

Automated Lip Reading: Exploring the potential for Accessibility Measures in XR

Hrishikesh Mulay, Prof. Sam Redfern, Prof. Eleni Mangina

Organisers:



Under the auspices of:



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 the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Introduction

- Immersive technologies will influence the pedagogical perspectives and practices
 - Virtual tools for real-world learning, training and development
 - Interactive classroom experience
 - Safer training options for vocational trades, manufacturing assembly lines, supply chain etc.
 - Global platform for Art Education
- Inclusivity - lack of accommodations for differently abled people
- Automated Lip-reading system
 - Speech Recognition Enhancement
 - Assistive Communication
 - Forensics
- Aim – advancement of accessibility of XR environments

Related Work

- Advancement of ALR systems – from alphabets and digits to sentences ‘in the wild’
 - Initially researchers focused on recognition of simple utterances such as digits/alphabets
 - Progress in digit/alphabet recognition made way for recognition of simple day-to-day phrases such as ‘Thank you!’, ‘Excuse me’ etc.
 - Modern ALR research focuses on recognition of full sentences in controlled and uncontrolled environments
- Identification of the modern ALR process that involves Lip Localization/ROI detection, Feature Extraction, Classification
- The digital transformation of ALR has taken place over the last decade due to the development of the deep learning algorithms
 - Emphasis in variations of CNN, RNN, and other complex machine learning algorithms
 - Utilization of common evaluation metrics – Accuracy, WER, CER, WRR etc.
 - Results suggest that ALR is improving

Related Work

- Modern deep learning algorithms exhibit near-perfect accuracies for easier content such as digits, alphabets etc.
- New database creation techniques – TV broadcasts and online content
- Sentence-level ALR – still a long way to go!
- Some algorithms are performing better than humans
 - Humans – 30-40%
 - LipNet – 52.3%
 - AV-HuBERT - 75% higher than other audiovisual speech recognition systems

XR Accessibility Analysis

- Lack of comprehensive standards for XR
- Suggestions – captions/subtitles, mono audio, visual/haptic cues, and exploration of sign language

- ✓ Simple and Intuitive
- ✓ Perceptible Information
- ✓ Tolerance for Error
- ✓ Low Physical Effort
- ✓ Size and Space for approach and use

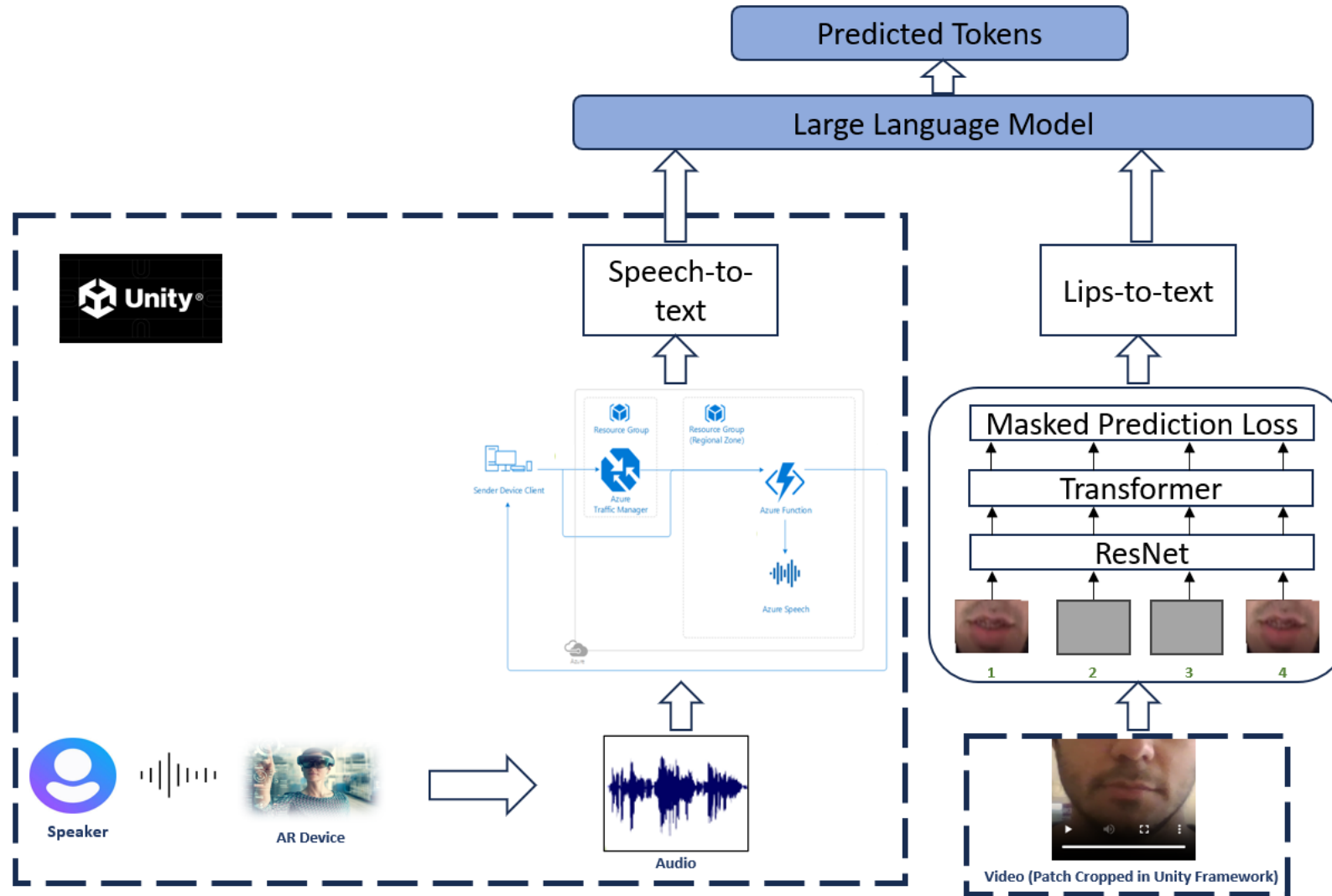
- Design Parameters

- | | | | |
|-----------------------|--------------------------|------------------------|----------------------------|
| ✓ Audio/Video Capture | ✓ Lip Reading Algorithms | ✓ Real-Time Processing | ✓ Caption Display |
| ✓ Customization | ✓ User Interaction | ✓ Testing | ✓ Accessibility Guidelines |

- Evaluation Criteria

- | | | |
|------------|-------------------------------|--------------|
| ✓ Accuracy | ✓ Time/Space Complexity | ✓ Robustness |
| ✓ Ethics | ✓ Usability and Accessibility | |

System Architecture



XR Accessibility Analysis

- Lack of comprehensive standards for XR
- Suggestions – captions/subtitles, mono audio, visual/haptic cues, and exploration of sign language

- ✓ Simple and Intuitive
- ✓ Perceptible Information
- ✓ Tolerance for Error
- ✓ Low Physical Effort
- ✓ Size and Space for approach and use

- Design Parameters

- | | | | |
|-----------------------|--------------------------|------------------------|----------------------------|
| ✓ Audio/Video Capture | ✓ Lip Reading Algorithms | ✓ Real-Time Processing | ✓ Caption Display |
| ✓ Customization | ✓ User Interaction | ✓ Testing | ✓ Accessibility Guidelines |

- Evaluation Criteria

- | | | |
|------------|-------------------------------|--------------|
| ✓ Accuracy | ✓ Time/Space Complexity | ✓ Robustness |
| ✓ Ethics | ✓ Usability and Accessibility | |

System Evaluation

- Technical Evaluation

Metric	Audio-Only Speech Recognition (ASR)	Automated Lip Reading (ALR)	ASR+ALR+LLM
WER	X	X	X
Latency	X	X	X
Time Complexity	X	X	X
Space Complexity	X	X	X

- Quality of Experience (QoE) Overview

- QoE measures user satisfaction and enjoyment of a service or application based on personal expectations, personality, and context.
- Influencing factors include context, culture, system performance, user satisfaction, socio-economic and psychological characteristics.
- QoE in AR systems is assessed through task performance, feedback, and various metrics.

System Evaluation

- **Participant Involvement**

- Participants include a diverse range of ages and AR experience levels to capture varied user experiences.
- Involves both novice and experienced AR users to ensure broad feedback and issue identification.

- **Study Design and Phases**

- Information Phase: Participants receive study details and complete a demographic questionnaire.
- Screening and Training Phase: Includes a reading comprehension test and basic AR headset training.
- Testing Phase: Participants use the AR system in quiet and noisy environments and complete task-based questionnaires to assess comprehension and usability.

- **Evaluation Metrics**

- Task Success Rates: Measures success in communication tasks, including message exchanges and conversations.

System Evaluation

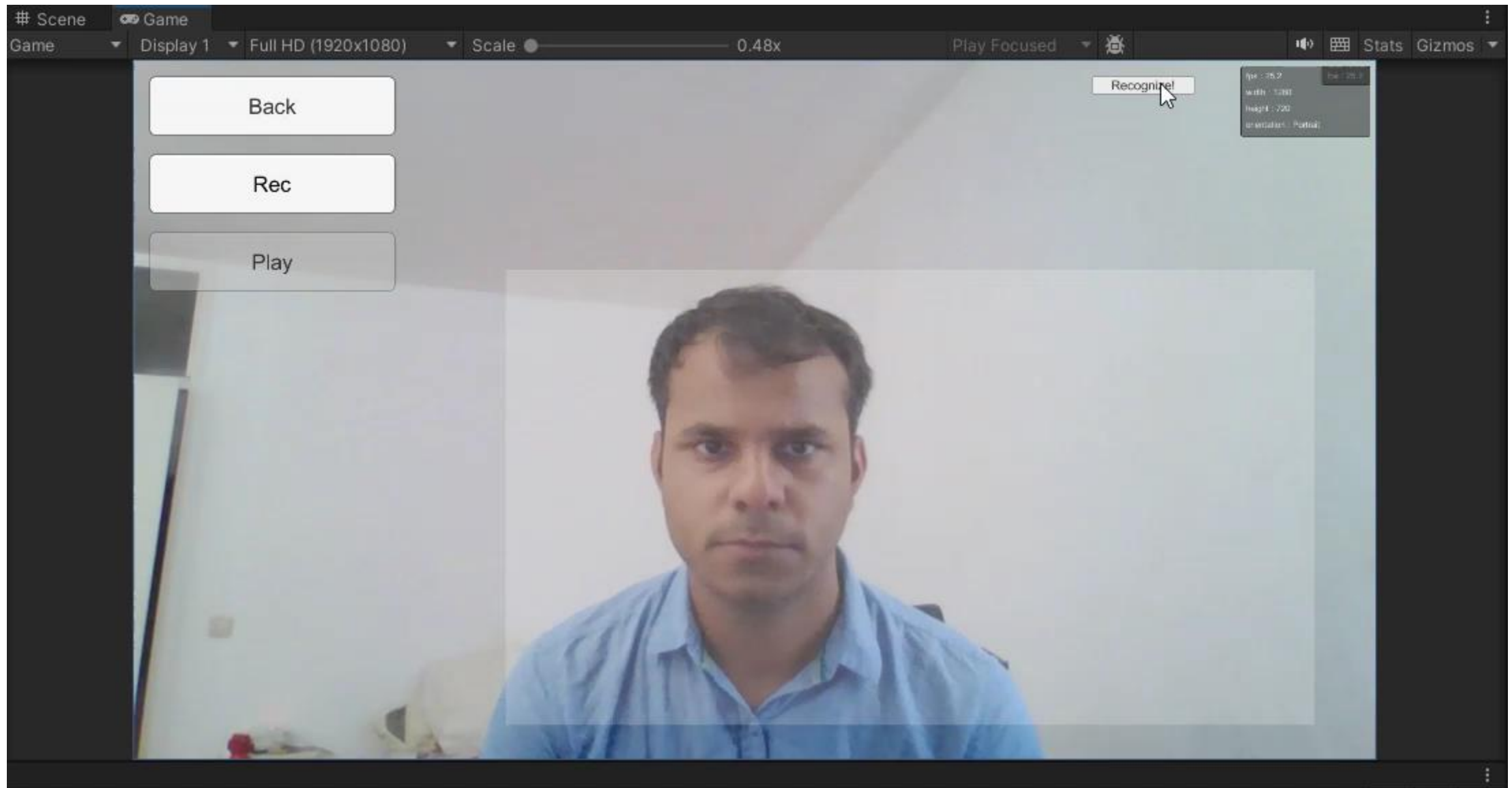
- Evaluation Metrics

- Task Success Rates: Measures success in communication tasks, including message exchanges and conversations.
- Customization Engagement: Assesses user interaction with customization options (font size, style, color, text position) for accessibility.
- User Satisfaction Surveys: Gathers feedback on system accuracy, usability, and user perceptions, identifying areas for improvement and additional feature requests.

Future Work

- LLM for Multimodal Fusion
 - Integrate a Large Language Model to compare outputs from Automatic Speech Recognition (ASR) and Automatic Lip Reading (ALR) to predict the most accurate sentence.
- Deployment on HoloLens 2
 - Integrate a Large Language Model to compare outputs from Automatic Speech Recognition (ASR) and Automatic Lip Reading (ALR) to predict the most accurate sentence.
- Edge Computing Approach
 - Use cloud servers for computational tasks due to HoloLens 2's limited processing power.
- User Studies
 - Conduct future evaluations to assess the system's effectiveness and suitability for real-world use.

Video Demo #1



Video Demo #1

