XR SALENTO 2025
International Conference on eXtended Reality
June 17-20, 2025 – Otranto, Italy

Enhancing Accessibility and Inclusivity in Business Meetings through AI-Driven Extended Reality Solutions

N.D. Tantaroudas, A. J. McCracken, I. Karachalios, and E. Papatheou
Institute of Communication and Computer Systems, 9 Ir. Polytechneiou, Zografou, 15773, Greece, https://www.iccs.gr/el/
DASKALOS-APPS, Rue de l'abbe Griffon, 01960, Peronnas, France, https://daskalos-apps.com
Department of Water Resources and Environmental Engineering, School of Civil, Engineering, National Technical University of Athens, Athens, Greece, http://www.hydro.civil.ntua.gr/en/
Exeter Small-Scale Robotics Laboratory, Engineering Department, University of Exeter, Exeter, EX4 4QF, UK, https://engineering.exeter.ac.uk/
nikolaos.tantaroudas@iccs.gr

Introduction and Related Work

Business meetings increasingly rely on teleconferencing platforms, yet most fail to support deaf, hard-of-hearing, and multilingual participants effectively. Existing systems lack reliable real-time captioning, multilingual translation, and sign language interpretation—creating barriers to participation [1,2,5]. Moreover, collaborative tools like virtual whiteboards or shared documents are often inaccessible, and few platforms incorporate emotional awareness to foster engagement [3,4,6]. INTERACT addresses these gaps by integrating AI-powered services into XR environments: real-time speech-to-text, multilingual translation, sentiment analysis, and International Sign Language (ISL) delivery via 3D avatars. By combining these tools into a single system, INTERACT offers inclusive, real-time, and emotionally responsive meeting experiences in immersive virtual spaces.

Methodology and System Components

Speech recognition is powered by Whisper AI [7], enabling multilingual transcription across English, French, and Greek. Facebook's NLLB [8] supports translation between language pairs, ensuring accurate multilingual interaction. For emotional context, we use a lightweight DistilRoBERTa transformer model [9], fine-tuned for multi-class emotion classification to extract sentiment from user utterances in real time (Fig. 1). ISL interpretation is powered by a custom gesture dataset built using 750 ISL videos processed with Google MediaPipe. These landmarks are mapped to avatar animations in XR, supporting dynamic, responsive sign language delivery.

```
(sentiment_env) root@DESKTOP-K6OCOR2:/home/ntanta/projects/SentimentAnalysis# python3 sentiment.py
config.json: 100%|| 1.00k/1.00k [00:00<00:00, 8.17MB/s]
pytorch_model.bin: 100%|| 329M/329M [00:57<00:00, 5.76MB/s]
tokenizer_config.json: 100%|| 294/294 [00:00<00:00, 2.97MB/s]
vocab.json: 100%|| 798k/798k [00:00<00:00, 1.99MB/s]
merges.txt: 100%|| 456k/456k [00:00<00:00, 2.96MB/s]
tokenizer.json: 100%|| 1.36M/1.36M [00:00<00:00, 2.13MB/s]
special_tokens_map.json: 100%|| 239/239 [00:00<00:00, 2.21MB/s]
Device set to use cpu
model.safetensors: 100%|| 329M/329M [00:57<00:00, 5.67MB/s]
model.safetensors: 100%|| 329M/329M [00:57<00:00, 5.34MB/s]

Sentiment Analysis Result:
Label: neutral, Score: 0.5495
(sentiment_env) root@DESKTOP-K6OCOR2:/home/ntanta/projects/SentimentAnalysis# python3 sentiment.py
Device set to use cpu
Enter the text for sentiment analysis: i am very dissapointed with the outcome of this project.

Sentiment Analysis Result:
Label: sadness, Score: 0.3571
(sentiment_env) root@DESKTOP-K6OCOR2:/home/ntanta/projects/SentimentAnalysis# python3 sentiment.py
Device set to use cpu
Enter the text for sentiment analysis: hello, how are you? what a nice day we have today!

Sentiment Analysis Result:
Label: joy, Score: 0.8232
(sentiment_env) root@DESKTOP-K6OCOR2:/home/ntanta/projects/SentimentAnalysis#
```

Figure 1: Real-time sentiment analysis using a fine-tuned DistilRoBERTa model.

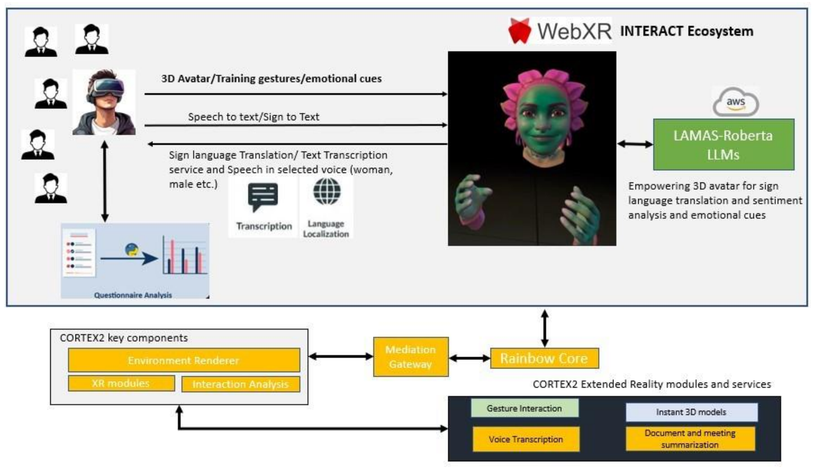

Figure 2: Envisioned Scenario and System Components of INTERACT

Sign Language Model Development and Gesture Animation Pipeline

To support real-time International Sign Language (ISL) translation, we developed a custom **text-to-sign pipeline** that combines deep learning-ready data preparation with XR avatar animation. A curated dataset of over **750 ISL gesture clips** was extracted from publicly available and recorded video materials. Each video was manually trimmed and annotated to isolate unique sign representations for high-frequency business-related vocabulary. Using Google MediaPipe's Holistic model and OpenCV, we extracted **33 pose** and **42 hand landmark points per frame**, capturing full-body motion across thousands of frames. The resulting data was serialized into structured JSON files, each representing temporal skeleton motion for a single word or phrase. These JSON files drive avatar animations in XR, enabling lifelike, responsive gesture rendering. A Unity-based animation handler reads the skeleton data and animates a rigged 3D avatar to reproduce each sign in sync with translated spoken input. This framework provides the backbone for a modular, reusable, and extensible ISL delivery system that can be scaled with new vocabulary as needed.

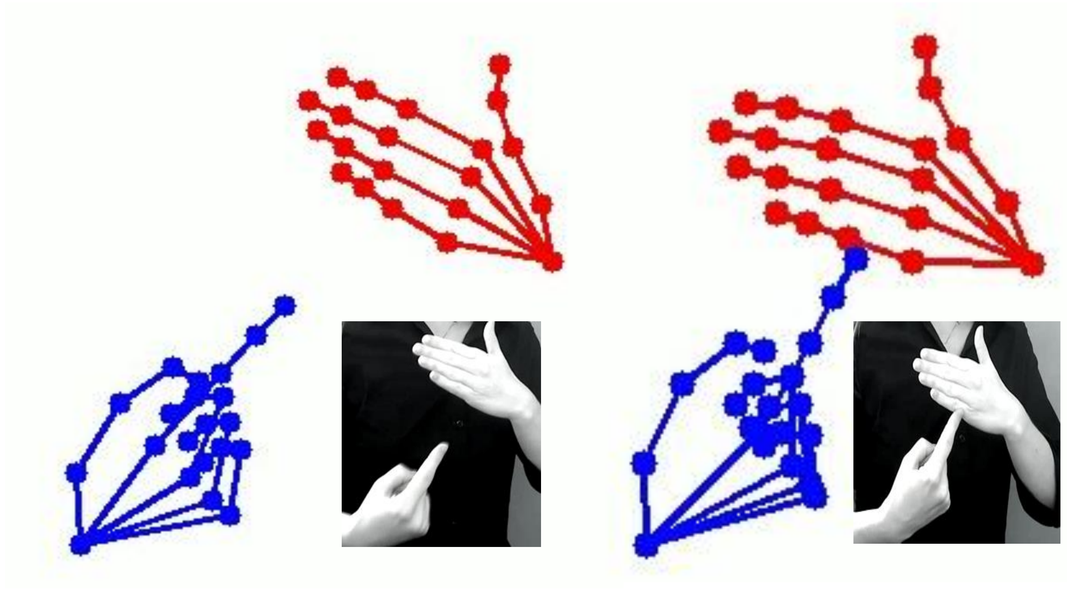

Figure 6: *A visual sequence of real-time avatar animation driven by extracted gesture landmarks from the ISL dataset.*

User Interface and System Architecture

The platform is implemented using Unity and the Meta Quest 3 headset, integrating conferencing tools via the Rainbow SDK and scalable back-end routing through the CORTEX2 Mediation Gateway [10,11]. Fig. 4 shows the system pipeline, with the headset as the main interaction point and the avatar mediating between speech, sign, and emotion. A visual dashboard overlays translated text, gestures, and emotional cues in real-time, tailored to each participant's preferences.

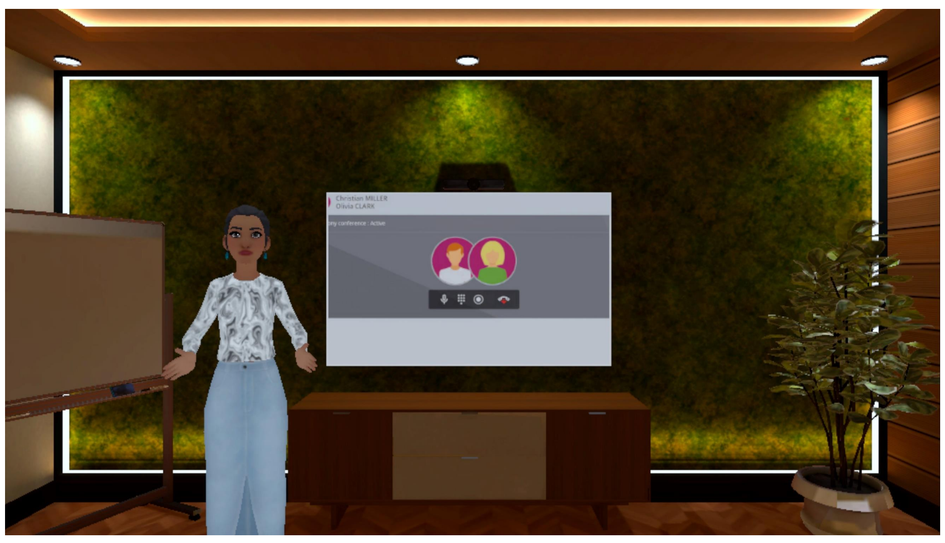

Figure 4: Unity implementation with conferencing and 3D avatar modules shown in the XR interface.

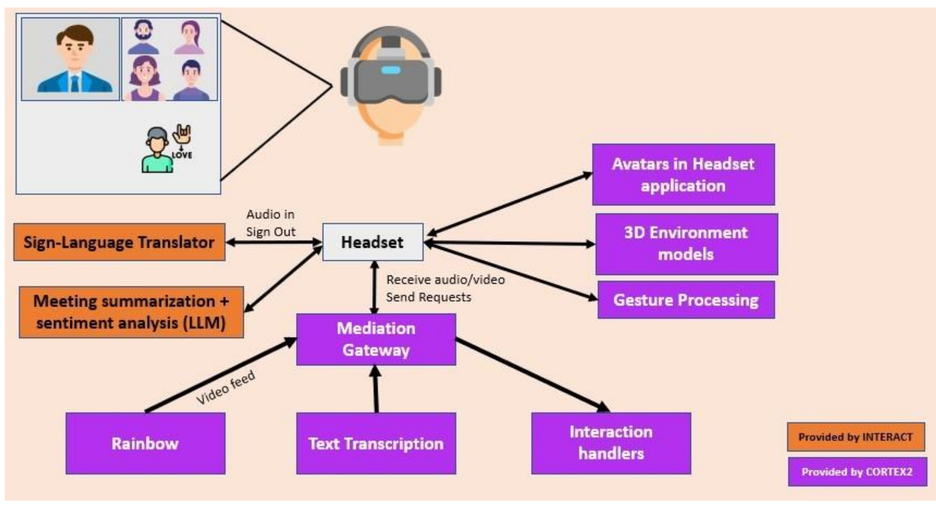

Figure 3: Real-time data flow among core modules, including Whisper/NLLB, sentiment, sign translator, and conferencing API.

Evaluation and User Feedback

A pilot study with five users tested the system's usability, avatar clarity, transcription accuracy, and overall experience. Users interacted with the XR interface and submitted in-environment questionnaires derived from SUS and UEQ. Feedback highlighted satisfaction with avatar responsiveness and multilingual communication. Minor gesture sync issues and sign clarity during fast speech were reported, leading to refinement of timing and visuals. Future testing will broaden the participant base and test across device types and languages to ensure accessibility and scalability.

Future Work and Studies

INTERACT represents the first known XR-based platform to integrate real-time ISL translation, multilingual transcription, and emotion-aware avatars into one conferencing solution. Future work will include expanding the ISL dataset to improve gesture expressiveness, refining latency for translation pipelines, and increasing language coverage in Whisper/NLLB. Further studies will test cross-cultural accessibility, adaptive UI personalization, and network performance under varied conditions. Ethics and fairness remain priorities: all data is anonymized, avatars are being audited for cultural bias in sign delivery, and accessibility settings are being designed to adapt in real time to user profiles. These enhancements will support equitable collaboration in hybrid and global business settings.

References
[1] Alford, Andrea D. et al. "Is the Window of Learning Only Cracked Open? Parents' Perspectives on Virtual Learning for Deaf and Hard of Hearing Students." American Annals of the Deaf 168 (2023): 17 – 28
[2] Franceschini, Dario et al. "Removing European Language Barriers with Innovative Machine Translation Technology." IWLTP (2020)
[3] Serafin, S., Adjorlu, A., and Percy-Smith, L.M. (2023). A Review of Virtual Reality for Individuals with Hearing Impairments. Multimodal Technol. Interact., 7,36
[4] Chen, Weisi et al. "An Event-Based Framework for Facilitating Real-Time Sentiment Analysis in Educational Contexts." 2022 11th International Conference on Educational and Information Technology (ICEIT) (2022): 57 -61
[5] Hosseinkashi, Yasaman et al. "Meeting Effectiveness and Inclusiveness: Large-scale Measurement, Identification of Key Features, and Prediction In Real-world Remote Meetings." Proceedings of the ACM on Human-Computer Interaction 8 (2023): 1-39
[6] Seraji, Farhad et al. "Comparing Two Forms of Spatial Contiguity Principle in Student Learning: 'Text Linked to Image' versus 'Text in Image Adjacency'." (2020)
[7] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR
[8] NLLB Team et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672. https://arxiv.org/abs/2207.04672
[9] Hartmann, J. (2022). Emotion English DistilRoBERTa Base: A Fine-Tuned Model for Emotion Classification. Hugging Face Transformers. https://huggingface.co/jh-artmann/emotion-english-distilroberta-base
[10] CORTEX2Project–https://cortex2.eu
[11] Rainbow SDK – https://developers.openrainbow.com

This research is supported by FSTP Funding from the Horizon Europe under grant agreement no. 101070192 (CORTEX2). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.