Is Speech Recognition Software a Viable Future for Dysarthric Speakers? A Critical Review

Sean Lewis
M.CLSc SLP Candidate
University of Western Ontario: School of Communication Sciences and Disorders

The idea of speech-recognition software, in its purest form opens the possibilities for individuals with dysarthria to bypass one of their biggest barriers, being difficulties with speech clarity. This critical review explored the relationship between speech to text software (STT), (specifically speaker-dependent and speaker-adaptive speech to text software), and its ability to accurately decipher the speech of individuals with dysarthria. Speech to text (STT) is defined as "software that lets the user control computer functions and dictates text by voice" (Das et al., 2015). Speaker-dependent STT is developed to understand the speech of a single individual, compared to, speaker-adaptive STT which is designed to understand the speech of different/multiple individuals (Yampolsky and Rosen, 2000). The studies analyzed included: 4 within-group studies, 1 single-participant case study, and 1 literature review. Overall, this critical review suggests individuals with dysarthria experience better accuracy with speaker-dependent software compared to speaker-adaptive for the more "severe" speakers.

Key Words: Speaker-dependent, Speaker-adaptive, dysarthria, speech recognition software, Speech to Text (STT)

Introduction

Dysarthria is a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system (Darley et al, 1969). Dysarthria causes problems in oral communication due to paralysis, weakness, or in coordination of the speech musculature (Darley et al, 1969). Over the last 20 years, amongst the abundance of research done on the many aspects of dysarthria, little has been done surrounding speech-to-text (STT) and its suitability for dysarthric speakers. Speech to text (STT) being "software that lets the user control computer functions and dictates text by voice" (Das et al., 2015).

In the world of speech-recognition, there are currently 2 recommended practices; 1.) use a speaker-dependent system; which is trained to understand a particular individual, or 2.) use a speaker-adaptive system; designed to recognize the speech of many people (Yampolsky and Rosen, 2000).

Speaker-dependent STT matches an incoming speech signal to templates created from an individual's speech (Yampolsky and Rosen, 2000). That is, with the speaker-dependent system, the user's speech is compared to templates of his/her pronunciation of a variety of words, phrases, sentences, and so on. A large benefit to this type of system is the software is less reliant on one's speech to be "typical" and more reliant on their speech being consistent (Coleman and Meyers, 1991).

On the other hand, speaker-adaptive STT software is designed to adjust to a new user without the need to train every word in the system (Yampolsky and Rosen, 2000). Like speaker-dependent systems, speaker adaptive systems also rely on templates of speech which the incoming speech signal is compared. However, the speech templates in speaker-adaptive systems are constantly updated in accordance with the incoming signals (Huang & Lee, 1991).

In both systems accuracy and success is determined based by how similar the incoming speech signal is compared to the templates that have been saved by the system. The further the incoming signal deviates from the template the greater the likelihood that the signal would be considered an "inaccurate" response.

Objective

The primary objective of this critical review was to determine whether or not STT software was effective, accurate, suitable, and practical for use by dysarthric speakers. Secondary and tertiary objectives include: 1.) determining whether or not there is merit in research in this area, and 2.) exploring whether or not STT software could be a practical application for speech and language pathologists to recommend to patients with dysarthria as a supportive communication tool to help ease the barriers of speech clarity.

Methods

Search Criteria:

Online databases: Google Scholar, PubMed, Elsevier, and JSTOR, were searched using the following terms: ((("speaker-adaptive") OR ("speaker-dependent")) OR ("dysarthria")).

Selection Criteria:

Studies were included into the critical review if they discussed and compared the effectiveness and accuracy of both speaker-dependent and speaker adaptive STT software in patients dysarthria. Patients in each of the studies had to have been identified as having dysarthria as diagnosed or laid out by a health care professional (i.e., doctor, or speech-language pathologist).

Data Collection:

Results of the literature review yielded 6 articles: four were level 2a research evidence studies; all of which were within-group experimental studies. One single case-study control (level 2b evidence), and one literature review (level 4 evidence).

Results

Within-Subjects Design:

Raghavendra, Rosengren and Hunnicutt (2001) wanted to test the feasibility of speaker-adaptive and speaker-dependent STT software as an input method for speakers with varying levels of dysarthria. This study followed a within-group design and contained 4 participants with dysarthria and one normal speaker (to act as a control). One participant had "mild" dysarthria, one had "moderate" dysarthria, and two had "severe" dysarthria. Participants were recruited by contacting local speech-language pathologists and asking for possible referrals. Prior to testing, all participants were administered the Swedish Dysarthria Test by a registered speech-language pathologist to get their baseline severity levels. Dependent variables in this study included: accuracy (% words correct) with independent variables being vocabulary size (how many words the system could remember) and severity level.

Overall, the study's results support the idea that participants with higher-levels of dysarthria (those who are more "severe") generally show more success with the speaker-dependent software. Using the speaker adaptive STT software, the "severe" participants had an accuracy rate of 38% and 26% respectively, whereas when they used the speaker-dependent systems, they achieved 70% and 28%. Alongside these results, the study also showed an increase in accuracy rates with decreased vocabulary size programmed into the STT software. That is, if the number of templates that the software needs to learn and compare the incoming speech signal is reduced, the software will be more likely to produce correct and accurate responses. When using the speaker-dependent STT software and a reduced vocabulary size (less words for the system to keep "templates" of), the participant with severe dysarthria who's accuracy score was 28% jumped to 62%. For the control, "mild" and "moderate" conditions, the speaker-adaptive system performed better on average however, the difference is rather small.

This study like many others has a clearly laid out purpose and methods sections; allowing for future replications to confirm the validity of the results. Disadvantages to this study include having a reduced sample size, reducing the overall reliability, and the possible presence of participant "fatigue". That is, over the different trials the participant's clarity may have decreased over time due to the extra work to produce speech as a consequence of their dysarthria and as a result the system's accuracy may have shown a decrease as well. With the most relevant issue of the study being the reduced participant pool, the conclusive evidence is deemed to be "suggestive" in nature.

The researchers propose that next steps include evaluating accuracy levels of STT upon adjusting for specific phonemic transcriptions in the software's lexicon made by speech variability. For example, participants who may repeat a syllable in a word, have a reduced rate of speech, or have some sort of hesitation in their speech could affect the system's accuracy of the user's message.

Rudzicz, (2007) evaluated the strengths of each type of STT software across dysarthric speakers. The author compared the accuracy (% of words correct) of speaker-dependent STT and speaker-adaptive STT software with individuals with ranging severity levels of dysarthria. The study was a within-group design which used speech from 11 dysarthric speakers (4 participants with "severe", 4 participants with "moderate", and 3 participants with "mild") as well as one control speaker with "normal" speech. Participant's speech was all taken from an online speech database called "Nemours". Participants' severity levels were determined prior to the study, by running their speech through a recognition software designed to process "standardized" transcripts. The Nemours data base contained complete and correct annotations alongside the participant speech samples which would be used cross reference the baseline system's accuracy.

Overall, the study provides evidence for speaker-dependent STT being more successful and beneficial for the "severe" dysarthric speakers compared to speaker-adaptive. It is important to note that the difference in accuracy rates for the speaker-dependent was only marginally higher for the "severe" condition compared to that of the speaker-adaptive software. This is interesting when compared to Raghavendra et al., (2001) who found a more pronounced difference between the speaker-dependent and the speaker-adaptive software's accuracy rates for the "severe" condition. However, in conjunction with Raghavendra et al., (2001) the speaker-adaptive software provided higher word accuracy rates for the "mild" and "moderate" dysarthric speakers, and the control speaker. The authors propose that moving forward, future steps are to look at the effectiveness of these speech recognition models across manipulations of consonant variations (i.e., phonemic substitutions and deletions). By this, the researchers propose that the researchers manually go into the software and make some changes so that when those phonemic errors occur it is able to "adjust" accordingly.

Structurally, the purpose of the study and the methods section have been clearly defined for the reader; the benefit being the study would be easy to replicate and validate in the future. However, there is one large disadvantage to the study which would be the rather small sample size which greatly hinders the reliability of the results. This study may have benefitted from using additional speech from other participants on the "Nemours" speech database or using participant's speech samples from other speech samples containing individuals with dysarthric speech. If there were more participants included in the study, then the conclusive evidence could be seen at a rating of "strong", but as of right now it is to be evaluated at "suggestive".

Sharma and Hasegawa-Johnson (2010) investigated the effect of modifying speaker-dependent and speaker-adaptive STT software to produce more accurate responses in individuals with spastic dysarthria. Specifically speaking, their definition of "modification" involved small changes regarding how the system software compared the incoming signal to its templates for speech. The variables evaluated were accuracy (% words correct) compared across the different software modifications. This study followed a within-group study design and evaluated the speech of 7 participants who had dysarthria. Participant's speech was taken from an online speech database called the "UA-Speech Database". Baseline severity levels were taken from averaged from the scores of a series of "unfamiliar" listeners. All participant's speech samples were evaluated to have spastic dysarthria as indicated by an informal evaluation conducted by a speech-language pathologist.

Interestingly, the results provide evidence that speaker-dependent STT software worked better before any specific modifications were made. Specifically, the original version of the speaker-dependent software had a percent word correct score of 52%, compared to the "modified" version of the speaker dependent software which had a percent word correct of 47%. Secondly, the results showed that overall, speaker-adaptive software had higher percent words correct across all severity levels (being on average 6-10% more accurate) when previous studies have indicated otherwise.

Similar to the above-mentioned articles, this study has been quite clear with its methodology in order for future researchers to easily replicate the study and provide increased validity to the results. On the contrary, since this study was pulling their participant data from an online database, they could have had an increase in the number of participants used in the study (and if not, they could have pulled from other online databases). The issue here being a decrease in the study's overall reliability of their results due to a drastic decrease in their participant pool. It is for this reason that this article's conclusive evidence be rated no higher than "suggestive".

Christensen et al., (2012) investigated the effect of using language strategies (i.e., the level of explicit teaching you program into the speech-recognition software) with speaker dependent and speaker adaptive systems in order to improve the accuracy in response of dysarthric speakers. The study followed a within-group design and contained 15 participants with dysarthria (4 male and 11 female) all of whom had varying levels of dysarthria severities (and intelligibility scores). Participants were selected from the UA-Speech Database. Identification of impairment was not specifically outlined in the paper. Baseline severity levels were taken from the average scoring of "unfamiliar" listeners. The variables evaluated are "accuracy" (% words correct) across level of explicit program instruction and type of software (speaker-dependent vs. speaker-adaptive).

Across both speaker-dependent and speaker-adaptive STT, the study provides evidence for increased accuracy rates associated with an increase in explicit teaching done to the program. Between speaker-dependent and speaker-adaptive STT software, it appears that the speaker-adaptive software had higher accuracy rates on average for all severity levels following explicit instruction. Depending on the client, speaker-adaptive software was about 2-10% higher with their word accuracy rates. Interestingly, the data presented by this article show the speaker-dependent software having higher accuracy rates for select participants in the "mild" and "moderate" severity levels after explicit instruction. This is directly contrary to the articles by Raghavendra et al. (2001) and Rudzicz (2007) who found speaker-adaptive software to be more effective with the "mild" and "moderate" severity levels and vice versa for the "severe" dysarthrics.

The validity of the article's results is quite "strong" with a very clear and well-defined methodology section, however there are some concerns with the reliability of the results. Firstly, there remains the common issue of a reduced sense of reliability in response to the reduced participant size. However, with that being said, it is nice to see the authors include as many participants as they did (i.e., more than double the number of participants in the Sharma and Hasegawa-Johnson (2010)). Secondly, a concern that has been brought forward by the authors of the article mention of certain "practice" effects as a result of certain programs in the study getting more usage than others. The issue here is with an unfair increase in usage will inevitably come an increase in that particular system's accuracy ratings. With all that being said, the conclusive evidence of this article still come as "suggestive".

Case Study design:

Bonilla-Enriquez and Caballero-Morales (2012) investigated the effectiveness of speaker-dependent and speaker-adaptive STT software in producing accurate responses in Mexican-Spanish dysarthric speakers. The authors decided to conduct a single-participant case study design, with a 64-year-old male with mild-moderate dysarthria. The dependent variable in this study was the percent accuracy of words correct, compared against the different types of speech recognition software (i.e., speaker dependent vs. speaker-adaptive).

The results of this study indicate that the speaker-adaptive STT software was able to demonstrate higher accuracy rates than the speaker-dependent software. The speaker-adaptive software was able to reach accuracy rates upwards of 96-98%, whereas the speaker-dependent software was only reaching accuracy rates of about 75%. These findings follow the trend of Raghavendra et al. (2001), and Rudzicz (2007) which demonstrate higher accuracy rates for speaker-adaptive software for those with "mild" to "moderate" dysarthria. However, interestingly, the vocabulary or "templates" used in this study were made sure to contain all the different sounds in the language across a variety of environments. With this being accounted for, the accuracy rates of the speech recognition software increased, albeit speaker-adaptive software being more successful. The authors propose that future research investigate ways to mitigate the phonemic and articulation errors in speech and their effect on speech-recognition software.

A natural disadvantage to this type of study design is the rather small sample sizes. This will no doubt reduce the reliability of the results, but it is for this reason that further studies (both replication and novel) need to be conducted to boost the reliability of the research findings in this article. Like the other articles in this critical review, the results of this article are to be seen as "suggestive".

Literature Review:

Young and Mihailidis (2010) conducted a literature review surrounding the effects of dysarthria on the performance of speech recognition software specifically in the geriatric population. A total of 23 papers were reviewed: with eleven discussing dysarthria and STT, three discussing older adults and STT, and three describing the link between speech-recognition software and its user's perspectives. Finally, six articles reviewed speech recognition software and its functionality across communication disorders or disabilities. The literature review evaluated variables / commonalities between the different articles including: "levels of intelligibility", "human speech perception vs. speech recognition", "perceptual ratings vs. speech consistency", "speech variability", "fatigue", "voice misuse/abuse", "personal factors", "system and user voice training" and "system usability". It is important to note here that this literature did not look specifically for "speaker-dependent vs. speaker-adaptive" but more on a general basis. However, this literature review does bring to light important information in regards to the secondary and tertiary objectives of this paper.

In line with the indications made by the previously discussed articles, the biggest factors that influenced speech recognition performance with individuals with dysarthria were the participant's level of fatigue, their type of system (speaker-dependent vs. speaker-adaptive), and amount of user and system training (both "amount of time used" and "amount of explicit teaching performed"). This literature review makes a similar claim to that of Raghavendra et al. (2001) which urges further literature to pursue ways to reduce the effect of speech variability in dysarthric speakers on the accuracy rates of speech recognition software.

Interestingly, this literature mentions that a clinician should be cautious not to provide certain speech recognition software based on a client's speech consistency, but to rather allow them to trial the software themselves. This has been indicated to increase the user's comfortability and compliance with the software and speech-recognition software as prescribed supportive communication tool.

One final note surrounding the evidence and results brought forward by this literature review. It was quite apparent that many of the articles discussed in this literature review were providing contrary results to each other. This is not to dig into the validity of these results, but to present the argument that even in the twenty-three articles that were reviewed, there still remains large gaps in the current literature. More research is needed in order to smooth out the apparent reliability issue that presents itself in many of these articles that research this topic.

Discussion

According to literature discussed in this critical review, the general consensus (as echoed in Raghavendra et al. (2001), Rudzicz (2007), Bonilla-Enriquez and Caballero-Morales (2012), and Young and Mihailidis (2010)) is speaker dependent STT systems may be more beneficial for those with a higher severity level of dysarthria (and or a lower intelligibility speech rating). Indicated through the articles by Raghavendra et al. (2001), the best available option for severely dysarthric speakers may be a speaker dependent system with a modified vocabulary. Research indicated that both speaker-dependent and speaker-adaptive STT software show beneficial success rates for the "mild" and "moderate" levels of severity, however in general there is a clearly defined better performance by the speaker-adaptive speech recognition systems.

Overall, there is a significant lack of research available investigating the efficacy of STT software with dysarthric speech. This sentiment is echoed by all the research articles evaluated in this paper. With that being said, the current literature does suggest possible benefits for dysarthric speakers resulting from the use of speech recognition software. It is clear that further research needs to be conducted in this area to understand how STT speech recognition software can be better leveraged for individuals with dysarthria. It is in response to the lack of available research that provides merit in further increasing the currently available research pool.

Moving forward, it seems that the next steps are to look at the different ways in which the impact of phonemic, articulation, and speech errors can be mitigated in speech recognition software. Specifically, is there an increase in type of error with an increase in severity levels? and if so, how can this be accounted for in speech recognition software moving forward?

Clinical Applications

There is promise to the practical application of STT software for speech- language pathologists working with patients with dysarthric speech, however further research is certainly warranted. However, there is one major issue stopping clinicians from freely administering speech recognition software for individuals with dysarthria. Primarily, the biggest hurdle of this research domain, is the clear lack of reliability obtained in the results of these papers. Of all the studies evaluated as part of this critical review, the highest participant pool noted consisted of 15 individuals. It would be wise to encourage collaboration between researchers, clinicians, organizations, and communities to allow for bigger access to possible participants.

Under a critical review the overall level of evidence of these studies is to be considered "suggestive" at best. In order to make clinical recommendations as a registered speech-language pathologist, it would be in the profession's best interest to wait for more research to become available and shine a more reliable light on this topic.

References

Bonilla-Enriquez, G., & Caballero-Morales, S., (2012). Communication Interface for Mexican Spanish Dysarthric Speakers.. Acta Universitaria, 22( ),98-105.[fecha de Consulta 6 de Marzo de 2021]. ISSN: 0188-6266. Disponible en: https://www.redalyc. org/articulo.oa?id=416/41623190014

Christensen, H., Cunningham, S., Fox, C., Green, P., & Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *Thirteenth Annual Conference of the International Speech Communication Association.*

Coleman, C., Meyers, L.S., (1991). Computer Recognition of the Speech of Adults with Cerebral Palsy and Dysarthria. *Augmentative and Alternative Communication.* 7; 34-42.

Das, P., Achajee, K., Das, P., Prasad, V. (2015). Voice recognition system: Speech-to-text. *Journal of Applied and Fundamental Sciences*, *1*(2), 191.

Dayley, F.L, Aronson, A.E, Brown, J.R., (1969). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research.* 12; 246-269.

Huang, X. D., Lee, K.F., (1991). On Speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *The Institute of Electrical and Electronic Engineers.* 2(1); 877-880.

Raghavendra, P., Rosengren, E., Hunnicutt, S., (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems, Augmentative and Alternative Communication, 17:4, 265-275, DOI: 10.1080/aac.17.4.265.275

Rosen, K., Yampolksy, S., (2000). Automatic speech recognition and a review of its functioning with dysarthric speech, Augmentative and Alternative Communication, 16:1, 48-60, DOI: 10.1080/07434610012331278904

Rudzicz, F. (October, 2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility* (pp. 255-256).

Sharma, H. V., Hasegawa-Johnson, M., NAACL HLT. (2010). Workshop on Speech and Language Processing for Assistive Technologies. *Association for Computational Linguistics.* Pg. 72-79.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, *22*(2), 99