An Extended Reality Solution for Mitigating the Video Fatigue of Online Meetings

Cornelius Glackin, Nigel Cannings, Vigneswaran Poobalasingam, Julie Wall, Saeed Sharif, and Mansour Moniri

1 Company Description

Intelligent Voice is a global leader in the development of proactive compliance and eDiscovery technology solutions for voice, video, mixed reality (MR), and other media. The core business of the company is speech recognition and natural language processing technology, providing complex analytic capabilities of speech audio. Its clients include government agencies, banks, securities firms, contact centres, litigation support providers, international consultancy, advisory businesses, and insurers, all involved in the management of risk and meeting of multi-jurisdictional regulation. Fundamental to its success, its patent-pending and patented technologies are developed by a team of dedicated researchers and system

C. Glackin (✉) • N. Cannings • V. Poobalasingam
Intelligent Voice, London, UK
e-mail: neil.glackin@intelligentvoice.com

J. Wall • S. Sharif • M. Moniri
University of East London, London, UK

T. Jung, M. C. tom Dieck (eds.), *XR-Metaverse Cases*, Business Guides on the Go,
https://doi.org/10.1007/978-3-031-30566-5_5

engineers based in the UK. The company leads the market and maintains its strengths in the areas of thought leadership, innovation, R&D, and providing solutions to its clients.

This project was supported by an Innovate UK Knowledge Transfer Partnership (KTP) in collaboration with the University of East London (UEL). UEL provided expertise in virtual reality (VR), augmented reality (AR), 3D tele-immersion, video processing, animation and personalisation of avatars, and audio adaptation and reconstruction.

2 Project Summary

COVID-19 and the related global lockdowns meant online meetings were no longer just an option (Richter, 2020). This new normal was necessitated by a movement from synchronous in-person meetings to synchronous online meetings (Davison, 2020). Post-pandemic, there has been a permanent shift towards this online communication (Guyot & Sawhill, 2020). However, the new normal has seen the rise of a new phenomenon, "Zoom fatigue", also known as video fatigue (Fosslien & Duffy, 2020). This is due to the lack of naturalistic cues being shared in the current tele-conferencing technology, together with the feeling of being observed all the time. During in-person communication, participants follow the 7-38-55 rule to decipher the meaning behind what's being said; 7% verbal, 38% tone of voice, and 55% body language (Mehrabian & Wiener, 1967). Video calls take away most body language cues, but because the person is still visible, the brain still tries to compute that non-verbal language. It means that participants are working harder, trying to achieve the impossible. This impacts data retention and can lead to participants feeling unnecessarily tired.

At the least, part of the answer could lie in the growing world of VR gaming, where the action is driven by the user's actions, usually through a gamepad or keyboard. The visual representation of the person is an avatar. This allows the human operator to have a degree of distance from their online presence. So, if they need to scratch their nose, the avatar is not mimicking them. On a video call, all actions are immediately transmitted and seen by other participants and that puts extra pressure on

everyone involved. But of course, participants in online meetings interact differently to gamers and, hence, will need a different type of control. This is where the voice becomes so important.

To avoid unnecessary user actions, it is desired to be able to turn off the camera completely. The actions of the avatars need to be driven by what the user is saying. The speech and emotion recognition expertise of the project partners allow understanding of what is being said, and how it is being said (Iyer et al., 2022). If the tone is light and friendly, the avatar will relax and smile. If the tone is aggressive, it may lean forwards to make a point. Virtual participants can sit naturally in a conference room or sit on the deck of a yacht to sip cocktails. The setting is irrelevant. The point is that it simplifies the information that the brain needs to compute, preventing "Zoom fatigue". Additionally, moving away from standard video conferencing reduces the bandwidth requirements of the application.

This project transformed the way online meetings happen enabling virtual shared experiences with cutting-edge AR/VR technology, speech-to-text and emotion recognition technologies, and sub-real-time hardware acceleration using high-performance computing (Ali et al., 2019).

3 Project Details

The technology presented here, called iVXR, uses both the Android and iOS developer kits to build an immersive meeting experience (see Fig. 1). The product can support each one of the following modes: *3D Mono*, standard 3D rendering in a standard display, supported by Windows 11, Windows 10, Android (version 7.0 or above with ARM64 architecture), macOS, and iOS (including iPadOS); *3D Stereo*, also known as Virtual Reality (VR) on the Oculus Quest 2; *AR Mono*, AR rendering in a standard display, supported by Android with ARCore support and iOS (including iPadOS) with ARKit support; and *MR Stereo*, "Passthrough" mode on the Oculus Quest 2 and HoloLens 2. The technology also supports the inexpensive cardboard options with split screen stereo AR or VR options. The cross-platform multi-mode functionality ensures that meetings can be conducted with any combination of application modes for the participants.

Fig. 1 AR Mono application mode, where the computer-generated imagery (CGI) is augmented onto the actual real-world video stream and rendered in a standard display

Fig. 2 iVXR UI, providing authentication and a configurable 3D avatar system shown in portrait and landscape view on a smartphone (iPhone)

This project utilises the Unity game engine as the basis of the application and user interface (UI) (see Fig. 2), a proven solution for massively multi-user applications, and this together with the preferred inexpensive

MR technology is intended to provide a useful application for a wider community of potential users (http://unity3d.com). The user's avatar can be configured and personalised using authentication supported by Google Firebase and is also stored locally in their profile (https://firebase.google.com/).

The application offers fine-grained personalisation of the avatars, around appearance, gender, ethnicity, and clothing. Avatars are animated in real time. Their facial expression system is based on the Facial Action Coding System (FACS) (Farnsworth, 2019) (see Fig. 3), providing eye movements of blinking, narrowing, open and closing, squinting, up and down, and left and right. Facial emotions include neutral, happy, sad, surprise, fear, anger, disgust, and contempt. The head can turn left to right, up, and down, and tilts forward and back. Avatar bodies and fingers are animated using an animation sequence clip system. Real-time communication of avatar movements across meeting participants in the shared experience is implemented using a third-party application

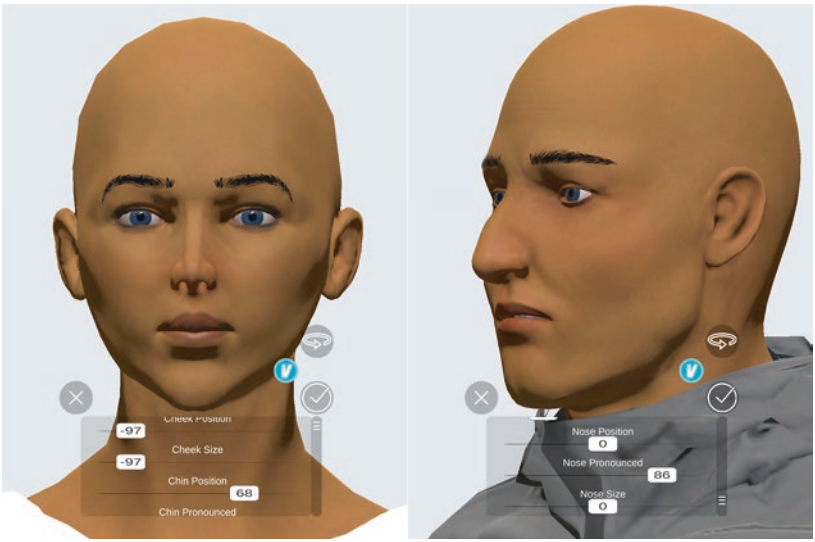

Fig. 3 Avatar facial geometry customisation options, up to 255[30], also supporting male/female genders, 35 skin colours, 6 hairstyles with 22 different hair colours, 20 eye colours, and a limited selection of clothing and shoe colours

programming interface (API) called VIVOX, a subsidiary of Unity Technologies (https://unity.com/products/vivox).

The application provides near real-time subtitling of the speech audio using the company's Automatic Speech Recognition (ASR) engine, which employs NVIDIA's Riva deep learning framework (https://developer.nvidia.com/riva), which serves the streaming ASR models, and a GStreamer bridge to the Unity-based application (https://gstreamer.freedesktop.org/), as shown in Fig. 1. The incorporated ASR technology was benchmarked on the company's proprietary speech test corpus, achieving an accuracy of correct words at 97.4% and a word error rate (WER) of 5%. Speaker-separated Smart Transcripts are produced by iVXR and automatically emailed, copied, or saved after each meeting, serving as a permanent record (Glackin et al., 2019). Real-time audio and chat communication are supported by VIVOX.

For the headset devices, Oculus Quest 2 and Hololens 2, hand tracking is in operation to interact with the UI, enabling the user to control the environment without hand-held controllers in VR and MR modes (see Fig. 4). Users switch the controlling hand by pinching the thumb and ring fingers on the respective hand. The pointer will move to indicate the active hand, as with the hand-held controller. Scaling, positioning, and rotation of the avatars and avatar seating are available in MR mode (see Fig. 5). Rotation is limited to the Y-axis, whereas positioning uses the

Fig. 4 Left: Hand tracking in VR mode with the Oculus Quest 2. Here, the right hand has the pointer, which means it is controlling, active, or dominant. Right: Hand tracking in MR mode with the Oculus Quest 2. Here, the camera feed of the actual hands are visible, and the right hand has the pointer, which means it is controlling, active, or dominant. This is a render buffer capture using SideQuest software

Fig. 5 In MR mode, the user's avatar and chairs are scaled, positioned, and rotated in the Y-axis next to the computer screen. Positioning and rotation are done using the hand-held controller's position and rotation. In this case, the right-hand controller is visible in the background. It is possible to observe that the hand is bending to rotate the controller for avatar and chair rotation in the Y-axis. This is a render buffer capture using SideQuest software

entire 3D space. The user's avatar and chair placement initialise on the top of the 3D overlaid controller.

4 Feedback from End Users

Regular user testing sessions have taken place amongst the project partner's staff to elicit feedback on and to support iterative refinement of the application. In terms of usability and user experience, the workshops focused on the ability of the users to join the meeting from the various supported platforms; ease of personalisation of the avatar and configuration of the virtual environment; technical feedback on latency, responsiveness, sound quality, and transcription accuracy; and the naturalness of the shared experience. Our users have expressed high levels of satisfaction, whilst identification of practical issues around usability have informed ongoing development, whilst feedback on user experience has

indicated positive responses to feelings of engagement with iVXR's virtual environment. Increasingly more productive meetings amongst the project partners using iVXR are taking place, reflecting the improvement through this ongoing consultation process. A broader user study is in preparation, to collect more specified user feedback from students and staff at the University of East London.

5 Future Outlook

Attempts have been made during the pandemic to try to change the current 2D imagery of video meetings and make it more accessible. Teams "Together" mode is an example. But these endeavours do not resolve the underlying problems of "presence". There are numerous competing solutions in various stages of development in the marketplace, such as Spatial, Horizon Workrooms, MeetinVR, Glue, Mozilla Hubs, BigScreen, ENGAGE, Rumii, AltspaceVR, Rec Room, and FrameVR. However, it is difficult to separate the conceptual from the real implementations.

In this project, the focus has been on providing an immersive meeting experience, with reduced bandwidth by concentrating on the audio channel, cross-platform support, speech and emotion recognition, and natural language processing technology. Building on the company's patented privacy preserving technology for audio communication, the ongoing plans for this solution are to support secure communication, making iVXR a security-conscious technology that integrates with the daily business workflow. The company is also investigating other use cases for this technology, such as VR for education and telemedicine.

6 Conclusion

This chapter presented a summary of the efforts and technological enhancements that we have made in order to address the phenomenon of video fatigue. The presented technology is audio-based solution for immersive online meetings, which reduces communication bandwidth and provides an interactive, portable, searchable, and speaker-separated

transcript of the meeting. Inexpensive technologies like the Oculus Quest 2 provide a vehicle for this application to reach a wider audience. As more AR applications become available for the end user, the hardware to support these will continue to reduce in price and become increasingly accessible. Having started this project pre-pandemic, the application described here is timely and provides a viable alternative approach to standard videoconferencing, promoting more effective and efficient meetings.

Acknowledgements This work was supported by an Innovate UK Knowledge Transfer Partnership (KTP) Grant No. 011056.

References

Ali, A., Glackin, C., Cannings, N., Wall, J., Sharif, S., & Moniri, M. (2019). A framework for augmented reality based shared experiences. *Immersive Learning Research Network-iLRN*.

BigScreen. https://www.bigscreenvr.com/

Davison, R. M. (2020). The transformative potential of disruptions: A viewpoint. *International Journal of Information Management, 55*, 102149.

Engage. https://engagevr.io/

Facebook Technologies, LLC. "Meta Quest Workrooms." https://www.oculus.com/workrooms/

Farnsworth, B. (2019). *Facial Action Coding System (FACS)—A visual guidebook*. Boston.

Fosslien, L., & Duffy, M. W. (2020). How to combat zoom fatigue. *Harvard Business Review, 29*.

Frame. https://framevr.io/

Glackin, C., Dugan, N., Cannings, N., & Wall, J. (2019, September). Smart Transcription. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (pp. 134–137).

Glue. https://glue.work/

Google Firebase. https://firebase.google.com/

GStreamer. https://gstreamer.freedesktop.org/

Guyot, K., & Sawhill, I. V. (2020). *Telecommuting will likely continue long after the pandemic*. The Brookings Institution.

Hubs, Mozilla. https://hubs.mozilla.com/

Iyer, S., Glackin, C., Cannings, N., Veneziano, V., & Sun, Y. (2022). A comparison between convolutional and transformer architectures for speech emotion recognition. In *IEEE International Joint Conference on Neural Network Proceedings*. IEEE.

MeetinVR. https://www.meetinvr.com/

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology, 6*(1), 109.2.

Microsoft inc, AltspaceVR. https://altvr.com/

Microsoft Teams Together Mode. https://techcommunity.microsoft.com/t5/microsoft-teams/enabling-together-mode-in-ms-teams/m-p/1698285

NVIDIA Riva. https://developer.nvidia.com/riva

Rec Room. https://rec.net/

Richter, A. (2020). Locked-down digital work. *International Journal of Information Management, 55*, 102157.

Spatial Systems. https://spatial.io/

Unity. http://unity3d.com

Vivox. https://unity.com/products/vivox