The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance

Ming Tu, Alan Wisler, Visar Berisha, and Julie M. Liss

Citation: The Journal of the Acoustical Society of America **140**, EL416 (2016); doi: 10.1121/1.4967208
View online: https://doi.org/10.1121/1.4967208
View Table of Contents: http://asa.scitation.org/toc/jas/140/5
Published by the Acoustical Society of America

The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance

Ming Tu,[1,a] Alan Wisler,[2] Visar Berisha,[1,b] and Julie M. Liss[1]
[1]*Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287, USA*
[2]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*
mingtu@asu.edu, awisler@asu.edu, visar@asu.edu, jmliss@asu.edu

Abstract: State-of-the-art automatic speech recognition (ASR) engines perform well on healthy speech; however recent studies show that their performance on dysarthric speech is highly variable. This is because of the acoustic variability associated with the different dysarthria subtypes. This paper aims to develop a better understanding of how perceptual disturbances in dysarthric speech relate to ASR performance. Accurate ratings of a representative set of 32 dysarthric speakers along different perceptual dimensions are obtained and the performance of a representative ASR algorithm on the same set of speakers is analyzed. This work explores the relationship between these ratings and ASR performance and reveals that ASR performance can be predicted from perceptual disturbances in dysarthric speech with articulatory precision contributing the most to the prediction followed by prosody.

[DDO]
Date Received: October 14, 2015 **Date Accepted:** September 28, 2016

1. Introduction

Producing clear and intelligible speech requires coordination among many subsystems, including articulation, respiration, phonation, and resonance. Individuals with disruption to any of the involved physical or neurological processes required in speech production suffer from a corresponding degradation in the quality of the speech they produce. This condition is called *dysarthria* and it can manifest itself in a variety of ways such as hypernasality, atypical prosody, imprecise articulation, poor vocal quality, etc. It is known that these perceptual degradations directly impact intelligibility; as a result, intervention strategies by speech-language pathologists (SLPs) focus on correcting these disturbances as a means of improving intelligibility in patients.

While a great deal of research has been devoted to characterizing the relationship between perceptual inconsistencies and intelligibility judgments by listeners, this is a topic that has yet to be addressed for automatic speech recognition (ASR) engines. In this paper we are interested in exploring the relationship between perceptual speech quality and performance of a state-of-the-art ASR engine. An accurate understanding and modeling of this relationship can have a significant impact in two areas. First, these models can help algorithm designers customize ASR strategies such that they perform well under conditions where users exhibit a great deal of variability in speech production. A deeper understanding of this relationship can also result in objective outcome measures for SLPs based on ASR performance. Reliable evaluation of dysarthric speech is a long-standing problem. While evaluations are traditionally done by SLPs, studies have found that the biases inherent in subjective evaluations result in poor inter- and intra-rater reliability.[1,2] As a result there is strong motivation for the development of improved methods of objective speech intelligibility evaluation. To that end, in this paper we explore the relation between the word error rate (WER) of an ASR system[3] (an objective measure) and subjective perceptual assessment along four perceptual dimensions (nasality, vocal quality, articulatory precision, prosody) and a general impression of the dysarthria severity. Below we describe related work in this area, our methodology, the results of this work, and discuss implications of the findings.

---

[a]Author to whom correspondence should be addressed.
[b]Also at: School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287.

2. Related work

There are two lines of research that are related to the work presented in this paper: automated measures of pathological speech intelligibility and ASR methods customized for dysarthric speech. A number of related objective methods have been used for intelligibility assessment of dysarthric speech.[4–6] Objective measures from the telecommunications literature, such as the speech intelligibility index[7] or speech transmission index,[8] have been applied to pathological speech; however, application of these methods is difficult due to the lack of a clean reference signal against which to test. With the recent improvements to ASR systems, an obvious approach to speech intelligibility estimation is to replace the human listener with an ASR system. Studies have found that ASR-based methods can be used to effectively estimate the intelligibility deficits caused by tracheoesophageal speech,[9] cleft lip and palate,[10] cancer of the oral cavity,[11] head and neck cancer,[12] and laryngectomy.[13]

Intelligibility estimates by themselves, however, have very limited utility in a clinical setting. It is often the case that clinicians are interested in evaluating speech along different perceptual dimensions to construct a complete and comprehensive understanding of the speech degradation and to customize treatment plans.[14,15] While some research has been done to model the relationship between these perceptual dimensions and intelligibility scores,[16] prior work has focused on scores from human listeners and does not generalize to ASR. In fact, while the WER of ASR systems has been found to be a useful tool for evaluating speech intelligibility, its efficacy for measuring degradations in other perceptual dimensions outside of speech intelligibility is largely unstudied.

Recently some work has been done on improving ASR performance for dysarthric speech. Christensen *et al.* carried out a study on applying training and adaptation methods for improved recognition of dysarthric speech.[17] They concluded that while there was perceptual variation among dysarthric speakers, adaptation following training improved the performance of the system to some extent. Similarly, Sharma and Hasegawa-Johnson proposed a new acoustic model adaptation method for dysarthric speech recognition.[18] While this method yielded improvement compared to standard speaker adaptation techniques, its defect is the large variability among different dysarthric speakers. These studies imply that if we can somehow characterize the perceptual variability exhibited by dysarthric speakers, we can customize ASR strategies based on this information. For example, different ASR strategies would likely be required for a dysarthric speaker who exhibits a rapid speaking rate than for a dysarthric speaker who exhibits imprecise articulation. As a first step, we must first understand the relationship between these disturbances and ASR performance.

3. Methodology

3.1 Data acquisition

The dysarthric speech database we used was recorded in the Motor Speech Disorders Lab at Arizona State University as a part of a larger ongoing study. We used 32 dysarthric speakers and clinically four categorizations of their diseases. The four categorizations are ataxic, mixed spastic-flaccid, hyperkinetic, and hypokinetic. Different categorizations have different perceptual symptoms of speech degradation; for example, speakers diagnosed with hypokinetic can have a rapid articulation rate and rushes of speech while other categorizations may not have.

Each speaker produced five sentences as described in Liss *et al.*[19] Following data acquisition, 15 students from the ASU Master's SLP program (second year, second semester), rated each speaker along five perceptual dimensions: severity, nasality, vocal quality, articulatory precision, and prosody on a scale from 1 to 7 (from normal to severely abnormal). Severity represents the annotator's judgment about the general quality of the produced speech. Nasality refers to the ability to control oral-nasal separation during speech production. Vocal quality refers to the presence of noise in the voicing. Articulatory precision refers to how well vowels and consonants are produced. Prosody refers to pitch variation, speaking rate, stress, loudness variation, and rhythm.[20] Each annotator's ratings were combined into a single set of ratings. We used the Evaluator Weighted Estimator (EWE) to integrate ratings from the 15 students into a single set of ratings by calculating a mean rating for each perceptual dimension, weighted by individual reliability.[21] The result was a final set of ratings where, for each speaker, we had ensemble ratings for each of the five perceptual dimensions considered here.

3.2 ASR for dysarthric speech

For each of the 32 dysarthric speakers we have 5 sentences with word-level transcription and perceptual ratings for each of the 5 dimensions. In lieu of constructing a custom speech recognition engine, we elected to use a representative and robust ASR algorithm for the study, the Google ASR engine.[3] Google provides an Application Programming Interface to interface with their algorithm. For all sentences of each of the 32 speakers, we calculated the WER based on the results of the Google ASR engine and ground truth transcription. The WER is calculated by the following formula:

$$\mathrm{WER} = \frac{S + I + D}{N}, \tag{1}$$

where S is the number of substitution errors, I is the number of insertion errors, D is the number of deletion errors, and N is total number of words in a sentence. The WER of one speaker is the average WER of his or her five sentences.

3.3 Statistical analysis

Following data acquisition, we initially analyzed the reliability of the combined EWE ratings. For reliability, we use the ratings from a randomly sampled subset of L evaluators, and the ratings from a *different* subset of eight evaluators. We then calculate the EWE of each subset of evaluators and find the mean absolute error (MAE) between each set of ratings. In this reliability analysis, we treat the ratings from the combined eight listeners as a "gold standard" against which we compare. It allows us to evaluate the error between a single evaluator, the EWE of two evaluators, the EWE of three evaluators, etc., against the EWE of the gold-standard.

The MAE is interpretable on a 7-point scale—for example, an MAE of 1 means that two sets of ratings fall within 1 of each other on a 7-point scale. We estimate the MAE for increasing values of L from 1 evaluator to 7 evaluators. Since we have a total of 15 evaluators, the largest that L can be is 7 because we require two subsets of different evaluators.

We also analyzed the Pearson correlation coefficient between perceptual ratings in each dimension and the WER rates. Specifically, after processing all 160 sentences through the ASR engine, we obtained the WER for each speaker. We then calculated the correlation coefficient between the WERs and perceptual ratings for all five dimensions of the 32 dysarthric speakers.

To further investigate the mapping from perceptual ratings to WER and which perceptual dimension contributes the most to WER, we built an ℓ_1-norm-constrained linear regression model with the values of the four perceptual ratings (nasality, vocal quality, articulatory precision, and prosody[22]) as input and the WER as output.[23] We changed the value of the regularization coefficient such that the model selects a different number of features ranging from 1 to 4. Leave-one-speaker-out cross validation was used to find the weights of the linear regression model. In this model, we normalized the independent variables (ratings) to zero mean, unit variance, and calculated the final feature weights by averaging the absolute value of weights of all 32 models. We computed the correlation coefficients of predicted WERs for the test samples of the 32 folds and the WERs from the Google ASR engine.

4. Results and discussion

4.1 Data description

Detailed information for all the speakers in the dataset is included.[30] For each speaker, we list the dysarthria subtype, gender, age, an average rating on a 1–7 scale for each perceptual dimension, and a listing of perceptual symptoms. The perceptual symptoms

Table 1. Correlations among five perceptual dimensions. "S," "N," "VQ," "AP," and "P" are abbreviations of the five perceptual dimensions.

	S	N	VQ	AP	P
Severity	1.00				
Nasality	0.79	1.00			
Vocal quality	0.91	0.69	1.00		
Articulatory precision	0.91	0.83	0.75	1.00	
Prosody	0.84	0.61	0.73	0.73	1.00

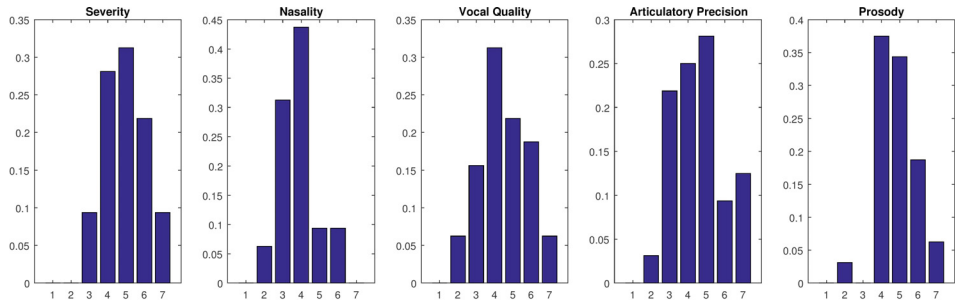


Fig. 1. (Color online) Histograms of ratings for the five perceptual dimensions. The X axis is the perceptual ratings from 1 to 7 and the Y axis is the percentage of speakers falling into each bin.

for each speaker were annotated by a speech language pathologist carefully listening to each speaker and listing the most degraded perceptual qualities of the resulting speech based on her judgement.[24]

In Table 1 we show the correlations between each of the five dimensions and in Fig. 1 we show a histogram of ratings for each perceptual dimension. The 32 speakers used in this study were sampled from a much larger dysarthria database. We aimed to sample the evaluation scale for each perceptual dimension such that the distribution of ratings for each perceptual dimension was approximately the same. As Fig. 1 shows, most of the samples come from the middle of the rating scale so as not to bias the observed correlations.

4.2 Data reliability

It is well accepted in the literature that the intra-rater agreement for auditory perceptual evaluation can be low.[1,2] Although the tasks are fundamentally different,[25] this can be observed when comparing the descriptions the SLP provided to the EWE ratings from the evaluators. For example, for speakers M10 and M11, the SLP noted irregular prosody; however M10's prosody was rated a 5.5, and M11's prosody was rated a 2.4.

Average ratings from multiple listeners are a common way to reduce variability.[21] In Table 2 we show the MAE for an increasing number of raters. We see that the combined EWE ratings yield significantly lower MAE values when compared against individual ratings. In fact, the MAE was reduced by almost a factor of 3 (to a value of 0.51). It is important to note that we actually combine 15 ratings when exploring the relationship between the perceptual dimensions and the WER, therefore the MAE is likely to be lower than the MAE for the 7 raters shown in Table 2.

4.3 Relationship between WER and perceptual dimensions

We first show the scatter plots and correlation coefficients of WERs and evaluators' perceptual ratings in Fig. 2. The results demonstrate that articulatory precision has the highest correlation coefficient with WERs while nasality and vocal quality have the lowest correlation coefficient. This is not surprising since subjective assessment of the articulatory precision provides a measure of the deviation from standard pronunciation on which the ASR engine is trained. Since it is often the case that ASR engines de-emphasize voicing information, it also makes sense that vocal quality degradation does not correlate as strongly with WER. Indeed, degradation of vocal quality does not greatly impact traditional ASR features such as Mel Frequency Cepstral Coefficient.[26] Among the other three perceptual dimensions, severity has the highest correlation coefficient (close to articulatory precision). This too makes sense since severity encompasses perceptual information in different dimensions, including articulatory precision.

Table 2. The reliability of the combined EWE ratings: The average MAE for increasing numbers of evaluators with $1-\sigma$ confidence.

1 evaluator	1.45 ± 0.48
2 evaluator	0.93 ± 0.21
3 evaluator	0.82 ± 0.22
4 evaluator	0.71 ± 0.20
5 evaluator	0.59 ± 0.12
6 evaluator	0.55 ± 0.13
7 evaluator	0.51 ± 0.11

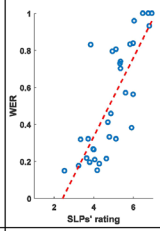
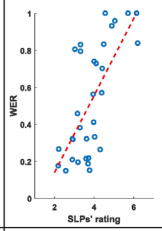
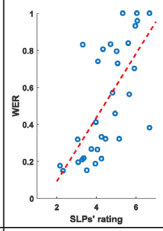
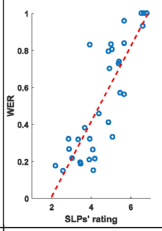
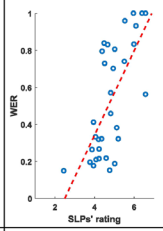
Perceptual dimension	Severity	Nasality	Vocal Quality	Articulatory Precision	Prosody
Scatter plot					
Correlation coefficient	0.799	0.688	0.680	0.841	0.694

Fig. 2. (Color online) Scatter plots and correlation coefficients of WERs and evaluators' perceptual ratings.

In analyzing Fig. 2 notice that there are some outliers. For example, in the severity plot we see a speaker with a subjective severity score below 4 but with an unusually high WER of over 0.8. The high WER is largely because the speaker stops and repeats himself multiple times. Human listeners do not perceive this as problematic during subjective evaluation; however the ASR engine has difficulties in dealing with this. In the prosody dimension there is also an outlier where the prosody rating is high, but the WER is lower than expected. In listening to this speaker, we notice that he has very good articulatory precision and long periods of normal rhythm, followed by short bursts of increased speaking rate. This gives the impression that overall prosody is greatly disturbed; however the ASR engine is able to correctly identify the words when the rate is not rapidly changing because the articulation is so clear.

In addition to the correlation analysis, we also constructed a regression model to find a mapping from the normalized evaluators' ratings (zero-mean, unit variance) along the four perceptual dimensions to the WER. We construct four different models with a varying number of perceptual dimensions considered for each model (we refer to these as features in the model). For each model, ℓ_1-weighted regression is conducted using four features: nasality, vocal quality, articulatory precision, and prosody; however, the value of the regularization coefficient is set such that only the desired number of features is selected (between 1 and 4). For each model, we determine the averaged absolute value of the four weights and the correlation coefficient between predicted WERs and true WERs. The result is shown in Fig. 3. The absolute values of the regression weights provide an estimate of the relative importance of a perceptual dimension to predicting the WER. This analysis further confirms our previous finding that articulatory precision is clearly most important in predicting ASR performance. Following articulatory precision, prosody is the second most important (since it is the second feature selected).

The correlation coefficient between predicted WERs using a regression model and the true WERs remains almost constant when using different numbers of features.

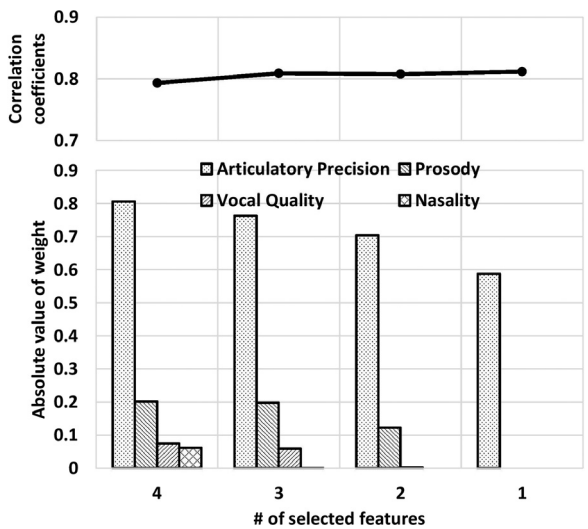


Fig. 3. Absolute weights of the regression models and the correlation coefficient between predicted and actual WERs.

This is consistent with the highest correlation between WERs and articulatory precision rating. Here, articulatory precision also dominates the prediction of WERs. Other studies have revealed similar relationships for dysarthric speech. Mengistu *et al.* showed that there exists a relationship between acoustic measures and ASR accuracy for dysarthric speech.[27] However, they draw this conclusion from a relatively small database of only nine speakers. De Bodt *et al.* investigated the linear relationship between overall intelligibility (by humans) and assessment of different perceptual dimensions of dysarthric speech.[16] Consistent with our results, they also found that articulatory precision contributes the most to overall intelligibility. Our previous studies[19,28] have shown the importance of different aspects of prosody information (such as rhythm, speaking rate) for understanding dysarthric speech. Also, the work by Nanjo and Kawahara demonstrated that when speaking rate information is considered, the performance of an ASR system can be improved.[29]

5. Conclusion

In this paper, we explored the relationship between subjective perceptual assessment of five perceptual dimensions and the WER of Google's ASR engine on dysarthric speech produced by 32 dysarthric speakers of varying dysarthria subtype and of varying severity. There are two principal contributions in this paper. First, we revealed the potential of using ASR performance as a proxy for assessing articulatory precision since the correlations between that dimension and ASR performance is high. Second, we showed that ASR performance can be predicted from perceptual disturbances in dysarthric speech. Understanding this relationship is a first step to accurately adapting ASR strategies for dysarthric speech. A natural extension of this work is to consider the subjective assessments on finer resolution subjective dimensions (e.g., rate, pitch variability, loudness variability instead of prosody). These dimensions would allow us to assess directionality instead of simply using the scale considered here. For example, we could consider a scale that ranges from "very slow" to "very fast" for rate. A more specific subjective evaluation would also allow us to assess finer resolution ASR errors, including insertion errors, deletion errors, and substitution errors.

Acknowledgment

This research was supported in part by the National Institutes of Health, National Institute on Deafness, and Other Communicative Disorders Grants Nos. 2R01DC006859 (J.M.L.) and 1R21DC012558 (J.M.L. and V.B.).

References and links

[1]S. A. Borrie, M. J. McAuliffe, and J. M. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies," J. Speech Lang., Hear. Res. **55**(1), 290–305 (2012).

[2]J. M. Liss, S. M. Spitzer, J. N. Caviness, and C. Adler, "The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria," J. Acoust. Soc. Am. **112**(6), 3022–3030 (2002).

[3]J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition* (Springer, New York, 2010), pp. 61–90.

[4]V. Berisha, S. Sandoval, R. Utianski, J. Liss, and A. Spanias, "Characterizing the distribution of the quadrilateral vowel space area," J. Acoust. Soc. Am. **135**(1), 421–427 (2014).

[5]S. Steven, V. Berisha, R. Utianski, J. Liss, and A. Spanias, "Automatic assessment of vowel space area," J. Acoust. Soc. Am. **134**(5), EL477–EL483.

[6]V. Berisha, J. Liss, S. Steven, R. Utianski, and A. Spanias, "Modeling pathological speech perception from data with similarity labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (2014), pp. 915–919.

[7]ANSI S3.5-1997, *American National Standard: Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York, 1997).

[8]T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," Acta Acust. Acust. **25**(6), 355–367 (1971).

[9]M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski, "Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (2005), pp. 61–64.

[10]A. Maier, E. Nöth, E. Nkenke, and M. Schuster, "Automatic assessment of children's speech with cleft lip and palate," in *Proceedings of the 5th Slovenian and 1st International Conference on Language Technologies* (*IS-LTC* 2006) (2006), pp. 31–35.

[11]A. K. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *INTERSPEECH* (2007), pp. 1206–1209.

[12]A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, and M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," EURASIP J. Audio, Speech, Music Process. **2010**, 1 (2010).

[13]M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski, "Intelligibility of laryngectomees' substitute speech: Automatic speech recognition and subjective rating," Euro. Arch. Oto-Rhino-Laryngol. Head Neck **263**(2), 188–193 (2006).
[14]M. R. McNeil, *Clinical Management of Sensorimotor Speech Disorders* (Thieme, New York, 2009).
[15]K. Bunton, R. D. Kent, J. R. Duffy, J. C. Rosenbek, and J. F. Kent, "Listener agreement for auditory-perceptual ratings of dysarthria," J. Speech, Lang., Hear. Res. **50**(6), 1481–1495 (2007).
[16]M. S. De Bodt, M. E. Hernández-Díaz Huici, and P. H. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," J. Commun. Disorders **35**(3), 283–292 (2002).
[17]H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH* (2012).
[18]H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," Comput. Speech Lang. **27**(6), 1147–1162 (2013).
[19]J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," J. Speech, Lang., Hear. Res. **52**(5), 1334–1352 (2009).
[20]A. E. Aronson and J. R. Brown, *Motor Speech Disorders* (WB Saunders Company, St. Louis, MO, 1975).
[21]M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," Speech Commun. **49**(10), 787–800 (2007).
[22]Severity is not included because it strongly correlates with other features.
[23]K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).
[24]The prompt for the SLP was "Please describe the most degraded perceptual qualities of the speech from each speaker."
[25]We asked the student evaluators to evaluate all five perceptual dimensions on a scale of 1–7, whereas the SLP was asked to describe the dimensions they deemed most degraded.
[26]M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," Speech Commun. **49**(10), 763–786 (2007).
[27]K. T. Mengistu, F. Rudzicz, and T. H. Falk, "Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers," in *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011, pp. 75–78.
[28]Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(9), 1421–1430 (2015).
[29]H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*) (2002), Vol. 1, pp. 1–725.
[30]See supplementary material at http://dx.doi.org/10.1121/1.4967208 for all speaker information in the dataset.