Application of Automatic Speech Recognition Technology for Dysphonic Speech Assessment

Presenters: Hannah Li, Theresa Murphy, Emily Heuck, Emily Demick, and Keiko Ishikawa, PhD

Department of Speech and Hearing Science, College of Applied Health Science, University of Illinois at Urbana-Champaign

ILLINOIS

INTRODUCTION

Dysphonia (AKA voice disorders): a broad term that encompasses any individual with a voice quality that varies from the norm based on their demographics (American Speech-Language Hearing Association, n.d.b).

- Affects 3-9% of the U.S. population, although many people with dysphonia do not seek treatment (Ramig & Verdolini, 1998; Roy, Merrill, Gray, & Smith, 2005)
- **Causes**: Abnormal vocal fold structure and function due to injury and/or growth on the vocal folds and neurological disorders
- **Symptoms**: rough, strangled, hoarse, or gurgly voice qualities that result in decreased intelligibility

Intelligibility: how well a speaker can be understood

- Very important in assessment, because the foundation of communication is to understand and be understood (Kent, Miolo, & Bloedel, 1994)
- Can be used in assessment to evaluate the need for intervention (ASHA, n.d.a)
- Current intelligibility assessment methods (Kent et al., 1994)
  - Use of pictures or words on cards, which the client reads/names and the listener judges and scores
  - Conversation or speech sample that is scored based on percentage of intelligible utterances
- Should be a major part of a dysphonic speaker's assessment. However, intelligibility is not routinely measured. Transcribing unintelligible speech manually is an expensive, time-consuming process which discourages regular use (Bazillon, Esteve, & Luzzati, 2008).

Automatic speech recognition (ASR): receives acoustic input and produces a text output

- ASR could provide a more consistent and efficient way to evaluate dysphonic speakers.
- Assisted transcription with the use of an automatic speech recognition (ASR) system can be up to four times faster than manual transcription of prepared speech (Bazillon et al., 2008)

Potential solution: ASR as a more efficient transcription tool for clinical use in assessing intelligibility of dysphonic speakers

AIM

- **Goal:** to evaluate the feasibility of ASR for dysphonic speech assessment. To do this, we examined the accuracy of an ASR system to transcribe normal vs. dysphonic speech
- **Hypothesis**: dysphonic speech transcription will have a lower confidence level, greater number of alternative words, and higher error rate, and as compared to normal speech.

University of CINCINNATI

METHOD

Participants

- 53 female adult participants--30 speakers with normal voice and 23 speakers with dysphonic voice as diagnosed by a speech language pathologist and laryngologist
- All native speakers of American English with no other communication disorders, including hearing loss

Instrumentation

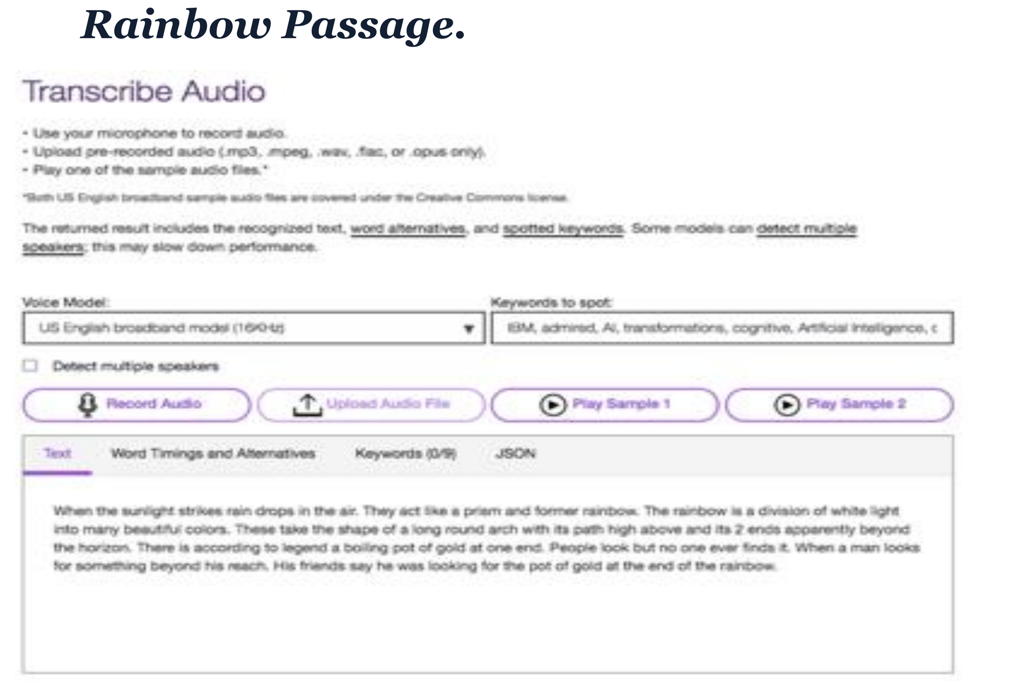
- IBM Watson: speech-to-text service (IBM Watson, n.d.)
- We chose this specific software because it allows transcription of uploaded, pre-recorded audio files
- This allowed the speech samples to be recorded in a controlled environment, and that exact sound file could be transcribed, eliminating many discrepancies between speakers.
- Alternative software: Google Cloud Speech Application Programming Interface (API)--this does not allow transcription of uploaded, pre-recorded audio files. It only transcribes live audio.

Measures

- Confidence level
  - IBM Watson's estimation that the transcribed word is correct (IBM Cloud Docs, n.d.)
- Number of alternative words
  - Gives a hypothesis for acoustically similar words to the audio input (IBM Cloud Docs, n.d.)
- Error rate (number of incorrect words divided by the total number of words)

Procedures

- **Speech recording**
  - Participants were recorded using a unidirectional microphone in a soundproof room.
  - The microphone was placed at a distance of 15cm away from the mouth at a 45 degree angle.
  - Each speaker was recorded while stating the ***Rainbow Passage.***



IBM Watson transcription of the Rainbow Passage sound file (Figure 1)

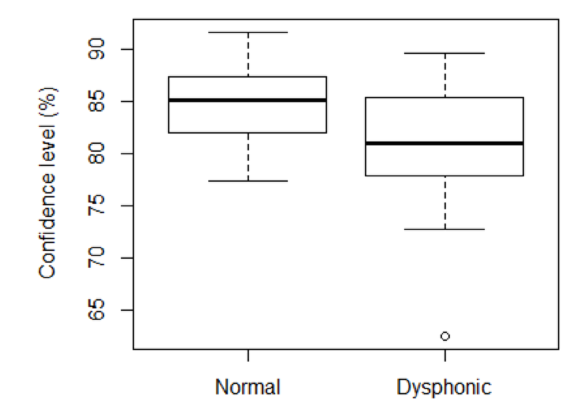
IBM Watson transcription showing word timings and alternatives (Figure 2)

- **Transcription**
  - Each sound file was transcribed through IBM Watson Speech to Text Service, producing a text transcription, alternatives of each word, as well as the percent likelihood of each alternative.
  - Two experimenters worked on every sound file to minimize human error and determine if software transcribed speech consistently

RESULTS

Confidence level

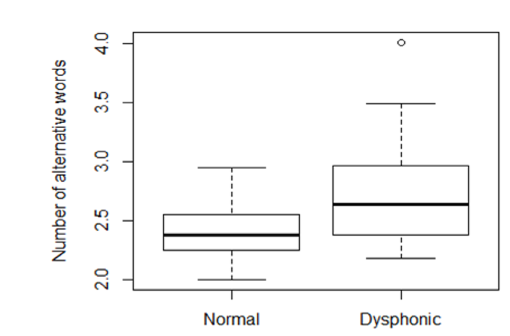
- In comparison to normal speech, the confidence level of transcribed words is significantly lower in dysphonic speech (p = 0.028)



Box-plot showing the confidence level of normal vs. dysphonic speakers (Figure 3)

Number of alternative words

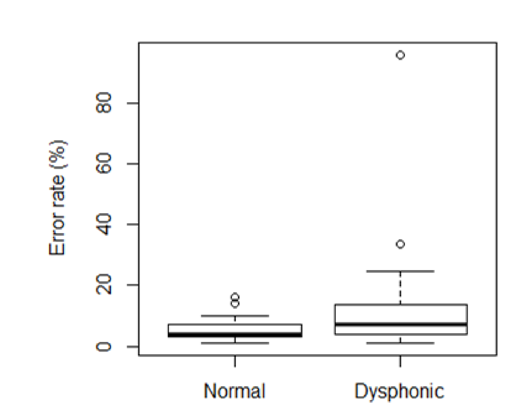
- The number of alternative words is significantly greater in dysphonic speech in comparison to normal speech (p = 0.008)



Box-plot showing the number of alternative words for normal vs. dysphonic speakers (Figure 4)

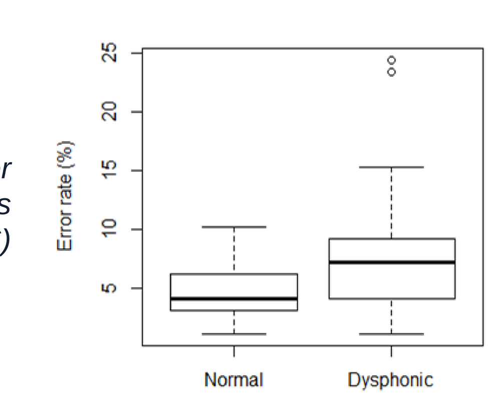
Error rate

- Error rate both with (p = 0.058) and without (p = 0.066) outliers showed no significant difference between normal and dysphonic speech.



Box-plot showing the error rate with outliers (Figure 5)

Box-plot showing the error rate without outliers (Figure 6)



DISCUSSION

Our hypothesis was partially correct.

- The confidence level for dysphonic speakers was lower, and the number of alternative words for dysphonic speakers was higher, as we predicted.
- However, there was no statistically significant difference between the error rate of dysphonic and normal speakers. If we had a larger sample size, we might have had a statistically significant difference since our p-value was very close to 0.05.
- The explanation for this could be that the software learns as it goes.
  - The Rainbow Passage is fairly long (98 total words), giving the software time to adjust.
  - Watson appeared to generate fewer alternative words in the second half of the transcription. The number of alternative words chosen for both dysphonic and normal speakers decreased significantly in the second half (31 alternatives in the first half to just 2 in the second for the dysphonic speaker DAF03; 43 alternates in the first half to just 19 in the second for normal speaker NAF07).

Overall, our study demonstrated that difference in dysphonic and normal speech can be described partially by the ASR-based measurement.

- Based on the differences seen in our results, we conclude that transcription of dysphonic speech was more challenging for the Watson speech-to-text software
- This challenge may reflect human perception of dysphonic speech (i.e. lack of intelligibility), and if so, Watson speech-to-text API would be a good platform for an automatic clinical speech analysis tool.

Limitations: our study only included data from adult women. Our research did not test the transcription abilities of ASR on adult men or children.

Future Directions:

- Evaluate performance of the program with a more diverse population.
- Examine correlation between listener's rating of intelligibility and the ASR-based measures.

REFERENCES

American Speech-Language Hearing Association. (n.d.a). Speech sound disorders-articulation and phonology. Retrieved February 25, 2018, from https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935321&section=Assessment

American Speech-Language Hearing Association. (n.d.b). Voice disorders. Retrieved February 25, 2018, from https://www.asha.org/practice-portal/clinical-topics/voice-disorders/

Bazillon, T., Esteve, Y., & Luzzati, D. (2008). Manual vs assisted transcription of prepared and spontaneous speech. *International Conference on Language Resources and Evaluation*, 1067-1071. Retrieved February 26, 2018, from http://lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf

IBM Watson. (n.d.). Speech to Text. Retrieved October 26, 2017, from https://speech-to-text-demo.ng.bluemix.net/

IBM Cloud. (n.d.). Retrieved April 12, 2018, from https://console.bluemix.net/docs/services/speech-to-text/output.html#word_confidence

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology, 3*(2), 81-95. doi:10.1044/1058-0360.0302.81

Ramig, L. O., & Verdolini, K. (1998). Treatment efficacy: Voice disorders. *Journal of Speech, Language, and Hearing Research, 41*(Suppl.), S101–S116.

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *The Laryngoscope, 115,* 1988–1995.