The American Otological Society

CLINICIAN SCIENTIST AWARD 2020-2023

"Use of Multisensory Input and Deep Learning Techniques to Develop a Next Generation Listening Device to Improve Speech Perception in Noise for Individuals with Hearing Loss"

Gavriel D. Kohlberg, MD
Assistant Professor, University of Washington
Department of Otolaryngology – Head and Neck Surgery

AMOUNT AWARDED BY AOS: $240,000.00

ONGOING FUNDING: N/A

PUBLICATIONS: Assessing the Relative Benefit of Real Time Captioning for Speech in Noise Benefit Gavriel D. Kohlberg, MD, Yi Shen, PhD, Adrian KC Lee, ScD, Jay T Rubinstein MD, PhD, Les E Atlas, PhD, Richard A Wright, PhD (in preparation for IEEE/ACM Transactions on Audio, Speech, and Language Processing)

RESEARCH SUMMARY: Difficulty understanding speech in background noise is a common problem that affects millions of people, including those with hearing loss and older adults. A potential solution to improve speech perception in background noise is the use of communication devices that provide computer-generated real-time captioning of the speech that can be read by the listener. Commercial automated speech recognition programs that convert speech into text are now readily available in mobile devices. The long-term goal of the project is to assess such communication devices to evaluate how they will be used and who they will benefit.

OUTCOMES: To improve speech perception in noise, the acoustic speech information can be supplemented with visually displayed speech text derived from the output of an automated speech recognition (ASR) program. It is not known if ASR generated speech text improves speech perception for listeners in more realistic settings. We evaluated speech perception in 10 normal hearing listeners and 15 subjects with hearing loss under three conditions: during auditory information alone, during speech text generated from an ASR alone, and in the combined condition with auditory information and text. At signal to noise ratios of +6, +4 and +2 dB, we found that listeners with hearing loss performed significantly better in the combined condition compared to the auditory condition alone at these noise levels *(Figure 1)*

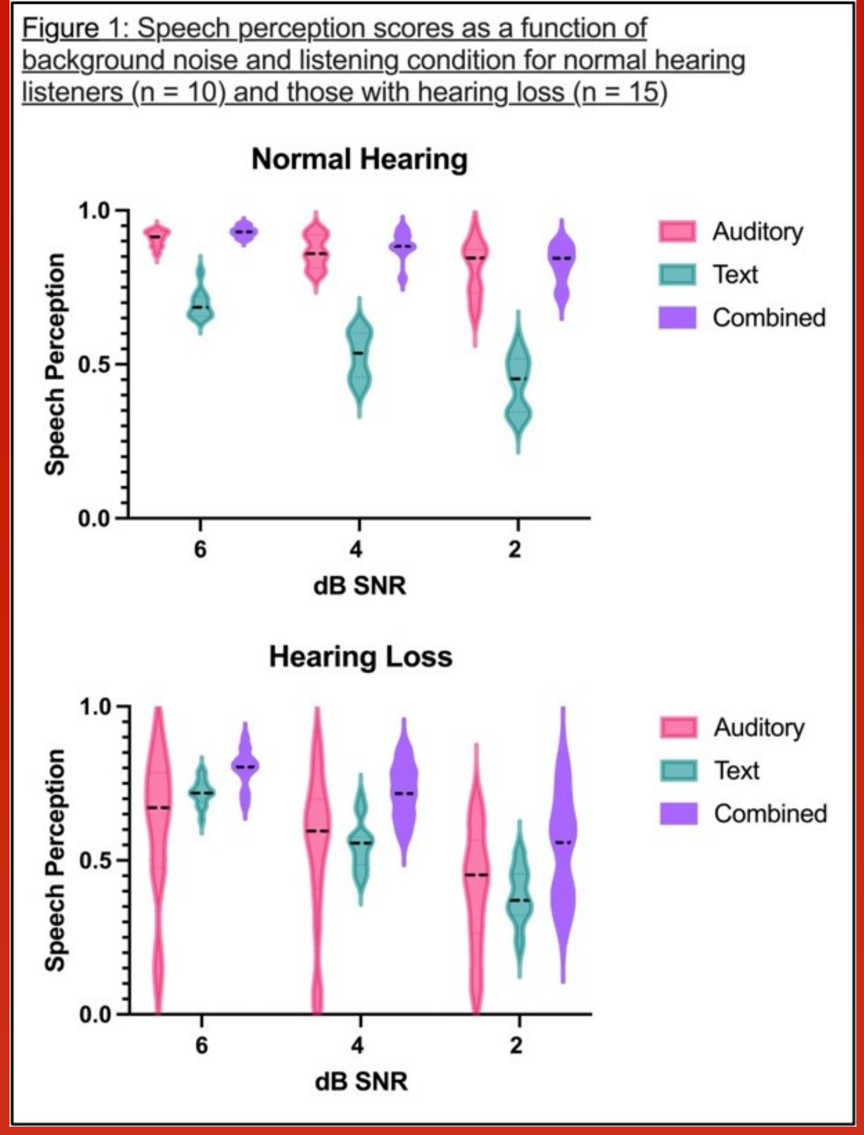


Figure 1: Speech perception scores as a function of background noise and listening condition for normal hearing listeners (n = 10) and those with hearing loss (n = 15)

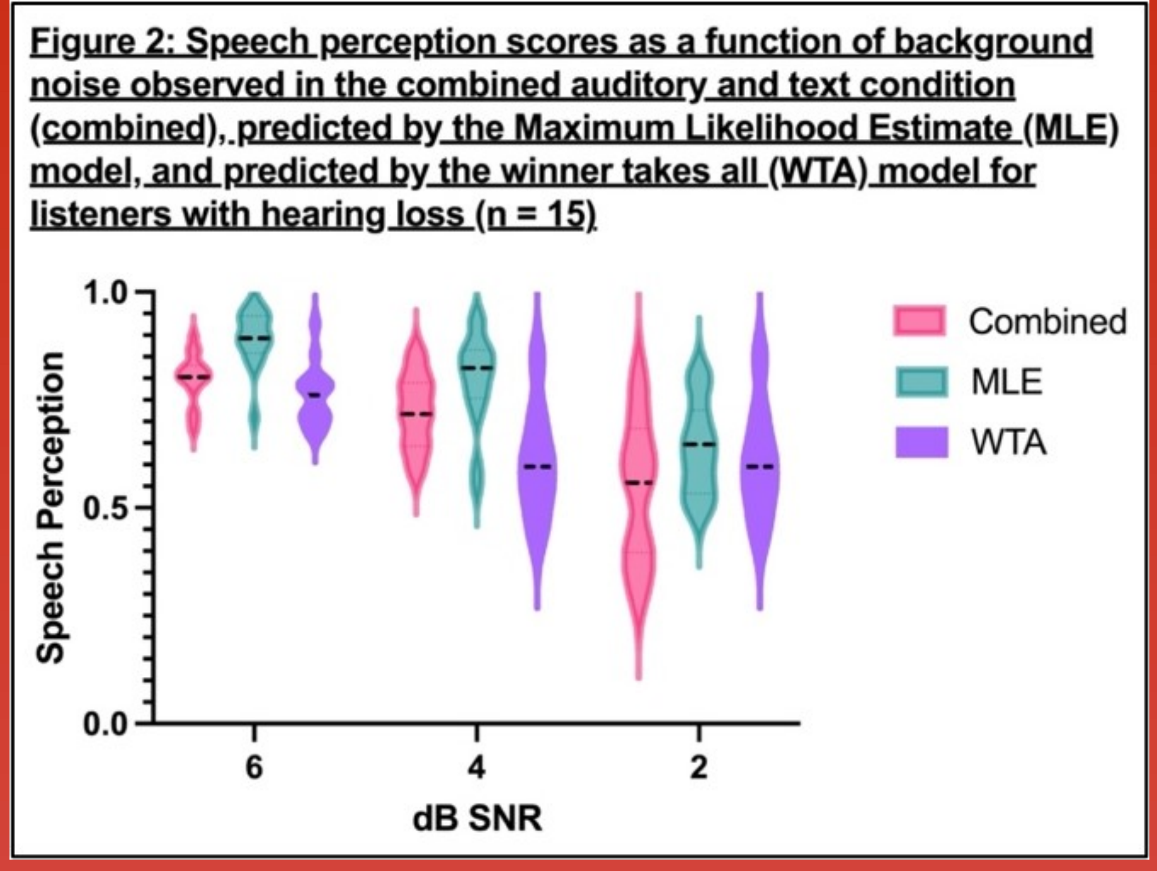


Figure 2: Speech perception scores as a function of background noise observed in the combined auditory and text condition (combined), predicted by the Maximum Likelihood Estimate (MLE) model, and predicted by the winner takes all (WTA) model for listeners with hearing loss (n = 15)

It is unknown how listeners combine these two modalities of speech information to achieve improved speech perception. For instance, it is not clear how much of a synergistic effect there is from presenting information simultaneously through these two modalities. The maximum likelihood estimation (MLE) model assumes a synergistic effect while the winner takes all (WTA) model assumes that the listener only utilizes information from the more reliable information stream. We evaluated 15 listeners with hearing to see if MLE or WTA models predicted their performance on the combined condition *(Figure 2).* We found that MLE overestimated their performance, and that WTA underestimated their performance at two noise levels and overestimated their performance at the noisiest level.

FURTHER FUNDING HAS ENABLED US TO EXPAND OUR RESEARCH TO: Funding has been applied for but not yet achieved for the following study : The ability to perceive speech in background noise decreases as the noise increases. Either access to facial cues or access to accurate captioning improves speech perception in noise. The question arises as to how listeners can integrate two competing visual cues – facial cues and text captioning – in particular – only partially ASR generated speech text. In addition, gaze behavior can be evaluated with eye tracking technology to capture where the listener is looking – at facial cues or captioning. Studying gaze behavior in this trimodal condition (auditory, facial cues and speech text) has the potential to elucidate how listeners handle competing visual information sources. Figure 3 illustrates the experimental set-up with a video of the speaker's face and captioning simultaneously displayed on a screen for a listener. Figure 4 shows the gaze behavior of the study subject.

Figure 3: An example image from video with facial cures and speech text with blue boxes marking regions of interest of the face and the text captioning.

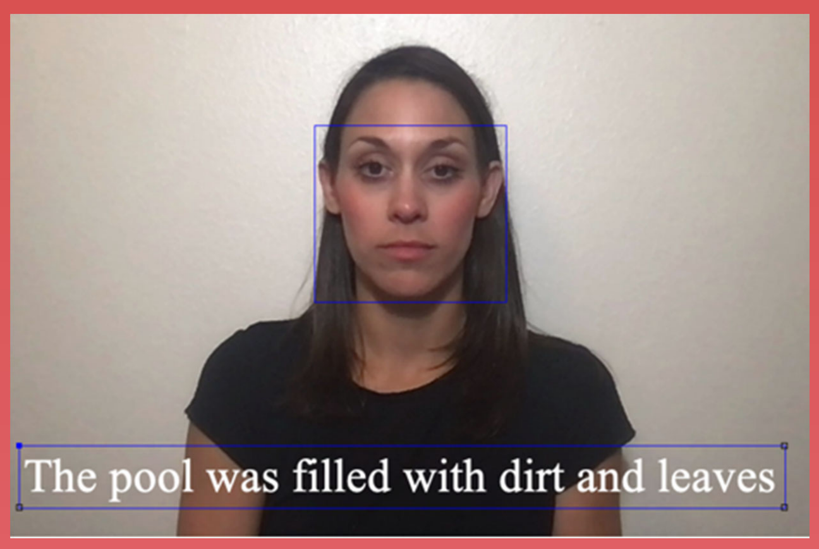
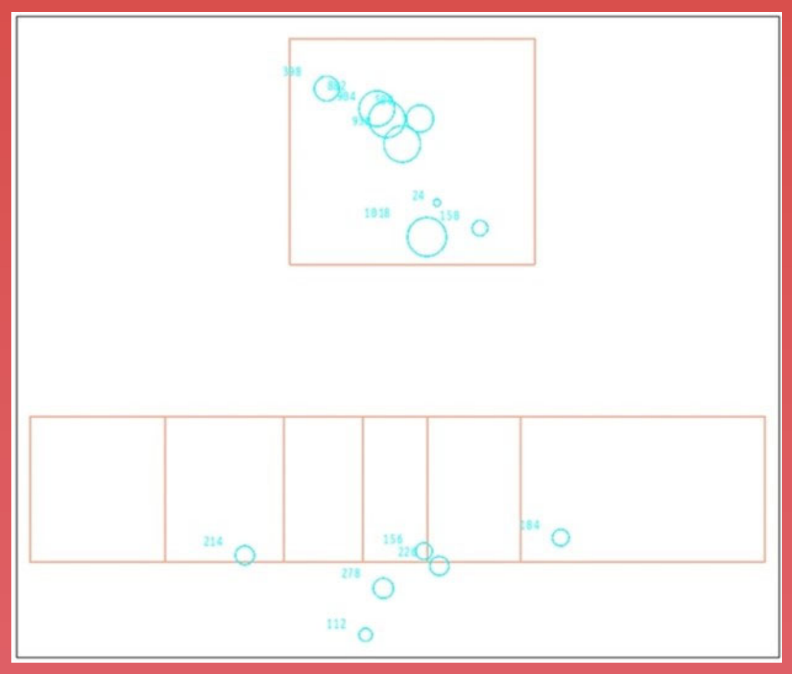


Figure 4: Example gaze data is shown for a sentence in which facial cues and partially accurate speech text were presented in -3 dB SNR. Each blue circle is a single gaze fixation, and the diameter of the blue circle represents duration of the fixation.



LAY SUMMARY OF FINDINGS AND IMPLICATIONS OF THIS RESEARCH: Difficulty understanding speech in background noise is a common problem, especially for the more than 30 million people who suffer from hearing loss in the United States. Real-time captioning provided by automated speech recognition programs has the potential to provide significant benefit to listeners in background noise. The research proposed here will investigate which listeners will benefit from captioning and how listeners integrate such captioning with auditory information and facial cues from the speaker.