Speech in Noise Recognition by Voice to Text Applications

Rebekah Havens, B.S. and Linda Thibodeau, Ph.D.

The University of Texas at Dallas

ABSTRACT

Because of the COVID-19 pandemic and the mask mandate, persons with hearing loss struggle to hear in their everyday listening environments due to the absence of speech reading cues. The purpose of this project was to obtain results from three voice-to-text (VTT) applications (Live Transcribe, Otter, and Dragon Anywhere) that can be run on an iPhone to determine their accuracy in background noise. The applications performed similarly in quiet but varied in noise, with the Otter application providing the greatest accuracy.

INTRODUCTION

- The benefits for speech recognition from visual cues on the face has been studied for many years. Multiple studies have shown visual cues enhance speech recognition in difficult listening situations (Reisberg, McLean, and Goldfield, 1987; Sumby and Pollack, 1954; Walden et al, 1975; Thibodeau et al 2021).
- It has been recommended persons with hearing loss utilize VTT applications to assist with speech understanding.
- VTT applications may also be applied in speech recognition research during a pandemic when safety protocols restrict data collection with humans.

PURPOSE

The purpose of this study was to assess the accuracy of three VTT applications: Live Transcribe (LT), Otter (OT), and Dragon Anywhere (DA) in three conditions

1) Quiet
2) Speech noise
3) Babble noise

METHODS

Each application was tested with the HINT sentence list #4 in three conditions:

Condition 1. Quiet
Condition 2. Speech noise
Condition 3. Babble noise

Each condition was run three times with each VTT application
Responses were scored by percent word correct

The equipment and smartphone applications utilized in the study are shown in Figure 1 and Figure 2.

EQUIPMENT

Figure 1. iPhone XR and Lenovo laptop

(a) iPhone XR

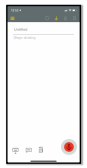
(b) Lenovo laptop

EQUIPMENT

Figure 2. Smartphone applications

Live Transcribe

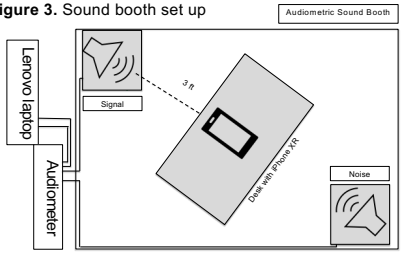
Otter



Dragon Anywhere

SET UP

Figure 3. Sound booth set up

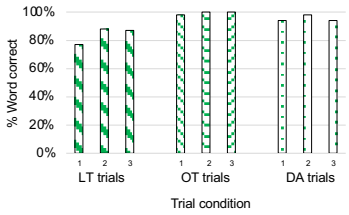


As shown in Figure 3, the physical sound booth set up is depicted. The HINT sentences are delivered 0 degrees azimuth and the noise is delivered 180 degrees to the smartphone.

The level of the HINT sentences (signal) was presented at 80 dB HL for each condition. The speech noise was presented at 55 dB HL and the babble noise was presented at 70 dB HL for each condition.

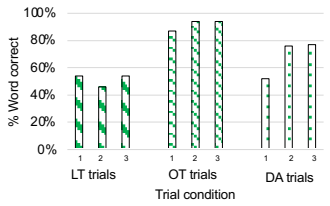
RESULTS

Figure 4. Quiet Condition



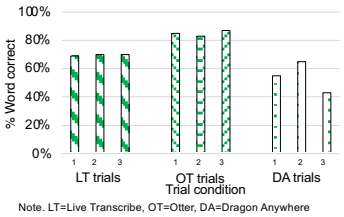
Note. LT=Live Transcribe, OT=Otter, DA=Dragon Anywhere

Figure 5. Speech Noise Condition



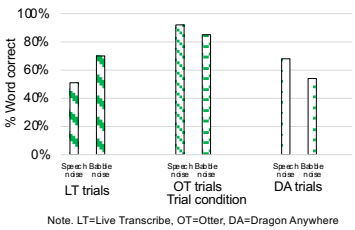
Note. LT=Live Transcribe, OT=Otter, DA=Dragon Anywhere

Figure 6. Babble Noise Condition



Note. LT=Live Transcribe, OT=Otter, DA=Dragon Anywhere

Figure 7. Noise Comparison



Note. LT=Live Transcribe, OT=Otter, DA=Dragon Anywhere

DISCUSSION

- The accuracy of 3 voice to text (VTT) applications that run on smartphones using the HINT sentences presented in quiet and in noise: Live Transcribe (LT), Otter (OT), and Dragon Anywhere (DA) is shown in Figures 4-7.
- In Quiet, OT and DA converted the speech to text with nearly 100% accuracy. LT was near 80% accuracy.
- In competition, all three apps did better with continuous speech noise compared to babble. OT showed the best overall performance, reaching 99% accuracy for speech noise. For the babble condition, OT achieved an 85% accuracy, compared to 54% and 27% for DA and LT, respectively.

CONCLUSION

- Due to the COVID-19 pandemic and the mask mandate, given the use of face masks during the pandemic which restrict audio and visual cues for the listener, the use of VVT apps as a viable solution is supported by this research.
- Of the three VTT applications, Otter yielded the highest accuracy and could be used to facilitate communication.
- These results also suggest the VTT smartphone apps could be used in speech recognition research.
- Rather than a person responding to the speech presented, the Otter application could provide the response.
- More research is needed to determine if results with this VTT application would correlate with human performance across speech recognition conditions that may vary by signal-to-noise ratio or assistive technology type.

ACKNOWLEDGEMENTS

This research is supported by National Institutes of Health (NIH)-National Institute on Deafness and Other communication Disorders (NIDCD) under award number R01DC015430. The content of this website is solely the responsibility of the contributors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (p. 97–113). Lawrence Erlbaum Associates, Inc.

Sumby W, Pollack I. (1954) Visual contributions to speech intelligibility in noise. J Acoust Soc Am 26:212–215.

Thibodeau, L., Nielsen-Thibodeau, R., Tran, C., & Jacobs, R. (2021). Communicating during COVID-19: The effect of transparent masks for speech recognition in noise. Submitted for publication.

Walden BE, Prosek RA, Worthington DW. (1975) Auditory and audiovisual feature transmission in hearing- impaired adults. J Speech Hear Res 18:272–280.

For more information please email: rebekah.havens@utdallas.edu