Research Note

Objective Intelligibility Assessment by Automated Segmental and Suprasegmental Listening Error Analysis

Yishan Jiao,^a Amy LaCross,^a Visar Berisha,^{a,b} and Julie Liss^a

Purpose: Subjective speech intelligibility assessment is often preferred over more objective approaches that rely on transcript scoring. This is, in part, because of the intensive manual labor associated with extracting objective metrics from transcribed speech. In this study, we propose an automated approach for scoring transcripts that provides a holistic and objective representation of intelligibility degradation stemming from both segmental and suprasegmental contributions, and that corresponds with human perception.

Method: Phrases produced by 73 speakers with dysarthria were orthographically transcribed by 819 listeners via Mechanical Turk, resulting in 63,840 phrase transcriptions. A protocol was developed to filter the transcripts, which were then automatically analyzed using novel algorithms developed for measuring phoneme and lexical segmentation

errors. The results were compared with manual labels on a randomly selected sample set of 40 transcribed phrases to assess validity. A linear regression analysis was conducted to examine how well the automated metrics predict a perceptual rating of severity and word accuracy.

Results: On the sample set, the automated metrics achieved 0.90 correlation coefficients with manual labels on measuring phoneme errors, and 100% accuracy on identifying and coding lexical segmentation errors. Linear regression models found that the estimated metrics could predict a significant portion of the variance in perceptual severity and word accuracy.

Conclusions: The results show the promising development of an objective speech intelligibility assessment that identifies intelligibility degradation on multiple levels of analysis.

Although perceptual evaluation is central to the differential diagnosis of motor speech disorders (see Preferred Practice Patterns for the Profession of Speech-Language Pathology; American Speech-Language-Hearing Association, 2004), abundant evidence suggests that perceptual estimates of speech intelligibility are inherently biased, unreliable, and particularly unsuitable for the tracking of speech change secondary to intervention or disease progression by the treating clinicians (Liss, Spitzer, Caviness, & Adler, 2002; McHenry, 2011; Sheard, Adams, & Davis, 1991). Despite that, the most commonly used clinical method for characterizing changes in speech intelligibility is the treating clinicians' informal perceptual estimation of their patients' speech (Duffy, 2013; Miller, 2013; King,

Watson, & Lof, 2012). Survey results suggest that speech-language pathologists (SLPs) highly value subjective perceptual assessment and feel comfortable with the common practice of informal estimation of intelligibility, and they regard objective metrics as a "nice to have" rather than a "must have" for clinical practice (Lee & FitzGibbon, 2008; Miller, 2013).

One of the important contributing factors to such preference is that most objective metrics are estimated from acoustic signals, but the quality of speech is ultimately judged by a human listener in a real-life speech communication process. For example, previous work has derived objective intelligibility measures by using high-dimensional acoustic features (Laaridh, Kheder, Fredouille, & Meunier, 2017) or an relying on an automated speech recognition/alignment system (Kim, Kim, & Kim, 2015; Middag, Martens, Van Nuffelen, & De Bodt, 2009; Middag, Van Nuffelen, Martens, & De Bodt, 2008). Although these metrics have been shown to correlate with perceptual intelligibility, the speech engineering features and nonlinear models result in clinical uninterpretable representations. Furthermore, they do not identify treatment targets or predict intervention outcomes.

^aDepartment of Speech and Hearing Science, Arizona State University, Tempe

^bSchool of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe

Correspondence to Yishan Jiao: yishan.jiao@asu.edu

Editor-in-Chief: Bharath Chandrasekaran

Editor: Kate Bunton

Received March 18, 2019

Revision received May 29, 2019

Accepted June 21, 2019

https://doi.org/10.1044/2019_JSLHR-S-19-0119

Disclosure: The authors have declared that no competing interests existed at the time of publication.

Another approach to obtaining an objective intelligibility measure is based on perceptual error analysis. Listener transcriptions are assessed for word accuracy—percentage of words that the listener's responses match the speaker's intentions, thus providing a measure of intelligibility of a given speaker (Kent, Weismer, Kent, & Rosenbek, 1989; Yorkston & Beukelman, 1981). However, a single-word accuracy score is usually not informative in clinical practice as it does not provide insight into the reasons why a word or words are unintelligible and thus cannot be utilized as a basis for treatment targets. Understanding the mechanisms underlying intelligibility breakdowns is more helpful for clinicians to identify speech problems and select treatment targets.

Previous studies (Liss et al., 2002; Liss, Spitzer, Caviness, Adler, & Edwards, 1998, 2000) have demonstrated that, when phonemic information in the acoustic signal becomes degraded, listeners rely on more robust prosodic cues, such as the contrasts between stressed and unstressed syllables, to segment continuous acoustic streams into word-sized frames to facilitate speech comprehension. Cutler and Carter (1987) found that 73% of English words begin with a stressed syllable. Therefore, listeners tend to identify a word boundary before a stressed syllable, which may result in lexical segmentation errors. These studies suggest that speech perception is a multilevel processing where perceptual errors can occur at phoneme recognition stage or at lexical segmentation stage. Based on this theory, speech intelligibility can be quantified by measuring phoneme recognition errors and word boundary identification errors.

Manual assessment of phoneme accuracy and word segmentation accuracy is not only prone to subjective variation but is prohibitively time-consuming and impractical for a clinical setting. Various automatic assessments of transcript scoring have been developed as a means of addressing this problem (Borrie, Barrett, & Yoho, 2019; Le, Licata, Persad, & Provost, 2016). However, these studies only measured word or phoneme errors. Regarding lexical segmentation, although there are studies using automated linguistic segmentation from an acoustic signal (Stolcke & Shriberg, 1996), to the best of our knowledge, no existing literature addresses the problem of inferring lexical segmentation errors from listener transcripts.

In this research note, we present a family of algorithms to extract not only phoneme recognition errors but also word segmentation errors by automatically analyzing listener transcripts. These error metrics form a set of objective intelligibility measures, and we investigate their relationship with perceptual ratings of severity and word accuracy.

Method

Participants

Two groups of individuals participated in this study with approval from the Arizona State University (ASU) Institutional Review Board: (a) 73 individuals with dysarthria provided the speech samples, and (b) 819 anonymous individuals provided the listener transcripts via Amazon

Mechanical Turk (MTurk). The 73 individuals with dysarthria have been described in our previous articles (Liss et al., 2009). Briefly, the dysarthria subtypes included ataxic dysarthria secondary to cerebellar degeneration ($N = 16$), hyperkinetic dysarthria secondary to Huntington's disease ($N = 6$), mixed spastic-flaccid dysarthria secondary to amyotrophic lateral sclerosis ($N = 14$), and hypokinetic dysarthria secondary to idiopathic Parkinson's disease ($N = 37$). These speakers were selected so that severity was deliberately varied among the speech samples, thus allowing us to elicit a wide range of listening strategies, error patterns, and intelligibility decrements.

All speech samples were recorded either in the Motor Speech Disorders Laboratory at ASU or Institutional Review Board (IRB)-approved auxiliary research sites (Liss et al., 2009). Participants were fitted with a head-mounted microphone, seated in a sound-attenuating booth or in a quiet room, and read stimuli from visual prompts on a computer screen. Participants were encouraged to produce speech in their typical voice.

To collect speech transcripts, 819 listener participants were recruited via MTurk. Though no explicit data regarding sex and age were collected, the average demographics reported by MTurk (<https://www.mturk.com>) indicate approximately 70% of MTurk workers are women and 30% are men. Ages range from 18 to 68 years, but the majority (54%) are 21–35 years old. Different from experienced SLPs, participants on MTurk are not likely to be familiar with dysarthric speech, the speakers, or the reading materials. By using this population, we mitigate familiarization bias (Liss et al., 2002) and simulate a more realistic listening scenario.

Stimuli

The speech stimuli consisted of 80 phrases, which were designed especially for the study of intelligibility and lexical boundary errors in dysarthric speech and have a rich phonemic inventory. These phrases have been described in detail previously (Liss et al., 1998, 2002, 2009). Briefly, they are composed of six syllables that create three to five mono- and di-syllabic words, which form grammatical phrases with low interword predictability. The phrases were designed to alternate strong and weak syllabic stress in either trochaic or iambic stress patterns to facilitate lexical boundary error analysis. Some examples of the phrases are presented in Table 1. It is important to note that though stress patterns in stimuli were carefully balanced and designed, it was expected that the acoustic signal generated by a dysarthric speaker would vary as a function of severity. It is precisely these variations that are of interest in eliciting a variety of perceptual errors.

Transcript Collection

To facilitate recruitment of a large number of nonexpert listener transcripts, we solicited participants via MTurk. MTurk is an online crowd-sourcing platform, which enables individuals (workers) and organizations (requesters) to

Table 1. Examples of speech stimuli.

Example no.	Phrase	Stress pattern
1	Address her meeting time	WS W SW S
2	Bolder ground from justice	SW S W SW
3	Beside a sunken bat	WS W SW S
4	Cool the jar in private	S W S W SW

Note. S = strong; W = weak.

coordinate the use of human intelligence to perform tasks and has been used successfully in the past speech perception research (Byun, Halpin, & Szeredi, 2015; Lansford & Borrie, 2017). In addition to the efficiency of subject recruitment and the ability to exclude certain demographic profiles, the advantages of using MTurk compared to the conventional participant recruitment method are that such nonexpert listeners represent a realistically diverse perceptual audience for transcribing pathological speech. Under such unsupervised environment, it is possible to collect less reliable data. To improve data reliability, we recruited approximately 10 different listeners to transcribe each phrase, and for each speaker, we calculated the proposed phoneme and lexical segmentation errors from all transcripts of all speech samples.

We developed a website for hosting the task, to which listeners were directed from MTurk. Participants were instructed to listen to each spoken phrase using headphones, and to type in what they thought was said as an orthographic transcription. They were asked to make their best guess if they were unsure. To screen inattentive participants, four easy-to-understand phrases read by two healthy speakers were randomly embedded in the task. To avoid familiarization with the speakers and speech materials, we randomized all speech samples and observed the following protocol: (a) Each participant listens to at most 84 different phrases (80 target phrases plus four screening phrases); (b) among the 80 phrases, there are no more than two phrases from the same speaker; (c) each phrase can be played no more than twice.

Using a survey, nonnative English speakers and listeners with hearing loss, head injury, psychiatric disorders, attention deficit disorder, attention deficit hyperactivity disorder, or obsessive-compulsive disorder were identified, and their data were excluded from analysis. Additionally, only the data from listeners who transcribed all four easy-to-understand phrases correctly were included for analysis. After filtering, 498 out of 819 listeners (60.8%) and 33,969 out of 63,840 transcriptions (53.2%) remained for analysis. Although this resulted in a large number of discarded transcripts, each of the 73 dysarthric speakers still had 465 transcripts on average left for error analysis.

Automated Transcript Analysis

The automated transcript analysis compares listener transcripts with the target phrases and automatically calculates three types of phoneme recognition errors and four types of lexical segmentation errors. Figure 1 shows the procedures of the algorithms. Figure 2 provides a schematic example.

The error metric calculation is rule-based, and our automated algorithm simply codifies these rules. In the following section, we will describe these rules in detail with the aid of the example in Figure 2.

Preprocessing of the Transcripts

Because listener responses were not constrained in any way, the collected data contain misspellings, non-English words, acronyms, and other unanalyzable responses. Although a protocol for manually assessing these entries could be undertaken, it defeats the purpose of an automated assessment measure. Therefore, these transcripts were subject to a series of automated corrections: (a) All special characters were removed except for single quotes and hyphens since they are allowed in the dictionary, (b) an automated spell checking and correction tool was used to correct any misspelling and typos, and (c) the transcript was discarded if there were any out-of-dictionary word in it. Here, we used the Carnegie Mellon University English pronouncing dictionary (CMUdict, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>); this is one of the largest open-source English pronouncing dictionaries containing over 134,000 words and is actively maintained and expanded.

Phoneme Recognition Error Analysis

Phoneme errors include phoneme insertion, deletion, and substitution errors. An example of the phoneme error analysis is shown in the left box of Figure 2. To calculate phoneme errors, we aligned the target and transcript using the ALINE algorithm, which was developed for phonetic sequence alignment by considering the articulation similarity between phonemes (Kondrak, 2003). We used the CMUdict to generate phonemic sequences from each word. The output of the alignment algorithm is a list of aligned phoneme pairs. There are four types of pairs from the output, and we define them as follows: (a) Two identical phonemes are aligned to each other, for example, (b, b), which is a correct transcription; (b) two distinct phonemes are aligned to each other, for example, (æ, ε), which is counted as a substitution error; (c) the target phoneme cannot be aligned to any transcribed phoneme, for example, (n, -), which is counted as a deletion error; and (d) the transcribed phoneme cannot be aligned to any target phoneme, for example, (-, λ), which is counted as an insertion error. For each speaker, the number of errors was then normalized by the total number of phonemes in the target phrase.

Lexical Segmentation Error Analysis

Lexical segmentation errors include four subtypes: (a) an inserted word boundary before a strong syllable (IS), (b) an inserted word boundary before a weak syllable (IW), (c) a deleted word boundary before a strong syllable (DS), and (d) a deleted word boundary before a weak syllable (DW). In the example shown in Figure 2, instead of perceiving the first word in the target phrase as “balance,” the participant perceived it as “bell is.” Thus, a lexical boundary was wrongly inserted before the unstressed second syllable of “balance,” resulting in an IW error.

Figure 1. The procedures of automated transcript analysis. LBE = lexical boundary error; IS = an inserted word boundary before a strong syllable; IW = an inserted word boundary before a weak syllable; DS = a deleted word boundary before a strong syllable; DW = a deleted word boundary before a weak syllable.

Figure 2. An example of automated transcript analysis. LBE = lexical boundary error; S = strong; W = weak.

To automate the calculation, we developed the algorithms as follows. For listener transcript, we first used a placeholder “X” to represent a syllable, and a space to represent the lexical boundary (see the right box of Figure 2). By comparing the location of the lexical boundaries with the target stress pattern, we counted the number of the four possible lexical segmentation errors automatically. Taking the transcript in Figure 2 as an example, the word boundary locations of the target phrase are before the third syllable and before the fourth syllable, whereas the boundary locations of the transcript are before the second syllable, before the third syllable, and before the fourth syllable. Therefore, there is an insertion error before the second syllable, and the second syllable in the target phrase is an unstressed syllable. Thus, it is an IW error. There were no other types of errors for this example. It should also be noted that, at this stage of development, the algorithm performs best on transcripts, which match the target in terms of syllable numbers; therefore, all analyses included are composed of transcript and targets with six syllables.

For each speaker, the four lexical segmentation errors were calculated as the total number of each error normalized by the total number of the corresponding error opportunities in the target phrases. Taking the sample in Figure 2 as an example, the error opportunity of IS, IW, DS, and DW is 0, 2, 2, 1, respectively.

Validation of the Automated Transcript Analysis

To assess the validity of the algorithms for transcript analysis, the estimated metrics were compared to the results of the gold standard: manual scoring. Two research assistants enrolled in the SLP master program of ASU were trained by an expert (A. L.) to independently analyze phoneme errors and lexical segmentation errors of a randomly selected subset of 40 samples from the collected data. They were given 40 pairs of a target phrase and the corresponding listener transcription. The target–transcript phoneme alignment generated by the alignment tool was also provided for their reference. They were trained to transform the text to phoneme sequences using CMUdict and align the target and transcript phonemes based on their knowledge. The research assistants were also instructed to manually code each lexical segmentation error as occurring before strong or weak syllables (IS, IW, DS, DW).

Regression Analysis With Perceptual Ratings Related to Intelligibility and Word Accuracy

To examine the ability of the estimated metrics to predict intelligibility, we performed a linear regression analysis using the three phoneme errors (phoneme substitutions, insertions, and deletions) and two lexical segmentation errors (IS and DW) as independent variables. The reason we only used two lexical segmentation errors is that these two errors are theoretically most commonly produced in English (Cutler & Carter, 1987), and the other two errors (DS and IW) are strongly correlated with these two. For the

dependent variable, we used the perceptual rating of severity as a proxy of the perceptual intelligibility and word accuracy as a traditional objective intelligibility measure derived from listener transcripts. The severity ratings of the 73 dysarthric speakers were obtained from 15 master students enrolled in the SLP program of ASU in a previous study (Tu, Berisha, & Liss, 2017). They were instructed to listen to five sentences from each speaker and provide ratings for severity on a 1–7 scale (*typical* to *severely atypical*). Their ratings were integrated using the evaluator weighted estimator. The word accuracy was calculated as the number of correctly transcribed words over the total number of words in the target phrase. Mean value was calculated for each speaker. The Statistical Package for the Social Sciences was used for this analysis.

Results

Validity of the Automated Algorithms

During the preprocessing of the transcripts, a total number of 2,645 transcripts (7.8% of the 33,969 transcripts) were discarded because they contained out-of-dictionary words. The remaining data were used to calculate phoneme errors and lexical segmentation errors. Among them, there were 15.4% transcripts consisting more or fewer syllables than the target phrases, and, thus, were not used for lexical segmentation error analysis. As a result, the numbers of transcripts for analyzing phoneme errors and lexical segmentation errors were 31,324 and 26,500, respectively.

The results of comparing the calculations between the two research assistants and between each research assistant and the algorithm are shown in Table 2. We use the Pearson correlation and mean absolute error (MAE) to evaluate the reliability of the algorithm relative to the reliability of the two raters.

For phoneme error analysis, the algorithm is strongly correlated with the results provided by the individual research assistants, but the MAE is larger. This is because the ALINE algorithm aligned phrases based on the calculated phonological distances between phonemes, whereas the human raters used internal strategies based on their linguistic knowledge. For example, when the target was “used” and the transcript was “good,” the research assistants aligned the first phoneme “j,” with “g” (one substitution error), whereas the algorithm did not align them together based on their phonological distance (one deletion and one insertion error). However, these differences are systemically observed for all speakers.

In contrast to the high correspondence between the coding of the two research assistants for phoneme errors, the lexical segmentation error analysis yielded low interrater reliability. This is not unexpected because coding consistency is a function of experience, which is why it is common protocol in segmentation error studies to include multiple coders, including at least one with coding expertise, to resolve discrepancies among coders (Liss et al., 1998). Also, as expected, the algorithm achieved higher correlation

Table 2. Validity evaluation results of the automated transcript analysis.

Evaluation measure	Interraters		Algorithm-rater 1		Algorithm-rater 2	
	Phoneme errors	LBEs	Phoneme errors	LBEs	Phoneme errors	LBEs
Correlation	0.97	0.59	0.89	0.90	0.90	0.69
Mean absolute error	0.15	0.12	0.38	0.03	0.34	0.09

Note. LBEs = lexical boundary errors.

coefficients and lower MAEs with each rater. The algorithm was more stable because decision rules are clearly defined instead of relying on the inevitable degree of human interpretation to which coding protocols are subject. As per manual coding protocol, the team's expert coder (A. L.) coded the segmentation errors on the same set as those coded by the research assistants and calculated the correlation and MAE between the algorithm and the expert's codings. This achieved 1.0 correlation coefficient and 0.0 MAE on 34 of the selected phrases. For the other transcriptions, the algorithm could not analyze the data due to the different number of syllables between the transcript and target, but the expert could.

Relationship With Intelligibility

Tables 3 and 4 show the coefficients of the metrics when predicting dependent variables. From the results, we can see that the models fit the data well with significance < .001. The R^2 for predicting severity is .667, which indicates that the metrics are reliable features to predict perceptual intelligibility. From the standardized coefficients, we can see one deviation in phoneme substitution errors account for a change of .679 in severity (on a 7-point scale), whereas one deviation in IS errors account for a change of .714. The R^2 for predicting word accuracy is .973, which indicates that 97.3% variations in word accuracy can be explained by the measured phoneme errors and lexical segmentation errors. In other words, these error metrics are

a nearly complete representation of the traditional intelligibility measure, word accuracy.

Discussion

The automation of scoring to capture the various contributors to speech intelligibility should have dramatic implications for both clinical practice in speech-language pathology and for moving forward communication sciences. However, there exists a natural tension in this goal because human perception is the final arbiter of speech goodness. Automated algorithms are only useful to the extent they can be designed to extract perceptually meaningful aspects from the speech signal. In this report, we presented a method for achieving this goal by focusing on measures that tap two fundamental components of speech intelligibility—the listener's ability to identify constituent phonemes and to segment the acoustic stream at its word boundaries. For the former, previous studies have also predicted intelligibility from phoneme recognition errors (Kim et al., 2015; Le et al., 2016; Middag et al., 2009; 2008), although they rely on automatic speech recognition/automatic speech alignment systems instead of listener transcripts. However, they did not consider the contributions of suprasegmental perceptual strategies to speech intelligibility, such as rhythm and intonation variations. Liss et al. (1998, 2000, 2002) have investigated the relationship of lexical segmentation errors with intelligibility, but there has not been any study attempted to automatically extract these errors. Borrie et al. (2019) developed a

Table 3. Linear regression analysis results with the proposed metrics as independent variables and the perceptual ratings of severity as a dependent variable.

Model*	Coefficients				
	Unstandardized coefficients		Standardized coefficients		Sig.
	B	SE	Beta	t	
(Constant)	2.367	0.194		12.177	.000
Phoneme insertion	−11.852	11.939	−.440	−0.993	.324
Phoneme deletion	−0.975	4.684	−.059	−0.208	.836
Phoneme substitution	12.734	7.305	.679	1.743	.086
IS	5.957	2.557	.714	2.330	.023
DW	−4.970	3.935	−.139	−1.263	.211

Note. IS = an inserted word boundary before a strong syllable; DW = a deleted word boundary before a weak syllable; Sig. = significance.

* R^2 = .667, p < .001.

Table 4. Linear regression analysis results with the proposed metrics as independent variables and the perceptual ratings of word accuracy as a dependent variable.

Model*	Coefficients				
	Unstandardized coefficients		Standardized coefficients		Sig.
	B	SE	Beta	t	
(Constant)	0.898	.008		113.935	.000
Phoneme insertion	0.681	.484	.177	1.406	.164
Phoneme deletion	−0.403	.190	−.171	−2.120	.038
Phoneme substitution	−1.853	.296	−.690	−6.254	.000
IS	−0.254	.104	−.212	−2.449	.017
DW	−0.699	.160	−.136	−4.379	.000

Note. IS = an inserted word boundary before a strong syllable; DW = a deleted word boundary before a weak syllable; Sig. = significance.

* $R^2 = .973$, $p < .001$.

tool to automatically score listener transcripts, but only word accuracy was calculated. In this study, we developed a family of automated algorithms to extract both phoneme errors and segmentation errors from listeners' orthographic transcriptions.

Although the automated phoneme measures were highly correlated with the manually coded measures, there was a systemic bias from human coding results. The automated lexical segmentation error coding was superior to trained human coders in its consistent application of operational definitions. This level of stability matched the coding of an expert coder, but unlike the expert, the algorithm could not navigate coding of transcribed phrases that exceeded or did not include the target six syllables. Within these parameters, the automated coding yielded a high correlation with the experts.

The algorithms were also largely successful in generating perceptually and clinically meaningful data in their correspondence with perceptual intelligibility and word accuracy. The overarching goal of this research is to deliver objective intelligibility assessments that are reliable and clinically meaningful. By scoring listener transcripts from segmental (phoneme recognition) and suprasegmental (lexical segmentation) levels, clinicians could have a clearer view of how one patient's speech challenged the listener's ability to comprehend it differently from another patient even if they may have the same level of overall intelligibility. In that, it can help them develop more personalized intervention strategies for different patients. For example, for a patient whose speech leads listeners to make many phoneme errors but fewer lexical segmentation errors, clinicians can recommend interventions focusing on articulation, whereas for a patient whose speech leads listeners to make many word segmentation errors but fewer phoneme errors, clinicians can recommend interventions focusing on rhythm control.

Limitations and Future Work

MTurk provides a platform to recruit a large number of participants with a relatively low cost and requires that

safeguards be in place to identify and discard unreliable or invalid data that are collected in the unsupervised experimental setting. The safeguards in this study resulted in discarding 48.3% of the collected transcripts, which was much greater than we initially expected. In the future, listener selection will be refined to include a prescreening survey to ensure only listeners that meet our criteria can participate in the study.

Additionally, the current lexical segmentation error analysis algorithm could only process transcripts with the same number of syllables as the target phrases (i.e., six syllables), whereas humans are capable of coding transcripts with fewer or more syllables at most times. In the future, we will improve the algorithm to enhance its capability. For example, we will study how human experts handle these cases and develop more specific rules, and we will develop syllable-level alignment algorithms to identify the extra and missing syllables in the transcriptions and adjust the calculation of segmentation errors accordingly. Moreover, the current phoneme errors are calculated as the number of errors, which is relatively coarse. In the future, we will develop measures considering the phonological distances between phonemes and make the metrics finer.

Acknowledgments

The authors report funding from the Foundation for the National Institutes of Health under Grant 2 R01 DC006859-11 (awarded to PI: Julie Liss and MPI: Visar Berisha).

References

- American Speech-Language-Hearing Association. (2004). *The preferred practice patterns for the profession of speech-language pathology*. Retrieved from <https://www.asha.org/policy/PP2004-00191/>
- Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, 145(1), 392–399.

- Byun, T. M., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*, 53, 70–83.
- Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, 2(3–4), 133–142.
- Duffy, J. R. (2013). *Motor speech disorders-e-book: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier Mosby.
- Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4), 482–499.
- Kim, M. J., Kim, Y., & Kim, H. (2015). Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4), 694–704.
- King, J. M., Watson, M., & Lof, G. L. (2012). Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. *Journal of Medical Speech-Language Pathology*, 20(1), 1–17.
- Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37(3), 273–291. <https://doi.org/10.1023/A:1025071200644>
- Laaridh, I., Kheder, W., Fredouille, C., & Meunier, C. (2017, August). *Automatic prediction of speech evaluation metrics for dysarthric speech*. Paper presented at Interspeech, Stockholm, Sweden.
- Lansford, K., & Borrie, S. (2017). Use of crowdsourcing platforms to examine listener perception of disordered speech. *The Journal of the Acoustical Society of America*, 141(5), 3911.
- Le, D., Licata, K., Persad, C., & Provost, E. M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), 2187–2199.
- Lee, A., & FitzGibbon, O. (2008). *Speech-language pathologists' experience on perceptual evaluation of speech disorders: A survey*. Ireland: ASHA Convention.
- Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, 112(6), 3022–3030. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12509024>
- Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *The Journal of the Acoustical Society of America*, 104(4), 2457–2466. <https://doi.org/10.1121/1.423753>
- Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (2000). Lexical boundary error analysis in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, 107(6), 3415–3424. <https://doi.org/10.1121/1.429412>
- Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research*, 52, 1334–1352.
- McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Journal of Speech-Language Pathology*, 20(2), 119–123.
- Middag, C., Martens, J. P., Van Nuffelen, G., & De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, 2009(1), 629030.
- Middag, C., Van Nuffelen, G., Martens, J. P., & De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)* (pp. 1745–1748). Baixas, France.
- Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, 48(6), 601–612.
- Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research*, 34(2), 285–293.
- Stolcke, A., & Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. *Proceeding of Fourth International Conference on Spoken Language Processing* (pp. 1005–1008).
- Tu, M., Berisha, V., & Liss, J. (2017, January). Interpretable objective assessment of dysarthric speech based on deep neural networks. *Proceedings of the Interspeech, Stockholm, Sweden*, 1849–1853.
- Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders*, 46(3), 296–301.