A Novel Telephony Recognizer for Dysarthric Speech

Hemanshu Bhargav [0000-0002-6412-6521]

Toronto Metropolitan (Formerly Ryerson) University, Toronto, Ontario, Canada
hbhargav@ryerson.ca

Keywords: ASR, CTC, Attention, Keyword Spotting, Quadriplegia, Dysarthria

1 Motivation

Single-switch users tend to have severe mobility impairments such as quadriparesis, and often suffers from motor-speech disorders such as dysarthria as well [1-2]. Such users can have difficulty answering phone calls or operating phone menu systems, because these are time-sensitive operations and single-switch scanning remains a slow process– due to its iterative or grid-based selection mechanism [3-4].
Operations such as answering a call must be done within a number of rings, and because elements are scanned orderly, the allotted time lapses [3]. The problems of frequent missed calls and menu loops are caused by a mismatch between the number of options available and the speed at which single-switch users can access them. Todman [5] reported that predictive single-switch text entry does not exceed five words per minute, so reducing keystrokes in time-sensitive operations is imperative. Swiffen et al., [6] and Matiasek et al., [7] suggest that using speech recognition and word prediction can speed up text entry by 69% for these users. Automatic Speech Recognition (ASR) can thus replace scanning technology, and Wake Word Detection (WWD), a subset of Keyword Spotting (a subset of ASR), can improve accuracy by limiting the speech vocabulary to specific telephony voice commands only [3, 8]. However, because speech production is inconsistent in dysarthric speakers, standard ASR/WWD approaches may not be usable [9]. As a result, researchers propose the use of hybrid encoder-decoder Deep Neural Network (DNN) architectures, in which two techniques, Connectionist Temporal Classification (CTC) and Attention, are combined to recognize speech [10].

1.1 Research Questions

1. What are the factors influencing recognition of dysarthric speech and limited-vocabulary language models in offline recognizers?
2. How can ASR and DNNs specific to dysarthric speech be combined to increase the selection efficiency and user satisfaction for access to smartphone telephony functions for single switch users?

2 Methodology

To accomplish this task, I have (according to my knowledge) acquired the largest collection of research-based dysarthric speech data to date (five datasets) and am leveraging the PyTorch-based toolkit SpeechBrain, a community and corporation

sponsored open-source tool which makes the entire ASR pipeline accessible and configurable for developers [11]. After training the amassed data (including new data requested by my team), I will conduct experiments on variations of Convolutional, Long Short Term Memory, Gated Recurrent Unit (GRU) and Light GRU architectures to see which combination is most accurate at recognizing telephony speech [12-15]. To my knowledge, there have not been any attempts to use a modern, hybrid encoder-decoder DNN or attention-based transformer on the volume of dysarthric speech data I have collected, regardless of its applicability to the problems of missed calls and menu loops described earlier. Although robust ASR systems are plentiful, such as Google's Project Relate, such systems are not dedicated to recognizing telephony dysarthric speech and are not optimized for the time-sensitive operation required [16]. Further, such systems do not make use of two-stage WWD, a technique I propose, which should prevent false activations, which are likely to occur due to the diffuse nature of dysarthric speech [17-18]. Once this system is complete and functional, I intend to carry out user studies with five single-switch users.

To ensure the needs of the study participants are met and that relevant factors are included in the system design, a participant researcher who is familiar with single switch interfaces and issues with dysarthric speech will advise this research (they have already approved the methodology). A user study will be carried out to evaluate the usability using the System Usability Scale [19] and post-study interview questions designed to collect satisfaction and opinions; usefulness; and efficiency, measured using time to complete each task and error; of the system. Participants will be recorded as they carry out typical telephone tasks with a smartphone.

The main outcome of my research will be one possible and novel solution to circumvent the challenges of speed and complexity found when using single-switch interfaces for telephony functions. The hybrid DNN model will be integrated into an iOS application, developed using Inclusive Design, thereby increasing equitable access to telephony functions for single-switch users. A second contribution will be the refinement and addition of new dysarthric speech data and a portable algorithm.

References

1. H. H. Koester and S. Arthanat, "Text entry rate of access interfaces used by people with physical disabilities: A systematic review," *Assistive Technology*, vol. 30, no. 3, pp. 151–163, Sep. 2018, doi: 10.1080/10400435.2017.1291544.
2. C. Havstam, M. Buchholz, and L. Hartelius, "Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control," *Logopedics Phoniatrics Vocology*, vol. 28, no. 2, pp. 81–90, Jan. 2003, doi: 10.1080/14015430310015372.
3. H. Bhargav, I. Ahmed, M. Whitfield, R. Shankar, M. Meza, and D. I. Fels, "Tecla Sound: Combining Single Switch and Speech Access," in Computers Helping People with Special Needs, Cham, 2020, pp. 348–354. doi: 10.1007/978-3-030-58805-2_41.
4. H. H. Koester and R. C. Simpson, "Method for enhancing text entry rate with single-switch scanning," *Journal of Rehabilitation Research and Development*, vol. 51, no. 6, pp. 995–1012, 2014.

5. J. Todman: Rate and quality of conversations using a text-storage AAC system: Single-case training study. Augmentative and Alternative Communication. 16, 164–179 (2000). https://doi.org/10.1080/07434610012331279024
6. A. Swiffin, J. Arnott, J. A. Pickering, and A. Newell, "Adaptive and predictive techniques in a communication prosthesis," *Augmentative and Alternative Communication*, vol. 3, no. 4, pp. 181–191, 1987, doi: 10.1080/07434618712331274499.
7. J. Matiasek, M. Baroni, and H. Trost, "FASTY — A Multi-lingual Approach to Text Prediction," *Lecture Notes in Computer Science Computers Helping People with Special Needs*, pp. 243–250, 2002.
8. G. M. Bohouta, "Improving Wake-Up-Word and General Speech Recognition Systems," Ph.D., Florida Institute of Technology, United States -- Florida. Accessed: Nov. 01, 2021. [Online]. Available: http://www.proquest.com/docview/2468704152/abstract/E7E91356B8A74538PQ/1
9. F. Rudzicz, "Production knowledge in the recognition of dysarthric speech," Ph.D., University of Toronto, Canada, 2011.
10. S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017, doi: 10.1109/JSTSP.2017.2763455.
11. M. Ravanelli et al., "SpeechBrain: A General-Purpose Speech Toolkit," arXiv:2106.04624 [cs, eess], Jun. 2021, Accessed: Jun. 30, 2021. [Online]. Available: http://arxiv.org/abs/2106.04624
12. S. Hochreiter and J. Schmidhuber, "LSTM can Solve Hard Long Time Lag Problems," in *Advances in Neural Information Processing Systems*, 1997, vol. 9. Accessed: Dec. 18, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/1996/hash/a4d2f0d23dcc84ce983ff9157f8b7f88-Abstract.html
13. K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078 [cs, stat]*, Sep. 2014, Accessed: Nov. 15, 2021. [Online]. Available: http://arxiv.org/abs/1406.1078
14. M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 92–102, Apr. 2018, doi: 10.1109/TETCI.2017.2762739.
15. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
16. J. Green *et al.,* Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. 2021, p. 4782. doi: 10.21437/Interspeech.2021-1384.
17. F. Rudzicz, G. Hirst, and P. van Lieshout, "Vocal Tract Representation in the Recognition of Cerebral Palsied Speech," *Journal of Speech, Language and Hearing Research (Online)*, vol. 55, no. 4, pp. 1190-1207A, Aug. 2012, doi: http://dx.doi.org/10.1044/1092-4388(2011/11-0223).
18. M. Wu et al., "Monophone-Based Background Modeling for Two-Stage On-Device Wake Word Detection," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2018, pp. 5494–5498. doi: 10.1109/ICASSP.2018.8462227.
19. B. Klug, "An Overview of the System Usability Scale in Library Website and System Usability Testing," Weave: Journal of Library User Experience, vol. 1, no. 6, 2017, doi: https://doi.org/10.3998/weave.12535642.0001.602