INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech intelligibility of dysarthric speech: human scores and acoustic-phonetic features

Wei Xue[1], *Roeland van Hout*[2], *Fleur Boogmans*[2], *Mario Ganzeboom*[2], *Catia Cucchiarini*[1], *Helmer Strik*[1,2]

[1]Centre for Language and Speech Technology (CLST), Radboud University, The Netherlands
[2]Centre for Language Studies (CLS), Radboud University, The Netherlands

w.xue@let.ru.nl, r.vanhout@let.ru.nl, fleurboogmans@live.nl, m.ganzeboom@let.ru.nl, c.cucchiarini@let.ru.nl, w.strik@let.ru.nl

Abstract

We investigated speech intelligibility in dysarthric and non-dysarthric speakers as measured by two commonly used metrics, ratings through the Visual Analogue Scale (VAS) and word accuracy (AcW) through orthographic transcriptions. To gain a better understanding of how acoustic-phonetic correlates could be employed to obtain more objective measures of speech intelligibility and a better classification of dysarthric and non-dysarthric speakers, we studied the relation between these measures of intelligibility and some important acoustic-phonetic correlates. We found that the two intelligibility measures are related, but distinct, and that they might refer to different components of the intelligibility construct. The acoustic-phonetic features showed no difference in the mean values between the two speaker types at the utterance level, but more than half of them played a role in classifying the two speaker types. We computed an acoustic-phonetic probability index (API) at the speaker level. API is moderately correlated to VAS ratings but not correlated to AcW. In addition, API and VAS complement each other in classifying dysarthric and non-dysarthric speakers. This suggests that the intelligibility measures assigned by human raters and acoustic-phonetic features relate to different constructs of intelligibility.

Index Terms: dysarthric speech, speech intelligibility, acoustic-phonetic features, speaker classification

1. Introduction

Speech intelligibility is an important construct in speech pathology that is employed for diagnosis and for determining whether speech therapy has been effective. A common definition in the clinical practice of speech therapy is that proposed by Hustad [1] "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener". In line with this definition, intelligibility has been measured by asking listeners to make orthographic transcriptions (OTs) of what they hear [2, 3]. Percentages of words correctly transcribed are then used as a measure of intelligibility as in the Sentence Intelligibility Test [4].

In the clinical field, intelligibility has also been measured by collecting scalar ratings from human judges [5-9]. For instance, by asking raters to indicate the degree of intelligibility on an equal-appearing interval scale (or Likert scale; e.g. [7]), or a visual analogue scale (VAS), (placing a point on a horizontal line; e.g. [10]). It is common practice to check the reliability of these kinds of ratings before these can be used for research purposes [11].

In a previous study [12], scalar ratings and OTs were used to obtain intelligibility scores of disordered speech at three different levels of granularity: utterance, word, and subword level. Utterance level evaluations were obtained using subjective rating scales (VAS and Likert scale); word and subword level evaluations were obtained from orthographic transcriptions, using automatic alignment and conversion methods. The phoneme scoring thus obtained turned out to be feasible, reliable, and provided a more sensitive and informative measure of intelligibility. The results showed that the intelligibility measures at the different levels of granularity were fairly highly correlated, but performed differently. The orthography-based measures revealed higher intelligibility scores than the scalar rating measures. This appeared to be in line with previous research [8, 13], suggesting that in scalar ratings experts tend to underestimate the degree of intelligibility. Even in the case that raters understand every word, they may still judge intelligibility as less than perfect because of imprecisions and deviations that make the speech difficult to understand.

An important aspect of the study presented in [12] was that all measurements were based on subjective scores provided by human judges, which of course is in line with the definitions of the constructs themselves. However, it is known that these human-based procedures are subjective, error-prone, and extremely time-consuming, thus making intelligibility measurement extremely problematic in practice. For these reasons, researchers have been studying alternative ways of measuring intelligibility that do not rely on human judgments. Many have employed ASR algorithms to obtain automatic measures of pathological speech quality [14-16]. However, it is not clear how exactly these ASR-based measures are related to speech intelligibility and to properties of pathological speech that can be addressed in the therapy.

Previous studies have shown that pitch [17-19], intensity [17, 19, 20-22] and formant frequencies [17, 18, 23] are related to intelligibility and can contribute to distinguishing dysarthric speech from healthy speech. For example, Parkinson's patients have limited pitch and loudness variability [20] in their voices. Their intelligibility can be improved by the Lee Silverman Voice Treatment [21], in which speakers are instructed to speak louder. Bunton et al. [22] found that a restricted intensity range tended to be associated with reduced speech intelligibility in amyotrophic lateral sclerosis speakers with moderate intelligibility. Xue et al. [24] investigated the usability of the eGeMAPS feature set,

http://dx.doi.org/10.21437/Interspeech.2021-1189

which contains the three mentioned features, for predicting speech intelligibility at phoneme level. Their results indicated that this feature set is potentially usable and revealed important differences between dysarthric speech and non-dysarthric speech. The authors also suggested that the correlations between the features and intelligibility measures at higher levels, e.g., at word or utterance levels, should be investigated since intelligibility measures at different levels of granularity turned out to perform differently, as also found in [12].

In the present research we investigated speech intelligibility in dysarthric and non-dysarthric speakers as measured by VAS ratings and OTs and compared these metrics to acoustic-phonetic correlates to determine to what extent these a) can provide a basis for more objective evaluations of speech intelligibility and b) contribute to speaker classification.

2. Method

2.1. Datasets and speakers

We used two datasets collected within the CHASING project [25] [1] which was aimed at developing a serious game for conducting research on speech disorder treatment through ASR-based technology. The two datasets [25] contain speech of thirteen dysarthric speakers (10 male and 3 female) and of five non-dysarthric speakers (4 male and 1 female). The dysarthric speakers were aged between 53 and 75 (M = 64.2, SD = 6.4), ten of them had Parkinson's and three had had a Cerebral Vascular Accident (CVA). All non-dysarthric speakers were aged between 61 and 69 (M = 65.0, SD = 3.4).

2.2. Speech materials

From the datasets described in Section 2.1, we extracted, for each speaker, eight recordings covering three types of speech materials: four meaningful sentences, two semantically unpredictable sentences and two three-word lists. The four meaningful sentences were selected from the text "Papa en Marloes" ("Papa and Marloes" in English). We made sure that all the recordings were made before the game to avoid any impact of the treatment.

2.3. Raters, intelligibility measures and experimental setting

We recruited eleven speech therapists as raters to participate in the listening experiment. All the selected raters are native speakers of Dutch, have normal hearing and vision and do not have the Attention Deficit Disorder or problems with fine motor skills (typing). They were all graduated as speech therapists being either all-round speech therapists or specialized in neurorehabilitation. Seven of them had experience in dysarthria. In detail, four of them followed the master speech-language pathology of Radboud University, two had training in Parkinson's and dysarthria, and one had followed a minor in neurorehabilitation. The number of years after they had graduated as speech therapists varied between 0.5 and 15 years (M = 3.6, SD = 5.0). They were all female and were aged between 23 and 38 years (M = 26.9, SD = 4.7).

In the listening experiment, each rater assigned a score through a Visual Analogue Scale (VAS) ranging from 0 (0% intelligible) to 100 (100% intelligible) and completed an Orthographic Transcription (OT) of the utterance allowing only existing words. Based on the transcriptions, we computed word accuracy (AcW), which is the percentage of correctly transcribed words [2]. Therefore, for each utterance, we obtained two intelligibility measures, a rating (VAS) and a word accuracy score (AcW).

During the experiments, the recordings to be assessed were presented with a loudness of 60 decibels through headphones. The researcher who collected the data of the intelligibility measures was always present to make sure that the recordings were presented with the same sound intensity and the experiments took place in a low-noise environment.

2.4. Acoustic-phonetic features

The acoustic-phonetic features were calculated using Praat [26]. The features extracted are duration, minimal pitch (pitchMin), maximal pitch (pitchMax), mean value of pitch (pitchMean), standard deviation of pitch (pitchStd), the mean of the absolute values of the pitch slope (pitchSlopeMean), minimal intensity (intensityMin), maximal intensity (intensityMax), mean value of intensity (intensityMean), standard deviation of intensity (intensityStd), formants 1 to 4 (F1, F2, F3 and F4), and center of gravity (centerGravity).

2.5. Statistical analysis

We first explored the two intelligibility measures with respect to their distribution and interrater reliability at utterance level. Interrater reliability was calculated by applying Generalizability Theory [27] with a fully crossed model design, where all utterances from all speakers were assessed by all raters. The Phi coefficient (D-coefficient) was then calculated as the reliability.

We calculated the means and standard deviations of the acoustic-phonetic features for the two types of speakers separately and conducted a t test to establish whether the mean values of the acoustic-phonetic features were significantly different for the two types of speakers.

In the next step, we averaged our two intelligibility measures over utterances and raters to obtain scores at the speaker level. We computed their reliabilities at the speaker level using Intraclass Correlation Coefficient (ICC) and their correlation. We made a scattergram distinguishing the dysarthric and non-dysarthric speakers.

The acoustic-phonetic features were submitted to a stepwise logistic regression [3] to predict speaker type (set '0' for 'non-dysarthric' and '1' for 'dysarthric') at the utterance level. The Akaike Information Criterion (AIC) was used in the stepwise procedure to decide which acoustic-phonetic features to include in the final regression model. The predicted speaker type probabilities at the utterance level were averaged over the eight utterances per speaker as the overall acoustic-phonetic probability index (API). According to the above speaker type setting, high API scores of test items refer to 'dysarthric' and low scores refer to 'non-dysarthric'. The next step was to expl-

[1] http://hstrik.ruhosting.nl/chasing/

[2] We used the asr-evaluation python module provided in https://github.com/belambert/asr-evaluation to calculate the word accuracy.

[3] We obtained similar models in forwards and backwards regression.

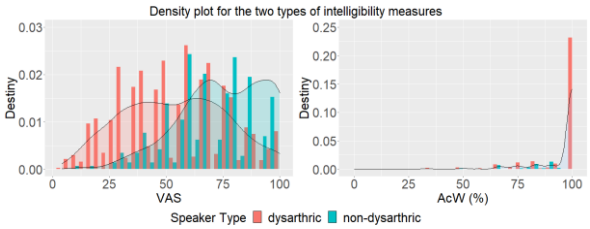

Figure 1: *Density plots for the two intelligibility measures.*

ore and interpret the correlations and scattergrams of the API and the two intelligibility measures.

We used the packages *gtheory* [28], *stats* [29], *psych* [30], *MASS* [31], and the *ggplot2* [32] for performing the analyses and making plots in RStudio [33] with R version 4.0.1 [29].

3. Results

3.1. Intelligibility measures at the utterance level

Figure 1 shows the density plots for VAS and AcW, with different colors for the two speaker types. VAS scores have a broad distribution (M = 59.6, SD = 23.6), the distribution of the non-dysarthric utterances having higher frequencies on the right part of the scale. AcW scores have an extremely skewed distribution (M = 95.7, SD = 11.3) with a very high concentration of maximum scores of 100. Both plots show an overlap between the two speaker types. The reliability (D-coefficient) values at the utterance level are 0.90 for VAS and 0.47 for AcW.

3.2. Acoustic-phonetic features at the utterance level

No significant difference in the mean values of the acoustic-phonetic features was found between the two types of speakers. The detailed results regarding the mean and standard deviation of the features for each speaker type can be found in Table 1.

3.3. Dysarthric and non-dysarthric speakers

3.3.1. Intelligibility measures

The reliability values using ICC (2, k) (items and raters random with k = 11) are 0.93 for VAS and 0.85 for AcW.

Table 1: *Mean (Standard Deviation) of the acoustic-phonetic features for the two types of speakers.*

Feature	Non-dysarthric	Dysarthric
duration (ms)	3575.38 (2115.36)	3640.38 (2090.28)
pitchMin (Hz)	86.11 (15.57)	92.6 (17.63)
pitchMax (Hz)	289.98 (173.51)	279.5 (156.68)
pitchMean (Hz)	131.11 (27.45)	139.14 (22.82)
pitchStd (Hz)	40.58 (30.81)	34.41 (25.92)
pitchSlopeMean (Hz)	361.91 (353.02)	341.66 (236.26)
intensityMin (dB)	19.59 (10.41)	14.49 (36.15)
intensityMax (dB)	77.19 (5.9)	81.09 (4.3)
intensityMean (dB)	66.82 (6.83)	71.22 (4.52)
intensityStd (dB)	15.99 (4.97)	16.64 (5.99)
F1 (Hz)	756.92 (472.12)	805.55 (482.52)
F2 (Hz)	1793.3 (487.57)	1902.21 (466.82)
F3 (Hz)	2915.68 (402.85)	2934.33 (412.66)
F4 (Hz)	3913.72 (418.45)	3952.35 (431.59)
centerGravity (Hz)	685.12 (243.93)	533.5 (186.43)

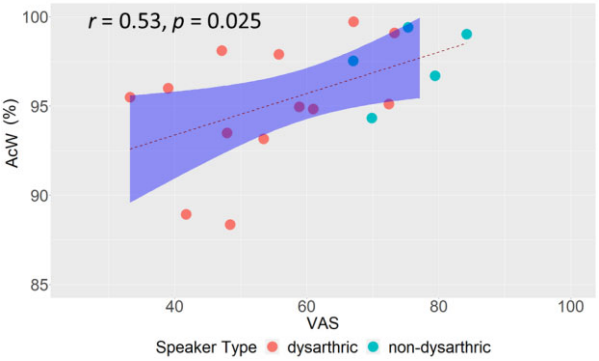

Figure 2: *Scattergram of VAS and AcW (%), including the correlation r, p value, the regression line and confidence intervals (95%).*

Figure 2 shows the scattergram between VAS and AcW, as well as the regression line. It can be seen that the points for non-dysarthric speakers are located close to the top-right corner, while those for dysarthric speakers are relatively scattered to the bottom-left. However, the ranges between the two measures differ considerably. The VAS scores range between 52 and 82, while the AcW scores range between 88 and 99. The VAS distinguishes dysarthric and non-dysarthric speakers reasonably well, with an overlap in the higher regions of the VAS scores. AcW is not successful in distinguishing the two types of speakers. The correlation between the measures (.53) is significant and moderate.

3.3.2. Acoustic-phonetic features

Table 2 shows the summary of the final model of the stepwise logistic regression. All variables in Table 2 were relevant according to the AIC criterium. These results indicate that the chances of being dysarthric increase as pitchSlopeMean, intensityMax, F1 and F2 increases and decrease as pitchStd, intensityMin, intensityStd, F3 and centerGravity increase. Six of the variables are significant in the classification. The final model also reported a mean classification accuracy of 79% with 50% for non-dysarthric and 88% for dysarthric speech.

3.3.3. Correlating the intelligibility measures and the acoustic-phonetic probability index (API)

As explained in Section 2.5, the predicted speaker type probabilities at the utterance level were averaged over the

Table 2: *Summary of the final model in stepwise logistic regression with selected predictors' coefficients (Coef.), standard errors of the coefficients (Std. Err), the z value, the p value and the significance levels (Sig. level; 0.001 – '***'; 0.01 – '**'; 0.05 – '*'; 0.1 – '.'; 1 – ' ').*

Predictors	Coef.	Std. Err	*z* value	*p* value	Sig. level
pitchStd	-0.039	0.0145	-2.074	0.012	**
pitchSlopeMean	0.0049	0.00157	3.135	0.00685	**
intensityMin	-0.011	0.0372	-2.943	0.00172	**
intensityMax	0.283	0.0648	4.375	0.00326	***
intensityStd	-0.2158	0.0768	-2.812	0.005	**
F1	0.00124	0.000765	1.624	0.104	
F2	0.00092	0.000631	1.458	0.145	
F3	-0.001627	0.000859	-1.893	0.058	.
centerGravity	-0.00363	0.0013	-2.784	0.00537	**

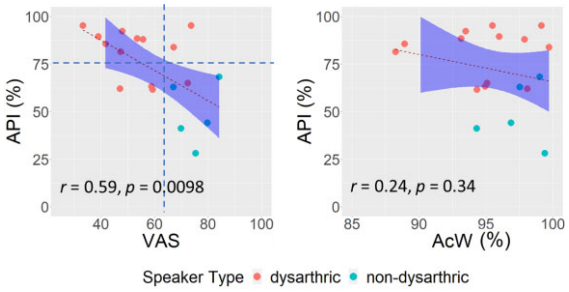

Figure 3: *Scattergrams of VAS and AcW (%) with the API (%), including the correlation r, p value, the regression line and confidence intervals (95%). VAS-API plot is split into four quadrants by dashed lines.*

eight utterances per speaker to compute the API. Figure 3 shows the scattergrams between the API, VAS and AcW, with different colors for the two speaker types. It can be seen that the API makes a distinction between the two types of speakers, but that there is an obvious overlap as well. The correlation of API with VAS is significant and moderate, while the correlation with AcW is non-significant.

Interestingly, we can split the VAS-API plot into four quadrants along a vertical and a horizontal dashed line, as shown in the left panel of Figure 3. Most of the speakers are classified correctly as located in the top-left quadrant for 'dysarthric' (with high API scores and low VAS scores) and bottom-right quadrant for 'non-dysarthric' (with low API scores and high VAS scores) with only one red spot (as 'dysarthric') in the bottom-right quadrant. Divergent results were found for several dysarthric speakers. Specifically, we found three dysarthric speakers in the bottom-left quadrant. These speakers had relatively low VAS scores, but were not classified as 'dysarthric' according to the low API scores. The opposite applies to two dysarthric speakers in the top-right quadrant, with high VAS scores and high API scores.

4. Discussion

In this study, speech of dysarthric and non-dysarthric speakers was evaluated on intelligibility by human raters through VAS scalar ratings and word accuracy (AcW) based on Orthographic Transcriptions (OTs). Acoustic analyses were then conducted to study how acoustic-phonetic features are related to intelligibility measures and to what extent they contribute to classifying speakers as either dysarthric or non-dysarthric.

We observed high interrater reliability for the VAS ratings, which is in line with previous findings [6, 9, 12, 34]. On the other hand, for AcW we found relatively low interrater reliability at the utterance level, which contrasts with previous findings [6, 34, 35]. We further explored the data to gain insight into the possible causes of this low reliability index. We noted that most of the word accuracy scores indicate a perfect intelligibility of 100, which implies that the variability in the scores was very limited. This seems to be plausible, as the speakers had mild dysarthria and thus, displayed relatively little variability in their speech. In addition, only existing words were allowed in the transcriptions, which in turn further reduced the degree of variability. So, this would seem to suggest that the raters did their job properly and transcribed the speech correctly. We consequently found that the correlation between the two intelligibility measures is significant, but not strong, partly due to this low variability.

The results presented above suggest that the two intelligibility measures investigated in this study are related, but have distinct qualities, and that they might refer to different constructs of the concept of intelligibility, even for the same speech material. Similar findings in the field of L2 pronunciation instruction, where speech intelligibility has also received considerable attention, led researchers [36] to draw a distinction between speech intelligibility defined as "the extent to which a speaker's message is actually understood by a listener" and comprehensibility, which stands for the degree of ease with which L2 speech can be understood. Accordingly, different operationalizations were adopted for the two constructs. To measure intelligibility raters had to "write out carefully in standard orthography what they heard" [36], while for comprehensibility raters had to assign scale ratings.

In the clinical field, the term comprehensibility has also been used, but with different definitions like "contextual intelligibility" [37] or a listeners' ability to recover the meaning of pathological speech utterances [38]. The results presented in this paper might be seen as indications that also in clinical practice two constructs should be distinguished: intelligibility refers to the degree of actual understanding as measured by writing down what has been said, and comprehensibility refers to the ease of understanding as measured through rating scales.

The mean values of the acoustic-phonetic features did not differ significantly for the two types of speakers, suggesting that no single feature is highly related to speaker type. The logistic regression outcomes show that the most important and promising point seems to be that the three groups of features play a role in the final regression model selected in distinguishing dysarthric and non-dysarthric utterances: pitch, intensity, and the formant frequencies. When we studied the correlations between the intelligibility measures on the one hand, and the overall acoustic-phonetic probability index (API) that a speaker be classified as either dysarthric or non-dysarthric, on the other, we found that the API showed moderate correlation with VAS and no correlation with AcW. With respect to the classification of speakers as either dysarthric or non-dysarthric, the results suggest that the intelligibility measures assigned by human raters and the probabilities computed through the objective procedure based on acoustic-phonetic features are partly complementary to each other, as also found by Bunton et al. [22]. These results are also in line with previous findings that acoustic-phonetic features have correlations to speaker types or to speech intelligibility [17, 22, 24], to a certain extent.

Future research could investigate the precise impact of the individual acoustic-phonetic features in more detail and with more utterances from speakers with varied dysarthria severity. Additionally, other relevant acoustic-phonetic features such as F2 slope [17, 39] and vowel space area [23], and more robust metrics of features such as percentiles could be considered as these might reveal different relations to speech intelligibility or may better contribute to classifying speaker types.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766287.

6. References

[1] K. C. Hustad, "The relationship between listener comprehension and intelligibility scores for speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 562–573, 2008.

[2] K. M. Yorkston and D. R. Beukelman, "Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate," *Journal of Speech and hearing disorders*, vol. 46, no. 3, pp. 296–301, 1981.

[3] J. M. Garcia and M. P. Cannito, "Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 4, pp. 750–760, 1996.

[4] K. Yorkston, D. Beukelman and R. Tice, "Sentence intelligibility test [Computer software]," *Lincoln, NE: Tice Technologies*, 1996.

[5] S. S. Barreto and K. Z. Ortiz, "Intelligibility measurements in speech disorders: a critical review of the literature," *Pró-Fono Revista de Atualização Científica*, vol. 20, no. 3, pp. 201–206, 2008.

[6] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language and Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013.

[7] K. M. Yorkston and D. R. Beukelman, "A comparison of techniques for measuring intelligibility of dysarthric speech," *Journal of Communication Disorders*, vol. 11, pp. 499–512, 1978.

[8] D. Abur, N. M. Enos, and C. E. Stepp, "Visual Analog Scale Ratings and Orthographic Transcription Measures of Sentence Intelligibility in Parkinson's Disease with Variable Listener Exposure," *American Journal of Speech-Language Pathology*, vol. 28, no. 3, pp. 1222–1232, 2019.

[9] K. Ishikawa, J. Webster, and C. Ketring, "Agreement between Transcription- and Rating-Based Intelligibility Measurements for Evaluation of Dysphonic Speech in Noise," *Clinical Linguistics and Phonetics*, pp. 1–13, 2020.

[10] C. Finizia, J. Lindstrom, and H. Dotevall, "Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy," *Laryngoscope*, vol. 108, no. 1, pp. 138–143, 1998.

[11] L. J. Beijer, A. C. M. Rietveld, M. B. Ruiter, and A. C. H. Geurts, "Preparing an E-learning-based Speech Therapy (EST) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers," *Clinical Linguistics and Phonetics*, vol. 28, no. 12, pp. 927–950, 2014.

[12] M. Ganzeboom, M. Bakker, C. Cucchiarini and H. Strik, "Intelligibility of disordered speech: global and detailed scores," in *Proc. INTERSPEECH*, pp. 2503–2507, 2016.

[13] K. C. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, 2006.

[14] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Noth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[15] C. Middag, J.-P. Martens, G. van Nuffelen, and M. de Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, 2009.

[16] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[17] K. Tjaden, and G. E. Wilding, "Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings," *Journal of Speech, Language and Hearing Research*, vol. 47, no.4, pp. 766–783, 2004.

[18] L. Feenaughty, K. Tjaden, and J. Sussman, "Relationship between acoustic measures and judgments of intelligibility in Parkinson's disease: A within-speaker approach," *Clinical linguistics and phonetics*, vol. 28, no. 11, pp. 857–878, 2014.

[19] K. Tjaden, and G. Wilding, "The impact of rate reduction and increased loudness on fundamental frequency characteristics in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4 pp. 178–186, 2011.

[20] J. Holmes, R., M. Oates, J., J. Phyland, D., & J. Hughes, A. Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders*, 35(3), pp. 407-418, 2000.

[21] M. P. Cannito, D. M. Suiter, D. Beverly, L. Chorna, T. Wolf, and R. M. Pfeiffer, "Sentence intelligibility before and after voice treatment in speakers with idiopathic Parkinson's disease," *Journal of Voice*, vol. 26, no. 2, pp. 214–219, 2012.

[22] K. Bunton, R. D. Kent, J. F. Kent, and J. C. Rosenbek, "Perceptuo-acoustic assessment of prosodic impairment in dysarthria," *Clinical Linguistics and Phonetics*, vol. 14, no. 1, pp. 13–24, 2000.

[23] G. Weismer, J. Jeng, J. S. Laures, R. D. Kent, and J. F. Kent, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," *Folia Phoniatrica et Logopaedica*, vol. 53, no. 1, pp. 1–18, 2001.

[24] W. Xue, C. Cucchiarini, R. van Hout, H. Strik, "Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech," in *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pp. 48-52, DOI: 10.21437/SLaTE.2019-9.

[25] M. Ganzeboom, M. Bakker, L. Beijer, T. Rietveld, and H. Strik, "Speech training for neurological patients using a serious game," *British Journal of Educational Technology*, vol. 49, no. 4, pp. 761–774, 2018.

[26] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer," Version 6.1.41, 2021. http://www.praat.org/

[27] R. L. Brennan, "Generalizability theory," Springer, 2001.

[28] C. T. Moore, "gtheory: Apply generalizability theory with R," 2016. https://CRAN.R-project.org/package=gtheory

[29] R Core Team, "R: A language and environment for statistical computing," 2014. http://www.R-project.org/

[30] W. Revelle, "psych: Procedures for Psychological, Psychometric, and Personality Research," Northwestern University, Evanston, Illinois. R package version 1.9.12. 2019.

[31] W. N. Venables and B. D. Ripley, "Modern Applied Statistics with S," Fourth edition, Springer, New York, 2002. https://www.stats.ox.ac.uk/pub/MASS4/

[32] H. Wickham, "ggplot2: Elegant graphics for data analysis," 2016. https://ggplot2.tidyverse.org

[33] Rstudio Team, "Rstudio: Integrated Development for R," 2020. http://www.rstudio.com/

[34] W. Xue, V. Mendoza Ramos, W. Harmsen, C. Cucchiarini, R. van Hout and H. Strik, "Towards a comprehensive assessment of speech intelligibility for pathological speech," in *Proc. INTERSPEECH*, 2020.

[35] K. Tjaden, and G. Wilding, "Effects of speaking task on intelligibility in Parkinson's disease," *Clinical Linguistics and Phonetics*, vol. 25, pp. 155–168, 2010.

[36] M. J. Munro and T. M. Derwing, "Foreign Accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[37] K. M. Yorkston, E. A. Strand, and M. R. Kennedy, "Comprehensibility of dysarthric speech: Implications for assessment and treatment planning," *American Journal of Speech-Language Pathology*, vol. 5, no. 1, pp. 55–66, 1996.

[38] K. C. Hustad and D. R. Beukelman, "Listener comprehension of severely dysarthric speech: effects of linguistic cues and stimulus cohesion," *Journal of Speech, Language and Hearing Research*, vol. 45, no. 3, 2002.

[39] Y. Chiu, K. Forrest, and T. Loux, "Relationship Between F2 Slope and Intelligibility in Parkinson's Disease: Lexical Effects and Listening Environment," *American Journal of Speech-Language Pathology*, vol. 28, pp. 887–894, 2019.