Utrecht University

Virtual Humans in Social XR

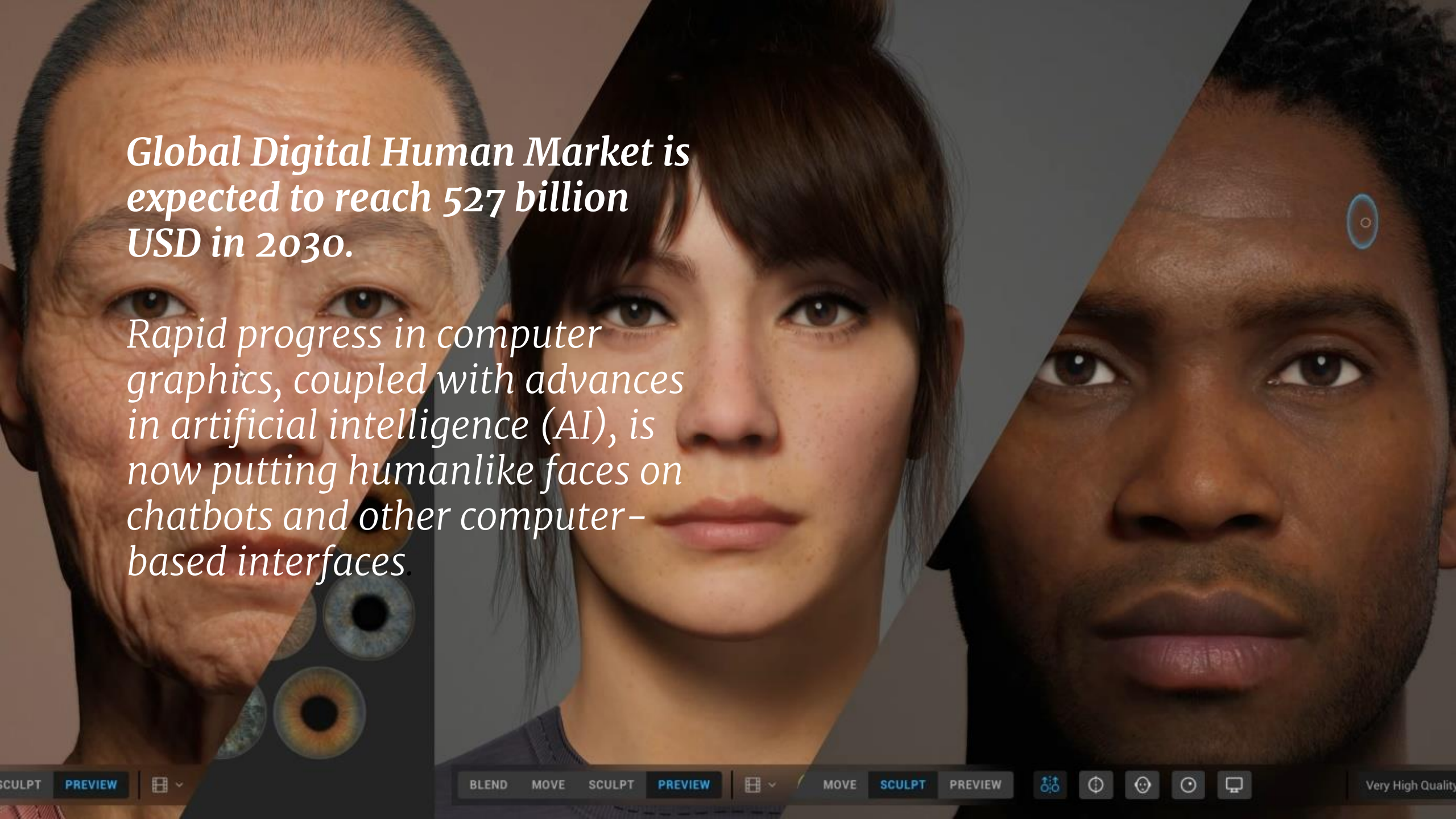
Zerrin Yumak

What are Digital (Virtual) Humans?

Metahumans Unreal Engine

They look like real humans and they have natural interaction capabilities similar to human beings

Sensing and analyzing users/environment Decision making and Dialogue Appearance and motion



Global Digital Human Market is expected to reach 527 billion USD in 2030.

Rapid progress in computer graphics, coupled with advances in artificial intelligence (AI), is now putting humanlike faces on chatbots and other computer-based interfaces.

Overview

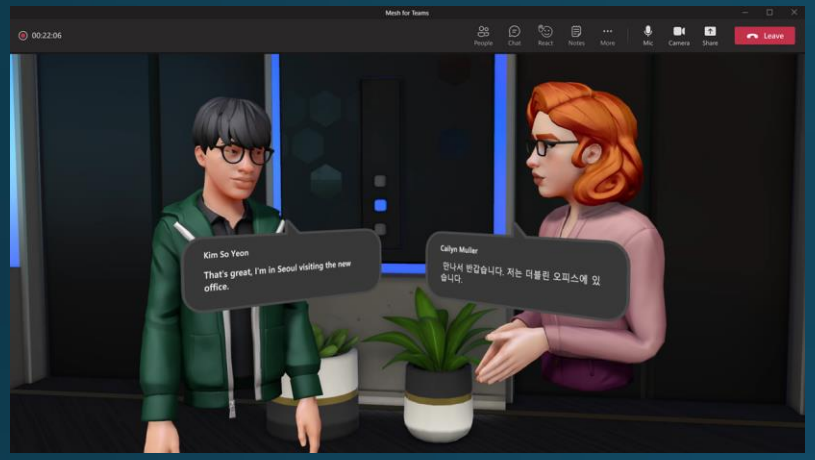
- Animation (3D Graphics)
  - How to model the characters and make them move?
    - Body and facial animation
  - Non-verbal behavior synthesis
    - Face, gaze, gestures
- Interaction (AI & HCI)
  - How to create conversational characters?
    - Sensing-decision making-acting loop
  - Social and emotional interaction
    - Modeling emotions & memory
    - Multi-party interaction
- Challenges & Conclusion

Realistic Virtual Humans in Games

HellBlade

Uncharted 4

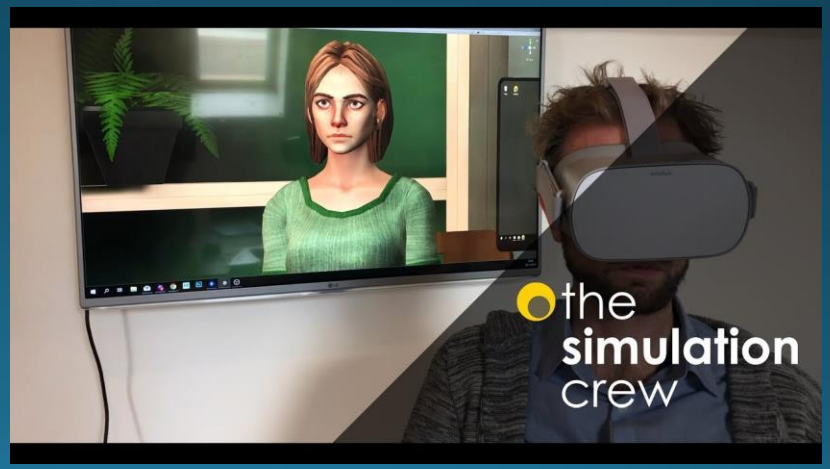
Virtual Humans in (Social) XR



Microsoft Teams and Mesh

Meta Horizon Workrooms

VR Training

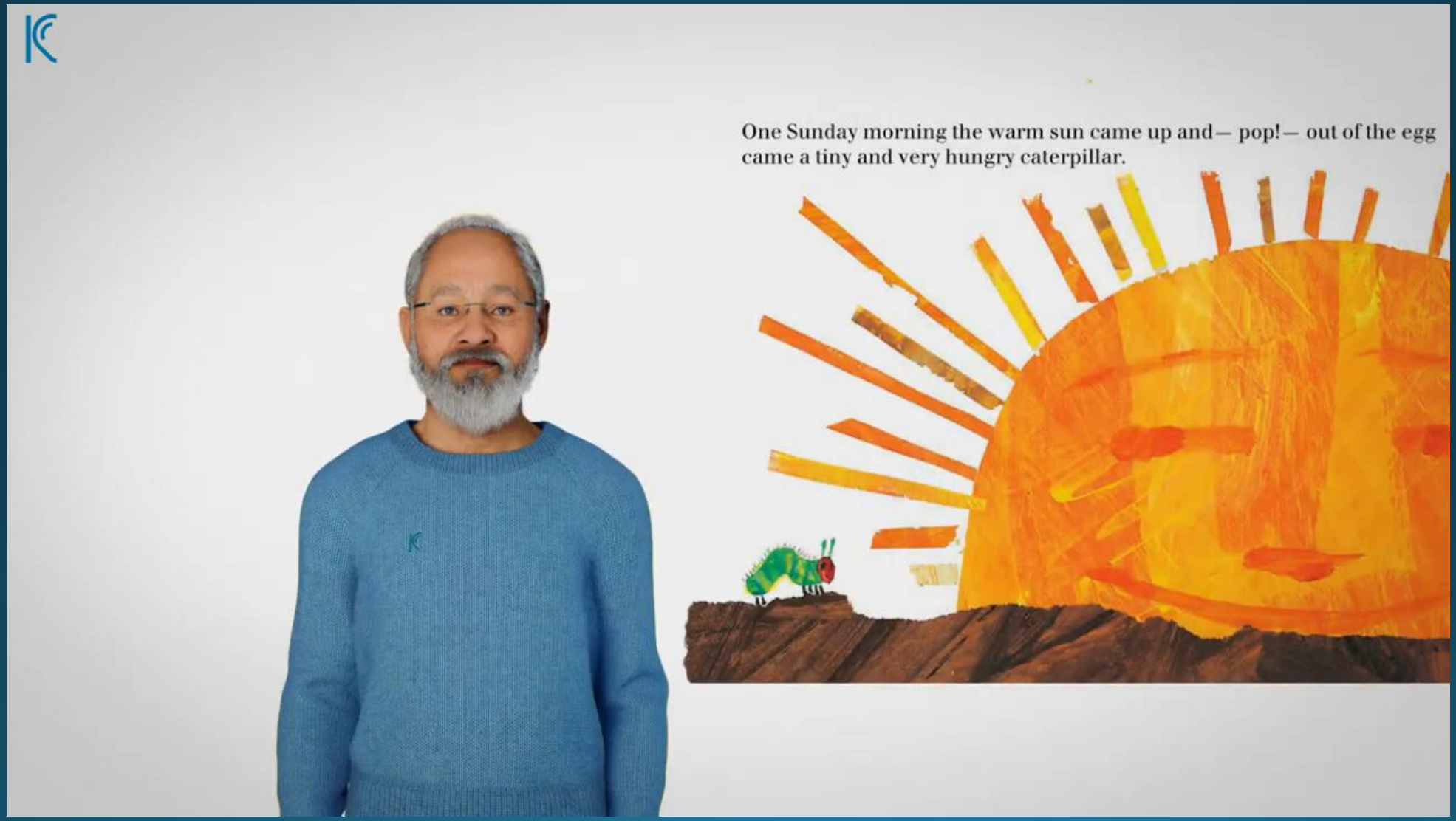


VR Training

Codec Avatars – Efficient Realistic Interaction in XR

S. Ma, et al., "**Pixel Codec Avatars**," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 64-73

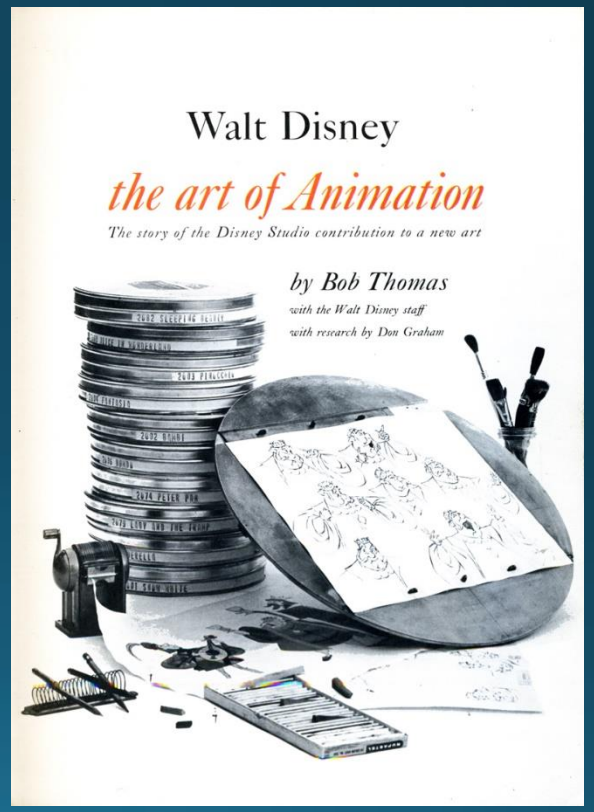
Sign Language Avatar



Alter Ego – TV Show

Animation

- Animate means "**to give life to**" according to Merriam-Webster
- Based on Latin word "**anima**", meaning "**breath**" or "**sprit**"
  - E.g. Animal
- **Animation** is the process of making the **illusion of motion** and **the illusion of change** by means of the **rapid display** of a sequence **of images** that minimally differ from each other.

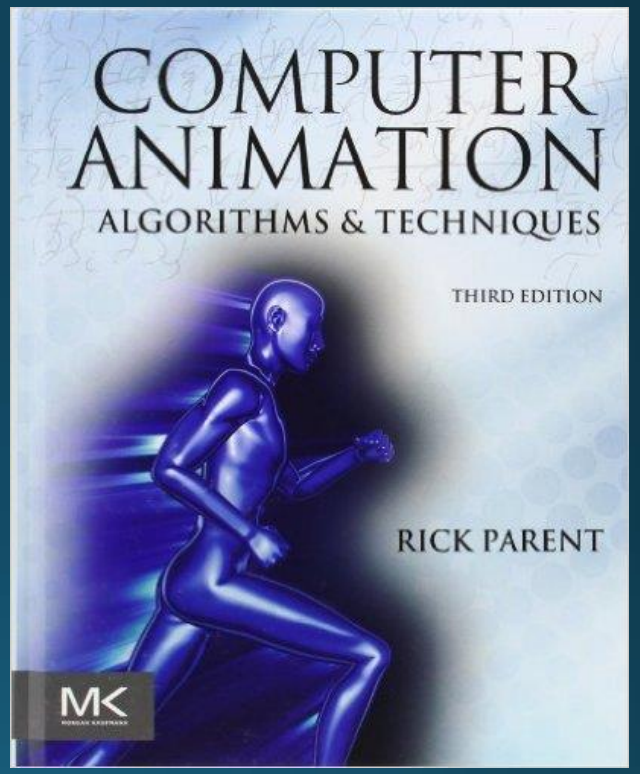


Pixar in a Box at Khan Academy - https://www.khanacademy.org/partner-content/pixar

Computer Animation

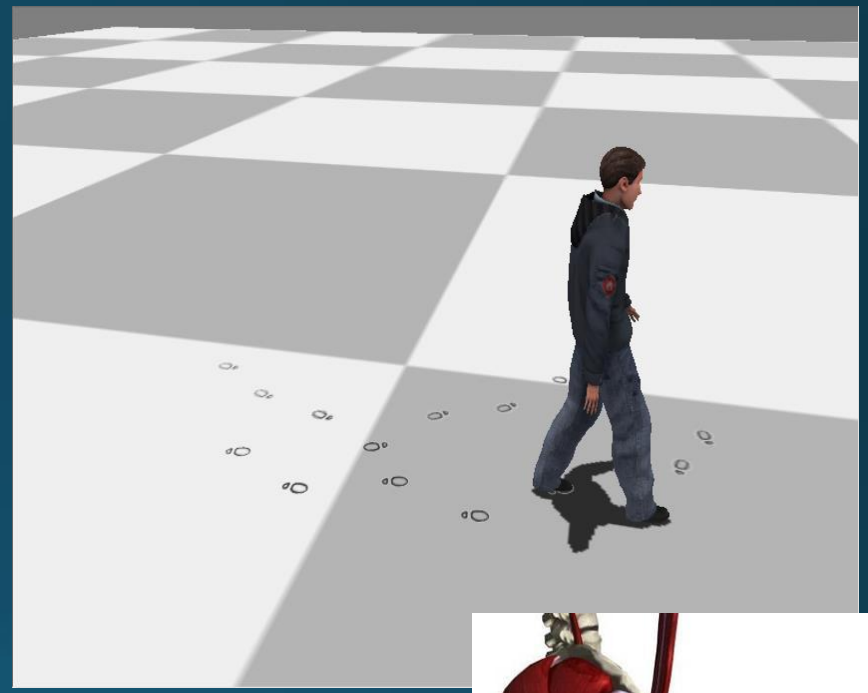
- **Computer-based computation** used in **producing images** intended to create the **perception of motion**
  - Algorithms and techniques that process 3D graphical data

- **Object's position and orientation**
  - But also shape, shading parameters, texture coordinates, light source parameters, camera parameters

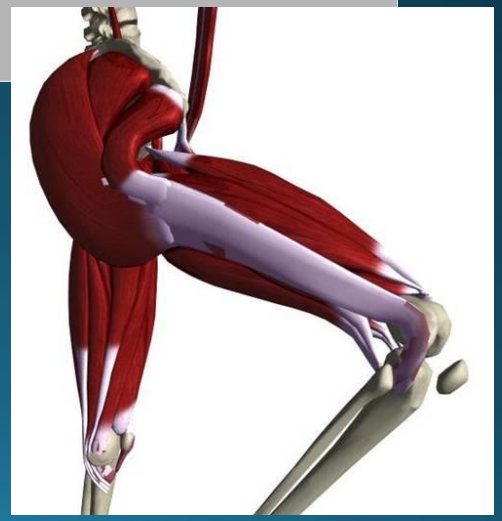


Three general approaches to computer animation

- **Artistic animation**
  - Animator crafts the motion
  - Keyframing and interpolation



- **Data-driven animation**
  - Digitizing live motion and applying to 3D objects
  - Motion capture and AI

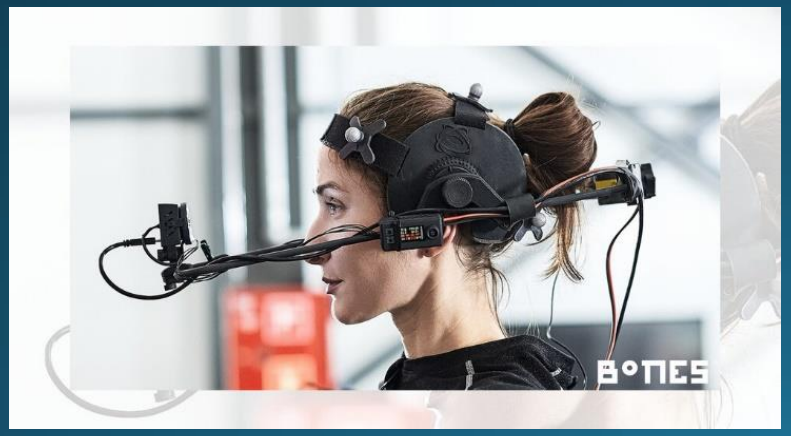


- **Procedural animation**
  - Computational models of motion
  - By setting initial conditions for physical and behavioral simulation

Van Welbergen, Herwin, et al. "Real time animation of virtual humans: a trade-off between naturalness and control." *Computer Graphics Forum*. Vol. 29. No. 8. Oxford, UK: Blackwell Publishing Ltd, 2010.

Motion Capture

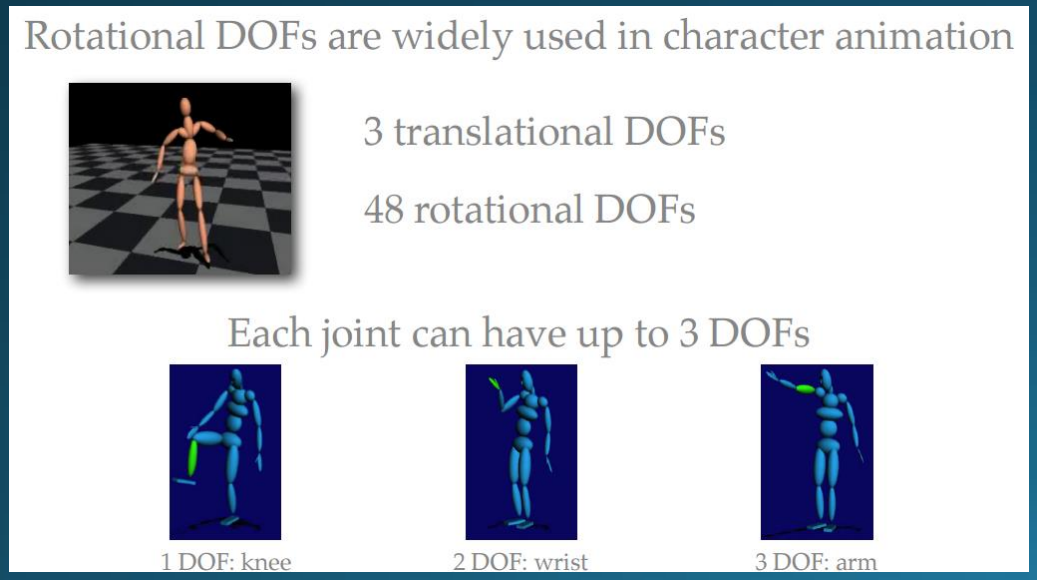
- Optical motion capture
  - Passive
  - Active
  - Markerless
- Inertial

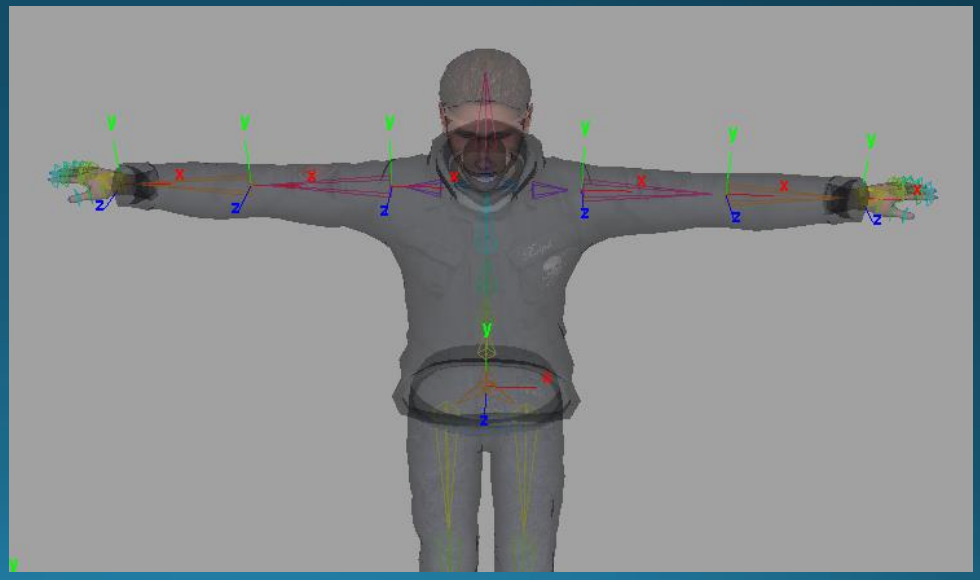


Representation of virtual humans

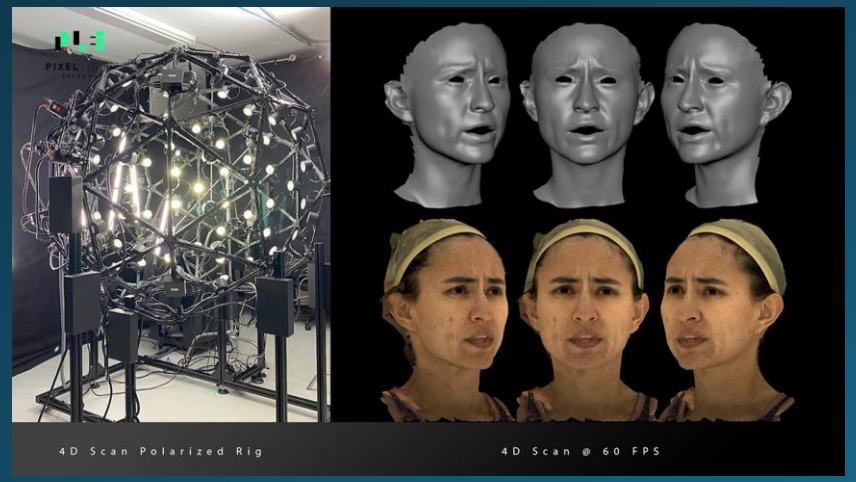
Skeletal model

- A virtual human is represented by a polyhedral model (or mesh)
- An underlying skeleton deforms this mesh
  - Joints, connected by bones
- A pose is defined by the rotations of the joints and the position of the root joint

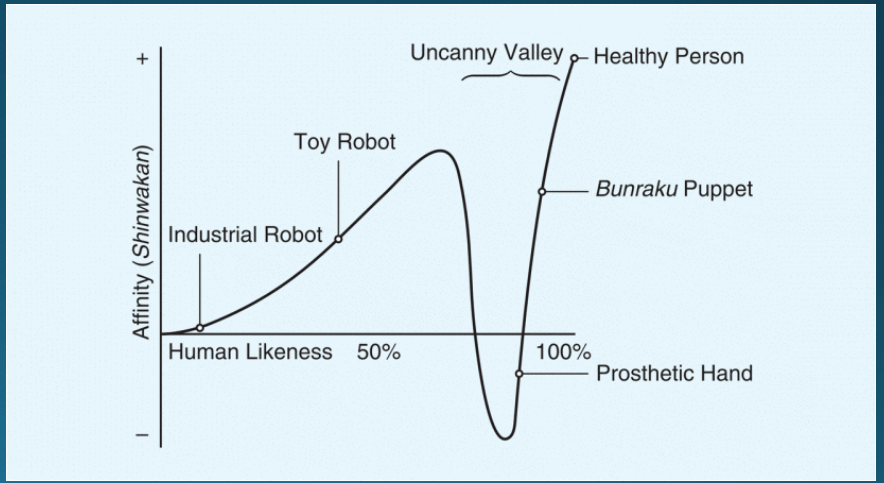




Facial animation



Reblika – Valentina 4D



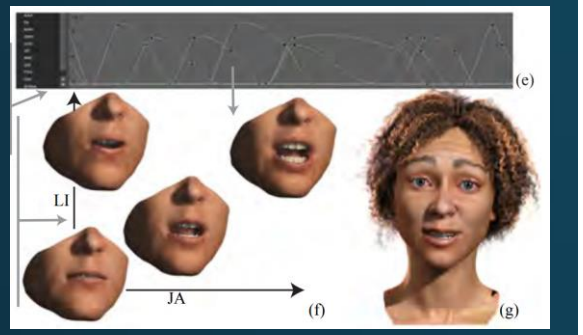
Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. **The uncanny valley**. IEEE Robotics and Automation Magazine 19 (2012), 98–100. Issue 2. https://doi.org/10. 1109/MRA.2012.2192811
Darragh Higgins, Donal Egan, Rebecca Fribourg, Benjamin R. Cowan, Rachel McDonnell . **Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny?** ACM Symposium on Applied Perception (SAP), Article No. 7, pp 1-5, 2021

Perception of Appearance and Animation

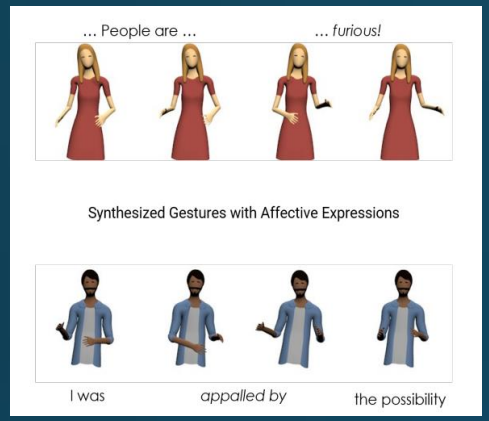
- Higher appearance realism and higher animation realism leads to higher social presence and higher attractiveness ratings
- Significant effects of animation realism on perceived realism and emotion intensity levels are found.
- ➢ **State-of-the-art photorealistic appearance lends itself better to highly realistic animations than lower appearance realism**
- ➢ **The study provides some evidence that photorealistic characters today don't trigger sense of uncanny anymore when matched with highly realistic animations**

N. Amadoue, K. Haque, Z. Yumak, **Effect of Appearance and Animation Realism on the Perception of Emotionally Expressive Virtual Humans**, ACM Intelligent Virtual Agents (IVA 2023)

Data-driven face and body animation

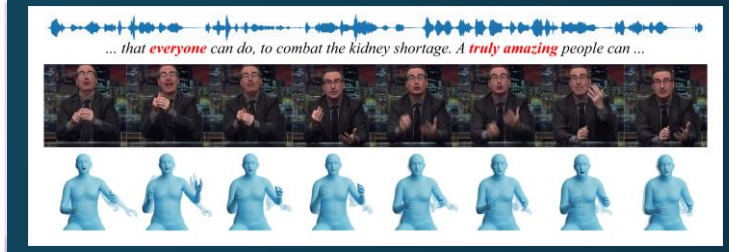


Edwards et al. 2016

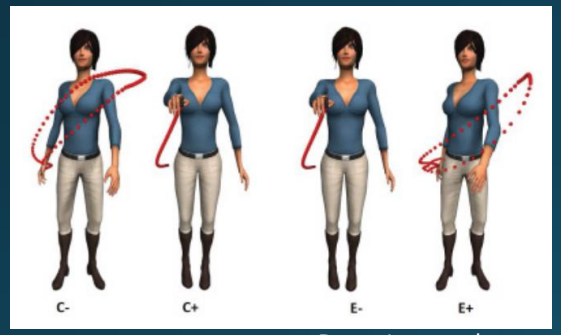


Bhattacharya et al. 2021

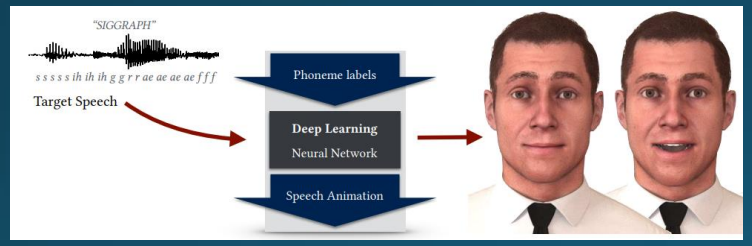
Alexanderson et al. 2020



Yi et al. 2023

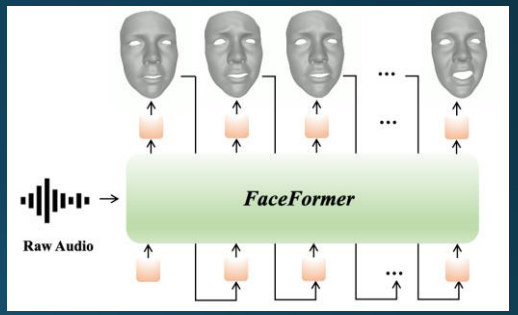


Durupinar et al. 2021

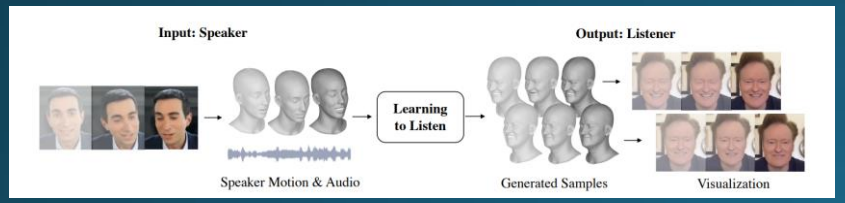


Taylor et al. 2017

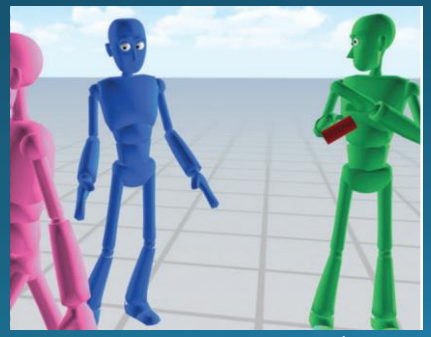
Ichim et al. 2015



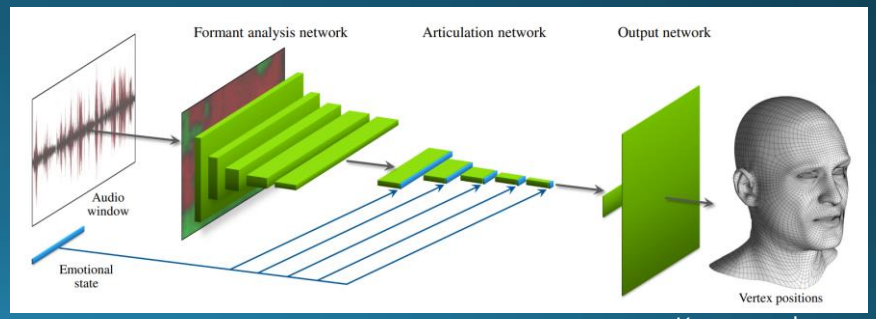
Fan et al. 2022



Ng et al. 2022

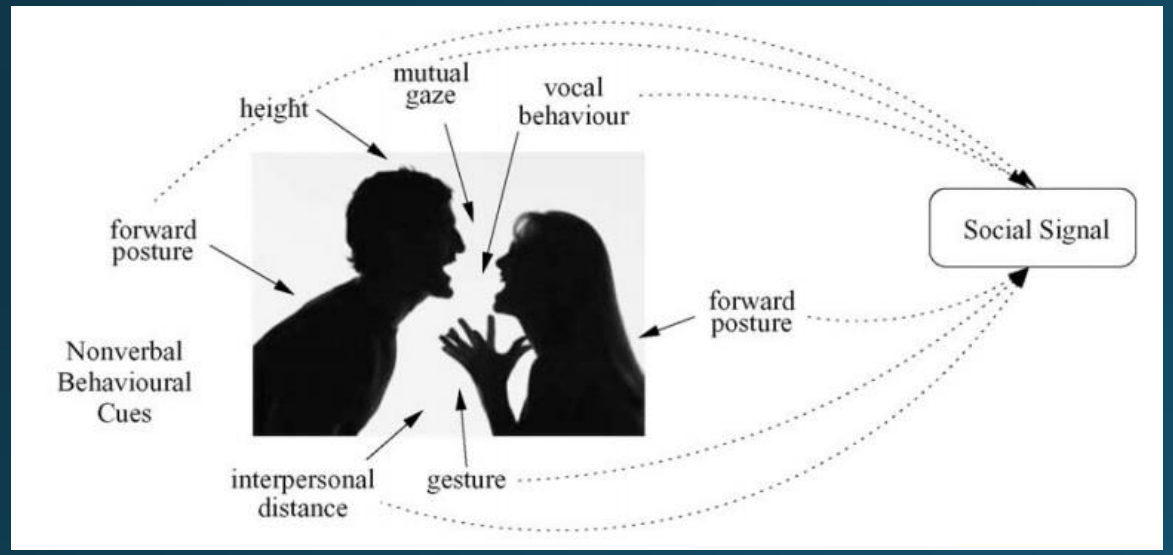


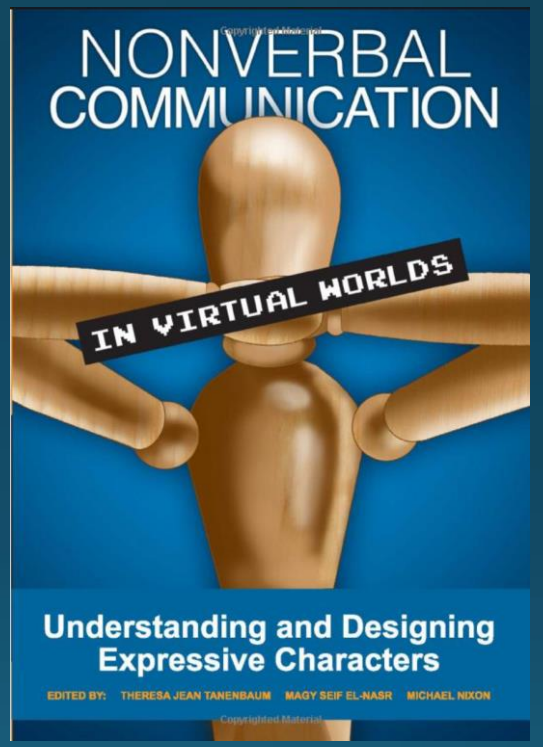
Pejsa et al. 2016

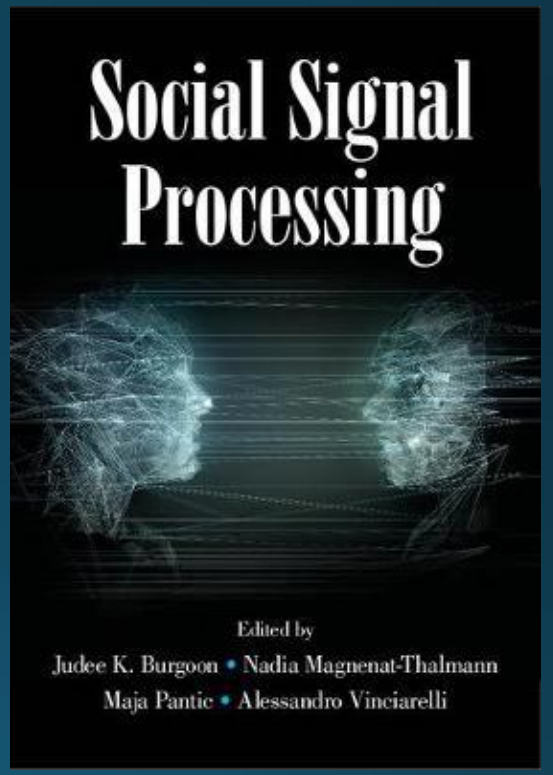


Karras et al. 2017

Non-verbal behaviors







A. Vinciarelli, M. Pantic, H. Bourlard. **Social Signal Processing: Survey of an Emerging Domain,** Image and Vision Computing Journal, Vol. 27, No. 12, pp. 1743-1759, 2009.
A. Beck, Z. Yumak, N. Magnenat-Thalmann. **Body movement generation for virtual characters and social robots**. Social Signal Processing, Cambridge University Press, 2016.

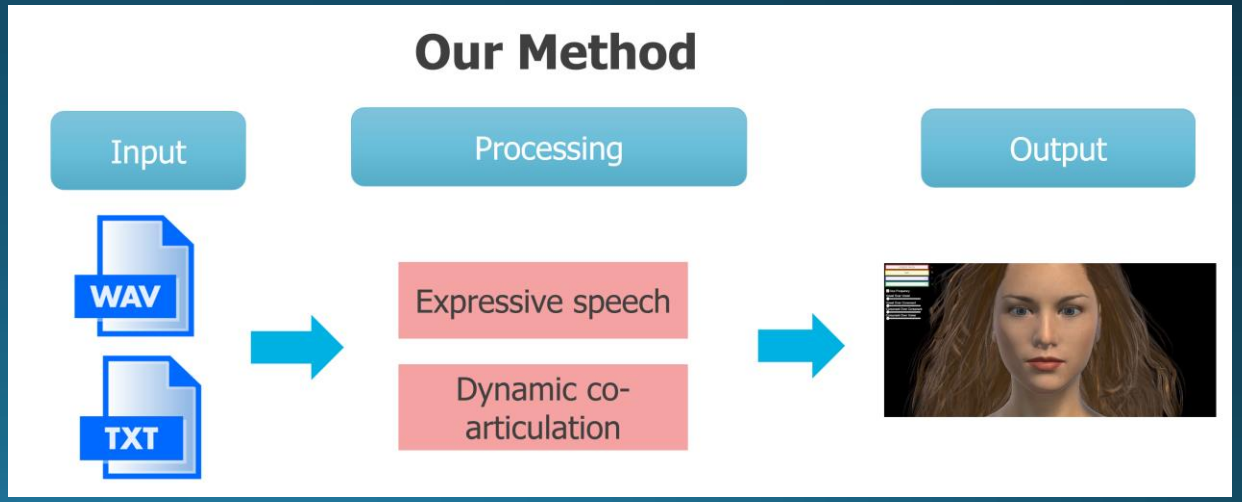
Procedural Speech Animation

- ✓ Control is **high priority**
- ✓ Naturalness is **desired**

Procedural approach taking into account the **effect of emotions on mouth movements**

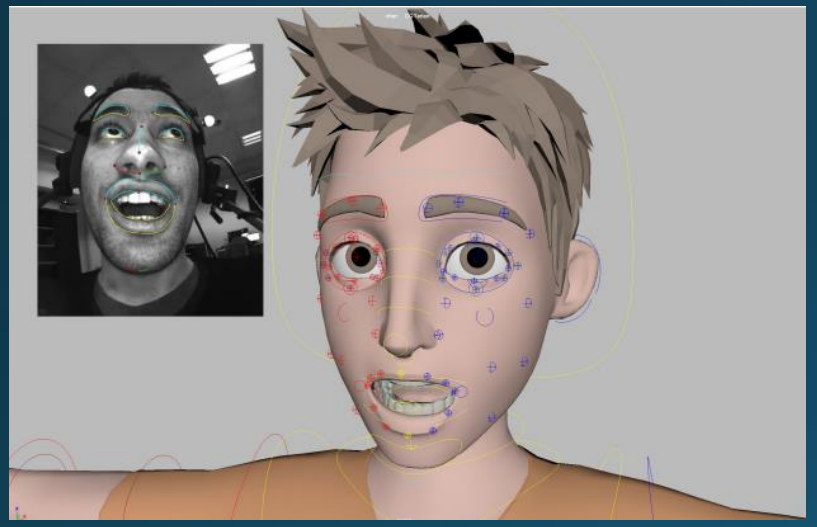
Comparison to FaceFX and RogoDigital

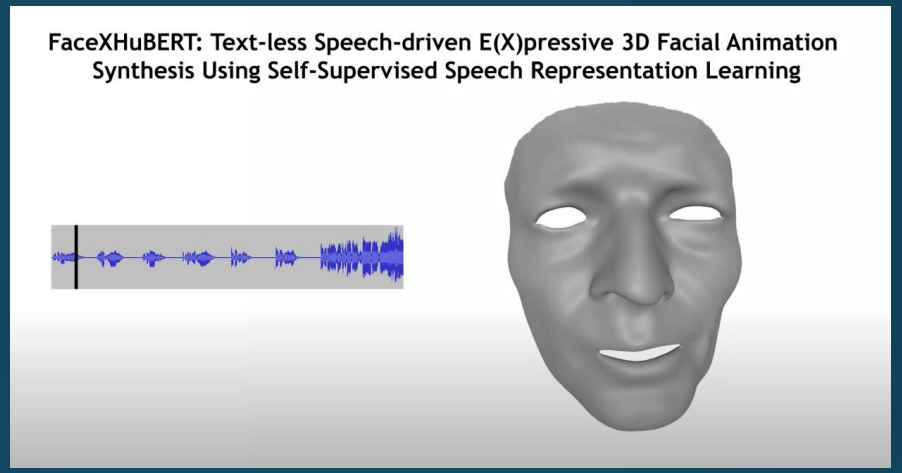
- ✓ **Ours (58%)** vs FaceFX (42%)
- ✓ **Ours (57%)** vs RogoDigital (43%)



C. Charalambous, Z. Yumak, A.F. van der Stappen. **Audio-driven emotional speech animation for interactive virtual characters**. Computer Animation and Virtual Worlds, June 2019

Data-driven Facial Animation





FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning, Kazi Injamamul Haque and Zerrin Yumak, 25th ACM International Conference on Multimodal Interaction, ACM ICMI 2023, Paris France

FaceDiffuser: Speech Driven 3D Facial Animation Synthesis using Diffusion, Stefan Stan, Kazi Injamamul Haque and Zerrin Yumak, ACM Siggraph Conference on Motion, Interaction and Games, MIG 2023, Rennes, France

Datasets

BIWI

VOCASET

Multiface

High dimensional
3D vertex

Low dimensional
parameterized data

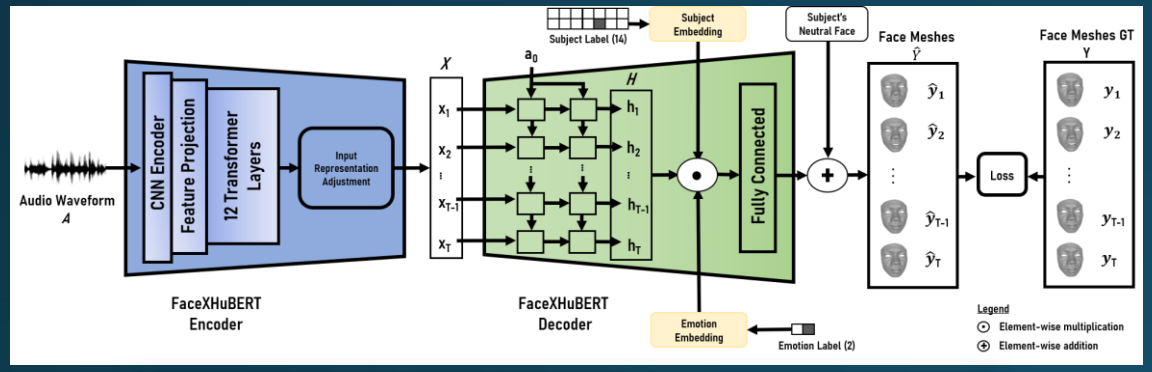
UUDaMM

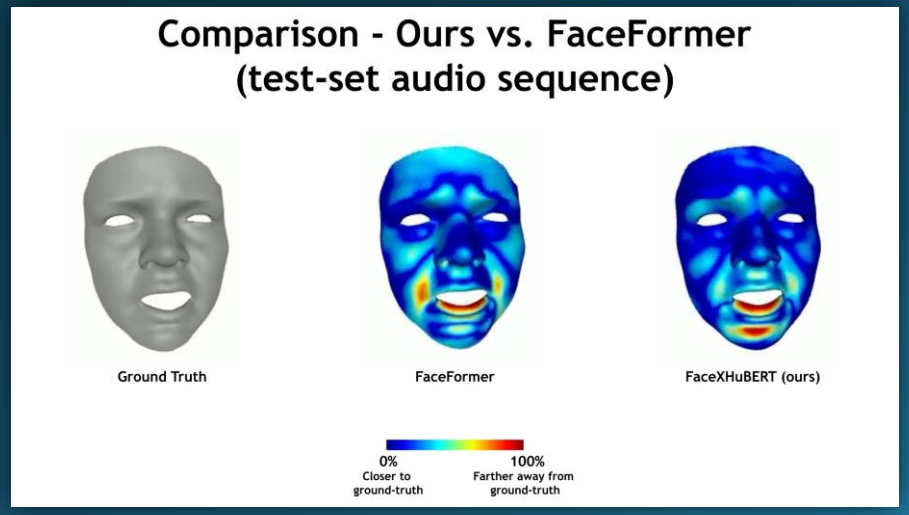
BEAT

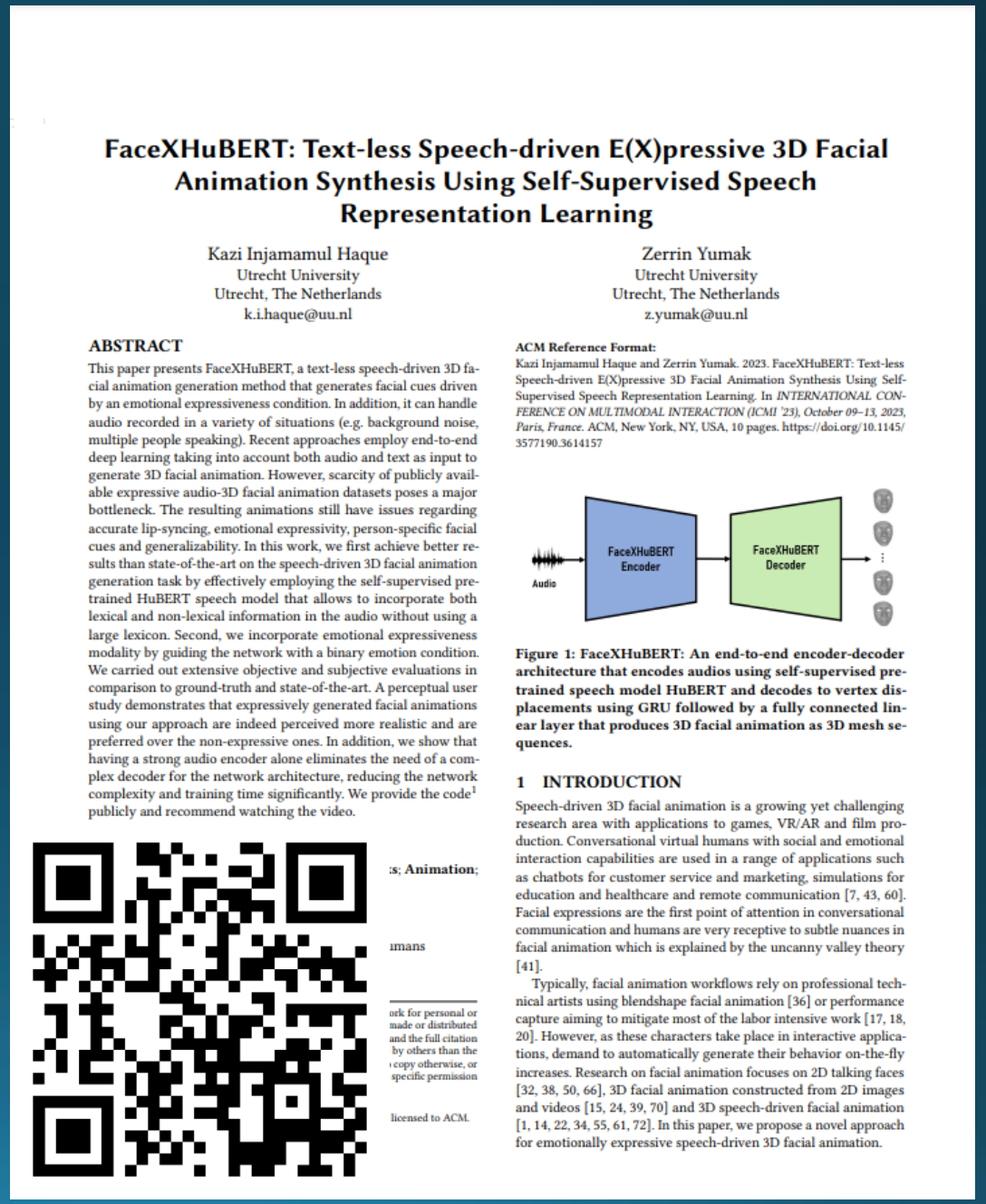
FaceXHuBERT

Outperforms state-of-the-art models: More natural animation with shorter training time

ACM ICMI 2023







FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning

Kazi Injamamul Haque
Utrecht University
Utrecht, The Netherlands
k.i.haque@uu.nl

Zerrin Yumak
Utrecht University
Utrecht, The Netherlands
z.yumak@uu.nl

ABSTRACT

This paper presents FaceXHuBERT, a text-less speech-driven 3D facial animation generation method that generates facial cues driven by an emotional expressiveness condition. In addition, it can handle audio recorded in a variety of situations (e.g. background noise, multiple people speaking). Recent approaches employ end-to-end deep learning taking into account both audio and text as input to generate 3D facial animation. However, scarcity of publicly available expressive audio-3D facial animation datasets poses a major bottleneck. The resulting animations still have issues regarding accurate lip-syncing, emotional expressivity, person-specific facial cues and generalizability. In this work, we first achieve better results than state-of-the-art on the speech-driven 3D facial animation generation task by effectively employing the self-supervised pre-trained HuBERT speech model that allows to incorporate both lexical and non-lexical information in the audio without using a large lexicon. Second, we incorporate emotional expressiveness modality by guiding the network with a binary emotion condition. We carried out extensive objective and subjective evaluations in comparison to ground-truth and state-of-the-art. A perceptual user study demonstrates that expressively generated facial animations using our approach are indeed perceived more realistic and are preferred over the non-expressive ones. In addition, we show that having a strong audio encoder alone eliminates the need of a complex decoder for the network architecture, reducing the network complexity and training time significantly. We provide the code[1] publicly and recommend watching the video.

ACM Reference Format:
Kazi Injamamul Haque and Zerrin Yumak. 2023. FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614157

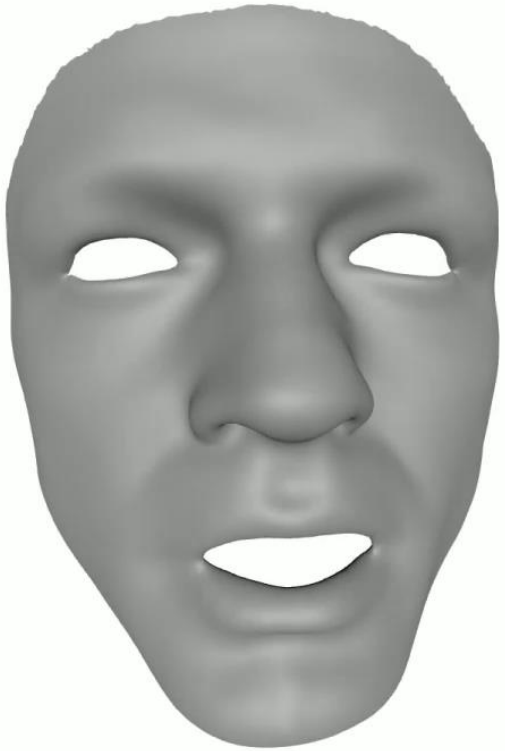
Figure 1: FaceXHuBERT: An end-to-end encoder-decoder architecture that encodes audios using self-supervised pre-trained speech model HuBERT and decodes to vertex displacements using GRU followed by a fully connected linear layer that produces 3D facial animation as 3D mesh sequences.

1 INTRODUCTION

Speech-driven 3D facial animation is a growing yet challenging research area with applications to games, VR/AR and film production. Conversational virtual humans with social and emotional interaction capabilities are used in a range of applications such as chatbots for customer service and marketing, simulations for education and healthcare and remote communication [7, 43, 60]. Facial expressions are the first point of attention in conversational communication and humans are very receptive to subtle nuances in facial animation which is explained by the uncanny valley theory [41].

Typically, facial animation workflows rely on professional technical artists using blendshape facial animation [36] or performance capture aiming to mitigate most of the labor intensive work [17, 18, 20]. However, as these characters take place in interactive applications, demand to automatically generate their behavior on-the-fly increases. Research on facial animation focuses on 2D talking faces [32, 38, 50, 66], 3D facial animation constructed from 2D images and videos [15, 24, 39, 70] and 3D speech-driven facial animation [1, 14, 22, 34, 55, 61, 72]. In this paper, we propose a novel approach for emotionally expressive speech-driven 3D facial animation.

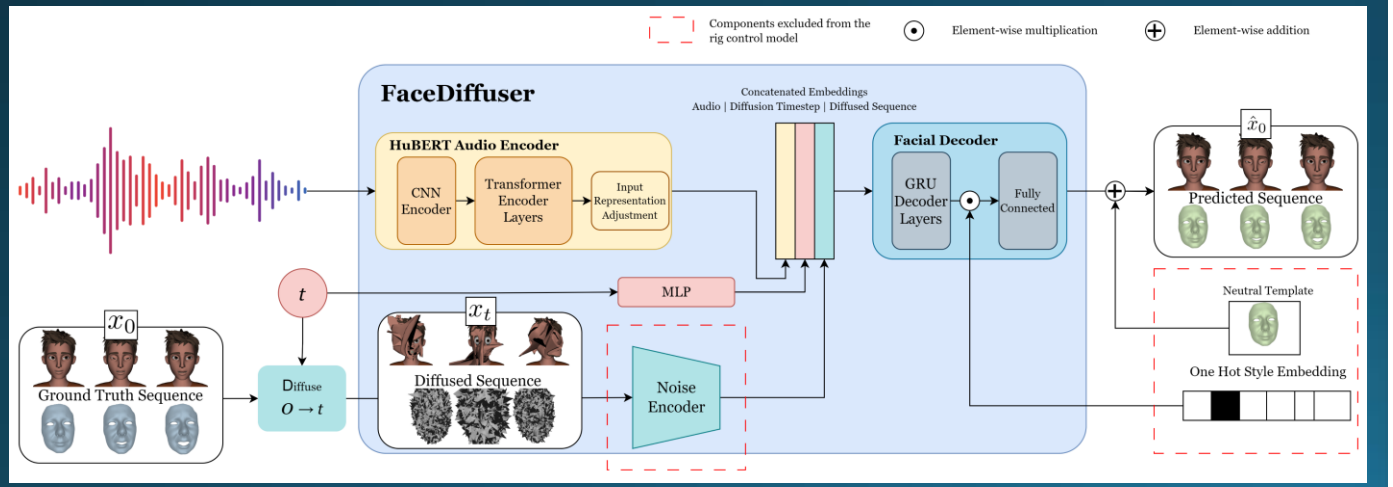
Qualitative Analysis



FaceDiffuser

First paper using diffusion models (non-determinism) for 3D facial animation for both vertex based and rigged characters

ACM Siggraph Motion, Interaction and Games 2023



FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion

Stefan Stan
Utrecht University
Utrecht, The Netherlands
st.stan96@gmail.com

Kazi Injamamul Haque
Utrecht University
Utrecht, The Netherlands
k.i.haque@uu.nl

Zerrin Yumak
Utrecht University
Utrecht, The Netherlands
z.yumak@uu.nl

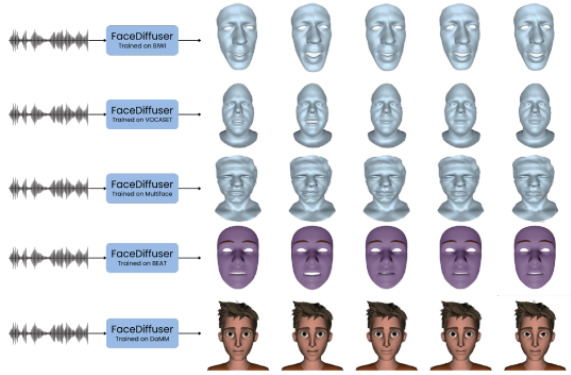


Figure 1: We present FaceDiffuser, an end-to-end non-deterministic neural network architecture for speech-driven 3D facial animation synthesis. Our proposed approach produces realistic and diverse animation sequences and is generalizable to both temporal 3D vertex based mesh animation datasets (top 3 rows) and temporal blendshape based datasets (bottom 2 rows).

ABSTRACT

Speech-driven 3D facial animation synthesis has been a challenging task both in industry and research. Recent methods mostly focus on deterministic deep learning methods meaning that given a speech input, the output is always the same. However, in reality, the non-verbal facial cues that reside throughout the face are non-deterministic in nature. In addition, majority of the approaches focus on 3D vertex based datasets and methods that are compatible with characters is fuser, a non-driven facial blendshape chnique and HuBERT to encode the audio input. To the best of our knowledge, we are the first to employ the diffusion method for the task of speech-driven 3D facial animation synthesis. We have run extensive objective and subjective analyses and show that our approach achieves better or comparable results in comparison to the state-of-the-art methods. We also introduce a new in-house dataset that is based on a blendshape based rigged character. The code and the dataset will be publicly available on the project page[1].

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Animation**; • **Human-centered computing** → *User studies*.

KEYWORDS

facial animation synthesis, deep learning, virtual humans, mesh animation, blendshape animation

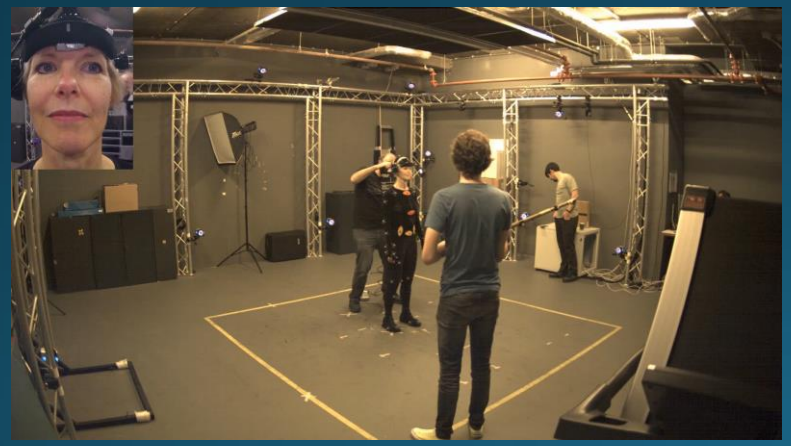
ACM Reference Format:
Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3623264.3624447

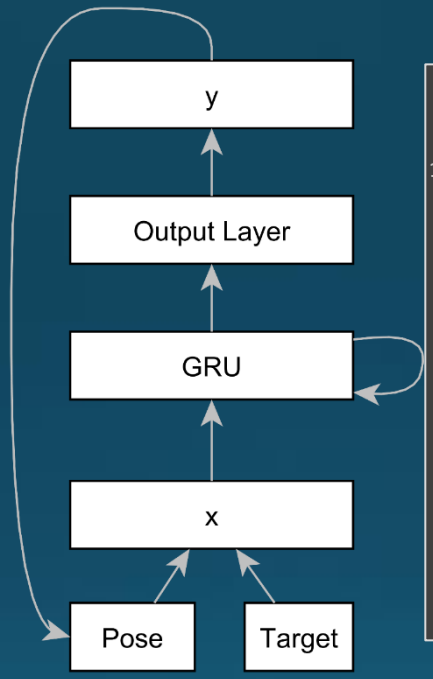
[1]https://uuembodiedsocialai.github.io/FaceDiffuser/

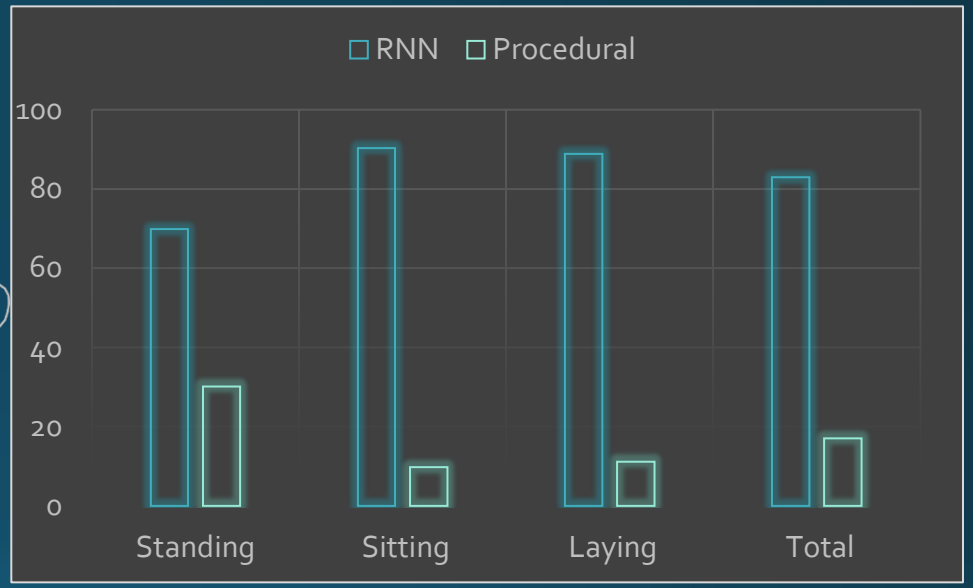
Results

FaceDiffuser
Trained on BIWI

Gaze Animation

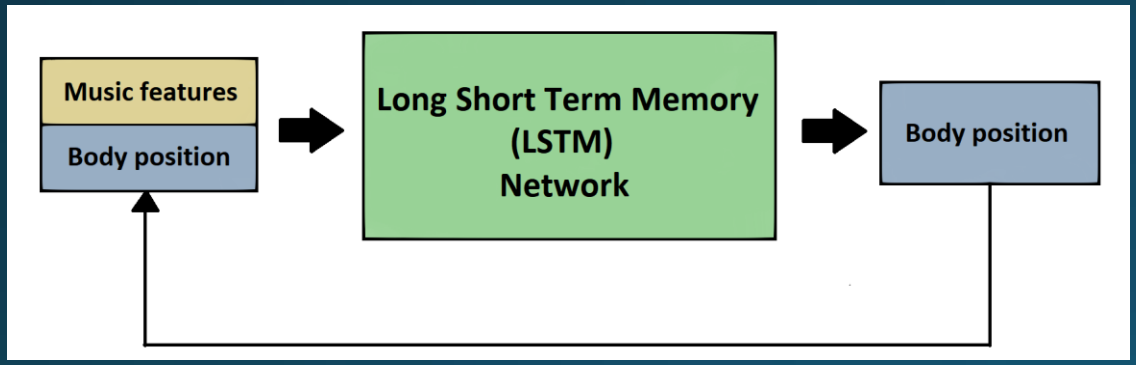






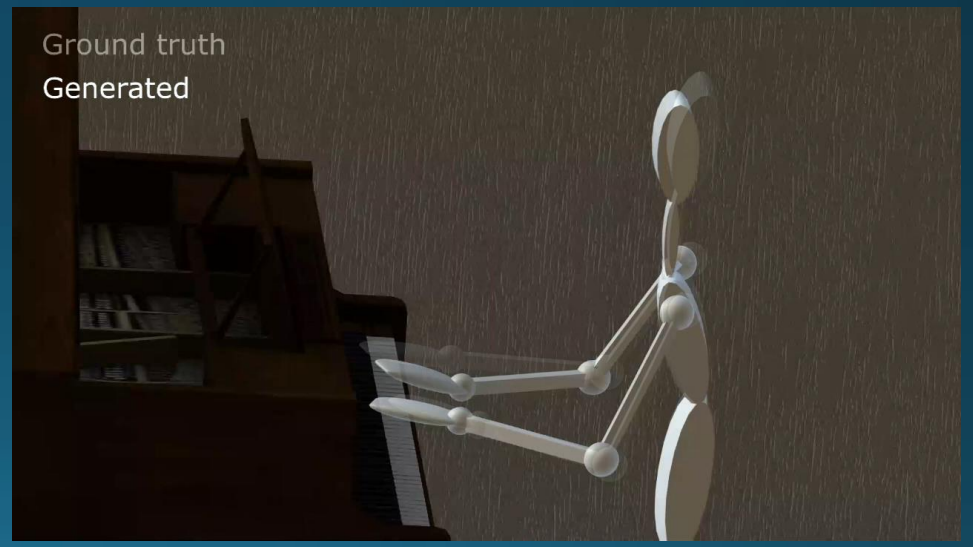
A. Klein, Z. Yumak, A. Beij and A. F. van der Stappen. **Data-driven Gaze Animation using Recurrent Neural Networks**. ACM Siggraph Conference on Motion, Interaction and Games, October 2019 **(Best Student Paper Award)**

Music-driven Expressive Gestures



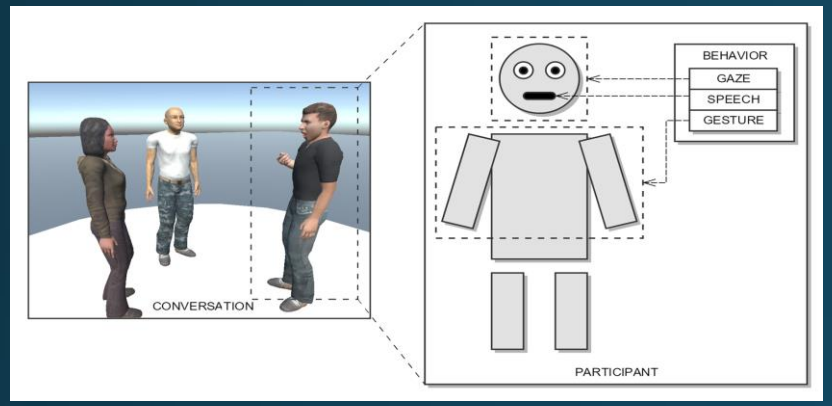
Features/Metric	MSE	#epochs	APE		Acceleration		Jerk	
			μ	σ	μ	σ	μ	σ
Pitch	2763.754	37	0.15126	0.03007	-2.19118	0.90039	18.59880	10.68545
Pitch+Beat	2837.244	22	0.15226	0.03064	-2.18774	0.89687	20.51324	14.38038
Pitch+RMS	2478.616	50	0.14495	0.04101	-2.13702	0.87939	11.22615	10.39768
Pitch+Beat+RMS	2675.961	35	0.14857	0.03512	-2.16969	0.87687	7.99591	9.56305
MFCC	2111.649	50	0.11749	0.03671	-2.14330	0.88659	12.32060	12.52234
All	2153.945	50	0.11790	0.03481	-2.12027	0.88340	1.18094	7.19398
Ground Truth	-	-	-	-	-2.12803	0.74442	2.62996	3.07753

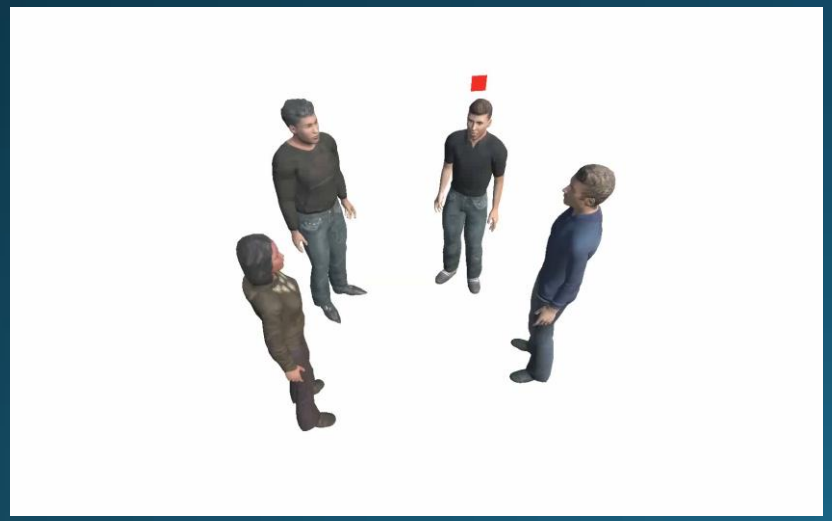
Table 1: The MSE, APE, acceleration and jerk of each condition over all generated animations.

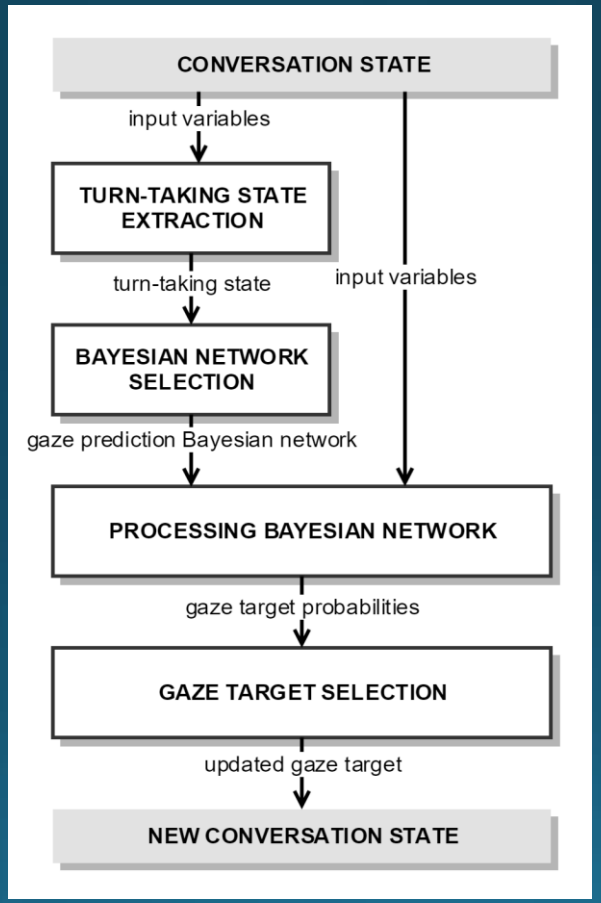


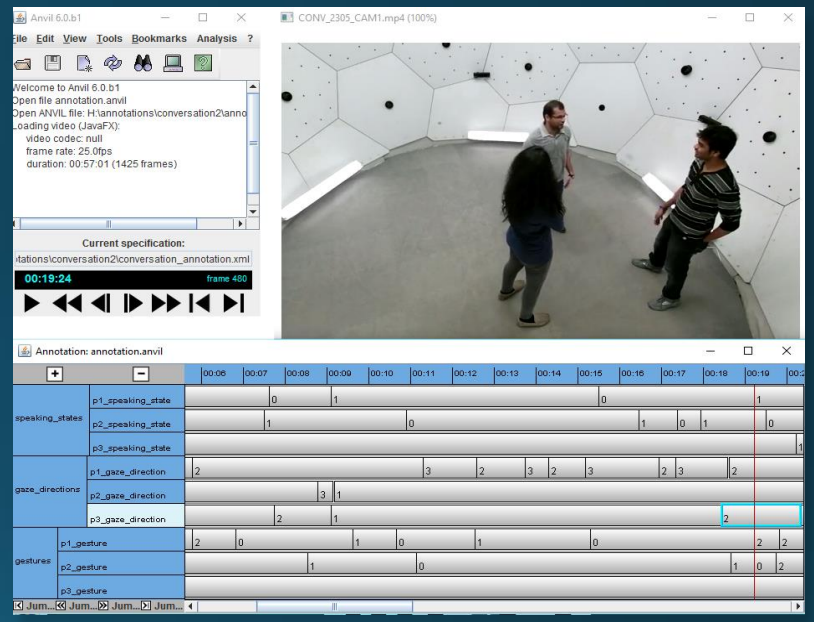
A. Bogaers, Z. Yumak, A. Volk. **Music-Driven Animation Generation of Expressive Musical Gestures**. ACM International Conference on Multimodal Interaction (ICMI 2020)

Non-verbal Behavior Generation during Group Conversations



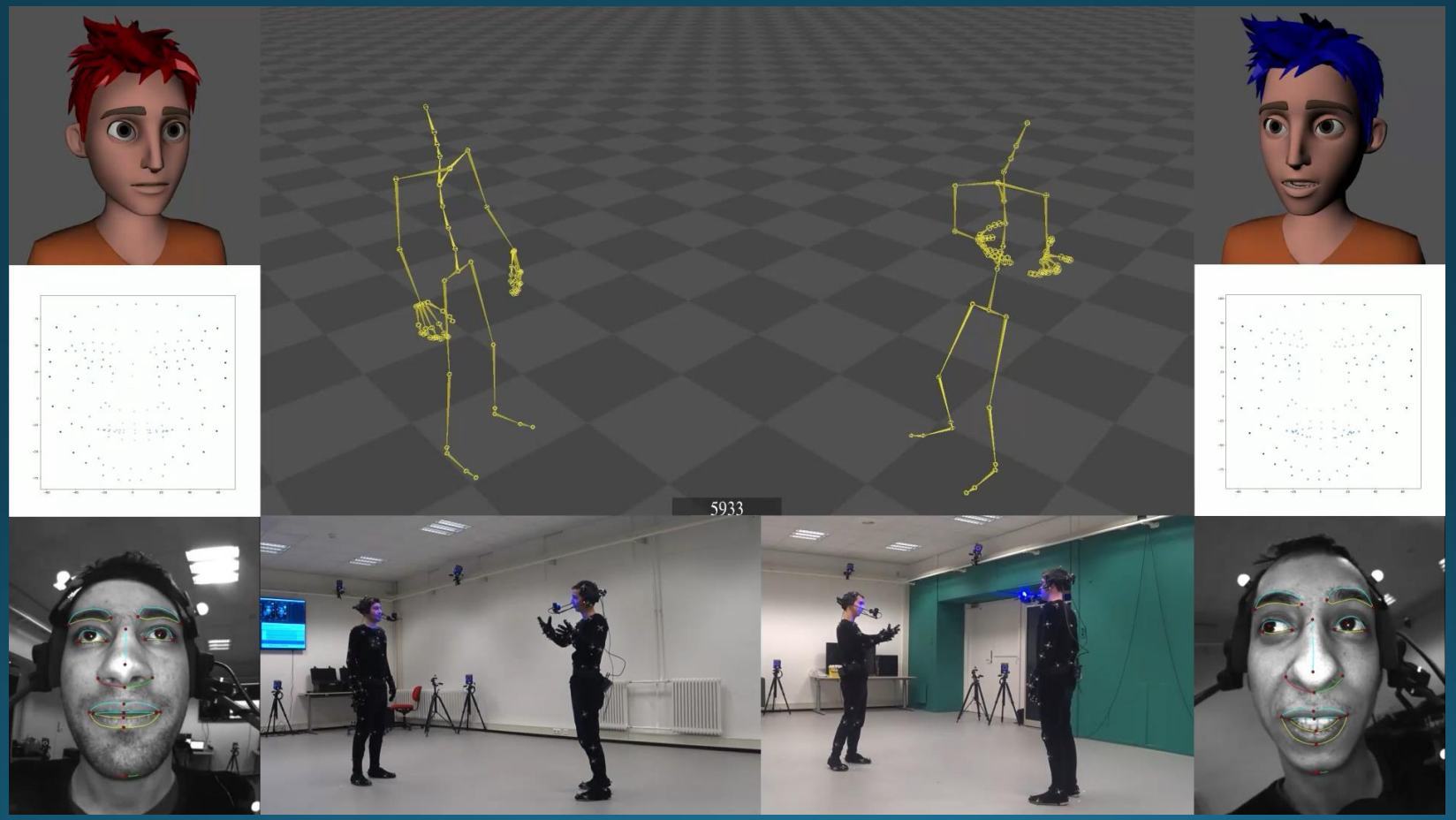






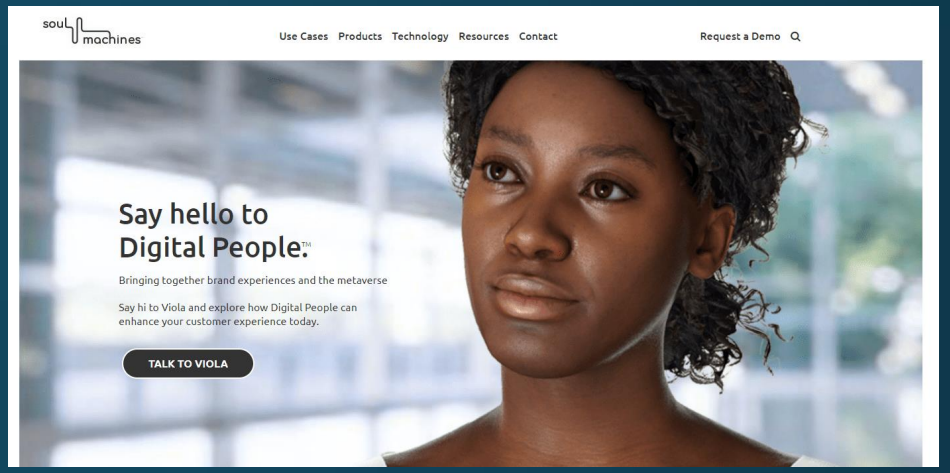
F. de Coninck, Z. Yumak, G. Sandino and R. Veltkamp. **Non-verbal Behavior Generation for Virtual Characters in Group Conversations**. 2nd IEEE Artificial Intelligence and Virtual Reality Conference (IEEE AIVR 2019) **(Honorable Mention Award)**

Multi-modal and Multi-party animations



Interaction

Samsung Neon

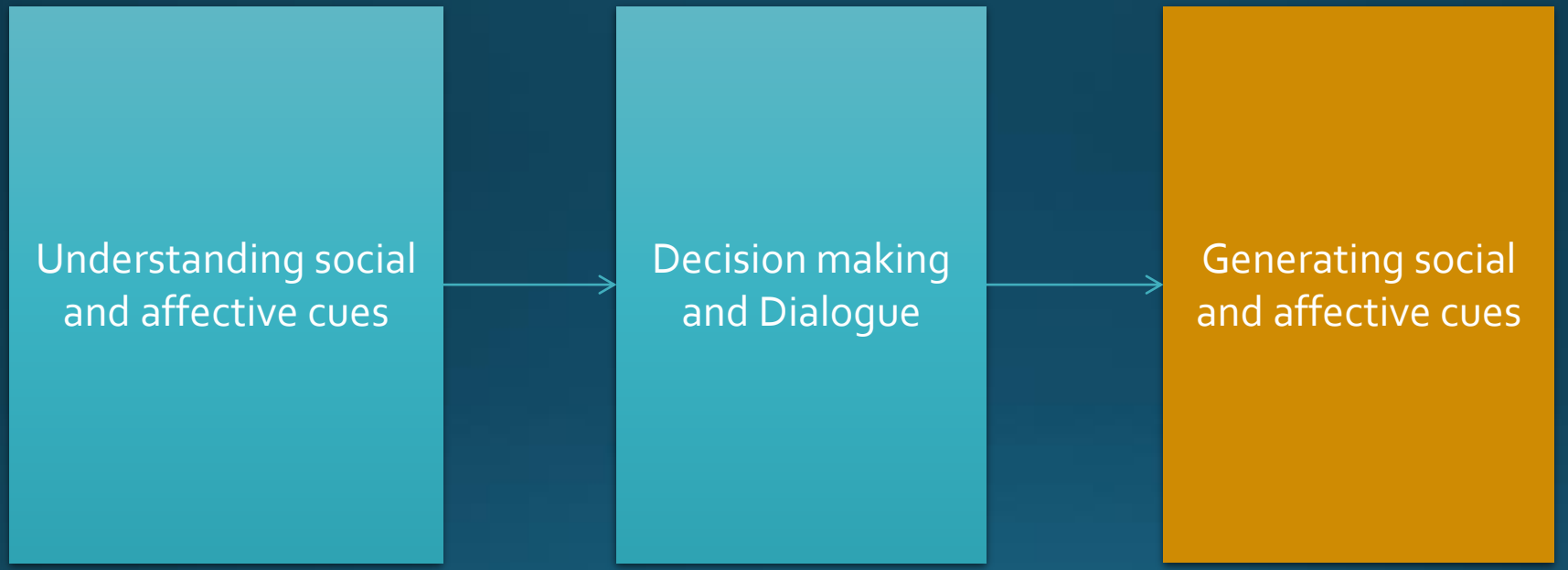


UneeQ

Nvidia Violet

Samsung Sam

Interaction with virtual humans: Overall steps

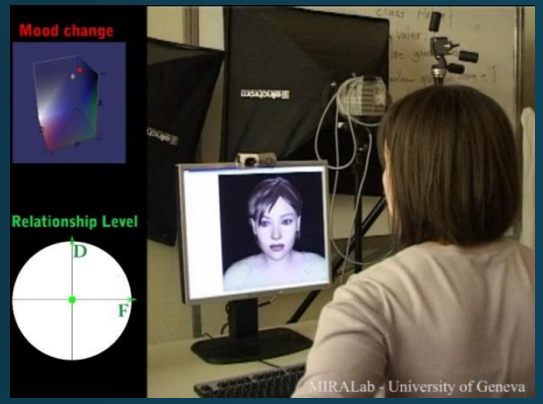


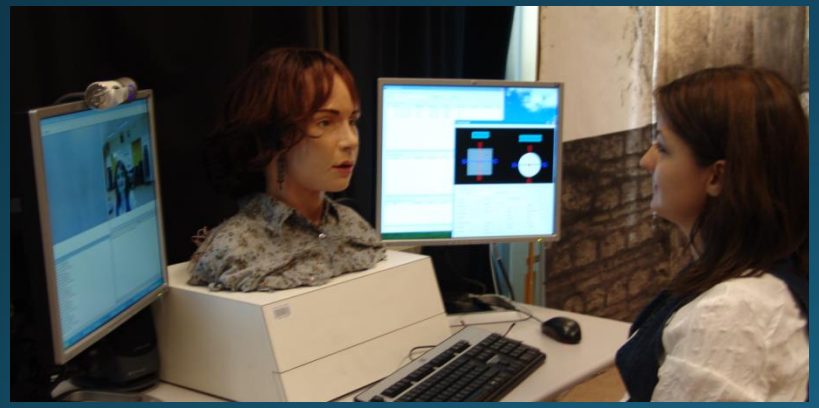
Required capabilities

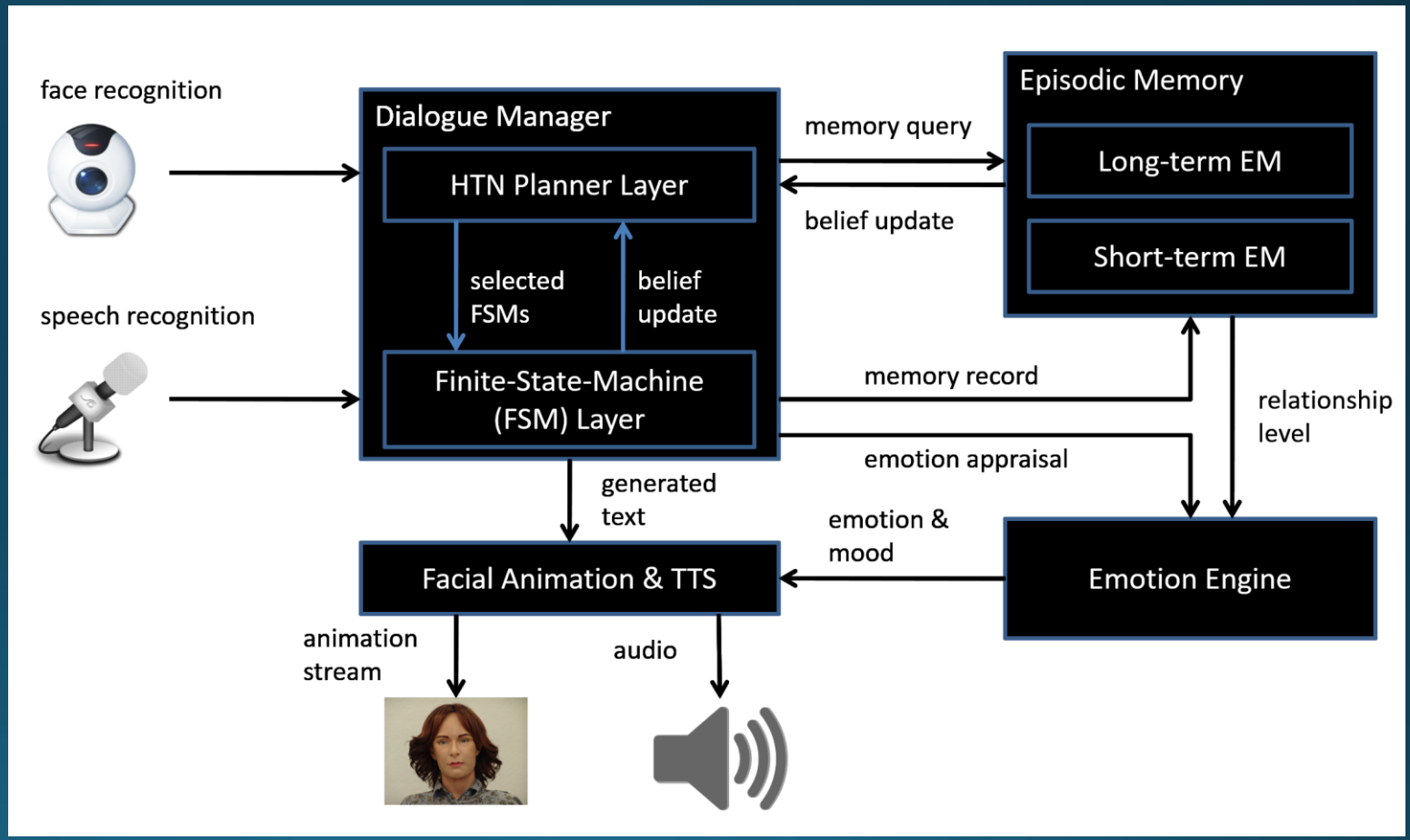
- Express and perceive emotions
- Communicate with high-level dialogue
- Use natural ways of communication
  - speech, facial expressions, gestures and gaze
- Establish/maintain social relationships
- Exhibit distinctive personality
- Learn/recognize models of others
- May learn/develop social competencies

T. Fong, I. Nourbakhsh, and K. Dautenhahn, **A survey of socially interactive robots**, Robotics and Autonomous Systems, vol. 42, pp. 143–166, 2003.

Modelling Emotions & Memory







Z. Yumak and N. Magnenat-Thalmann. **Building long-term relationships with virtual and robotic characters: The role of remembering**. The Visual Computer. 28(1). pp. 87-97. Springer-Verlag. 2012.
Z. Yumak, M. Ben Moussa, P. Chaudhuri, **N. Magnenat-Thalmann. Making them Remember-Emotional Virtual Characters with Memory**. IEEE Computer Graphics and Applications, Vol. 29, No. 2, pp. 20-29, 2009.

Multi-party interaction

Feature	Explanation
Change of distance and orientation	If the user is not only passing by but staying stable for a while
Greeting/ calling by name	If the user calls by name or greets explicitly
Waving hand	If the user waves hand
Distance	If the user is closer to the vh
Orientation	If the user configures his/her body towards the vh
Closeness to the center of FoV	If the user is closer to the center of the field of the view of the vh
Speaking	If the user is speaking
Smiling	If the user is smiling

Z. Yumak, B. van den Brink, A. Egges. **Autonomous Social Gaze Model for an Interactive Virtual Character in Real-Life Settings,** Computer Animation and Virtual Worlds, 2017.

Multi-party Dialogue

R. Harel, Z. Yumak, F. Dignum. **Towards a generic framework for multi-party dialogue with virtual humans**, CASA 2018, Beijing, China, May 2018

Challenges

- Data collection
- Sophisticated algorithms
  - Deep learning, black box algorithms, difficult to debug, computationally expensive
- Gap between subjective and objective metrics
  - Defining meaningful evaluation metrics
- Social Signal Processing in XR
  - how to interpret the social and emotional states of users in XR and how to generate automatic responses

IEEE VR 2023 and 2024MASSXR Workshop – Social and Affective Behavior Analysis and Synthesis in XR)

https://sites.google.com/view/massxrworkshop2023

https://sites.google.com/view/massxrworkshop2024/

Conclusion

- Virtual humans reached to a state where they can be modeled very realistically
- However, realism in movement and interaction is still missing
- Interactive applications (e.g. XR) requires on-the-fly generation of animations and behavior
- Perceptual studies to understand what works for which applications are needed.

Digital Humans are here to stay!

My vision > I/O Magazine > July 2023

DIGITAL HUMANS ARE HERE TO STAY

By Marysa van den Berg Image Bram Saeys

Technology for creating life-like virtual representations is being developed as we speak. And so we can best invest our time and energy in the beneficial applications of this technology for social good, according to Zerrin Yumak, assistant professor at Utrecht University.

'My research is about believable virtual humans and social robots. I use both motion capture and AI techniques to generate facial expressions, gestures and gaze movements using data-driven methods. We aim to push the boundaries of realistic human appearance and expression representation in the 3D digital world, and we are analysing the effect of these representations on the perception of users. As part of this research, we have set up a new AI lab called "Embodied AI Lab for Social Good".

There are worries about the fast development of AI. There was even a call for a six-month pause of Generative AI, particularly for large language models such as ChatGPT. Although I understand the concerns, I do not think such a pause will help solve the problem. Instead, we should focus on improving the regulations and develop new models to catch up with the fast pace of AI developments.

If digital characters and digital worlds are too life-like, that might elicit concerns for certain groups of people, like children and the elderly. We therefore need to carefully analyse and weigh up the pros and cons for each application. For example, the realism requirements for a children's educational game might not need to be that high to achieve the learning goals. On the other hand, a social skills training environment for business might be more effective when facial expressions and behaviours are modelled realistically.'

MEANINGFUL BENEFITS

'There are also discussions about the metaverse, the so-called next version of the internet where we all live and work in a connected 3D digital environment. I do not see one single metaverse platform emerging in the near future or even at all. However, I do think that 3D digital content will be increasingly used in daily life as it has a lot of expressive power. These digital humans can do a lot of the mundane work people are doing at the moment. The benefit is that they are less costly, more scalable and available anytime and anywhere.

Studies have even shown that some people prefer talking to digital humans more than to a real person, since they feel they are not judged and there is no bias. This means we need to think about where the technology could be beneficial and positively impact our lives. And in an EU and Dutch context, we should consider investing more in Generative AI technology and talent development instead of just focusing on the regulatory aspects.'

Zerrin Yumak is an assistant professor at the Human-Centered Computing Group, Department of Information and Computing Sciences at Utrecht University. She obtained her PhD in Computer Science from MIRALab, University of Geneva, and was a scientific collaborator at the HCI Group in EPFL, Switzerland. Subsequently, she was a research fellow at the Institute for Media Innovation, Nanyang Technological University in Singapore.

- Mis-match between technology level and expectations of users
  - Digital humans are still far from being capable

- How to design digital humans that can follow ethical and lawful guidelines
  - More research and development on Responsible AI/Digital Humans

Utrecht University