Peter 2.0: Building a Cyborg

Matthew P. Aylett
CereProc Ltd.
Edinburgh, UK
matthewaylett@gmail.com

Ari Shapiro
Embody Digital
Los Angeles, USA
ariyshapiro@gmail.com

Sai Prasad
Human & AI Systems Research Lab at Intel Corporation
Santa Clara, USA
sai.prasad@intel.com

Lama Nachman
Human & AI Systems Research Lab at Intel Corporation
Santa Clara, USA
lama.nachman@intel.com

Stacy Marsella
University of Glasgow
Glasgow, UK
stacy.marsella@glasgow.ac.uk
Northeastern University
Boston, USA
marsella@ccs.neu.edu

Peter Scott-Morgan
Scott-Morgan Foundation
Torquay, UK
peter@scott-morgan.com

ABSTRACT

Peter Scott-Morgan has MND/ALS. He is now paralyzed and depends on technology to keep him alive and communicate with others. In this paper we outline the design and creation of unique communication system driven by an open source eye-tracking interface (ACAT) which aims to preserve Peter's character through: 1. A cloned artificial voice, 2. An animated avatar. We describe the AI techniques adopted, the interface design and integration process for what is fundamentally an applied accessibility AI project. Finally we discuss Peter's use of the word "transitioning cyborg" to describe his take-up of AI support technology.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; **Accessibility systems and tools**; • **Computing methodologies** → *Artificial intelligence*; *Natural language processing*; *Graphics systems and interfaces.*

KEYWORDS

accessibility, eye-gaze interfaces, speech synthesis, avatars, artificial intelligence

ACM Reference Format:
Matthew P. Aylett, Ari Shapiro, Sai Prasad, Lama Nachman, Stacy Marsella, and Peter Scott-Morgan. 2022. Peter 2.0: Building a Cyborg. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529190.3529209

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3529209

1 INTRODUCTION

"It is true that AI technologies have the potential to dramatically impact the lives of people with disabilities. However, widely deployed AI systems do not yet work properly for disabled people, or worse, may actively discriminate against them." - Peter and Laura Smith [17]

When Peter Scott-Morgan got in touch with the authors in 2018 he had been diagnosed with MND (termed ALS in the United States, MND in the UK). With MND, motor neurons gradually stop sending messages to the muscles. This leads the muscles to weaken, stiffen and waste. There is no known cure. MND affects approximately 1 in 300 people. Although some people live with the disease for a considerable time many do not and there are approximately only 5000 MND sufferers alive in the UK. Peter was given 2 years to live in 2017. However *"This was to be Terminal Disease like no one had seen it before!"*[1] and Peter believed that AI technology could support him in his illness, not just to keep him alive, but to thrive. A crucial part of this support would be:

(1) Spontaneous Communication: For Peter to be able to communicate spontaneously to those around him, from his loved ones to his research colleagues after the muscles required to speak stopped working.
(2) Personality Retention: That Peter would not just communicate with others but would do so in a way which conveyed is unique personality and self.

In this paper we focus on personality retention.

Current systems used by MND patients to communicate are typically driven by eye-gaze and use a text to speech synthesizer (TTS) to produce words that are tirelessly and slowly authored by the user into speech. These systems are typically referred to as AAC (Alternative and Augmented Communication) systems.

Despite commercial voice cloning being available since 2014[2] there was no support for personalized voices in current AAC systems. Furthermore, Peter wanted to have an avatar to speak for him as his face and head muscles would be paralyzed. Despite such

[1] Peter Scott-Morgan, Interview at the Hay Festival 2021
[2] https://www.techerati.com/the-stack-archive/big-data/2016/09/13/the-future-of-voice-synthesis-after-google-wavenet-debut/, https://www.bbc.co.uk/news/business-57761873

avatar technology being widely used in films and digital games, no AAC system had integrated their system with a digital animation.

TTS and the technology required to drive an avatar with speech so that head movements appear natural and convey meaning, depends on applied AI, in particular machine learning and natural language processing (NLP). In addition, the planned system was further complicated by the requirement to build speech and avatar systems that mimic the behavior of an individual as well as offering appropriate control of this behavior for an AAC user.

This paper describes and discusses the custom system built for Peter Scott-Morgan. This involves an open source software (S/W) platform that enables full PC control through various sensors (eye-tracker in the case of Peter) which Intel Labs first created for Professor Stephen Hawking , a TTS system which clones Peter's voice and offers a rich set of expressive functionality created by CereProc Ltd. , and a avatar based on a movie quality 360 degrees body scan created by Embody Digital Inc. .

2 CONTRIBUTION

The main contribution of this paper is as a case study that shows how a set of diverse AI based components can be built into a world leading accessibility system. This is directly within scope of the PETRA themes *human centered computing* and *disability computing* but also directly relevant to *Intelligent assistive environments* in that this system will need to interact with a smart home as well as with other people. Finally the question of the extent AI support technology creates a *cyborg* and the ethical and social aspects of this process is an example of the *social impact of pervasive technologies*.

3 BACKGROUND

3.1 Eye tracking interface

Back in 2011, Intel Labs worked with Professor Stephen Hawking to develop a S/W platform to enable him to control his PC using minimal input. Professor Stephen Hawking wanted to have full control over typical applications and to communicate with others using this platform. After months of exploring his needs, observing his daily interactions, and researching existing systems out there, the team decided to build an open source S/W platform that can enable mapping any limited body movements (e.g. cheek movement, eye brow, mouth, etc) to a trigger signal that can be utilized to enable full control over applications (e.g. word processor, web browser, email, etc) as well as communicating with others through a TTS system. This software platform was named ACAT (Assistive Context Aware Toolkit), and released to open source a few years later).

3.2 Text-to-speech

TTS technology has recently experienced a paradigm shift away from concatenative synthesis [9], where units of speech are rearranged and spliced together, to systems based neural networks. The modern neural network approach (typically termed *Neural TTS*) is an extension of earlier work in parametric speech synthesis which aimed to model speech using machine learning (initially Hidden Markov Models [22], and later neural networks, e.g. [27]). These systems are trained on linguistics features to learn the spectral and prosodic parameters of the output speech given an input string of text. Originally, these parameters were then converted into a waveform using a vocoder – a digital signal processing algorithm which can change spectral specifications into a waveform. However the vocoding process reduced the perceived quality and naturalness of the speech output. By replacing the vocoding algorithm with a recurrent multi-level neural net, these systems are able to significantly improve the quality of parametric-based systems to the extent that the quality exceeds output from unit selection systems (e.g. [12]). In some cases the linguistic front-end can also be replaced by sequence-to-sequence neural net models [25].

This new approach offers enormous flexibility. It is possible to use 3rd party speech to prime and adapt models, meaning less source audio is required to copy or *clone* a voice. It is possible to add different voice styles to convey emotions and fine control of speech rate and intonation to control the way speech is rendered. The approach is so effective at mimicking a source voice that it has raised concerns on the possibility of malicious deep audio fakes. The dominant method for controlling expressive functionality in speech synthesis is through the use of XML mark-up. SSML[3] [19] is the dominant standard.

Current AAC systems do not normally support cloned voices, depending instead on a fixed set of pre-created voices, nor do they offer XML control. They send text and not mark-up meaning users cannot easily alter the *way* something is spoken.

3.3 Avatars

Humans are skilled at identifying individuals from their faces, as well as understanding verbal and nonverbal communicative signals from speech and facial expressions. Thus modeling the human face has become an important goal in computer graphics. Recent research has demonstrated the ability to generate faces synthetically through capturing facial imagery through high resolution cameras, then modeling the appearance and movement of skin, hair and eyes [23]. Earlier research has successfully replaced a human face within existing video footage [1], as well as created convincingly realistic-looking faces entirely using 3D technology [15, 24]. Some research seeks to create synthetic imagery that can pass for human [14]. A more difficult task is to synthetically replicate a specific person [4]. More recently, synthetic human imagery obtained from video footage has yielded highly convincing and realistic results [18] as well as the ability to puppeteer such imagery using one's own face to control the synthetic one [21] via facial tracking through a camera trained on the operator.

Our goal is to create an avatar that resembles Peter, and allow him to control it in real time through an eye tracking interface. Thus our challenge is to generate the body, facial and expressive movements that one would use while communicating using a low dimensional signal; the text of the utterance generated by Peter from the ACAT interface. This is in contrast to using camera tracking methods that are common to motion generation where actors wear motion capture suits that are tracked in real time by cameras or inertial sensors, and allow their face to be tracked by a camera worn on a helmet. In our case, neither body movements nor the facial movements can be copied from a puppeteer/actor. Past work in driving avatars with text include systems that can transform

[3] https://www.w3.org/TR/speech-synthesis11

text input into meaningful communicative animation. Early text-driven systems include BEAT [3], GRETA [13], and the Virtual Human Toolkit [8]. Interfaces that allow for direct control of avatar behaviors using predefined speech and nonverbal behaviors have also been created [6].

4 DESIGN AND IMPLEMENTATION

A standard AAC eye-tracking system would consist of an interactive module which displays information and allows selection and input using an eye tracker. This uses predictive text to speed up composition and also allows the rapid selection of common phrases. The text is built up and then, when complete, sent to a TTS system to produce audio.

A key difference between this and the system we present here is that the personalized avatar, rather than the TTS system, receives the marked-up text, then pushes it forward to the TTS system and, once it receives the speech audio, word and phone timings, synchronizes animation of the avatar to the speech and plays both the audio and resulting animation (See Figure 1). The system we describe here runs in windows on a standard desktop in real time but can also be used offline to prepare material for public speaking.

Each of the three main components (ACAT , TTS and Avatar) have been designed specifically for this system but with the view of making the set-up available to a wider range of users in the future.

4.1 ACAT

By separating out the specific movement from the trigger, users can be supported with various sensors as long as the sensor abstracted the trigger function. In the case of Stephen Hawking, the team built a proximity sensor that was mounted on his glasses and mapped his cheek movement to a trigger. In addition, many other sensors were developed to support different users including camera to capture facial gestures, EOG, EMG, accelerometer and many others. Contextually aware menus were created to enable interfacing with different applications through limited menu options to reduce the interaction time and avoid using mouse emulation due to its inefficiency. The team also integrated the presage word predictor and created the required language models in different languages. During the early interaction with Peter, we realized the need for an open platform for innovation and extended ACAT to support gaze tracking as well (the Tobii Eye Tracker 5), which was his modality of choice. Through an iterative design process with Peter, a customized keyboard layout optimized for gaze tracking was designed and integrated into ACAT . While other gaze tracking solutions use the standard QWERTY soft keyboard for typing, ACAT uses a circular keyboard layout that is optimized to reduce eye stress that may result from prolonged use. A QWERTY keyboard entails lateral eye movements whereas the circular keyboard layout in ACAT enforces circular eye movements which reduces stress on the eye muscles.

ACAT was key to the success of the project because most AAC systems are proprietary and it is very hard to alter the way they interface to 3rd party systems such as TTS and almost impossible to connect them to an avatar rendering system at present.

4.2 TTS

The TTS voice used in the system was built based on recording of Peter Scott-Morgan in the early stage of his illness before his vocal tract was removed. The recording process was co-designed with Peter to meet his communication needs. Recording (almost 10 hours of material) consisted of the following sections:

(1) Phonetic coverage material spoken in a neutral voice style (154 minutes).
(2) Additional voice styles recorded with appropriate text. Rather than wanting emotional styles - like happy or sad - the requirement was for styles that fitted specific communication contexts. These were: intimate, enthusiastic/presentational, conversational. After an iteration of the voice build we made the *intimate* voice style the default and moved neutral material into a voice style we called *serious* (376 minutes).
(3) Pre-recorded prompts. Most AAC systems will support common prompts and vocal gestures such as sighs and laughs. These can be pre-recorded into the TTS system and extracted using XML mark-up verbatim. A typical set might be between 100-200 prompts. In contrast, working with the user the number required was much higher than this and we recorded over 2000 pre-recorded prompts and vocal gestures (62 minutes).

This is an unusually large single speaker database for voice cloning and produces a commercial quality TTS voice (comparable or exceeding Amazon Polly neural TTS engine in quality). However voices built with an hour of data also produce good quality voices but with less emotional range.

The CereProc Ltd. TTS system, CereWave uses a recurrent neural network architecture to firstly produce prosody targets, and then produce an intermediate acoustic feature set. After predicting the acoustic features, it uses a custom neural vocoder to produce the final output waveforms similar to [12]. Bespoke XML mark-up is used to select voice styles or pre-recorded prompts. Because prosodic modeling is carried out separately it is possible to specify the duration and pitch at the phone level making it possible to graft specific intonation onto the utterance. Unlike most other Neural TTS systems, CereWave runs on device, preventing the privacy and security issues inherent in cloud based systems.

4.3 Avatar

The avatar was constructed by capturing 2D imagery of Peter using a neutral expression with high resolution RGB cameras then reconstructing a 3D model from those images using photogrammetric reconstruction. Individual expressions such as differing mouth, eyebrows, and eye positions were captured separately, see Figure 2.

Once the 3D avatar and expressions were created, they were then unified under a common vertex topology, and connected together as blendshapes, allowing for the interpolation of various different expression poses and lip shapes. Hair was synthetically designed to match Peter's hair type and coloring using individual strands. A high resolution image resulting from this capture and construction process could be used inside of a 3D modeling tool and rendered as a still image using synthetic lighting, termed the *high-resolution avatar*.

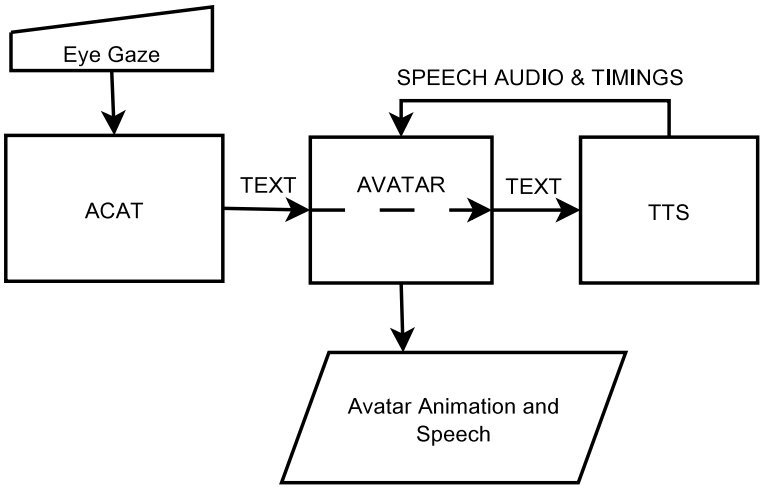

Figure 1: Avatar AAC system

Figure 2: Expression capture. Peter moves his face into different expressive poses which are then captured by cameras and subsequently transformed into a 3D model. Markers on the face provide a guideline for remeshing in order to obtain a set of meshes that share the same topology and thus can be interpolated from the neutral pose. Facial regions are then separated into separate meshes in order to provide local control, such as turning up the corner of the mouth or raising an eyebrow.

For interactive use, the avatar and rendering engine must be able to display the imagery at a rapid rate (30 to 60 fps). In order to do so, a *real-time avatar* was created from the original reconstructed data with fewer vertices and smaller texture resolutions than the *high-resolution avatar* in order to run on the real-time game engine [7].

4.3.1 Behavior and animation. The behavior engine animates the avatar by synchronizing the lip movements with various facial expressions, gaze, head movements, eye movements and other nonverbal gestures. The utterance to be spoken is first sent to the TTS system in order to retrieve the audio and phoneme timings, then the animated behavior for the character is constructed in real-time based on the content of the utterance and timings of the phonetic information in order to coordinate with the audio using an animation system [16, 20]. Utterances can be tagged with various emotional states, such as *happy, sad, angry*, as well as with word stress level.

Lip synchronization to speech is done by extracting the phonemes and timing information from the TTS component, then matching those phonemes against a database of animations that contained prescribed motion of pairs of phonemes in order to accommodate speech coarticulation (diphones). That animation data is then retimed to reflect the length between the two phonemes, then concatenated with neighboring animations of adjacent phonemes [26].

Nonverbal behavior, including facial expressions, head movements, and other gestures are driven by a rule-based system that extracts syntactical and semantic information from the utterance in combination with the emotion and stress, then converts that information into animation that drives various degrees of freedom (DOF) of the face and body. The behavior rules are generated both by extracting data from machine learning of human behavior, as well as expert knowledge of human communication [11]. The system implements over 200 rules based on the syntactical and semantic input, as well as additional metadata such as emotional state or word stress. Syntactical analysis is run to identify sentence boundaries, noun phrases, prepositions, and verb phrases, among others. Such syntactical information is used to trigger a set of rules that dictate nonverbal behavior for human speech, such as tilting the head before speaking to mimic the intake of breath before speech, or the raising of eyebrows when asking a question.

Semantic analysis is run to identify mental states or communicative intent. For example, if the utterance contains the words *very much* or *a lot* or *incredible*, then an *intensifier* semantic would be identified. Similarly, positive (*good, excellent, amazing*) or negative (*bad, terrible, disaster*) semantics are identified, along with object location (*around* or *in front of*). In all, over 50 semantic categories can be identified [10]. Each semantic is then inserted into the rule system, which then converts the semantics, along with the syntax, into a set of timed behavioral instructions which are then converted into animations which drive the characters various DOF. Behavioral instructions include the coordination of different DOF together, such as nodding the head while blinking. Behaviors movements include head movements, facial expressions, gaze location, arm gestures, and body posture. A preanimated gesture set contained approximately 30 different gestural types (including several types of metaphoric, deictic, emblem, beat, iconic) is played on a virtual body that is connected to the avatar's head and neck. Although the hand and body gestures cannot be seen on the digital model, the movements of the gestures on virtual body have an effect on the head and neck, which follows the movement of the body. Results from the system can be seen in Figure 3. Although only the avatar head and bust are shown, a virtual body including arms and legs is being controlled by the animation engine. The effect of movement of the virtual body on the avatar head can be can be seen as it leans, turns and responds to the body movement underneath.

Since numerous semantics are identified and create behavioral instructions accordingly, the behavioral descriptions are over-specified. In other words, there can be more behaviors specified to perform than the avatar could perform. The animation system then extracts a subset of behaviors that can be performed without violating the capabilities of the avatar, such as performing two different gestures at the same time. Various schemes allow for non-determinism in this selection, such as random choice, or priority to certain kinds of behaviors.

5 CASE STUDIES

https://www.youtube.com/watch?v=HKCbLfknYJs shows an example output from the system. It has generally been received well and the use of the avatar, in particular, has gathered significant media coverage. However, we encountered issues for day-to-day use and also discovered an important use case in public speaking.

5.1 Day to Day Use

The original vision for the design was to have a wheel chair, front mounted, screen, displaying the avatar in real time. This presented significant practical problems in terms of mounting the screen. In addition, in the current design the avatar added some extra latency before speech audio was produced, especially for longer utterances. Given the significant time it currently takes for a user to compose text using eye gaze the extra latency was unwelcome. However, the main negative finding to this original design was we discovered the primary use case required by the user was for video calls and this was not supported.

In addition, because of the dependence on the open source ACAT system, it was a challenge to find open source solutions that could always match the quality of proprietary systems. These professional systems are very expensive but for someone who is using an AAC system to communicate all the time, the opportunity to use more sophisticated systems is a requirement.

Figure 3: Real-time avatar while speaking using neutral (top) or expressive emotions (bottom). ©Scott-Morgan Foundation

Another main issue was the difficulty of controlling non-verbal avatar animation such as smiling. For the user being able to smile in character was very important and managing to model the smile and seamlessly incorporate it in the communication stream proved to be a significant challenge. On a positive note, the personality, humor and character of the user were preserved to a large extent.

5.2 Public Speaking

The project generated significant media attention. The system was originally designed for day to day use but it became clear that changes were required to use the system in public engagements.

In public speaking, speakers will typically rehearse and heavily script their contribution. Using default output without allowing the user to finely control and curate the output cannot guarantee the best performance and that is what is required. In addition, many media engagements are pre-scripted or pre-recorded, so a user has the opportunity to hone, tune and create the best performance in advance and pre-record it. Finally the latency caused by the time taken to compose text in real time using eye tracking is still too long to allow an acceptable media performance.

In order to support public speaking some extra design elements were incorporated into the system.

(1) For offline video generation, the *high resolution avatar* can be animated by executing an utterance on the *real-time avatar*, then extracting the motion data from it. This motion data can then animate the *high resolution avatar* which is then rendered offline, and that imagery is then collected into a single video in combination with the original audio.
(2) Using vocal puppetry [2], we made it possible to use a a third party voice to produce a default performance with emphasis and intonation closely following a simple mark-up created by the user.
(3) We used prompt editing software to allow the alteration of the performance, for example changing pauses, emphasis and speech rate. This allowed the user to give direct feedback based on pre-generated content.

The system has been used for 3 main stream TV performances, two corporate interviews, and the first ever address to the House of Lords in the UK parliament by a severely disabled person.

6 SOCIETAL IMPACT

This project is ultimately about giving a voice and a presence unconstrained by disability. As Peter Scott-Morgan points out:

"I realise that one of the greatest powers of having an avatar is circumventing the otherwise instantaneous preconditioned reaction to the "sadly pathetic cripple" that will soon be the only biological image I will be able to portray. Giving people like me their own avatar is going to break the rules of how people view us. I cannot express how humiliating it is to be instantly pigeonholed! We have the opportunity to turn that on its head." - Peter Scott-Morgan.

The scope for developing AI technology that can support the disabled is vast and mostly untapped.

7 DISCUSSION

Peter Scott-Morgan describes himself as a *"transitioning cyborg"*. Initially, this may seem a grandiose term for the technology that is currently keeping him alive. However, he is looking forward. Plans exist to add self-driving car technology to his wheelchair, to continuously increase the AI that is supporting his communication systems, to make it faster and more seamless until it will be hard to distinguish where the person begins and the AI ends. So taken in this context the term is appropriate.

Users like Peter are a massive asset for engineers wanting to develop AI support systems. Having a motivated human intelligence *"in the machine"* as it were, offers a smooth tractable route to develop increasingly sophisticated systems.

7.1 Design Guidelines

Peter Scott-Morgan has been using the system for nearly two years. In that time we have refined the system had can make some concrete design recommendations for AI based AAC system that intend to convey the users personality.

The default is not neutral:
In reality there is no such thing as a neutral voice style. Often what people mean when they use the term neutral is a measured reading style with news presenter's style of intonation. This has tended to become the default voice style for TTS voices. AAC users do not need, or want, this sort of voice style as a default. In Peter's case the warmer, more intimate style of speech was preferred. When designing a voice for AAC users its important to explore the voice style required.

AAC users desire a rich communication options:
Conventional system might include a small set of fixed prompts to speed up communication. In fact the number of high energy, verbal and non-verbal, prompts required can be very large (over 2000 in Peter's case). How to allow the selection from of a very large set of emotionally charged prompts using eye-gaze is a challenging design problem that requires novel interaction techniques. The scope of this communication is widened by the use of an avatar and must also includes expressions such appropriate smiling.

The system needs to support more than face to face communication:
One experience from the recent pandemic has been the increase in the use of mediated communication such as video calling. Modern AAC systems need to support these systems both with audio and video. In addition, the key requirement for AAC is to give the disabled and severely disabled a voice. Sometimes this is a rehearsed public voice.

Limitations in underlying technology have a big impact:
It can be hard to build prototype systems with state of the art components. Elements such as word prediction, that we take very much for granted on modern mobile phones, are proprietary. Using open source replacements which are not as sophisticated can severely limit the user experience, sometimes critically over a long period of time. A balance is required between using state of the art proprietary systems and flexible open source system that can be easily modified and developed but may perform less well.

We need to replicate a user's personality effectively:
This replication includes their voice, but also their facial expressions, gestures, responses, and other nuances that distinguish them from others. Currently, our behavioral system reflects a generic model of human movement, but not one specific to Peter. Recent research has shown that it is possible to replicate a person's gestural or presentation style [5], although that requires many hours of footage of that user.

A requirement to elevate the interaction level:
Currently the interactions happen at the letter level and a word predictor is utilized to do word completion and next word prediction. However, given advances in ASR and response generation systems, it would be more desirable for

the system to listen to the other speaker and recommend responses that the user can choose from and/or manipulate effectively to improve the spontaneity of the interaction. We are currently exploring this approach and plan to integrate with ACAT and test the feasibility and effectiveness of such an approach.

8 CONCLUSION

The technology discussed here can be regarded as a sub-set of the requirements for a social robot or animated agent. A robot/agent needs to output speech and control its visual appearance. Having a unique personality is often regarded as a requirement. The technology that will be added to Peter's system is also required by such systems: the ability to monitor conversation and suggest sensible responses, the ability to respond rapidly in a conversation, the ability to control conversational dynamics, interrupting, allowing interruption, giving feedback etc. Integrating sophisticated AI technologies for social robots/agents, developed by domain experts, can be challenging. For example, speech synthesis can be seen as a well described technology that takes text in and produces audio out. Researchers and developers often assume this use of the technology. But for both AAC use and for social robots/agents, this is not a suitable framework. When interaction becomes important, the way such synthesis integrates with movements and responds outside events becomes paramount. The way NLP techniques such as language generation support or undermine a personality design in the visual and audio rendering becomes important. A big challenge is to take, what are often rather insular technologies and integrate them together to produce something delightful. By doing so we may improve our social robots/agents but more important we can completely rewrite the future of disability. *"As a scientist, AND as a prototype, I'm VERY optimistic about the power of AI and robotics to transform our expectations of what it means to be old. Even in terms of becoming forgetful, or getting dementia. We're at the early dawn of escaping the fear of becoming infirm, of being powerless, of feeling trapped in an inadequate body."* – Peter Scott-Morgan, Fred Hood Memorial Lecture, Edinburgh Book Festival 2021.

REFERENCES

[1] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. Creating a photoreal digital actor: The digital emily project. In *2009 Conference for Visual Media Production*. IEEE, 176–187.

[2] Matthew P Aylett and Yolanda Vazquez-Alvarez. 2020. Voice Puppetry: Speech Synthesis Adventures in Human Centred AI. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*. 108–109.

[3] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[4] Hang Chu, Shugao Ma, Fernando De la Torre, Sanja Fidler, and Yaser Sheikh. 2020. Expressive Telepresence via Modular Codec Avatars. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 330–345.

[5] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[6] Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *International Conference on Intelligent Virtual Agents*. Springer, 201–215.

[7] John K Haas. 2014. A history of the unity game engine. (2014).

[8] Arno Hartholt, David Traum, Stacy C Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now. In *International Workshop on Intelligent Virtual Agents*. Springer, 368–381.

[9] Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1. IEEE, 373–376.

[10] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 243–255.

[11] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35.

[12] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[13] Isabella Poggi, Catherine Pelachaud, Fiorella de Rosis, Valeria Carofiglio, and Berardina De Carolis. 2005. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*. Springer, 3–25.

[14] Mark Sagar. 2015. Auckland face simulator. In *ACM SIGGRAPH 2015 Computer Animation Festival*. 183–183.

[15] Mike Seymour, Chris Evans, and Kim Libreri. 2017. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*. 1–2.

[16] Ari Shapiro. 2011. Building a Character Animation System. In *Motion in Games*, Jan M. Allbeck and Petros Faloutsos (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 98–109.

[17] Peter Smith and Laura Smith. 2021. Artificial intelligence and disability: too much promise, yet too little substance? *AI and Ethics* 1, 1 (2021), 81–86.

[18] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[19] Paul Taylor and Amy Isard. 1997. SSML: A speech synthesis markup language. *Speech communication* 21, 1-2 (1997), 123–133.

[20] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. 2008. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. 151–158.

[21] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*. Springer, 716–731.

[22] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, Vol. 3. IEEE, 1315–1318.

[23] Javier von der Pahlen, Jorge Jimenez, Etienne Danvoye, Paul Debevec, Graham Fyffe, and Oleg Alexander. 2014. Digital Ira and beyond: Creating Real-Time Photoreal Digital Actors. In *ACM SIGGRAPH 2014 Courses* (Vancouver, Canada) *(SIGGRAPH '14)*. Association for Computing Machinery, New York, NY, USA, Article 1, 384 pages. https://doi.org/10.1145/2614028.2615407

[24] Javier von der Pahlen, Jorge Jimenez, Etienne Danvoye, Paul Debevec, Graham Fyffe, and Oleg Alexander. 2014. Digital ira and beyond: Creating real-time photoreal digital actors. In *ACM SIGGRAPH 2014 Courses*. 1–384.

[25] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[26] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. 2013. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*. 131–140.

[27] Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 7962–7966.