Automatic Speech Recognition Services: Deaf and Hard-of-Hearing Usability

RIT | Abraham Glasser
Golisano College of Computing and Information Sciences (GCCIS), Rochester Institute of Technology

Motivation

There has been recent growth in the popularity of **Internet-connected devices**, e.g. personal assistants or home-speakers, that use Automatic Speech Recognition (ASR) to understand **verbal commands**.

However, these interfaces are **not accessible for Deaf and Hard-of-Hearing** (DHH) individuals. Even if a human listener believes that a DHH individual's speech is intelligible, ASRs generally have **high Word Error Rates (WERs) for DHH speech**.

Methodology

Our audio dataset consist of recordings of DHH individuals reading aloud a standard set of English sentences (Clarke Sentences, consisting of lists of 10 sentences with 10 syllables each).

- Send audio to ASR engines
  - IBM
  - MS
  - MSPPT (with custom language models)
- Evaluate ASR output. The National Institute of Standards and Technology (NIST) Scoring Toolkit (SCTK). SCTK's sclite v2.10 from SCTK 2.4.11 was used in this study.

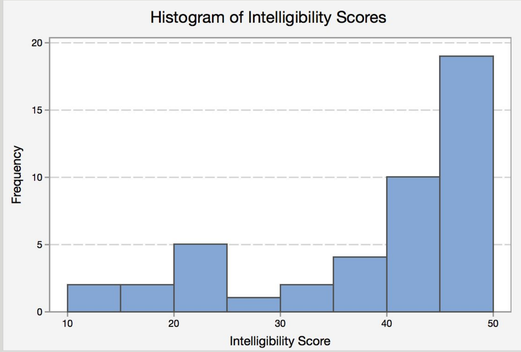

Figure 1: Distribution of Deaf and Hard-of-Hearing speakers

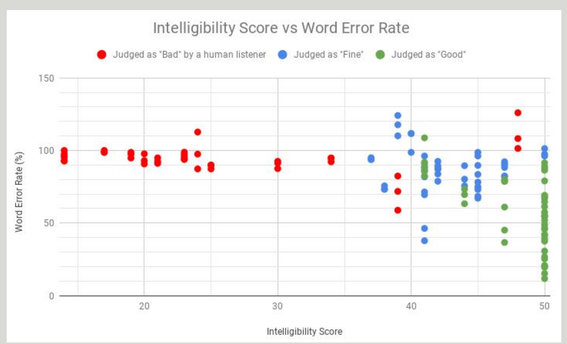

Figure 2: Word Error Rates for Deaf and Hard-of-Hearing speakers

Evaluations

Audio from individual speakers were labeled as "bad, fine, or good" based on judgements provided by a naive listener according to how understandable the speech was.

There was no significant difference (t-test, alpha = 0.05) in WER, for any of the ASR systems, when comparing the "bad" and "fine" audio categories. In Figure 3, it can be seen that these categories have a high WER and small variance, while the "good" category has a lower WER but huge variance.

- 1-sample T test gave the following 95% confidence intervals for the WER:
  - (91.338, 97.443) for "bad"
  - (82.109, 91.316) for "fine"
  - (51.288, 66.068) for "good"

Providing custom language models and context awareness for the ASR did lower WER however not significantly. The median WER for the "good" audio improved by about 10%, but the standard deviation was 20%. There was almost no improvement, if at all, between MS and MSPPT for the "bad" and "fine" audio categories..

- Lowest 2-sample T test P-value for MS vs MSPPT was .5472 for the "good" category.

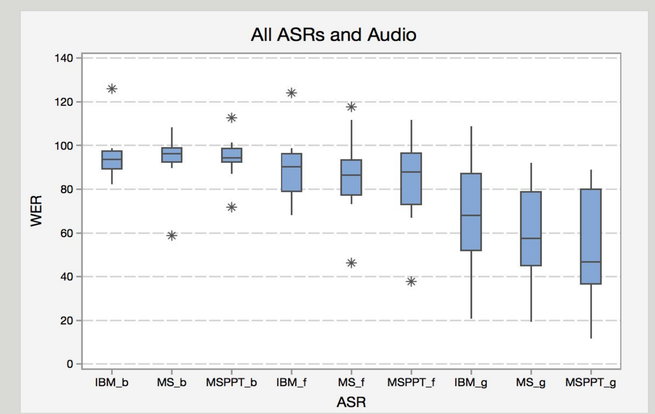

Figure 3: Side by side boxplots for all ASRs and all audio categories

Conclusions

- No significant difference if DHH speech is "bad" or "fine".
- Even if DHH speech is clear and sounds good to a naive listener, it will have unpredictable results.
- Context awareness (such as custom language models) did improve WER slightly, but it was not significant in this particular study.

With more data or training, it should be possible for ASRs to achieve consistent results (whether or not those results achieve low WERs). Once results are consistent, then opportunities for improvement via context awareness and custom dictionaries open up.

This research shows that you **cannot yet use ASRs with general DHH speech.**

This material is based upon work supported by the National Science Foundation under awards number 1822747 and REU Site 1460894.

abraham.glasser@mail.rit.edu
Contact: http://www.abrahamglasser.com