To appear in the IEEE Proceedings of 27th IEEE Virtual Reality Conference (VR '20), Atlanta, USA.

Addressing Deaf or Hard-of-Hearing People in Avatar-Based Mixed Reality Collaboration Systems

Kristoffer Waldow* Arnulph Fuhrmann †

Computer Graphics Group
TH Köln, Germany

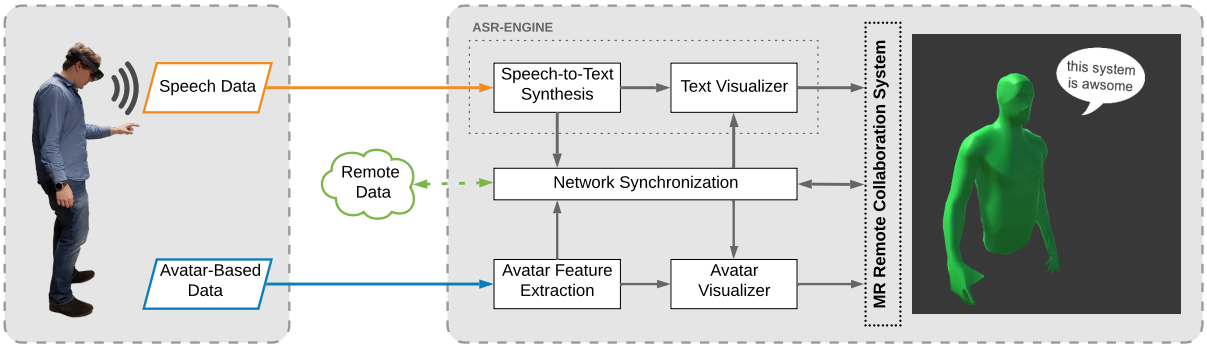

Figure 1: The abstract overview of our **MASR**-engine (Mixed Reality Audio Speech Recognition) embedded in our previously published MR remote collaboration system [6, 7]. Speech is recorded via the integrated microphone of the HMD and further processed to text, that is send over network to be visualized on a remote client. This enables deaf or hard-of-hearing people to participate in MR remote scenarios.

Abstract

Automatic Speech Recognition (ASR) technologies can be used to address people with auditory disabilities by integrating them in an interpersonal communication via textual visualization of speech. Especially in avatar-based Mixed Reality (MR) remote collaboration systems, speech is an important additional modality and allows natural human interaction. Therefore, we propose an easy to integrate ASR and textual visualization extension for an avatar-based MR remote collaboration system that visualizes speech via spatial floating speech bubbles. In a small pilot study, we achieved word accuracy of our extension of 97% by measuring the widely used word error rate.

Index Terms: Human-centered computing—HCI theory, concepts and models; Human-centered computing—Systems and tools for interaction design;

1 Introduction

More and more technologies and research arise that address people with disabilities in everyday applications [5]. Mixed Reality (MR) applications allow people to communicate and collaborate remotely in a natural way [7]. These remote collaboration systems connect people who are not physically able to do so. Considering the deaf and hearing impaired, another challenge emerges. They mostly communicate via sign language or lip-reading which is difficult to handle in MR remote scenarios. For natural communication a correct representation of gestures and facial expressions would be required in such remote scenarios. But often, the lack of accuracy in tracking does not allow to faithfully reproduce those communication channels. As a consequence, speech becomes the dominant information channel. As a result, recent studies show that Automatic Speech Recognition (ASR) technologies can be used to address people with auditory disabilities by visualizing the recognized words [1, 4, 5]. However, there exist specially designed ASR cloud-based technologies that make it easy to access and use such a recognition engine. But those services often cost money or do not support real-time recognition. On the other hand, many open source implementations are hard to integrate into already existing systems.

Therefore, we propose an avatar-based MR remote collaboration system that visualizes speech via spatial floating individual speech bubbles using an easy to integrate MR Automatic Speech Recognition extension called **MASR**. Thereby, we address deaf or hard-of-hearing people and let them participate in shared virtual environments while normal hearing people are not distracted by visual overload.

*e-mail: kristoffer.waldow@th-koeln.de
†e-mail: arnulph.fuhrmann@th-koeln.de

2 System

For our ASR system, we combined a visual textual representation system with an easy to integrate ASR engine. An MR collaboration system that uses avatar-mediated communication was used for integration [6]. It already contains a lightweight interface for network communication that can easily be accessed. For our MASR-engine, only events were synced across the network when speech is recognized, currently spoken or stopped. The data is then processed further for visual representation via floating speech bubbles in *Unity3D*.

2.1 Speech-to-Text Synthesis

The goal of our speech-to-text synthesis is to make it as easy as possible to use and integrate. The recognition quality is based on multiple factors, e.g. the audio receiving engine or the digital converter. That is why we followed a lightweight approach.

As a base, we used the Windows 10 OS integrated speech recognition engine. Most of the Head-Mounted-Displays (HMD) in VR and

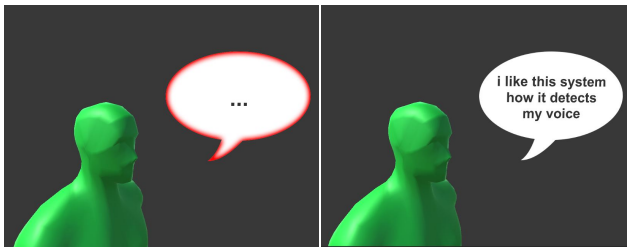

Figure 2: The speech bubbles. (Left) Recognizing state when speaking. (Right) Final recognized text.

AR are already connected to a Windows OS by reason of compatibility. A built-in microphone is integrated in most of the HMDs (HTC Vive, Valve Index, Oculus Rift S, Microsoft HoloLens). Hence, speech can easily be recorded. The *Windows SpeechRecognition*-library is able to record and interpret the speech automatically and notifies subscribed callbacks with the processed data and state. Additionally, *Unity3D* already offers a useful interface to handle this library with callbacks, like in our case the textual visualization.

2.2 Text Visualization

We decided to use a naïve strategy of textual representation in space: Floating Speech Bubbles (FSBs). This representation is easy to understand because people know the meaning and understand the context immediately. Our FSBs have the following behaviors:

- An FSB sticks to the position of an avatar and rotates planar towards the camera for best readability.
- When the remote ASR-engine is currently recognizing speech and it is not finally processed, the FSB is slightly pulsing in size and a red outline is represented as well as three dots (see Figure 2 left).
- After finishing the data processing, the pulsing and red outline disappears. A char-by-char animation is played that guarantees a visual flow. Therefore, people are visually more attracted (see Figure 2 right).
- Finally, after a certain amount of time (depending on the recognized text length), the FSB fades away and creates room for other FSBs without visually overloading the scene.

3 Evaluation

A common tool to evaluate ASR-engines is the word error rate (WER). It is based on comparing the recognized spoken sentence with a target sentence. Word substitutions (S), deletions (D) and insertions (I) are counted to determine an error rate based on the length of the sentence (N). We used a weighted calculation for D & I where substitution is more critical [3]:

$$W_{acc} = (1 - \frac{S + \frac{1}{2}D + \frac{1}{2}I}{N}) \cdot 100 \quad (1)$$

To guarantee a reliable system the word accuracy W_{acc} should be above 95.

We conducted a small pilot study ($n = 5$) in VR where people wore a Valve Index HMD connected to a Windows OS. In VR and through the HMD's integrated microphone they had to read ten pre-defined different sentences in their native language while the MASR engine hypothesized the results. By interpreting the individual results as one large text per subject, we achieved an $M_{Wacc} = 97.2$ ($SD = 5.8$).

4 Limitations, Future Work and Conclusion

As a drawback, our system only works with Windows OS. This can bee seen as a limitation but most MR systems already use this operating system. Nevertheless, a big advantage of using Windows in combination with the integrated MASR is that it automatically supports multiple languages out of the box (determined by the systems standard language, e.g. English, German, Spanish, Mandarin, etc.). In combination with the MR remote telepresence system, a variety of scenarios can be addressed.

Currently, communication is only possible in one direction. There are hearing impaired people who can still speak, but for those who cannot, new methods need to be evaluated. A solution could be, for example, prepared answers that only need to be selected.

Another problem is the readability of the text. The problem can be minimized with distance independent fonts. We chose a naïve approach of black text in front of a white background. Based on the pixel layout, some HMDs have a higher resolution for green content (e.g. Pentile Matrix). We made sure that the readability was constant throughout the pilot study.

But speech is not only an auditive representation of textual information. The intonation and sound level are as important as the textual information itself. That is why sign language is the most expressive and accurate way to communicate with deaf or hard-of-hearing people. It is faster to understand and is more expressive compared to a simple textual representation. Using avatars as mediators for synthesized sign language can increase communication quality significantly [2] and so the effectiveness of collaboration increases.

For future work, we plan to perform a user study that uses the extended system in a remote collaborative scenario in VR. In the experiment hearing impaired and normal hearing subjects will be compared while the floating speech bubbles are either enabled or not.

To conclude, by using our system, people with auditory disabilities can now be integrated into future evaluation of remote MR scenarios.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant number 02K16C232 as part of the project *Retail 4.0*.

References

[1] I. Dabran, T. Avny, E. Singher, and H. Ben Danan. Augmented reality speech recognition for the hearing impaired. In *2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*, pp. 1–4.

[2] S. Halawani and A. Zaitun. An avatar based translation system from arabic speech to arabic sign language for deaf people. *International Journal of Information Science and Education*, 2(1):13–20, 2012.

[3] M. J. Hunt. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336, 1990.

[4] M. R. Mirzaei, S. Ghorshi, and M. Mortazavi. Combining augmented reality and speech technologies to help deaf and hard of hearing people. In *2012 14th Symposium on Virtual and Augmented Reality*, pp. 174–181. IEEE, 2012.

[5] M. R. Mirzaei, S. Ghorshi, and M. Mortazavi. Audio-visual speech recognition techniques in augmented reality environments. *The Visual Computer*, 30(3):245–257, 2014.

[6] K. Waldow and A. Fuhrmann. Using MQTT for platform independent remote mixed reality collaboration. In *Proceedings of the Mensch und Computer 2019 Workshop on User-Embodied Interaction in Virtual Reality*, 2019.

[7] K. Waldow, A. Fuhrmann, and S. M. Grünvogel. Investigating the effect of embodied visualization in remote collaborative augmented reality. In *Virtual Reality and Augmented Reality. EuroVR 2019.*, pp. 246–262. Springer, 2019.