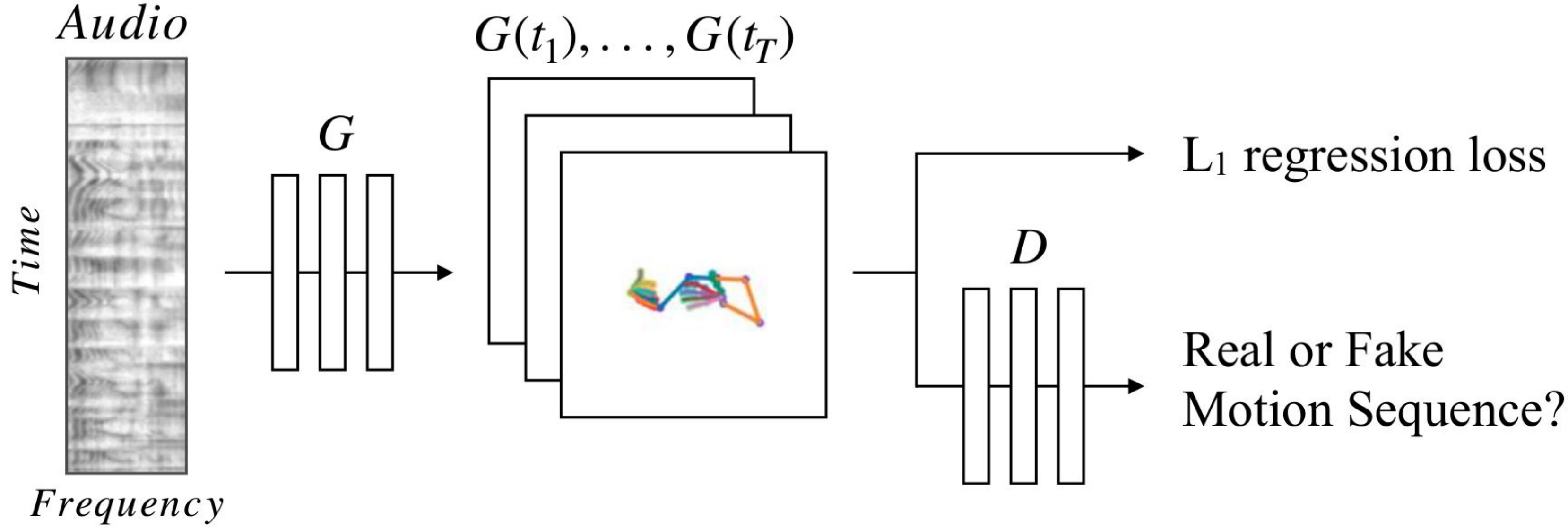
Recent related work



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

Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi "Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network." International Conference on Intelligent Virtual Agents. 2018.

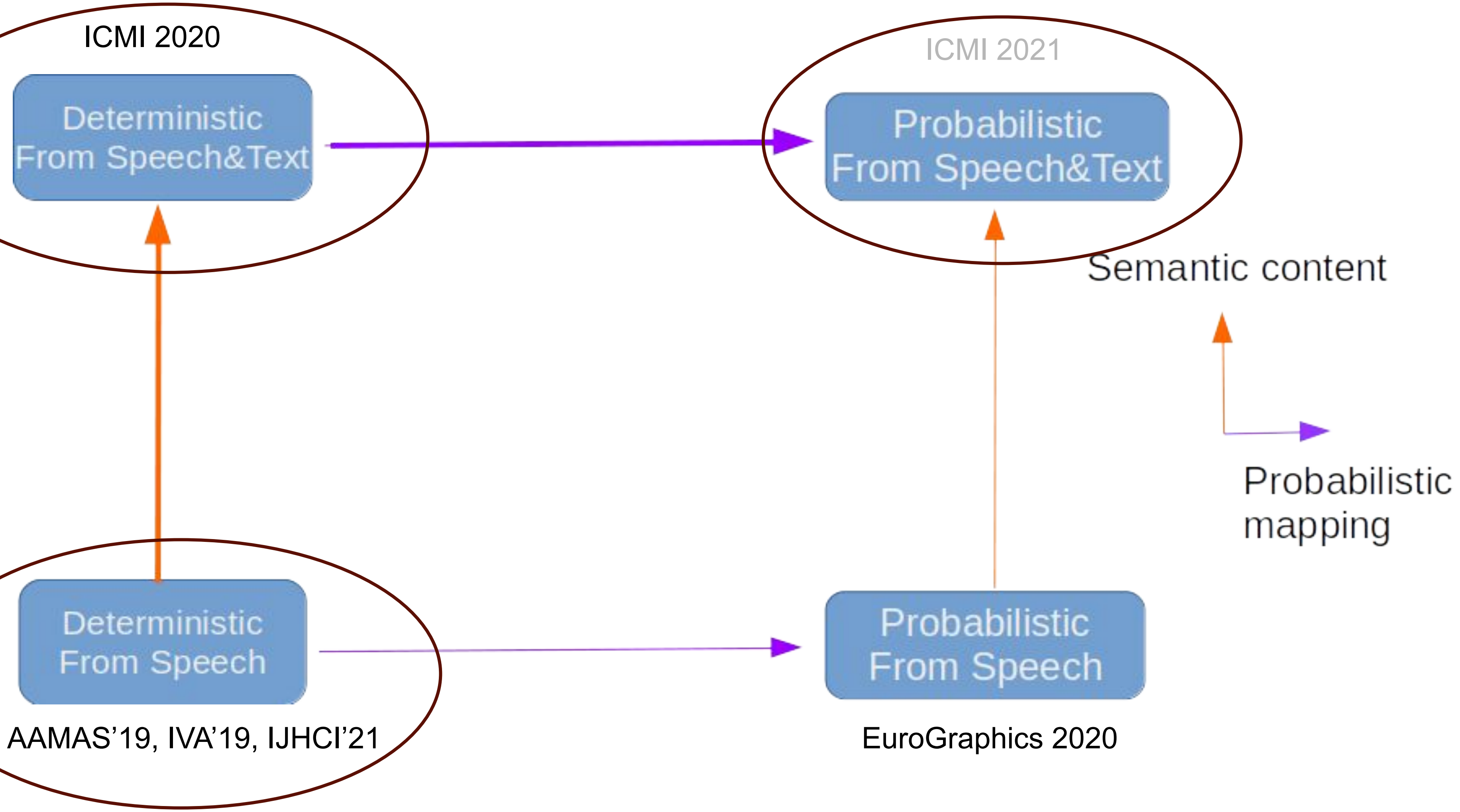
Recent related work



- Used Deep Learning
- GAN inspired loss
- Generated 2D motion

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, Jitendra Malik
"Learning Individual Styles of Conversational Gesture". CVPR. 2019

Approach to Gesture Generation


ICMI 2020

Deterministic From Speech&Text

ICMI 2021

Probabilistic From Speech&Text

Semantic content

Probabilistic mapping

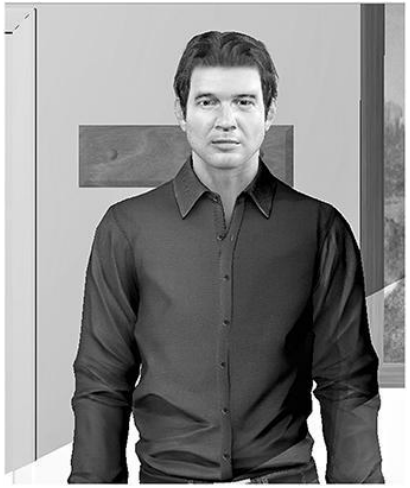
Deterministic From Speech

AAMAS’19, IVA’19, IJHCI’21

Probabilistic From Speech

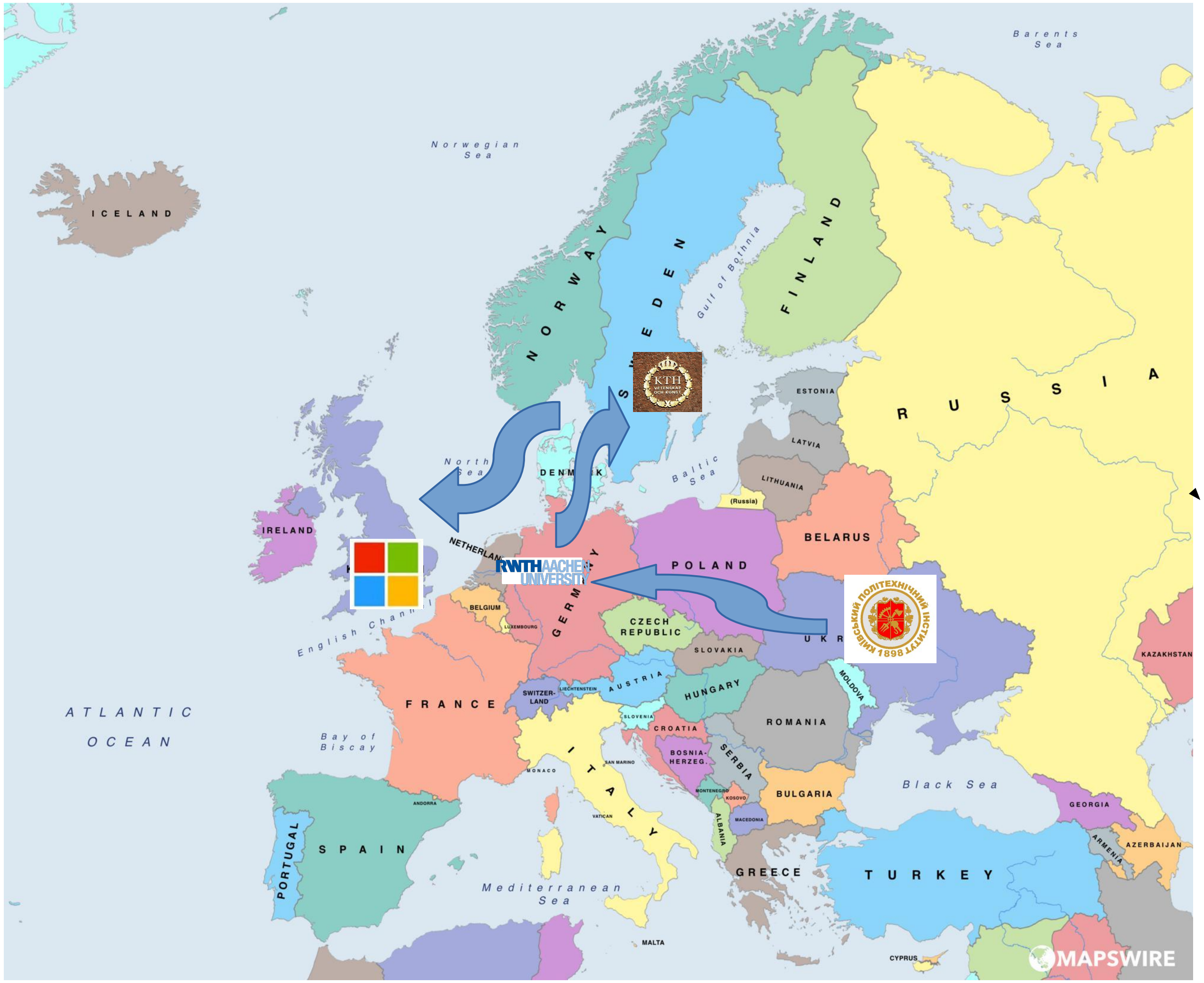
EuroGraphics 2020

Who are virtual agents?



www.splintercell.com

My journey


Barents Sea
Norwegian Sea
ICELAND
NORWAY
SWEDEN
FINLAND
Gulf of Bothnia
RUSSIA
KTH VETENSKAP OCH KONST
ESTONIA
LATVIA
LITHUANIA
(Russia)
North Sea
DENMARK
Baltic Sea
BELARUS
IRELAND
NETHERLANDS
RWTH AACHEN UNIVERSITY
GERMANY
POLAND
BELGIUM
LUXEMBOURG
CZECH REPUBLIC
SLOVAKIA
UKRAINE
КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ 1898
MOLDOVA
KAZAKHSTAN
English Channel
FRANCE
SWITZERLAND
LIECHTENSTEIN
AUSTRIA
HUNGARY
ROMANIA
SLOVENIA
CROATIA
BOSNIA-HERZEG.
SERBIA
ATLANTIC OCEAN
Bay of Biscay
MONACO
SAN MARINO
ITALY
VATICAN
MONTENEGRO
KOSOVO
ALBANIA
MACEDONIA
BULGARIA
Black Sea
GEORGIA
ARMENIA
AZERBAIJAN
ANDORRA
PORTUGAL
SPAIN
Mediterranean Sea
GREECE
TURKEY
MALTA
CYPRUS
MAPSWIRE

Body language

Big part of human communication is non-verbal

Importance of body language

- People read and interpret robots’ non-verbal cues, similarly to how people read non-verbal cues from each other

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In International Conference on Intelligent Robots and Systems (pp. 708–713).

- Interactions with virtual agents have shown to be more engaging when the agent’s verbal behavior is accompanied by appropriate nonverbal behavior

Salem, M., Rohlfing, K., Kopp, S., & Joublin, F. (2011, July). A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In 2011 Ro-Man (pp. 247-252). IEEE.

- Equipping robots with such non-verbal behaviors have also shown to positively affect people’s perception of the robot

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. International Journal of Social Robotics, 5 (3), 313–323.

Outline

1. Motivation
2. Related Work
3. Dataset
4. System
5. Results
6. Conclusions

Importance of body language

- We convey plenty of information using non-verbal behavior, such as intent, emotional state, and attitude

R. M. Krauss, Y. Chen, and P. Chawla, “Nonverbal behavior and nonverbalcommunication: What do conversational hand gestures tell us?,” in Advances in Experimental Social Psychology, vol. 28, pp. 389–450, 1996.

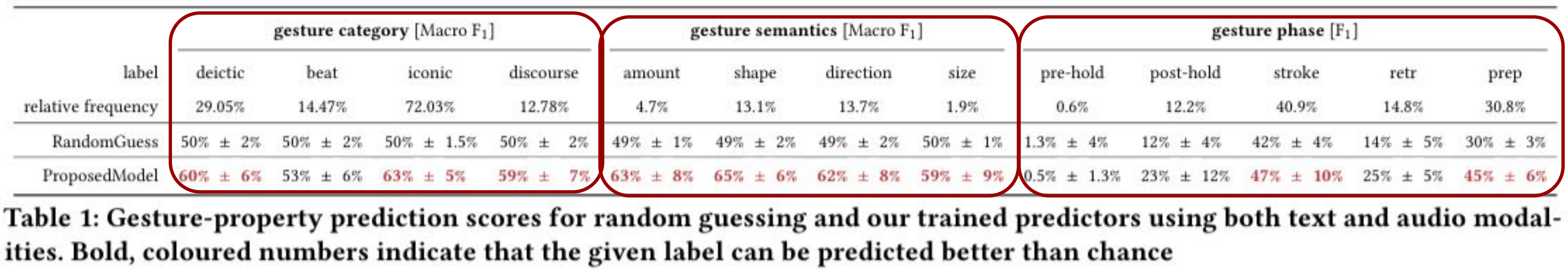
- Around 90% of spoken utterances in descriptive discourse are accompanied by gestures

S. Nobe, “Where do most spontaneous representational gestures actually occur with respect to speech,” Language and gesture, vol. 2, p. 186, 2000.

- Co-speech gestures can accompany the content of the speech – what is being said – on all levels, from partial word meanings to situation descriptions

S. Kopp, H. Rieser, I. Wachsmuth, K. Bergmann, and A. Lücking, “Speech-gesture alignment,” in Proceedings of the Conference of the International Society for Gesture Studies, 2007.

Speech2Prop results



	gesture category [Macro F_1]				**gesture semantics** [Macro F_1]				**gesture phase** [F_1]				
label	deictic	beat	iconic	discourse	amount	shape	direction	size	pre-hold	post-hold	stroke	retr	prep
relative frequency	29.05%	14.47%	72.03%	12.78%	4.7%	13.1%	13.7%	1.9%	0.6%	12.2%	40.9%	14.8%	30.8%
RandomGuess	50% ± 2%	50% ± 2%	50% ± 1.5%	50% ± 2%	49% ± 1%	49% ± 2%	49% ± 2%	50% ± 1%	1.3% ± 4%	12% ± 4%	42% ± 4%	14% ± 5%	30% ± 3%
ProposedModel	**60% ± 6%**	53% ± 6%	**63% ± 5%**	**59% ± 7%**	**63% ± 8%**	**65% ± 6%**	**62% ± 8%**	**59% ± 9%**	0.5% ± 1.3%	23% ± 12%	**47% ± 10%**	25% ± 5%	**45% ± 6%**

Table 1: Gesture-property prediction scores for random guessing and our trained predictors using both text and audio modalities. Bold, coloured numbers indicate that the given label can be predicted better than chance