Autoscore: An open-source automated tool for scoring listener perception of speech

Stephanie A. Borrie, Tyson S. Barrett, and Sarah E. Yoho

Citation: The Journal of the Acoustical Society of America **145**, 392 (2019); doi: 10.1121/1.5087276
View online: https://doi.org/10.1121/1.5087276
View Table of Contents: https://asa.scitation.org/toc/jas/145/1
Published by the Acoustical Society of America

Autoscore: An open-source automated tool for scoring listener perception of speech

Stephanie A. Borrie[a)]
Department of Communicative Disorders and Deaf Education, Utah State University, Logan, Utah 84322, USA

Tyson S. Barrett
Department of Psychology, Utah State University, Logan, Utah 84322, USA

Sarah E. Yoho
Department of Communicative Disorders and Deaf Education, Utah State University, Logan, Utah 84322, USA

(Received 12 October 2018; revised 26 November 2018; accepted 10 December 2018; published online 25 January 2019)

Speech perception studies typically rely on trained research assistants to score orthographic listener transcripts for words correctly identified. While the accuracy of the human scoring protocol has been validated with strong intra- and inter-rater reliability, the process of hand-scoring the transcripts is time-consuming and resource intensive. Here, an open-source computer-based tool for automated scoring of listener transcripts is built (Autoscore) and validated on three different human-scored data sets. Results show that not only is Autoscore highly accurate, achieving approximately 99% accuracy, but extremely efficient. Thus, Autoscore affords a practical research tool, with clinical application, for scoring listener intelligibility of speech.
 https://doi.org/10.1121/1.5087276

[DDO] Pages: 392–399

I. INTRODUCTION

Studies that examine speech perception in adverse conditions (i.e., disordered speech, accented speech, speech in noise), frequently rely on an objective measure of percent words correct to generate data regarding intelligibility of the acoustic signal (e.g., Bilger *et al.*, 1984; Liss *et al.*, 2000). To obtain these data, listener participants are often presented with audio files of spoken stimuli (target words or phrases) and asked to orthographically transcribe what they think is being said or, in some cases (particularly with older listeners, e.g., McAuliffe *et al.*, 2013), repeat what they think is being said aloud with study personnel orthographically transcribing the responses. Researchers then employ trained laboratory research assistants (or complete the task themselves) to hand score these orthographic listener transcripts for the number of words correctly identified, according to a set of previously determined rules regarding what is acceptable to be counted as correct. A measure of percent words correct, or speech intelligibility, can then be calculated by dividing the number of words scored as correct by the total number of words possible (Yorkston and Beukelman, 1980).

The process of scoring orthographic listener transcripts for words correct by hand appears to be largely accurate. While not all studies report on reliability, those that have demonstrated a high inter-rater agreement (e.g., Stilp *et al.*, 2010; Hustad, 2006; Huyck, 2018). There are, therefore, no substantial concerns about the accuracy of the current scoring approach. However, the process of scoring the transcripts by hand is notoriously time-consuming. In a recent study by Borrie and colleagues (2017a), the data set consisted of 160 listener transcripts, with each transcript consisting of 315 words. Scoring of words correct for each individual transcript took, on average, 15 min, so the total time for one rater to score the complete data set approximated 40 h. A second rater analyzed 20% of randomly selected transcripts for a measure of inter-rater reliability. Adding this time to the workload, the total time for scoring the data set for this study, including reliability checks, approximated 50 h.

A small handful of speech perception studies have reported the use of in-house computer software to automatically score listener transcripts for words correct (e.g., Allison and Hustad, 2014; Wild *et al.*, 2018). However, there are no well-documented, open-source software packages that can automatically score listener transcripts for this objective intelligibility measure. Furthermore, and importantly, there is a paucity of studies documenting the efficiency and accuracy of in-house automated methods for scoring words correct. The purpose of the current study was to build and evaluate an open-source computer-based application, Autoscore, which automates the scoring of orthographic listener transcripts for an objective measure of words correct. If this automated tool is to have practical application, it must not only be efficient, reducing time to score data sets, but also very accurate. Here, we apply Autoscore to three established human-scored data sets, each with different scoring rules, and evaluate its performance relative to the performance of human scores in terms of efficiency and accuracy.

A. Autoscore

The Autoscore tool is built on open-source software through the R statistical environment (R Core Team, 2018),

[a)]Electronic mail: stephanie.borrie@usu.edu

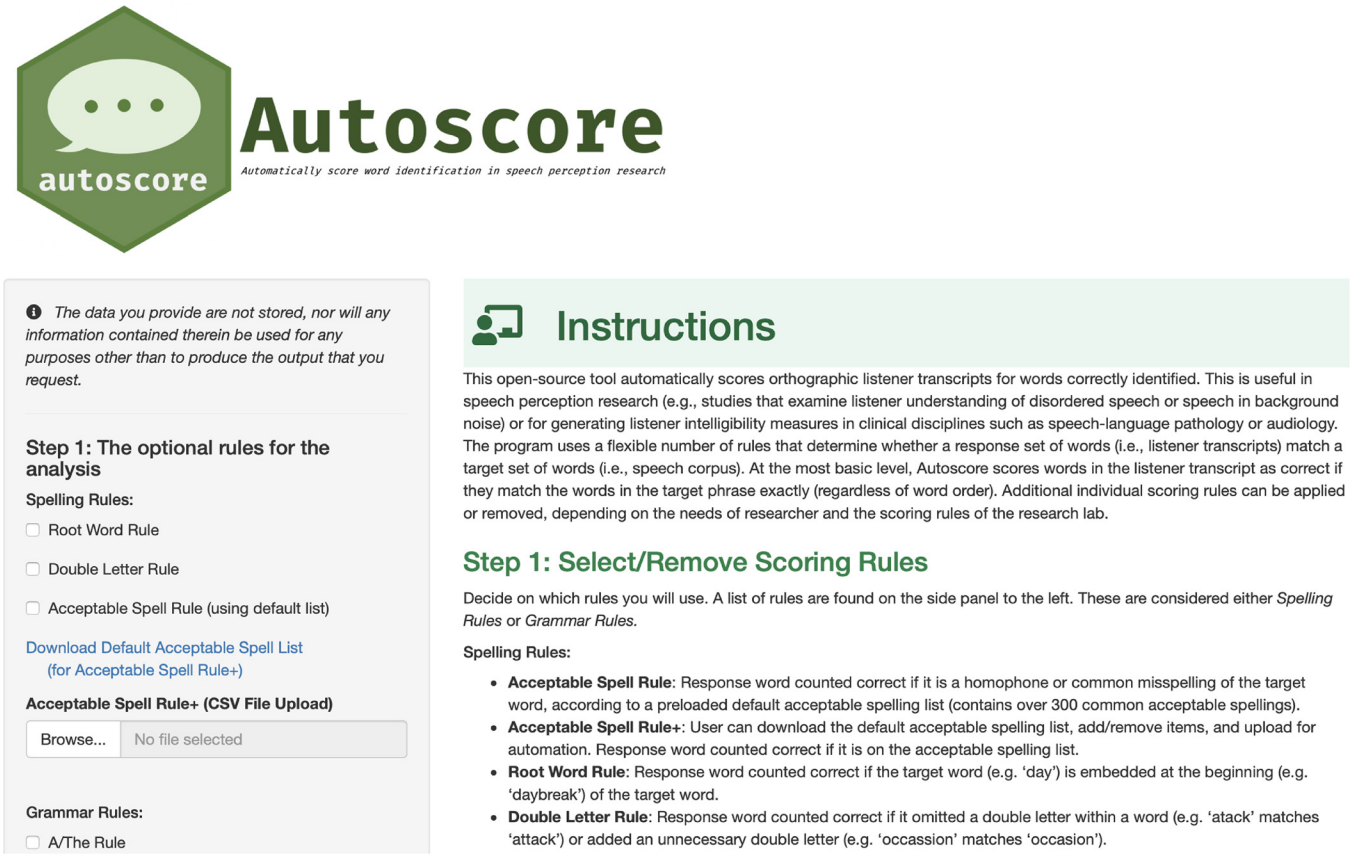

FIG. 1. (Color online) The interface of the online application.

using a variety of packages to build the specific algorithms (Bache and Wickham, 2014; Barrett and Brignone, 2017; Csárdi, 2017; Feinerer *et al.*, 2008; Henry and Wickham, 2018; Müller and Wickham, 2018; Wickham, 2018; Wickham *et al.*, 2018; Wickham and Henry, 2018). The tool is accessible via two media:

(1) An online version that uses the Shiny web tool that allows the user to interact with the R code via a point-and-click interface (see Fig. 1), and
(2) via the R statistical environment as a package for researchers with experience in coding.

Both systems use the same algorithms and will provide identical results. Herein, we emphasize the online tool as it is most accessible to users in the field.

The online tool is flexible to the needs of the user. This is particularly important given that no absolute standardization of the scoring rules for words correct has been adopted currently, even in the subdisciplines of speech perception research [see Hustad (2006) for a discussion in perception of dysarthric speech]. Table I highlights the various scoring rules that Autoscore can apply to the orthographic listener transcripts. These rules are stratified by whether they adjust the spelling that is considered correct (e.g., double letters, root words) or grammar particulars of the word (e.g., past-tense, plurals).

The online version of Autoscore does not save any data provided to it. Rather, it runs the analyses on a server, provides the output on the screen with options to download the data, and then once the user closes the browser, the server deletes the data used in that analysis. Further, Autoscore does not log the user's IP address. The only data stored is a log of the use of Autoscore which simply details the date and time of use.

B. Using Autoscore

To use the online version of Autoscore, the user can navigate to http://autoscore. usu.edu.[1] On the website, instructions for its use are described in detail via three steps. To highlight its use, we will walk through each step.

Step 1. The first step in using Autoscore is deciding the scoring rules that will be applied to the transcripts (see Table I for a list of scoring rules currently available). None of the rules are active by default; therefore, rules must be selected to be used. It is also in this initial step that a user can, optionally, update the default acceptable spell file (currently contains over 300 words; Acceptable Spell Rule) with any additional acceptable spellings (i.e., homophones, common misspellings) of target stimuli that should be scored as correct (Acceptable Spell Rule+). Thus, the default acceptable spell list is a CVS file that users can download, update, and upload to the application.

Step 2. Once the rules are selected, the next step is uploading the orthographic listener transcripts, CSV file(s) containing both the target and response words or phrases. Notably, multiple files can be uploaded simultaneously but should be formatted the same to reduce risk of errors. To

TABLE I. The optional scoring rules that researchers can use to adjust the types of response words that Autoscore counts as matches to the target words.

Scoring Rule	Function
Spelling Rules	
Acceptable Spell Rule	Response word counted correct if it is a homophone or common misspelling of the target word, according to a preloaded default acceptable spelling list (contains over 300 common acceptable spellings).
Acceptable Spell Rule+	User can download the default acceptable spelling list, add/remove items, and upload for automation. Response word counted correct if is on the acceptable spelling list.
Double Letter Rule	Response word counted correct if it omitted a double letter within a word (e.g., "atack" matches "attack") or added an unnecessary double letter (e.g., "occassion" matches "occasion")
Root Word Rule	Response word counted correct if the target word (e.g., "day"') is embedded at either the beginning (e.g., "daybreak") or end (e.g., "monday") of the target word
Grammar Rules	
Tense Rule	Response word counted correct if it differs from the target word by the addition or omission of "d" or "ed" (e.g., "assumed" matches "assume" and "jump" matches "jumped")
Tense+ Rule	Response word counted correct if differs by the target word by the addition (not omission) of "d" or "ed" (e.g., "jumped" matches "jump" but "jump" does not match "jumped")
Plural Rule	Response word counted correct if it differs from the target word by the addition or omission of "s" or "es" (e.g., "cats" matches "cat" and "echo" matches "echoes")
Plural+ Rule	Response word counted correct if differs from the target word by the addition (not omission) of "s" or "es" of the target word (e.g., "cats" matches "cat" but "cat" does not match "cats")
A/The Rule	Substitutions of "a" and "the" are scored as matches

format them, there is only a single need for the program to run—an "id" column (containing a listener identifier), a "target" column (containing the target words or phrases), and a "response" column (containing the listener's responses). Optionally, a "human" column (containing the prior scoring of human raters) can also be included. This human scoring column can be used for comparison between human- and computer-scored values. These columns can be in any order as long as they are labeled in the first row of the data sheet and labeled as described here. Note, column labels, target words, and response words are not upper/lower case dependent. With the data in this format, additional columns (i.e., columns indicating experimental conditions or subgroups) are completely acceptable and will not interfere with the analyses. If the user uploads multiple files, the output will end up as a combined data set with all information from all the files in the single output file. Thus, the need for the id column.

Step 3. Once the file(s) is(are) uploaded, Autoscore will then start the computations. If there are many files or each file is large, the tool will let you know that it is working on the computations and will output once it is finished. The output will be printed in an interactive table at the bottom of the screen. This table can be sorted by the individual columns for quick checks. Additionally, various ways to download the file will be shown at the top of the table (i.e., CSV, Excel).

II. METHOD

To perform a primary validation of Autoscore, we evaluated two data sets collected by the authors and previously scored by human raters using lab-specified scoring rules. Specifically, we assessed the *efficiency* (the time taken for scoring) and the *accuracy* (the amount of correct scoring as validated by two independent research assistants blind to whether the score was produced by a human or automation) of Autoscore as compared to the human scoring. For the evaluation of Autoscore, we applied the same rules, or as close as possible (see data set details below), used by the original raters to score transcripts for words correct across two data sets. To perform a secondary validation of Autoscore, we evaluated its accuracy on an independent data set published by Stilp and colleagues (2010).

A. Data set 1

Data set 1 was from a study on perceptual processing of neurologically degraded speech looking at the relationship between rhythm perception abilities and the ability to decipher the dysarthric speech (Borrie *et al*., 2017b). The data set consisted of 50 listener transcripts (listeners used a computer to type out what they thought the speaker was saying), each consisting of 80 phrases, for a total of 4000 phrases transcribed overall. The phrases were semantically anomalous, all six syllables, and ranging from 3 to 5 words in length. All words in each phrase were scored. A standard scoring procedure, developed by Liss and colleagues (1998) for studies on listener processing of dysarthric speech was used for hand scoring the orthographic transcripts. According to this protocol, raters were instructed to score words as correct if they matched the intended target precisely, or differed only by tense "ed" or plural "s" without adding another syllable (e.g., assume/assumed is counted correct but amend/amended is not). Substitutions between "a" and "the" were also regarded as correct, as were homophones and obvious misspellings. Words were counted as correct regardless of what order they were repeated in.[2] One trained rater scored all 50 transcripts. Twenty percent of the transcripts were then randomly selected and reanalysed by the original rater (intra-rater) and a second trained rater (inter-rater) to obtain reliability estimates for scoring of words correct. Discrepancies revealed high intra- and inter-rater agreement, with Pearson correlation r scores of

TABLE II. Types of errors made when scoring orthographic transcripts for words correct.

Error type	Description
UW	When Autoscore or Human does not score a response word as correct, even though it is (e.g., no point for response "pen" when target is "pen")
MW	When Autoscore or Human score a response word that does not match the target (e.g., a point for response "man" when target is "mean")
OM	When Autoscore or Human does not score a response word that is an OM of the target (e.g., no point for response "believe" when target is "believe")
AE	When Autoscore makes an error because of the automation process (e.g., a point for "see" when the target is "seed" because of the Tense Rule)
RW	When Autoscore or Human does not score a response word that differs from the target by the addition of an extra syllable (e.g., no point for response "enjoy" when target is "joy")
TE	When Autoscore or Human does or does not score a response word that differs from the target by "d" or "ed" (e.g., no point for response "assumed" when target is "assume")
PE	When Autoscore or Human does or does not score a response word that differs from the target by "s" or "es" (e.g., no point for response "trains" when target is "train")
TA	When Autoscore or Human does or does not score "the" and "a" as a match (e.g., no point for response "the" when target is "a")

0.98. Total time scoring the full data set, including scoring for reliability, was approximately 20 h.

B. Data set 2

Data set 2 was from a study of speech perception in noise looking at the effect of listener and talker sex on speech intelligibility (Yoho *et al.*, 2018). The data set included responses from 50 listeners who each heard 100 sentences, for a total of 5000 sentences heard overall. The sentences were from the Harvard IEEE speech corpus (IEEE, 1969), where each sentence contains five key words[3] to be scored. In this study, the listener repeated back as much of each sentence as they could, and the experimenter typed out what the listener repeated. The procedure for scoring the orthographic transcripts consisted of the following rules: the words were counted as correct if they were repeated back precisely, or if the listener added the tense "ed" or the plural "s." Tense and plural omissions were not scored as correct. Words were also counted as correct if a syllable was added to the word but not if a syllable was omitted without changing the pronunciation of the word (e.g., batman/bat is correct but assert/assertion is not). Homophones and obvious misspellings were also counted as correct. Words were counted as correct regardless of what order they were repeated in. Two trained raters scored all transcripts to ensure inter-rater reliability of scoring. An analysis of inter-rater reliability indicated that the two scorers agreed 98% of the time. Total time scoring the full data set, including checking for agreement, was approximately 30 h.

C. Independent data set

The independent data set was from a study of speech perception of temporally distorted sentences across a wide range of simulated speaking rates (Stilp *et al.*, 2010). The data set included a total of 12 195 sentences heard across 129 listeners. The sentences were from the Hearing In Noise Test corpus (Nilsson *et al.*, 1994). Three raters independently scored the typed responses according to guidelines listed in the published appendix.

D. Primary evaluation

To afford a broad evaluation of the accuracy and features of Autoscore, in-house data sets were processed by Autoscore according to multiple levels of rule application: (i) applying standard automated rules (basic-level), (ii) applying standard automated rules and the default acceptable spelling list (Acceptable Spell Rule; mid-level), and (iii) applying standard automated rules and employing the user option to download the default acceptable spelling list and add additional acceptable spellings (Acceptable Spell Rule+; full-level; only data set 2 was evaluated at this level, given the more subjective nature of the human scoring rules). The standard automated rules for data set 1 included the Double Letter rule, Tense Rule, Plural rule, and T/A rule. The standard automated rules for data set 2 included the Root Word Rule, Tense+ Rule, and the Plural+ Rule (see Table I for rule definitions).

Two trained research assistants, different from the original human rater, then coded all cases where there were discrepancies between human rater and Autoscore, identifying which of the two approaches produced the correct score for each target word. The research assistants also classified each discrepancy between human scorer and Autoscore for error type, according to the following six error categories: (1) Unmarked Correct Word (UW); (2) Marked Wrong Word (MW); (3) Obvious Misspelling (OM); (4) Automated Error (AE); (5) Root Word Error (RW); (6) Tense Error (TE); (7) Plural Error (PE); and (8) The/A Error (TA). Explanations and examples of the error types can be found in Table II.

E. Secondary evaluation

The secondary evaluation of the accuracy of Autoscore was carried out on an independent data set, namely Stilp *et al.* (2010). The independent data set was processed by using a mid-level analysis. The standard automated rules for processing this data set included the Double Letter rule,

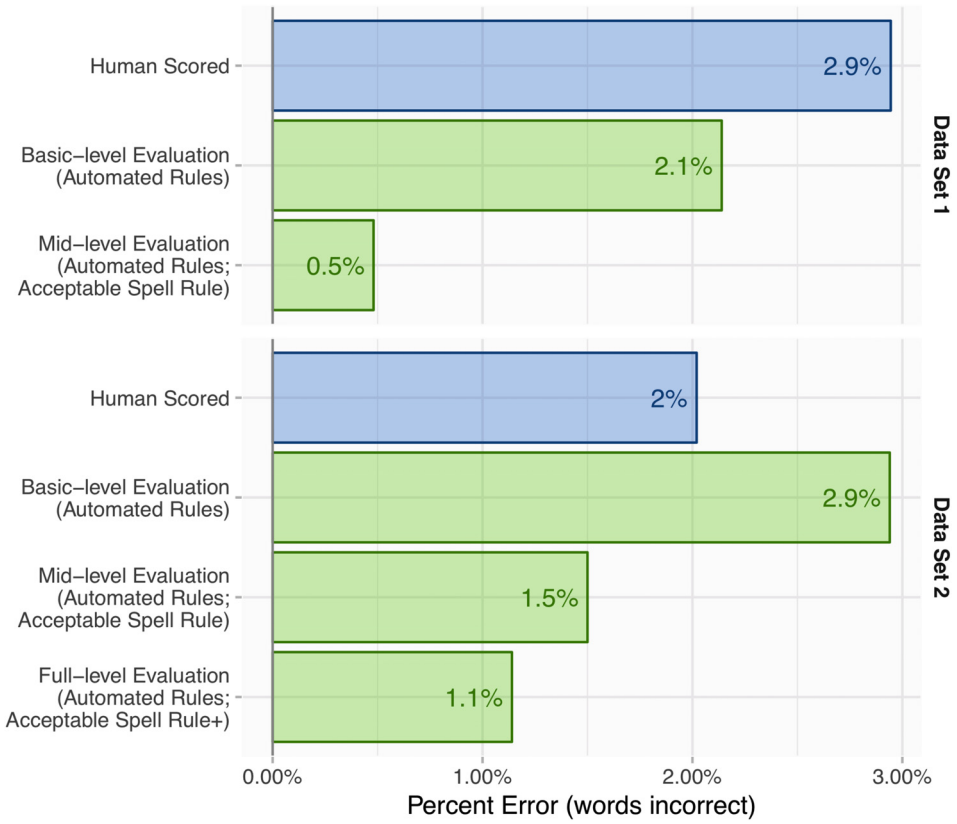

FIG. 2. (Color online) The error rates of scoring orthographic listener transcripts for a measure of words correct for Autoscore and human scorers for both in-house data sets.

Tense Rule, Plural rule, T/A rule, and the default acceptable spelling list.

III. RESULTS

A. Efficiency

Autoscore consistently scored the data sets for words correct faster on a scale of hours to seconds. For example, for the mid-level analysis, Autoscore took 30 s, on average, to score each of the data sets, approximately 2500 times faster than human scorers. All Autoscore benchmarks were performed on a 3.4 GHz Intel Core i5 processor with 8 GB of RAM.

B. Accuracy: Primary evaluation

Accuracies for both human scorers and Autoscore were high across both in-house data sets, where Autoscore, overall, slightly outperformed the human scorers for both data sets (see Fig. 2). In the basic-level evaluation using just the standard automated rules, Autoscore yielded accuracies of 97.9% and 97.1% for data sets 1 and 2, respectively. In the mid-level evaluation in which the default acceptable spelling list (Acceptable Spell Rule) was added to the standard automated rules, Autoscore accuracy increased to 99.5% and 98.5% for data sets 1 and 2, respectively. Finally, in the full-level evaluation in which the Autoscore feature of downloading and adding lab- and/or corpus-specific suitable spellings to the acceptable spelling list was added to the standard automated rules (Acceptable Spell Rule+), Autoscore accuracy increased to 99.1% for data set 2. Human raters were 97.1% and 98.0% accurate for data sets 1 and 2, respectively. Inter-rater reliability between the human raters and Autoscore were very high, with Cronbach alphas of 0.99 for both data sets.

Further, a chi-square analysis revealed that the ways in which the human raters and Autoscore made mistakes differed significantly ($\chi^2 = 312.82$, degrees of freedom $= 14$, $p < 0.001$). As reported in Table III, the vast majority of human rater errors were either an UW (the human scorer did not count a correct word), MW (the human rater marked a word as correct when it did not match), and TA (the human scorer did not match "the" and "a"). In contrast, the majority of the errors committed by Autoscore were OM errors, reflecting obvious spelling mistakes. The number of OM errors decreased substantially when the default acceptable spelling list was added to the analysis (mid-level) and further decreased when the researcher-supplied, lab- and corpus-specific suitable spellings were added to the list (full-level). Beyond the OM errors, Autoscore produced a few AE (from rule combining) and RW errors. Notably, Autoscore did not produce any UW, MW, PE, or TA errors.

C. Accuracy: Secondary evaluation

Accuracies for both human raters and Autoscore (applying a mid-level analysis) were high for the independent data set, achieving 98.3% and 95.3% correct, respectively. Error analysis revealed that upward of 80% of errors made by Autoscore on the independent data set were OMs. This suggests that the addition of a researcher supplied suitable spellings to the acceptable spelling list (allowing the application of a full-level analysis) could increase accuracy of Autoscore to approximately 99%. Thus, independent users

TABLE III. Errors committed by Autoscore and the human scorers on the two in-house data sets. Percentages, in parentheses, are for the proportions of total errors made.

	Autoscore			
Error type	Basic	Mid	Full[a]	Human
UW	0 (0%)	0 (0%)	0 (0%)	87 (38.2%)
MW	0 (0%)	0 (0%)	9 (15.8%)	70 (30.7%)
OM	200 (84.7%)	57 (60%)	25 (43.9%)	5 (2.2%)
AE	16 (6.8%)	18 (18.9%)	5 (8.8%)	0 (0%)
RW	20 (8.5%)	20 (21.1%)	17 (29.8%)	8 (3.5%)
TE	0 (0%)	0 (0%)	1 (1.8%)	15 (6.6%)
PE	0 (0%)	0 (0%)	0 (0%)	2 (0.9%)
TA	0 (0%)	0 (0%)	0 (0%)	41 (18%)
Total Errors	236 of 107 944 words	95 of 107 944 words	57 of 75 000 words	228 of 107 944 words

[a]Only data set 2 was used in the full-level evaluation of Autoscore.

may consider it well worth their time up front to develop a detailed list of acceptable misspellings (Acceptable Spell Rule+) for improved Autoscore accuracy, particularly if the speech target list will be used in multiple studies.

IV. DISCUSSION

We describe Autoscore, an open-source tool for scoring orthographic listener transcripts for an objective measure of speech intelligibility, words correct, and validate the tool in terms of efficiency and accuracy of scoring. Evidence of tool validation is observed in three key results. First, the time that Autoscore takes to score the transcripts cannot be matched by human raters. Autoscore is upward of 2500 times faster at scoring study data sets than human scorers. Thus, Autoscore greatly reduces the time taken to generate data, consequently affording substantial savings on lab resources in terms of human labor and compensation. Second, while the accuracy of having human raters hand score orthographic transcripts is not an issue, Autoscore achieved scoring accuracies as high or higher in the two in-house data sets and could achieve that for the independent data set if an acceptable spelling list was provided. Third, the reliability estimates between human raters and Autoscore were on par with the inter-rater reliability estimates that have been previously reported between human raters.

While accuracy between human and automated scoring is comparable, analysis of error types revealed some noteworthy differences in the way in which humans and Autoscore made mistakes. Human raters made errors such as UWs (i.e., no point given for a response that matches the target) and MWs (i.e., giving a point for a response that does not match the target). Such errors may reflect some level of decision-making fatigue, which is not surprising given the laborious task of hand scoring large data sets. Autoscore, on the other hand, made no errors of this nature. The most common errors committed by Autoscore were OM errors (i.e., no point given for a response that is an OM of the target). Beneficially, these OM errors were largely remedied by using the Acceptable Spell Rule, whereby a default list of homophones and misspellings of the target words was used by Autoscore, increasing the accuracy by approximately 2.4% (mid-level evaluation). Further, by adding researcher supplied suitable spellings to the acceptable spelling list for data set 2 via the Acceptable Spell Rule+, accuracy further increased by 0.4% (full-evaluation). Making use of the Acceptable Spell Rule+ for the independent data set would certainly have resulted in meaningful benefits to scoring accuracy of Autoscore. It is worth noting that the default list of homophones and acceptable misspellings was developed by the authors of this paper and thus is best suited to the targets encountered in the in-house data speech corpora, evident in higher mid-level accuracy scores relative to the independent data set. Indeed, the default list was so well suited to data set 1 that making any additions via the Acceptable Spell Rule+ was not necessary. Thus, whether or not users should add additional spellings to the default list will depend on the needs of the research lab and/or the uniqueness of the speech corpus, possibly based on a quick assessment of the responses.

A nuance of Autoscore is that it cannot perform some of the more subjective evaluations that trained human raters are able to make. This is apparent in the application of the Tense and Plural Rule to the perception of dysarthric speech in which, as noted for data set 1, these rules should only apply when they do not change the syllabic structure of the target word. Autoscore, of course, awards a point for the response word if it differs by "e," "ed," "s," "es," regardless of whether it adds or omits a syllable. Similarly, Autoscore may over-score for the Root Word Rule in studies such as data set 2, where that particular rule should only be successful when the addition of a syllable does not alter the pronunciation of the word. Another shortcoming of Autoscore is that it can, at times, combine rules in odd ways, resulting in the incorrect scoring of a response word (i.e., AEs). For example, scoring the response word "cold" as correct for the target word "cool." In this instance, Autoscore applied both the Double Letter and Tense Rule. Despite these nuances, Autoscore performed at an accuracy level comparable to that of trained human raters, suggesting that idiosyncrasies associated with automated scoring may be noncritical to overall study outcomes.

With multiple scoring rule options, Autoscore is highly flexible, and thus is appropriate for many subdomains of speech perception research, including but not limited to

perception of dysarthric speech (e.g., Hustad *et al.*, 2003; Liss *et al.*, 2002; McAuliffe *et al.*, 2013), speech in noise (e.g., Cooke *et al.*, 2013; Luce and Pisoni, 1998; Van Engen *et al.*, 2014), accented speech (e.g., Bradlow and Bent, 2008; Munro, 1998), noise-vocoded speech (e.g., Davis *et al.*, 2005; Guediche *et al.*, 2016), or speech perception by the hearing impaired (Healy *et al.*, 2013; Tye-Murray *et al.*, 2007). Indeed, no accepted standard set of scoring rules exists across studies in speech perception, yet such an ideal may not be warranted. For example, given the propensity for high-frequency hearing loss and resulting difficulty in identifying high frequency phonemes (e.g., Hogan and Turner, 1998; Turner and Cummings, 1999), studies that examine speech perception of listeners with impaired hearing may prioritize the Plural -s Addition Rule, in which a response word is scored as incorrect if it differs from the target word by the omission of the phoneme "s" or "es," but correct if it adds these phonemes. Conversely, studies that examine perception of dysarthric speech may not accept additions that alter the syllabic structure of target word (e.g., Liss *et al.*, 2002; Borrie *et al.*, 2012), given that such changes likely reflect disordered productions of the acoustic signal.

Further, Autoscore was specifically built to serve as an adaptable, open-source application, allowing users to program additional scoring rules if desired.[2] Helpfully, the "human scoring" column in the data output allows users to debug new rules by performing their own accuracy comparisons between Autoscore and human raters. Although Autoscore was developed for the English language, it would be relatively straightforward for a user to build on the core functionality to add rules specific to other languages. The Autoscore code can also be embedded in experiments to provide real-time feedback on listener performance during speech perception tasks. For example, many studies employ a measure of speech reception threshold, or SRT, as a sensitive means of listener performance under various speech in noise conditions (e.g., Festen and Plomp, 1990; Wang *et al.*, 2009). This measure requires adaptation of the signal-to-noise ratio over the course of the experiment based on performance, with an individual's SRT being determined as the signal-to-noise ratio required for them to achieve some predetermined level of intelligibility. Real-time scoring of words correct could increase both accuracy and efficiency of adapting stimuli parameters. Another application of the Autoscore code is in studies that exploit perceptual learning to train listeners to better understand degraded speech (i.e., dysarthric speech, e.g., Borrie *et al.*, 2012; Borrie *et al.*, 2017a,b). Research in perceptual learning of dysarthric speech has revealed large individual differences in the ability to adapt to the degraded signal (Borrie *et al.*, 2017b). Thus, the addition of real-time feedback regarding learning progress would allow perceptual training programs to be adaptive, modulating the amount and type of training depending on the progress of the learner.

Autoscore also offers a number of clinical applications in fields such as speech pathology and audiology. For example, speech-language pathologists frequently use orthographic transcription of the acoustic signal by listeners to obtain measures of speech intelligibility (percentage intelligibility) for clients with speech disorders such as dysarthria. This objective measure enables clinicians to quantify an overall measure of the understandability of speech and is often used as an index of severity (Strand and Yorkston, 1994) and to document treatment progress and effectiveness (Yorkston *et al.*, 1990). Audiologists also use a measure of percentage intelligibility to assess and manage speech-in-noise abilities of listeners with hearing loss (e.g., Killion *et al.*, 2001). Thus, a quick, reliable, and easy-to-use tool for scoring orthographic transcripts of the acoustic speech signal has implications beyond the needs of researchers.

In sum, we built an open-source tool for scoring orthographic listener transcripts of the acoustic signal for an objective measure of speech intelligibility, words correct. We validated the tool in its entirety on two unique data sets from two different speech perception research labs, demonstrating high levels of accuracy as well as a major advantage in terms of efficiency over traditional hand-scoring carried out by trained research assistants. We also validate the tool on a unique data set from an independent research lab. To our knowledge, an open-source tool for automated intelligibility scoring, validated in terms of accuracy and efficiency, does not exist. The flexible and simplistic nature of Autoscore allows the tool to be used *as is* by researchers and clinicians alike; and for those with programming experience, modified as desired.

ACKNOWLEDGMENTS

This paper was written with partial support from the National Institute of Deafness and Other Communication Disorders, National Institutes of Health Grant No. R21 DC 016084, awarded to S.A.B. We gratefully acknowledge Monica Muncy, research assistant in the Speech and Auditory Perception Lab at Utah State University for assistance with evaluation accuracy. We also gratefully acknowledge Christian Stilp at the University of Louisville for sharing his published data set, enabling us to carry out an independent evaluation of Autoscore.

[1]R package source code can be accessed at https://github.com/autoscore/autoscore.

[2]Autoscore, in its current automatic application format, is not equipped to score for word order. However, researchers with programming experience will find some code in R package as a starting place for integrating a word order rule to their own specific version of the program.

[3]If scoring of key words in a sentence is desired (e.g., he TOOK the MAN to see the FOREST), only list key words in the target column (e.g., TOOK MAN FOREST). In this way, even if listeners responses are full sentences, Autoscore will only score for target words.

Allison, K. M., and Hustad, K. C. (**2014**). "Impact of sentence length and phonetic complexity on intelligibility of 5-year-old children with cerebral palsy," Int. J. Speech Lang. Pathol. **16**(4), 396–407.

Bache, S. M., and Wickham, H. (**2014**). "magrittr: A forward-pipe operator for R," R package version 1.5. https://CRAN.R-project.org/package=magrittr (Last viewed December 1, 2018).

Barrett, T. S., and Brignone, E. (**2017**). "Furniture for quantitative scientists," R Journal **9**, 142–148.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (**1984**). "Standardization of a test of speech perception in noise," J. Speech Lang. Hear. Res. **27**, 32–48.

Borrie, S. A., Lansford, K. L., and Barrett, T. S. (**2017a**). "Generalized adaptation to dysarthric speech," J. Speech Lang. Hear. Res. **60**, 3110–3117.

Borrie, S. A., Lansford, K. L., and Barrett, T. S. (**2017b**). "Rhythm perception and its role in recognition and learning of dysrhythmic speech," J. Speech Lang. Hear. Res. **60**, 561–570.
Borrie, S. A., McAuliffe, M. J., Liss, J. M., Kirk, C., O'Beirne, G. A., and Anderson, T. (**2012**). "Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech," Lang. Cogn. Process. **27**, 1039–1055.
Bradlow, A. R., and Bent, T. (**2008**). "Perceptual adaptation to non-native speech," Cognition **106**, 707–729.
Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (**2013**). "Evaluating the intelligibility benefit of speech modifications in known noise conditions," Speech Commun. **55**(4), 572–585.
Csárdi, G. (**2017**). "crayon: Colored Terminal Output," R package version 1.3.4. https://CRAN.R-project. org/package=crayon (Last viewed December 1, 2018).
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.
Feinerer, I., Hornik, K., and Meyer, D. (**2008**). "Text mining infrastructure in R," J. Stat. Software **25**, 1–54.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736.
Guediche, S., Fiez, J. A., and Holt, L. L. (**2016**). "Adaptive plasticity in speech perception: Effects of external information and internal predictions," J. Exp. Psychol. Hum. Percept. Perform. **42**(7), 1048–1059.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**(4), 3029–3038.
Henry, L., and Wickham, H. (**2018**). "purrr: Functional programming tools," R package version 0.2.5. https://CRAN.R-project.org/package=purrr (Last viewed December 1, 2018).
Hogan, C. A., and Turner, C. W. (**1998**). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432–441.
Hustad, K. C. (**2006**). "A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners," Am. J. Speech Lang. Pathol. **15**, 268–277.
Hustad, K. C., Jones, T., and Dailey, S. (**2003**). "Implementing speech supplementation strategies: Effects on intelligibility and speech rate of individuals with chronic severe dysarthria," J. Speech Lang. Hear. Res. **46**, 462–474.
Huyck, J. (**2018**). "Comprehension of degraded speech matures during adolescence," J. Speech Lang. Hear. Res. **61**, 1012–1022.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Killion, M. C., Niquette, P. A., Revit, L. J., and Skinner, M. W. (**2001**). "Quick SIN and BKB-SIN, two new speech-in-noise tests permitting SNR-50 estimates in 1 to 2 min," J. Acoust. Soc. Am. **109**(5), 2502.
Liss, J. M., Spitzer, S. M., Caviness, J. N., and Adler, C. (**2002**). "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," J. Acoust. Soc. Am. **112**, 3022–3030.
Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., and Edwards, B. (**1998**). "Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech," J. Acoust. Soc. Am. **104**, 2457–2466.
Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., and Edwards, B. (**2000**). "Lexical boundary error analysis in hypokinetic and ataxic dysarthria," J. Acoust. Soc. Am. **107**, 3415–3424.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation Model," Ear Hear. **19**(1), 1–36.
McAuliffe, M. J., Gibson, E. M. R., Kerr, S. E., Anderson, T., and LaShell, P. J. (**2013**). "Vocabulary influences older and younger listeners' processing of dysarthric speech," J. Acoust. Soc. Am. **134**, 1358–1368.
Müller, K., and Wickham, H. (**2018**). "tibble: Simple data frames," R package version 1.4.2. https://CRAN.R-project.org/package=tibble (Last viewed December 1, 2018).
Munro, M. J. (**1998**). "The effects of noise on the intelligibility of foreign-accented speech," Stud. Second Lang. Acquist. **20**, 139–154.
Nilsson, M., Soli, S., and Sullivan, J. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
R Core Team (**2018**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (Last viewed December 1, 2018).
Stilp, C. E., Kiefte, M., Alexander, J. M., and Kluender, K. R. (**2010**). "Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences," J. Acoust. Soc. Am. **128**, 2112–2126.
Strand, E. A., and Yorkston, K. M. (**1994**). "Description and classification of individuals with dysarthria: A 10-year review," in *Motor Speech Disorders: Advances in Assessment and Treatment*, edited by J. A. Till, K. M. Yorkston, and D. R. Beukelman (Paul H. Brooks, Baltimore, MD), pp. 37–56.
Turner, C. W., and Cummings, K. J. (**1999**). "Speech audibility for listeners with high-frequency hearing loss," Am. J. Audiol. **8**, 47–56.
Tye-Murray, N., Sommers, M. S., and Spehar, B. (**2007**). "Audiovisual integration and lip reading abilities of older adults with normal and impaired hearing," Ear Hear. **28**(5), 656–668.
Van Engen, K., Phelps, J. E. B., Smiljanic, R., and Chandrasekaran, B. (**2014**). "Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker," J. Speech Lang. Hear. Res. **57**, 1908–1918.
Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**(4), 2336–2347.
Wickham, H. (**2018**). "stringr: Simple, consistent wrappers for common string operations," R package version 1.3.1. https://CRAN.R-project.org/package=string (Last viewed December 1, 2018).
Wickham, H., François, R., Henry, L., and Müller, K. (**2018**). "dplyr: A grammar of data manipulation," R package version 0.7.6. https://CRAN.R-project.org/package=dplyr (Last viewed December 1, 2018).
Wickham, H., and Henry, L. (**2018**). "tidyr: Easily tidy data with 'spread()' and 'gather()' functions," R package version 0.8.1. https://CRAN.R-project.org/package=tidyr (Last viewed December 1, 2018).
Wild, A., Vorperian, H. K., Kent, R. D., Bolt, D. M., and Austin, D. (**2018**). "Single-word speech intelligibility in children and adults with down syndrome," Am. J. Speech Lang. Pathol. **27**, 222–236.
Yoho, S. E., Borrie, S. A., Barrett, T. S., and Whittaker, D. (**2018**). "Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology," Atten. Percept. Psychophys. (in press).
Yorkston, K., and Beukelman, D. (**1980**). "A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility," J. Commun. Disord. **13**, 15–31.
Yorkston, K. M., Hammen, V. L., Beukelman, D. R., and Traynor, C. D. (**1990**). "The effect of rate control on the intelligibility and naturalness of dysarthric speech," J. Speech Hear. Disord. **55**(3), 550–561.