Towards More Robust Speech Interactions for Deaf and Hard-of-Hearing Users

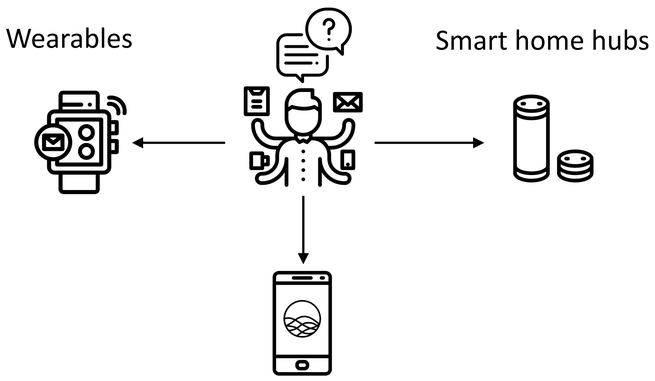
Raymond Fok, Harmanpreet Kaur, Skanda Palani, Martez E. Mott, Walter S. Lasecki

COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

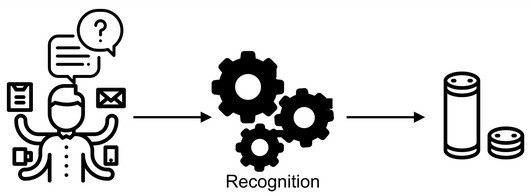
CROMA

Motivation

Technology with speech-based interactions has become more popular and is reinventing the way we think about productivity.



But these speech-based interfaces are inaccessible to users with speech impairments.



We study **deaf speech**, which has characteristics that vary greatly from that of general speech.

Differences:
- Pace of speech
- Tone and accents
- Articulation and clarity

Experiment Details

Datasets

Clarke Sentences

Alexa Commands

Metrics

$$WER = \frac{S + D + I}{N}$$ Word Error Rate

Latency

Each speaker in the Clarke Sentences dataset is given an intelligibility score from 0 – 50. We create a smaller experimental dataset by sampling clips at intelligibility levels 30, 40, and 50. The Alexa clips have no corresponding score.

Approaches

We evaluate automated and crowd-powered approaches for recognizing deaf speech.

Automated Speech Recognition

Google Cloud

We ran our dataset through Google's Cloud Speech-to-Text system.

50 **Ground Truth**: Mother gave Paul some money for a book
Transcription: mother gave her some money for the book

40 **Ground Truth**: Fourteen boys worked on a farm last summer
Transcription: what was James Bond's ranked number

30 **Ground Truth**: A puppy made a big hole in a shoe
Transcription: Muffin Man

Individual Crowd Worker

We recruited non-expert crowd workers from Amazon Mechanical Turk as individual human transcribers.

50 **Ground Truth**: A dog made a noise when the baby laughed
Transcriptions:
- The dog made a noise when the baby laughed
- What I made a noise when the baby laughed

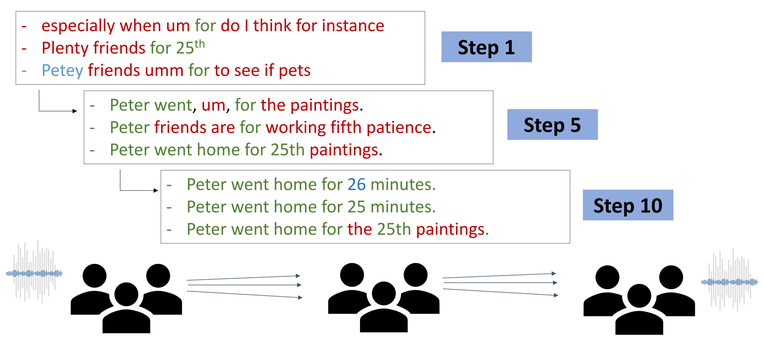
40 **Ground Truth**: Two little girls cleaned the room and played house
Transcriptions:
- She will go into the room and play house
- Three little girls cleaned their room and played house

30 **Ground Truth**: A puppy made a big hole in a shoe
Transcriptions:
- I'm backing may I'm backing two
- I'm happy about nothing at all

Complex Crowds - Iteration

We chained together crowd workers in a fully-connected iterative transcription workflow.

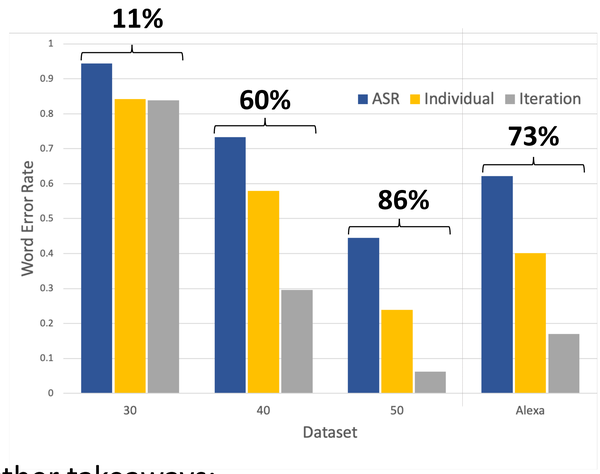
40 **Ground Truth**: Peter went home for twenty-five minutes



Results

Currently, non-expert crowd workers can individually outperform machines in recognizing deaf speech.

An iterative crowd workflow allows even further recognition over individuals alone.



Other takeaways:

- **Thematic context** helped crowd workers recognize more Alexa commands accurately (*i.e.* when they knew the clips were some command to a personal assistant device).
- **Semantic context** helped crowd workers recognize more Clarke sentences accurately (*i.e.* when they were given surrounding words to the part they had to recognize).
- Modifications in **playback speed** of the deaf speech clips had little to no effect on the recognition of that clip by crowd workers or ASR.

Future work may explore hybrid intelligent approaches for deaf speech recognition.

Periodically introduce external transcriptions (*e.g.* via ASR) to iterative flow to possibly escape local word error rate optima.

Use crowd-generated transcriptions as semi-accurate labeling feedback to improve the underlying machine learning models of ASR.