Research Article

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2019;71:286–296
DOI: 10.1159/000499156

Published online: May 22, 2019

Automated Speech Recognition in Adult Stroke Survivors: Comparing Human and Computer Transcriptions

Adam Jacks[a] Katarina L. Haley[a] Gary Bishop[b] Tyson G. Harmon[c]

[a]Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [b]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [c]Department of Communication Disorders, Brigham Young University, Provo, UT, USA

Keywords

Speech transcription · Automated speech recognition · Aphasia · Stroke · Assessment · Intelligibility

Abstract

Objective: Speech sound errors are common in people with a variety of communication disorders and can result in impaired message transmission to listeners. Valid and reliable metrics exist to quantify this problem, but they are rarely used in clinical settings due to the time-intensive nature of speech transcription by humans. Automated speech recognition (ASR) technologies have advanced substantially in recent years, enabling them to serve as realistic proxies for human listeners. This study aimed to determine how closely transcription scores from human listeners correspond to scores from an ASR system. ***Patients and Methods:*** Sentence recordings from 10 stroke survivors with aphasia and apraxia of speech were transcribed orthographically by 3 listeners and a web-based ASR service. Adjusted transcription scores were calculated for all samples based on accuracy of transcribed content words. ***Results:*** As expected, transcription scores were significantly higher for the humans than for ASR. However, intraclass correlations revealed excellent agreement among the humans and ASR systems, and the systematically lower scores for computer speech recognition were effectively equalized simply by adding the regression intercept. ***Conclusions:*** The results suggest the clinical feasibility of supplementing or substituting human transcriptions with computer-generated scores, though extension to other speech disorders requires further research.

Introduction

Speech sound errors negatively impact the communication of people with a wide range of communication disorders, sometimes precluding listeners from understanding a speaker's intended words. Speech intelligibility, the degree to which a person's oral communication is understood by a listener, is one of the primary measures that describe a person's functional communication, described by Kent as the "*sine qua non* of spoken language" [1, p. 9]. It differs from other measures of speech production by summarizing the overall effectiveness of message transmission rather than individual speech qualities. Abnormal features of voice, resonance, and articulation may

Initial results of this study were presented at the 2016 International Motor Speech conference in Newport Beach, CA, USA.

KARGER

E-Mail karger@karger.com
www.karger.com/fpl

Adam Jacks
Division of Speech and Hearing Sciences, University of North Carolina at Chapel Hill
3122 Bondurant Hall
Chapel Hill, NC 27599-7190 (USA)
E-Mail adam_jacks@med.unc.edu

have significant perceptual effects, yet may be minimally consequential to message transmission [2, 3]. For these reasons, speech intelligibility is frequently referenced as a measure of functional communication and used to justify treatment, establish baseline performance, and document treatment effects or disease progression.

In the area of speech disorders, much of the early intelligibility research focused on dysarthria [4–7]. Despite diverse constellations of symptoms across affected motor systems and dysarthria types, imprecise articulation is almost always a prominent feature of dysarthria and widely recognized as a negative influence on listeners' ability to understand the intended messages. Intelligibility measures therefore serve as an index of how speech impairment in speakers with dysarthria impact their ability to communicate.

Until recently, much less was known about speech intelligibility in aphasia. Speakers with both fluent and nonfluent aphasia profiles, with and without coexisting apraxia of speech, generate sound errors in their speech output. It is thought that some of these errors originate at a phonologic level of processing, others at a motor programming level, and yet others through some combination of the two. Regardless of the underlying impairment, speech sound errors have salient consequences for speech intelligibility. Our group has demonstrated that single-word intelligibility is useful for estimating severity and documenting change in individuals diagnosed with aphasia and apraxia of speech [8–10]. Furthermore, we have shown that monosyllabic single-word intelligibility can be completed by speakers with a wide range of impairments, allowing for meaningful comparisons across participants.

Options for Intelligibility Estimation

There are two general methods for quantifying intelligibility: scaling and word transcription. Scaling requires a listener to make an impressionistic judgment about intelligibility, often using verbal descriptions of severity (normal, mild, moderate, or severe) or numeric scales with equal-appearing intervals (e.g., 1–10). Impressionistic percentage estimates, where a clinician listens to a sample and estimates the percentage of words correctly understood, is the most commonly used form of intelligibility scaling procedure [11]. Unfortunately, psychophysical research suggests that human listeners are not well equipped to rate intelligibility using so-called equal-appearing interval scales and that impressionistic estimates of intelligibility are unreliable [4, 5, 12, 13].

Word selection or transcription methods are generally considered more valid and reliable for quantifying intelligibility than scaling methods (see Schiavetti [12] for a review and critique of intelligibility scaling). Word transcription procedures are derived experimentally through the following three steps: (1) a speaker produces speech stimuli (words or sentences, optimally of an unpredictable nature); (2) a listener transcribes the words perceived without knowledge of the stimuli; and (3) a comparison is made between the transcribed words and the intended target words, expressed as a percentage of words correctly understood. The Assessment of Intelligibility of Dysarthric Speech (AIDS [14]) and its computerized version, the Speech Intelligibility Test (SIT [15]), are commonly used to evaluate intelligibility in dysarthria, as is the Frenchay Dysarthria Assessment (FDA-2 [16]). The AIDS has also been applied to evaluate speech intelligibility in aphasia, with good results [17, 18]. Additionally, our group developed a test that is sensitive to consonant and vowel errors in aphasia and apraxia of speech, and robust to speaker and listener familiarity effects: the Chapel Hill Multilingual Intelligibility Test (CHMIT [19, 20]). Although the CHMIT can be used in varied clinical populations, it is particularly suited to people with aphasia or others who have difficulty reading (e.g., people with visual impairment or limited literacy), as it provides an auditory-verbal cue for each stimulus in addition to a written prompt.

Clinical Use of Intelligibility Measures

There is broad agreement among speech-language clinicians that speech intelligibility is a useful severity index [21], and that it should be a primary target for intervention [11, 22]. However, most clinicians restrict their use to informal ratings or impressionistic estimates. In a survey of speech-language pathologists across practice areas (i.e., those who do and those who do not treat patients with dysarthria [23]), only 12% of the clinicians reported regular use of the two most common intelligibility assessments (AIDS or FDA-2 [14, 16]). Among speech-language pathologists who treat patients with dysarthria, formal intelligibility testing is more common, but still only 35% reported using a formal test [11] – again, either the AIDS or the FDA-2. Clinical use of word transcription testing for aphasia and apraxia of speech is unknown.

What Are the Barriers to Formal Intelligibility Testing?

There are several reasons clinicians use impressionistic ratings in place of the more valid and reliable intelligi-

DOI: 10.1159/000499156

bility metrics. First, they may lack access to an appropriate measurement instrument, reportedly due to expense and insufficient financial resources [23]. Second, they may experience practical restrictions related to time and logistics, including making recordings and finding an unfamiliar person to complete transcriptions. Finally, they may perceive that informal ratings are equally useful and efficient [23].

Automated Speech Recognition instead of Perceptual Intelligibility Testing

Several of the challenges to completing intelligibility transcriptions – familiarity with the speaker, familiarity with the conversational topic, and limited time – might be addressed using automated speech recognition (ASR) technologies, with the added benefit of obtaining scores instantly. ASR is generally defined as computer-based systems that convert spoken language into text, and it has undergone several waves of innovation in nearly a century of development. Early ASR systems were *speaker dependent*, developed to detect digits spoken by specific, individual speakers [24], requiring speaker-specific training to achieve accurate word recognition.

More recent developments in computing power and processing strategies have allowed for *speaker-independent* ASR, such as voice input applications on personal computers, smartphones, and other personal computing devices. These are designed to be used by any speaker without the need for speaker-specific training, allowing for different voices and speech patterns within a language population. As of this writing, speech recognition engines on phones and in web-based applications (e.g., IBM Watson, Google Speech-to-Text) operate essentially in real time, relying on the computational power of server-based processing systems.

ASR as a Predictor of Intelligibility Scores for Speakers with Disorders

Several researchers have found that ASR systems are less accurate at identifying words produced by people with dysarthria compared to people with typical speech production, reducing their usefulness as accessibility tools [25–30]. Notwithstanding its poorer accuracy, recent research has shown that ASR may be useful for clinical purposes. Ballard et al. [31], for example, have shown that ASR can be effective in providing speakers with apraxia of speech and aphasia feedback on their speech productions. In particular, the authors found that items recognized by ASR corresponded to human decisions on correct/incorrect trials 75% of the time. Anecdotally, we are also aware of clinicians using web-based ASR systems as real-time feedback tools to encourage speakers to slow their rate and overarticulate.

In addition to the emerging use of ASR as a treatment feedback tool, others have observed that the systems could be used to estimate or predict speech intelligibility, noting relationships between speech intelligibility rated by human listeners and words transcribed with ASR [26, 32–34]. Ferrier et al. [32] compared computer recognition and human intelligibility scores for 10 speakers with dysarthria due to cerebral palsy. After 5 training sessions, word recognition scores were strongly correlated with intelligibility of words and sentences from the AIDS (*r*'s 0.86–0.92). Similarly, Thomas-Stonell et al. [26] found that word recognition of speakers with dysarthria by a trained ASR system was significantly correlated with intelligibility scores from 10 listeners ($r = 0.80$).

More recent studies have reported weaker relationships between human impressionistic intelligibility estimates and results from untrained, speaker-independent ASR systems. For example, Rosdi et al. [35] used a custom system to predict human intelligibility percentage estimates of Malay children with speech impairment, reporting a moderate but significant correlation between the measures ($r = 0.57$). Computer recognition was lower than human intelligibility ratings across severity levels (approx. half as accurate as humans). However, visual inspection of the data indicates that intelligibility for the human listeners was often at or near ceiling, which created a nonlinear relationship and suggests that the relationship between human and computer measures was stronger than reported. This ceiling effect may be owed in part to the use of impressionistic percentage estimates by the human listeners, which, as noted before, are ill-suited to human perceptual abilities.

The purpose of this study was to evaluate the potential of a commercial speech recognition engine for quantifying word production by stroke survivors with aphasia and/or apraxia of speech. The study is organized around three questions:

1. Is there a significant difference in transcription scores between 3 human listeners and a commercial speech-to-text engine?
2. What is the agreement and reliability among transcribers' scores?
3. Can human transcription scores be reliably predicted from computer-generated scores, and what is the relationship between them?

Table 1. Participant information

Participant	Age, years; months	Sex	Time after onset, years;months	Western Aphasia Battery		Speech diagnosis
				aphasia quotient	type	
P01	47;5	M	0;5	98	Anomic	MIN
P02	52;8	F	1;3	96	Anomic	APP
P03	64;1	M	0;2	86	Anomic	APP
P04	51;9	F	7;6	88	Anomic	APP
P05	59;5	F	1;2	69	Broca's	APP
P06	27;8	F	1;6	82	Anomic	AOS
P07	50;1	M	4;10	70	Conduction	AOS
P08	66;8	F	0;6	60	Broca's	AOS
P09	66;6	F	2;3	76	Broca's	AOS
P10	73;4	M	4;9	30	Broca's	AOS

MIN, minimal speech impairment; APP, aphasia with phonemic paraphasia; AOS, apraxia of speech.

Method

Speech materials for this study were sentences produced by 10 adults whose speech output was characterized by varying frequencies of speech sound errors due to aphasia with and without apraxia of speech. The recordings were collected as part of a related study testing the effects of different auditory feedback conditions (e.g., normal auditory feedback, masked auditory feedback) on measures of speech fluency [36]. Orthographic transcriptions also were completed to evaluate effects on speech production. As such, these transcriptions are used as a convenience sample. The speakers' characteristics are presented in Table 1; for convenience of data presentation, the participants are numbered in order of descending performance on the adjusted transcription scores, the primary dependent variable of this study.

The participants included 6 females and 4 males, with ages ranging from 27;8 to 73;4 years;months (median age 56 years). Nine had survived a left hemisphere stroke; 1 had a traumatic brain injury. All experienced some degree of language impairment, although 2 tested in the normal range of the Western Aphasia Battery, Revised (WAB-R [37]; median Aphasia Quotient = 79; range 30–98). Clinical assessment of speech impairment was completed by authors A.J. and K.L.H. using methods published elsewhere [10], with 1 participant classified as having minimal speech impairment, 4 participants as having aphasia with phonemic paraphasia, and 5 participants as having apraxia of speech.

Speaking Conditions

The participants produced sentences in a single-session treatment introduction and withdrawal paradigm, with a different auditory feedback condition for each phase. An ABA design was used, where auditory feedback was unobstructed for 20 sentences produced in the A phase, and masked auditory feedback (e.g., 85 dB pink noise) was presented for 20 sentences in the B phase and 20 sentences in a second A phase. Sentence stimuli were drawn from the Harvard sentences lists [38], with different sets of 20 sentences for each phase, resulting in speakers producing 60 different sentences in the recording session. To partially address listener familiarity with the sentences produced by multiple speakers, 4 different sentence sets were created; speakers were randomly assigned to one of the stimulus sets.

Stimulus presentation and recording were controlled using Alvin2 software [39] on a PC running 64-bit Windows 7. Audio was recorded with a headset microphone (C555L; AKG Acoustics GmbH, Vienna, Austria) and masking noise was delivered through foam-tipped earphones (ER-3A; Etymotic Research, Inc., Elk Grove Village, IL, USA) via an external USB soundcard (M-Audio Fast Track Ultra; Avid Technologies, Inc., Burlington, MA, USA). During the masking phase, pink noise was delivered binaurally at 85 dB SPL, calibrated using a Larson-Davis System 824 sound level meter (Depew, NY, USA) with a 2cc coupler (GRAS RA0038, Holte, Denmark). For each sentence trial, speakers were provided with an auditory example and allowed to view the written sentence while speaking.

Transcriptions

The listeners were 3 graduate students of speech-language pathology (pseudonyms: Daria, Elvia, and Fiona). They completed the tasks in fulfillment of a research experience requirement. Their instructions were to listen to each sentence, type the words they heard, and type "x" for each unintelligible syllable. Because sentences were repeated across some participants, there was potential for listeners to learn the sentences. To reduce familiarity effects, when a speaker clearly produced an error (e.g., word or sound substitution), the listeners were instructed to type it as they heard it, even if they knew it was incorrect and thought they knew what word was intended. Each listener transcribed 600 sentences, randomized across speakers and conditions in Alvin2 [39].

In addition to the human transcriptions, computer transcriptions were generated using the IBM® Watson Speech to Text service [40] implemented with the Bluemix/IBM Cloud programming platform [41]. Transcriptions were obtained from the system at 2 points in time, July 2015 (Watson 2015) and June 2018 (Watson 2018), to permit an evaluation of performance improvement over time. Parameters passed to the service for recognition included: (1) US English broadband model (default; base model in use at each time point); (2) continuous speech stream (i.e., don't stop at pauses); and (3) opt-out of request logging, preventing the system from saving audio files for service improvement. Output parameters included the word with the highest match likelihood, time stamps for each word, and confidence rates for a maximum of 5 alternative words.

Measures

For both human and computer transcriptions, the dependent measure was an adjusted transcription score, based on exact match for 5 keywords in each sentence plus an adjustment based on close similarity between transcription and target words. The rationale for using an adjusted score was to allow for a more fine-grained metric than possible with the raw score of 5 keywords (e.g., 1/5 = 20%, 2/5 = 40%, etc.). In addition, use of an adjusted score also reduces differences between human and computer listeners due to minor differences, such as suffixes (e.g., "liked" vs. "likes"). The score adjustment was computed algorithmically using the Levenshtein edit distance from the Natural Language Toolkit [42]. This procedure is commonly used to account for small differences in text strings (e.g., suffixes, as noted above). For example, for the target sentence "The birch canoe slid on the smooth planks," if the listener transcribed "The Bart came stead on the smooth plants," they would receive full credit for "smooth" and partial credit for "plants," which differed by only 1 character from "planks." The edit distance between "plants" and "planks" is 1 (i.e., 1 letter must be altered to change the word "plants" to "planks"). The adjusted score for that word is then computed as the proportional difference between the letter strings (e.g., [12 (sum of characters in the 2 text strings) – 1 (edit distance)]/12 [sum of characters in the 2 text strings]).

The mean of the adjusted transcription scores was obtained for each of 3 speaking conditions and used for each analysis.

Analysis

To answer the first question on the difference in transcription scores between human and computer transcribers, we used a mixed-effects model in JMP software [43], with transcribers and speakers as fixed effects, and speaking condition (e.g., normal feedback and masked auditory feedback) as a nested effect within speakers.

The second question, on the agreement and reliability among transcribers, employed intraclass correlations (ICCs) between each pair of listeners to assess agreement, and Spearman correlations for reliability. We selected the type 3 ICC analysis (two-way mixed), treating speakers as random effects and transcribers as fixed effects, and used the values of absolute agreement for single raters. ICCs and Spearman correlations were computed in R using the "psych" package [44, 45].

The third question, on the prediction of human transcription scores from computer-generated scores, was addressed using linear regression, with the mean of human scores as the dependent variable and computer-generated scores as the independent predictor. Sequential regression models were first run to compare the performance of Watson 2015 and 2018 in predicting human performance, using the extra-sum-of-squares principle and the R^2 change metric to evaluate significant differences between models. A simple linear regression was run with the single best predictor as an independent factor to generate the slope and intercept needed to predict human scores from computer results. These analyses were completed in JMP [43].

Results

The overarching goal of this study was to evaluate the potential usefulness of commercial ASR for quantifying speech production in clinical populations, in this case a convenience sample of stroke survivors with speech sound impairment consistent with aphasia and/or apraxia of speech. In the following pages, we begin by determining whether human listeners differ from each other in their transcriptions, and whether they differ from scores obtained by the IBM Watson system. Second, we address the reliability and agreement among the listeners and systems. Finally, we examine whether there is a predictable relationship between human-generated and computer-generated transcription scores.

Effect of Transcriber on Scores

Transcription scores varied widely across speakers, as shown in Figure 1, with mean adjusted scores ranging from 10% (P10) to 93% (P01). Transcribers also varied in their scores, with the human listeners recognizing more target words than the computer, in the following order from most to least accurate: Daria, Elvia, Fiona, Watson 2018, and Watson 2015. The mixed-effects model confirmed the significance of each of these factors ($F(9, 116) = 937.40$, $p < 0.0001$ for speaker; $F(4, 116) = 52.85$, $p < 0.0001$ for transcriber) as well as the condition[1] (masking vs. no masking; $F(20, 116) = 5.51$, $p < 0.0001$). Pairwise comparisons using Tukey's honestly significant difference showed significant differences between almost all transcriber pairs and most speakers. All transcriber pairs differed, except the scores from Watson in 2018 compared to 2015. Speakers P01, P02, and P03 were similar in their transcription scores, with no significant differences according to Tukey's honestly significant difference. P03's scores were also similar to those of P04; all other differences were significant.

[1] The observed main effect of "condition," although not important to the study purpose, deserves mention. The results of the original study [36] indicated that several speakers spoke faster with masking noise. This was a positive outcome, considering their baseline speech rate was very slow, but transcription scores were obtained to determine whether articulation was negatively impacted. Planned comparisons between the conditions indicated no statistically significant difference for 4 of the 10 speakers and significantly decreased transcription scores during masking for 3. Because speech samples from all recording sessions were included for all participants in the present study, effects of "condition" have no bearing on our analysis of transcriber effects.

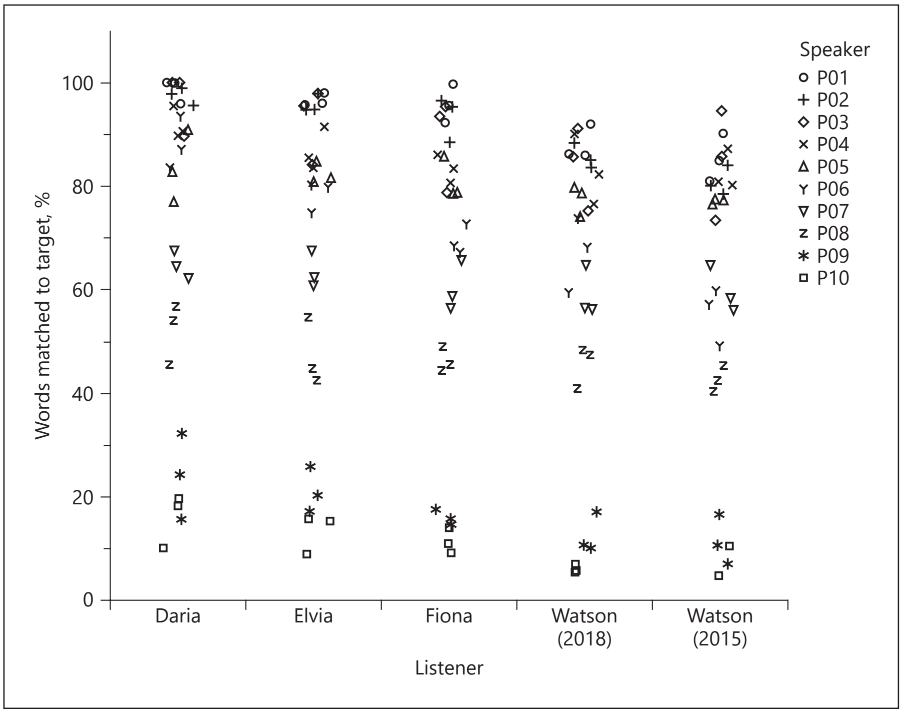

Fig. 1. Intelligibility scores by listener and speaker. The percentage of words correctly identified is plotted for each speaker, shown in markers with distinct shapes and colors, with separate measures shown for 3 human listeners and the computer transcription of Watson from 2015 and 2018. The results for each speaker include 3 separate values, the mean intelligibility for each of 3 conditions (normal auditory feedback, masked auditory feedback, and normal auditory feedback).

Table 2. Pairwise intraclass correlations (agreement)

	Elvia	Fiona	IBM18	IBM15
Daria	0.99	0.98	0.98	0.96
Elvia		0.99	0.99	0.97
Fiona			0.99	0.98
IBM18				0.99

Table 3. Pairwise Spearman correlations (reliability)

	Elvia	Fiona	IBM18	IBM15
Daria	0.96	0.97	0.95	0.93
Elvia		0.98	0.96	0.95
Fiona			0.96	0.95
IBM18				0.98

Agreement and Reliability

The metrics of agreement and reliability indicated a high correspondence among the transcribers. In particular, ICCs ranged from 0.98 to 0.99 among human listeners and for human-to-Watson 2018 comparisons (Table 2; excellent agreement [46]). ICCs between humans and Watson 2015 ranged from 0.96 to 0.98. Similarly, Spearman correlations were very high, from 0.96 to 0.98 among human listeners, though slightly lower (0.95–0.96) for human-to-computer comparisons (Table 3). Notably, reliability and agreement were poorer for the IBM system used in 2015.

Prediction of Human Scores from Computer-Generated Scores

Finally we come to the question of whether computer-generated transcription scores can predict human listener scores. For this question, we shifted from individual human listener scores to the mean of the 3 listeners, which is considered sound clinical practice to account for listener variability [47]. We completed a sequential regression analysis, with the mean of the human listener scores as a dependent factor, and with Watson's adjusted scores from 2015 and 2018 entered sequentially as independent predictors. The first model was highly significant, with Watson's 2015 scores accounting for 95% of the variance in human scores ($F(1, 27) = 1{,}390.01$, $p < 0.0001$, $R^2 = 0.952$). Adding Watson's 2018 scores to the model significantly improved the model, increasing the variance accounted for to 98% ($\Delta F(1, 27) = 43.29$, $p < 0.0001$, $\Delta R^2 = 0.029$). A further sequential analysis was completed to confirm that Watson's 2018 results alone are sufficient to predict human scores, entering the 2018 scores first, followed by the 2015 scores. The model with

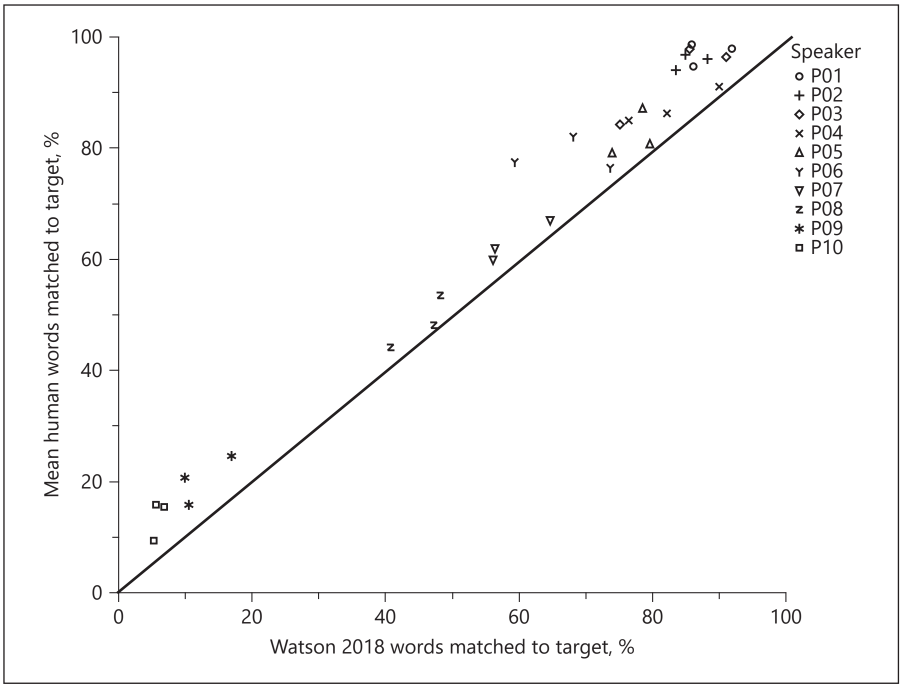

Fig. 2. Prediction of human intelligibility scores from automated results. The mean of the intelligibility scores from 3 human listeners is plotted against the scores from IBM Watson as obtained in 2018. Each speaker is represented in markers of distinct color. A line with a slope of 1 is shown in black, demonstrating that human listeners perceived words more accurately than the computer for each speaker.

2018 scores alone accounted for 98% of the human score variance ($F(1, 27) = 1{,}432.50$, $p < 0.0001$, $R^2 = 0.98$); adding the 2015 scores did not significantly improve the model fit ($\Delta F(1, 27) = 1.69$, $p = 0.20$, $\Delta R^2 = 0.001$).

Based on these analyses, we conclude that Watson's 2018 transcription scores predicted human listener scores with high accuracy (Fig. 2). As noted previously, however, each of the 3 human listeners identified words more accurately than Watson. If we were to use Watson's scores as a proxy for human listeners, an adjustment would be needed. This can be accomplished using the coefficients from the linear regression analysis with Watson 2018 as the sole predictor ($B_0 = 0.068$, $B_1 = 1.003$). In essence, for this sample, the mean of human transcription scores can be predicted by taking Watson's 2018 score and adding 0.068 (6.8%). Notably, if using Watson's 2015 scores, the correction would be 0.09 (9%).

Discussion

The findings from this study are promising for the use of free or very-low-cost ASR in the clinical assessment of speech production. The web-based IBM Watson Speech to Text Engine transcribed words in sentences produced by stroke and brain injury survivors with speech sound errors with accuracy nearing that of human listeners. Although the computer-based transcription scores were significantly lower than those of each of the human listeners, agreement and reliability between the listeners and the computer were nevertheless very strong. Furthermore, computer-generated scores very closely predicted the mean of the 3 human listeners' scores, suggesting a potential for clinical use in the future, though significant research is needed, particularly in broader clinical populations (e.g., speakers with dysarthria).

ASR Performance May Differ across Speaker Diagnoses

The accuracy of ASR may vary across different speaker populations, including those with different speech and language diagnoses. Early use of ASR software in people with dysarthria generally showed much lower accuracy than human listeners. For example, even after 5 training sessions, Doyle et al. [33] found that human listeners perceived 30% more words produced by speakers with dysarthria compared to an ASR system. In contrast, the present results for speakers with aphasia and apraxia of speech show that a current web-based ASR engine performed well relative to human listeners, recognizing approximately 7% fewer words than human listeners.

One possibility for the improved performance in this study relative to earlier work is that our participants did not have clinically significant dysarthria as was the case in prior studies of ASR in speech disorders. It is possible that current ASR significantly underperforms in identifying words of speakers with dysarthria because more speech subsystems may be impacted than in speakers with aphasia or apraxia of speech. In particular, speakers with dysarthria often have impairments involving phonation and nasal resonance, which weaken the signal-to-noise ratio and may potentially lessen the accuracy of ASR. While recent research has shown that Google's Speech-to-Text engine makes more errors with speakers with Parkinson's disease than with controls [30], there was no comparison to human listener performance with these speakers.

Ongoing ASR Development

Major advances in ASR technology in the past decade are the most likely explanation for the differences between our results and those of early ASR studies on speakers with dysarthria. Most of the research on speech recognition in speakers with dysarthria was completed in the 1990s, with software developed in the early 1990s. Although the exact details of the methodology used by IBM's current speech recognition engine are not available, major developments in the use of deep neural networks for ASR in the past decade [48] have resulted in great improvements, in comparison to Gaussian mixture models/Hidden Markov models used in the prior decades.

Commercially available ASR applications have great potential for clinical use, including estimation of intelligibility and tracking change over time for clinical documentation and self-monitoring by clients while practicing with strategies. The web-based technology used in this study is convenient and easily accessible, unlike custom research-based systems used for advanced development of ASR processing strategies. Demonstration versions are available via web browser for the IBM and Google services, and they allow users to obtain transcriptions for live recordings or uploaded audio files.[2] The services are generally free for small amounts of audio (e.g., currently up to 60 min of audio per month is free). Alternatively, clients who own smartphones may also use the voice recognition on their personal device to monitor their speech production during treatment sessions or in-home practice. One important caveat regarding the web-based services is that they are not intended to process – or to protect – personal data; thus, caution is needed to ensure personal information is not disclosed when using this application. Note that for technical users using the application programming interface, IBM provides an option so that recordings are not saved in the system.

There are disadvantages to relying on commercial systems from a researcher's and a clinician's perspective, related to lack of control over the processing system and its change over time. IBM and its competitors are no doubt constantly working to improve the performance of this product, and thus results obtained today will likely differ from those obtained 3 years from now. In fact, that is exactly what we discovered when we compared computer-generated results across time: Watson had significantly improved its transcription of our participants' sentences from 2015 to 2018, resulting in a more robust prediction of human listener results. While this is a very positive outcome from the perspective of the end user with speech impairment, it also means that the formula we generate to predict transcription scores will change over time and will need to be updated to be useful as a clinical tool.

To date, even the most advanced ASR applications are less accurate than human listeners. Current ASR technology is able to account for sentence context when decoding specific words, but human listeners likely still benefit from more general context (e.g., topic, etc.). We assume that humans will not continue to outperform computers indefinitely – there may come a time when ASR supersedes human listeners in accurately transcribing the intended words of people with speech impairments. If so, an open question is whether human transcription scores will still be predictable from the superior computer-based scores.

Clinical Applications

In our view, the expected change in ASR accuracy is welcome, but it must be considered with caution when used for clinical purposes, and safeguards should be used to prevent effects of commercial ASR development from influencing derived transcription scores. For example, if ASR is used to predict human transcription scores, then comparisons between recordings made at different time points should be processed at the same time, allowing a direct comparison of results. If a formula is to be used to predict human listener performance based on a statistical analysis of computer versus human scores, such as described in this paper, then the analysis must be rerun periodically to update the correction factor needed. It is possible for users to specify the language model IBM's speech recognition machine uses to identify words, allowing clinicians in the future to use an ASR version that has previously been validated against human transcription scores.

[2] https://www.ibm.com/watson/services/speech-to-text/; https://cloud.google.com/speech-to-text/.

However, this detail might easily be overlooked by users without programming experience, resulting in inconsistent use of this technology.

Limitations and Future Directions

As previously noted, the prior work is limited in its focus on speakers with speech sound errors with aphasia or apraxia of speech due to stroke or brain injury; future work is needed to determine whether commercially available ASR is as successful in predicting transcription scores of speakers with speech sound errors due to dysarthria. In addition, although the listeners were highly reliable among each other, the use of only 3 listeners to transcribe speech from 10 speakers – some of whom produced the same sentences – is a potential limitation of the study. Some studies using direct magnitude estimation to rate intelligibility have used 10 or more listeners [49–51]. However, in previous research using bootstrap sampling of various listener sample sizes, we have observed good reliability for only 3 listeners [52]. Furthermore, this number of listeners approximates a feasible clinical scenario. In fact, when transcription-based intelligibility is done in practice, it is likely more common to have only one clinician transcribe recordings of a peer clinician's patients, although this practice is risky and does not permit analysis of inter-listener variance.

The limited number of listeners also caused us to use a sentence transcription task differing from standard intelligibility tasks. Whereas standard intelligibility paradigms have listeners type the words they think the speaker is trying to produce, even if they are not produced correctly, we had listeners type the words as they heard them, even if they knew what the target word was, in order to counteract familiarity with the sentences. Future work would benefit from using more listeners with more standard intelligibility instructions.

In addition to the limited number of listeners, we note that speech-language pathology students are not representative of the general listening population. As with the number of listeners, this choice was meant to approximate a likely clinical scenario. Nevertheless, these students have training in phonetic variability and greater familiarity with impaired speech than the general population; thus, their transcription scores are likely higher than those of typical untrained listeners. Furthermore, graduate students are often highly motivated and exert more effort in the listening task than the general listener, which may impact intelligibility scores.

The present study examined only transcription of humans and IBM's speech recognition engine, with results adjusted for small variations using the commonly used edit distance algorithm. We did not address the types of errors produced by human or computer transcribers. Future analysis of transcription errors in this and other speaker populations may be instructive both for the development of ASR technologies and for guiding speakers and clinicians to productive strategies for improving intelligibility.

Conclusions

Several decades of advancement in ASR technology have resulted in free and accessible web-based speech recognition tools that approach the performance of human listeners. The present study indicates that ASR may be useful for predicting human transcription scores for stroke and traumatic brain injury survivors with speech sound impairment. Further research is needed to determine if similar relationships exist between computer and human transcription of dysarthric speech. Clinician-researcher partnerships are needed to determine how ASR might be integrated most effectively into clinical practice. However, given the current state of clinical intelligibility documentation for adults, there is significant potential for using ASR to strengthen prognostication and outcome evaluation for people with neurological communication disorders.

Acknowledgments

We gratefully thank the participants who gave their time to be in this research study, and the graduate students who completed the intelligibility transcriptions.

Statement of Ethics

All participants provided informed consent; the study was approved by the University of North Carolina's Institutional Review Board.

Disclosure Statement

The authors have no conflicts of interest to disclose.

Funding Sources

The study was funded in part by a grant from the National Institutes of Health (R03 DC011881).

References

1 Kent RD. Introduction. In: Kent RD, editor. Intelligibility in speech disorders: theory, measurement and management. Amsterdam, The Netherlands: John Benjamins Publishing; 1992.

2 Dagenais PA, Brown GR, Moore RE. Speech rate effects upon intelligibility and acceptability of dysarthric speech. Clin Linguist Phon. 2006 Apr-May;20(2-3):141–8.

3 Sheard C, Adams RD, Davis PJ. Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. J Speech Hear Res. 1991 Apr;34(2):285–93.

4 Yorkston KM, Beukelman DR. A comparison of techniques for measuring intelligibility of dysarthric speech. J Commun Disord. 1978 Dec;11(6):499–512.

5 Beukelman DR, Yorkston KM. Influence of passage familiarity on intelligibility estimates of dysarthric speech. J Commun Disord. 1980 Jan;13(1):33–41.

6 Tikofsky RS. A revised list for the estimation of dysarthric single word intelligibility. J Speech Hear Res. 1970 Mar;13(1):59–64.

7 Tikofsky RS, Tikofsky RP. Intelligibility measures of dysarthric speech. J Speech Hear Res. 1964 Dec;7(4):325–33.

8 Haley KL, Wertz RT, Ohde RN. Single word intelligibility in aphasia and apraxia of speech. Aphasiology. 1998 Jul;12(7-8):715–30.

9 Haley KL, Bays GL, Ohde RN. Phonetic properties of aphasic-apraxic speech: a modified narrow transcription analysis. Aphasiology. 2001 Dec;15(12):1125–42.

10 Haley KL, Jacks A, Cunningham KT. Error variability and the differentiation between apraxia of speech and aphasia with phonemic paraphasia. J Speech Lang Hear Res. 2013 Jun;56(3):891–905.

11 King JM, Watson M, Lof GL. Practice Patterns of Speech-Language Pathologists Assessing Intelligibility of Dysarthric Speech. J Med Speech-Lang Pathol. 2012 Mar;20:1–10.

12 Schiavetti N. Scaling procedures for the measurement of speech intelligibility. In: Kent RD, editor. Intelligibility in speech disorders: theory, measurement and management. Amsterdam, The Netherlands: John Benjamins Publishing; 1992.

13 Schiavetti N, Metz DE, Sitler RW. Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: evidence from a study of the hearing impaired. J Speech Hear Res. 1981 Sep;24(3):441–5.

14 Yorkston KM, Beukelman DR, Traynor C. Assessment of Intelligibility of Dysarthric Speech. Austin (TX): Pro-Ed, Inc.; 1984.

15 Yorkston KM, Beukelman DR, Hakel M, Dorsey M. Speech intelligibility test for windows. Lincoln (NE): Communication Disorders Software; 1996.

16 Enderby PM, Palmer R. Frenchay Dysarthria Assessment, Second Edition (FDA-2). Austin (TX): Pro-ed, Inc.; 2008.

17 Wambaugh JL, Nessler C, Cameron R, Mauszycki SC. Acquired apraxia of speech: the effects of repeated practice and rate/rhythm control treatments on sound production accuracy. Am J Speech Lang Pathol. 2012 May;21(2):S5–27.

18 Wambaugh JL, Wright S, Mauszycki SC, Nessler C, Bailey D. Combined aphasia and apraxia of speech treatment (CAAST): systematic replications in the development of a novel treatment. Int J Speech Lang Pathol. 2018 Apr;20(2):247–61.

19 Haley KL. Chapel Hill Multilingual Intelligibility Test. Chapel Hill (NC): UNC School of Medicine; 2011.

20 Haley KL, Roth H, Grindstaff E, Jacks A. Computer-Mediated Assessment of Intelligibility in Aphasia and Apraxia of Speech. Aphasiology. 2011;25(12):1600–20.

21 Miller N, Deane KH, Jones D, Noble E, Gibb C. National survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: therapists' practices. Int J Lang Commun Disord. 2011 Mar-Apr;46(2):189–201.

22 Collis J, Bloch S. Survey of UK speech and language therapists' assessment and treatment practices for people with progressive dysarthria. Int J Lang Commun Disord. 2012 Nov-Dec;47(6):725–37.

23 Gurevich N, Scamihorn SL. Speech-Language Pathologists' Use of Intelligibility Measures in Adults With Dysarthria. Am J Speech Lang Pathol. 2017 Aug;26(3):873–92.

24 Davis KH, Biddulph R, Balashek S. Automatic recognition of spoken digits. J Acoust Soc Am. 1952 Nov;24(6):637–42.

25 Sy BK, Horowitz DM. A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition. IEEE Trans Biomed Eng. 1993 Dec;40(12):1282–98.

26 Thomas-Stonell N, Kotler AL, Leeper H, Doyle P. Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. Augment Altern Commun. 1998 Jan;14(1):51–6.

27 Blaney B, Wilson J. Acoustic variability in dysarthria and computer speech recognition. Clin Linguist Phon. 2000 Jan;14(4):307–27.

28 Raghavendra P, Rosengren E, Hunnicutt S. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. Augment Altern Commun. 2001 Jan;17(4):265–75.

29 Mengistu KT, Rudzicz F. Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. In: Butz C, Lingras P, editors. Advances in artificial intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 291–300.

30 Dimauro G, Di Nicola V, Bevilacqua V, Caivano D, Girardi F: Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System. IEEE Access. 2017;5:22199–208.

31 Ballard KJ, Etter N, Shen S, Monroe P, Tian-Tan C. Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia [poster]. 19th Biennial Conference on Motor Speech, 2018 Feb 23, Savannah, GA, USA.

32 Ferrier L, Shane H, Ballard H, Carpenter T, Benoit A. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. Augment Altern Commun. 1995 Jan;11(3):165–75.

33 Doyle PC, Leeper HA, Kotler AL, Thomas-Stonell N, O'Neill C, Dylke MC, et al. Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. J Rehabil Res Dev. 1997 Jul;34(3):309–16.

34 Hattori M, Sumita YI, Kimura S, Taniguchi H. Application of an automatic conversation intelligibility test system using computerized speech recognition technique. J Prosthodont Res. 2010 Jan;54(1):7–13.

35 Rosdi F, Mustafa MB, Salim SS. Assessing automatic speech recognition in measuring speech intelligibility: a study of Malay speakers with speech impairments. In: 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). IEEE; 2017. p. 1–6.

36 Jacks A, Haley KL. Auditory masking effects on speech fluency in apraxia of speech and aphasia: comparison to altered auditory feedback. J Speech Lang Hear Res. 2015 Dec;58(6):1670–86.

37 Kertesz A. Western Aphasia Battery-Revised. San Antonio (TX): Pearson; 2006.

38 Rothauser EH, Chapman WD, Guttman N, Hecker MH, Nordby KS, Silbiger HR, et al. IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust. 1969;17:227–46.

39 Hillenbrand JM, Gayvert RT. Open source software for experiment design and control. J Speech Lang Hear Res. 2005 Feb;48(1):45–60.

40 IBM. IBM Watson Speech-to-Text [Internet]. IBM; 2018 [cited 2018 Jun 6]. Available from: https://www.ibm.com/watson/services/speech-to-text/.

41 IBM. IBM Cloud/Bluemix programming platform [Internet]. IBM Cloud [cited 2018 Jul 13]. Available from: https://console.bluemix.net/docs/.

42 Loper E, Bird S, Lippincott T. Levenshtein edit distance; Natural Language Toolkit [Internet]. Source code for nltk.metrics.distance 2018 [cited 2018 Jun 5]. Available from: https://www.nltk.org/_modules/nltk/metrics/distance.html.

43 SAS Institute Inc. JMP. Cary (NC): SAS Institute, Inc.; 2016.

44 R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.

45 Revelle W. psych: procedures for personality and psychological research. Evanston, IL: Northwestern University; 2017.

46 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016 Jun;15(2):155–63.

47 Miller N. Measuring up to speech intelligibility. Int J Lang Commun Disord. 2013 Nov-Dec;48(6):601–12.

48 Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process Mag. 2012 Nov;29(6):82–97.

49 Weismer G, Laures JS. Direct magnitude estimates of speech intelligibility in dysarthria: effects of a chosen standard. J Speech Lang Hear Res. 2002 Jun;45(3):421–33.

50 Tjaden K, Sussman JE, Wilding GE. Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. J Speech Lang Hear Res. 2014 Jun;57(3):779–92.

51 Yunusova Y, Weismer G, Kent RD, Rusche NM. Breath-group intelligibility in dysarthria: characteristics and underlying correlates. J Speech Lang Hear Res. 2005 Dec;48(6):1294–310.

52 Haley KL, Jacks A, Truong YK. Documenting intelligibility in speakers with aphasia: how many listeners are needed? Paper presented at American Speech-Language-Hearing Association Convention, 2013 November, Chicago, IL, USA.