# Displaying Confidence From Imperfect Automatic Speech Recognition For Captioning

R·I·T

Larwan Berke, GCCIS Ph.D. Student – larwan.berke@mail.rit.edu
Rochester Institute of Technology (RIT)
Advisor: Matt Huenerfauth, Associate Professor – matt.huenerfauth@rit.edu

## Abstract

As the accuracy and latency of Automatic Speech Recognition (ASR) technology improves over time, it may become a viable method for transcribing audio input in real-time for specific situations.

Such technology can provide access to spoken language for people who are Deaf or Hard of Hearing (DHH). However, ASR is imperfect and will remain in that state for a while, thus there is a need for users to cope with errors in the output.

My research focuses on how to best present captions which make use of the ASR system's word-level confidence. Some findings from a preliminary study is presented along with a description of the proposed solution.

## Motivation

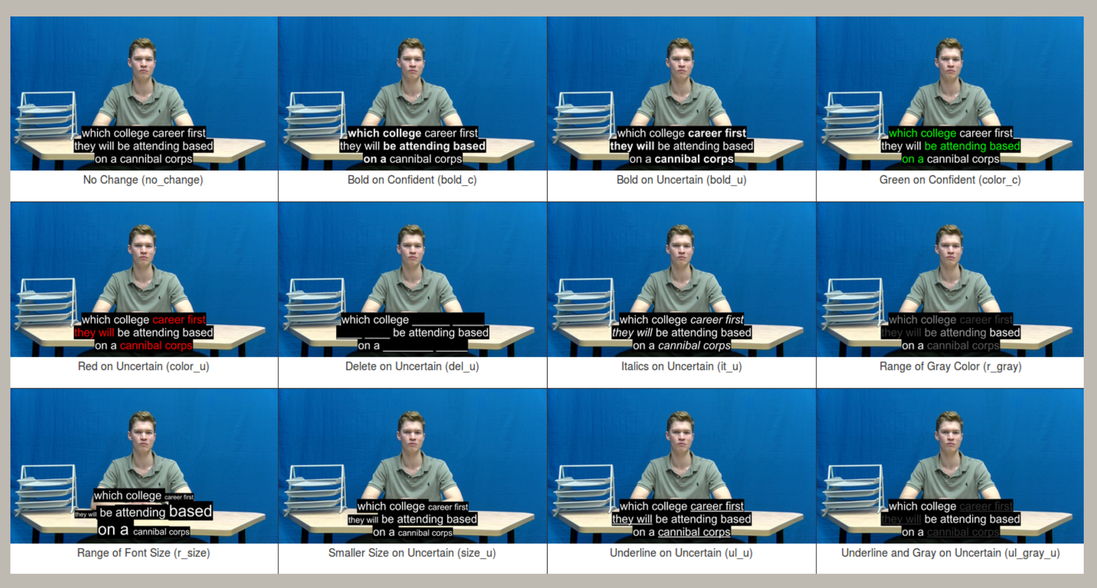
Advances in speech and language technology can benefit people who are DHH by enabling access to information in the form of spoken language. Many of those users currently benefit from a wide variety of services such as American Sign Language interpretation but in some situations DHH individuals would not be able to use interpreting due to their high cost or limited availability.

Several researchers have already investigated the application of ASR for DHH users and did not receive satisfactory results due to the high Word Error Rate (WER) in the output due to factors such as ambient noise and multiple speakers. In contrast, my proposed research selects a context which we suspect would work better: live one-on-one meeting between a DHH individual and a hearing person. My proposed solution is to display the ASR results on a mobile tablet device viewable by the DHH participant during the meeting.

## Preliminary Study

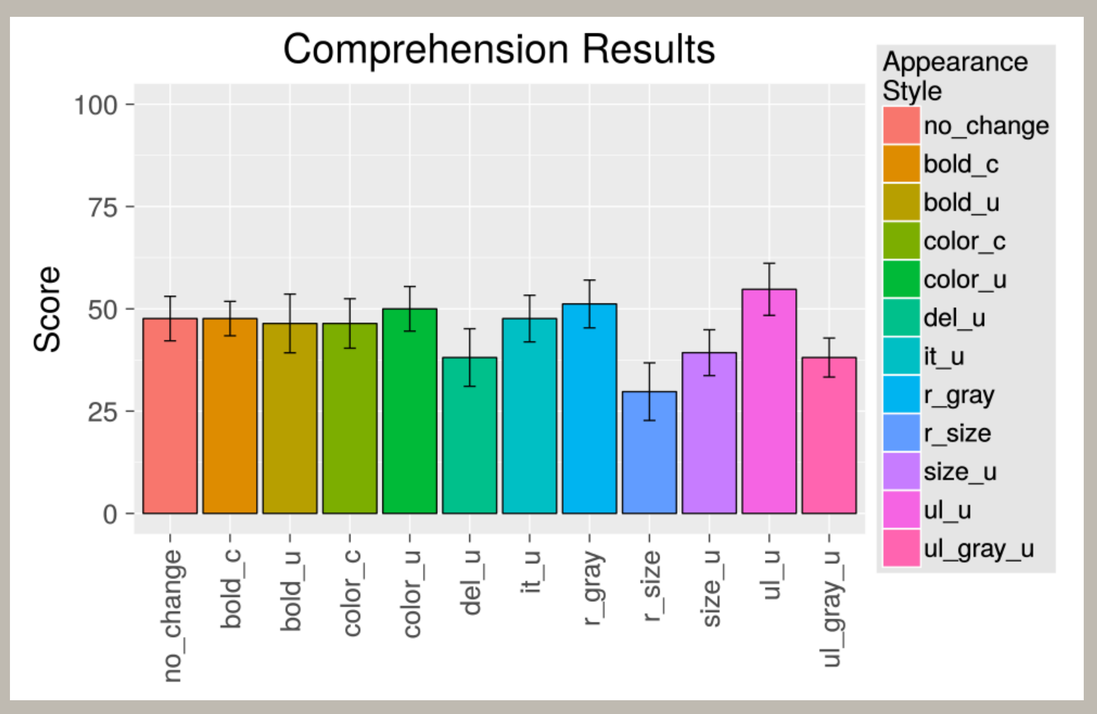
We built a prototype tool in order to explore this problem space and to answer our own research questions. After a literature review, we settled on a list of 12 caption-appearance conditions that we could apply to the output of the ASR engine based on the confidence value. We then designed and coded a PHP application that allowed us to display videos of a mock meeting with ASR captions. DHH participants viewed the videos on a laptop then answered comprehension and preference questions for the content and appearance style.

The 12 caption appearance styles that we implemented in our preliminary study:

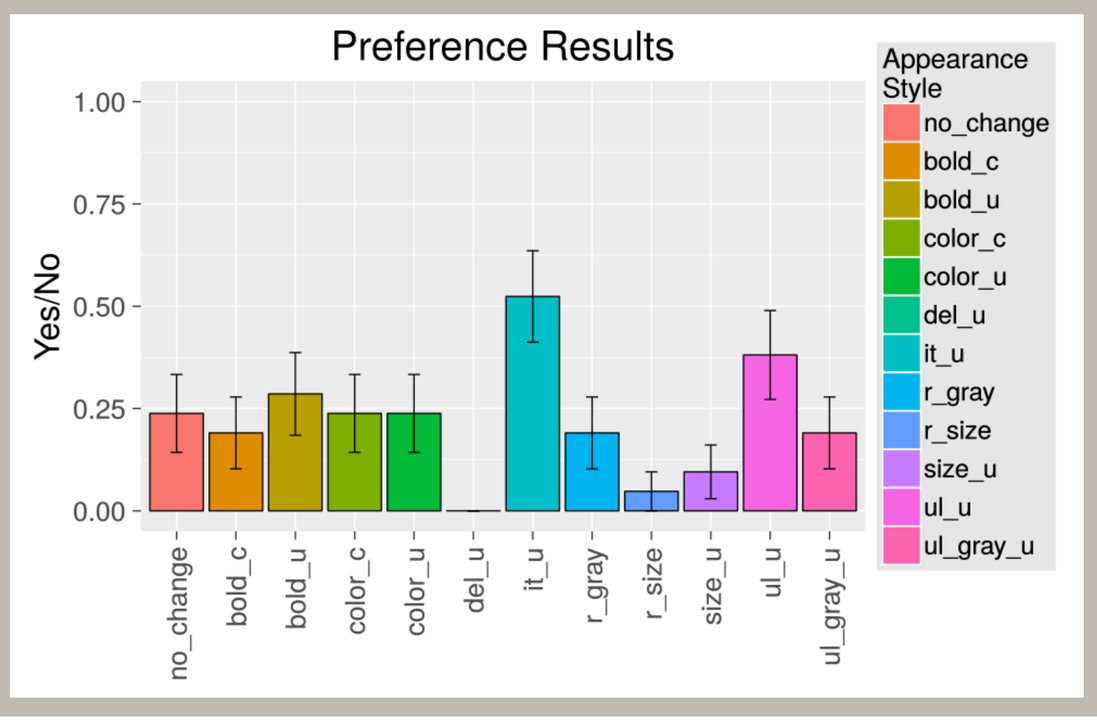



Actual script: "which college career fairs they will be attending based on the candidate requirements"

**Comprehension Results:** We executed a preliminary study on 21 DHH students from RIT by creating a fictional meeting script and breaking it up into 12 "paragraphs" so we could use a latin squares-style experiment. Each participant answered 4 questions per paragraph relating to the spoken content, and the mean scores (scaled to 100%) for each appearance style is displayed below.




**Preference Results:** The same 21 DHH users also answered several questions asking their opinion of each appearance style. One such question was "Did you like the style of captioning in this video clip?" and the mean scores (No=0/Yes=1) for each appearance style is displayed below.




CAIR Center for Accessibility and Inclusion Research

This material is based upon work supported by the National Technical Institute for the Deaf.

The LATLab Linguistic and Assistive Technologies Laboratory