The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Can Machines Learn Continuous Measures of Speech Severity from Ordinal Training Labels?

Alan Wisler,[1] Kristin Teplansky,[1] Jordan R. Green,[2,3] Sara G. Austin[4], Jun Wang[1,4]
[1]Department of Communication Sciences and Disorders University of Texas at Austin, Austin, TX, USA
[2]Department of Communication Sciences and Disorders MGH Institute of Health Professions, Boston, MA, USA
[3]Speech and Hearing Bioscience and Technology Program, Harvard University, Boston, MA, USA
[4]Department of Neurology, Dell Medical School, University of Texas at Austin, Austin, TX, USA

Abstract

In this study we examined the efficacy of machine learning general regression algorithms for predicting ordinal variables based on the acoustic speech signal. We were specifically interested in whether predictions that fell between ordinal levels (e.g. a predicted score of 3.2 instead of a true score of 3) contained meaningful information about the outcome variable. As a test case, we explored speech-based estimation of the Amyotrophic Lateral Sclerosis Functional Rating Scale - Revised (ALSFRS-R), a clinical measure for speech severity. The ALSFRS-R is a diagnostic tool that measures individual components of motor function for patients with ALS along a 5-point ordinal scale. Using artificial neural networks (ANN) we can generate continuous estimates of the speech component in ALSFRS-R. However, the degree to which the improved resolution of these estimates contains useful information related to patients' motor control has not been thoroughly studied. In this paper, we sought to answer this question by comparing the residuals in machine learning estimates of the ALSFRS-R speech score with patients' intelligible speaking rate (ISR), which is a more granular measure of speech motor control. Experimental results using speech data from 45 patients with ALS confirmed that ANN regression is effective at learning granular information even when trained with coarse ordinal labels.

1 Introduction

A growing area of research is focused on utilizing machine learning and speech processing in diagnostics and the development of assistive tools for individuals with neurological speech disorders (Berisha, Utianski, and Liss 2013; Benba, Jilbab, and Hammouch 2015; Williamson et al. 2015; Orozco-Arroyave et al. 2016; Hsu et al. 2017; Norel et al. 2018; An et al. 2018; Wang et al. 2018). A major challenge in this field is the development of tools that can quantify the degree to which aspects of speech, such as articulatory precision, are degraded as a result of a neurological disorder. Quantifying specific areas in which speech, or motor function more broadly, are impaired is critical for clinicians and doctors to both plan and effectively evaluate different treatment strategies. Unfortunately, existing evaluation strategies are often reliant on subjective assessments that can vary widely across different clinicians. These measurements can also be easily biased by the degree of familiarity between the clinician and their patients (Liss et al. 2002; Borrie, McAuliffe, and Liss 2012). The measurable bias and time-cost of these subjective assessments provide a strong motivation for the development of objective assessment methods.

Often in clinical situations, the variables we are trying to predict are ordinal rather than continuous or categorical. Ordinal variables are like categorical variables, but their different categories follow a natural order. A common example of this variable type are Likert scale measures in which respondents may be asked to state their feelings towards a given statement along a predefined scale such as: {strongly disagree, disagree, neutral, agree, strongly agree}. Ordinal variables are widely used in a number of fields, including communication disorders, because they provide concrete benchmarks that are easy to interpret.

Although there exist a number of supervised learning strategies designed specifically for the prediction of ordinal independent variables (Cheng, Wang, and Pollastri 2008; Chu and Keerthi 2007; Fu et al. 2018), these methods restrict their predictions to an ordinal scale. This restriction may not be ideal in all applications (such as clinical assessment). Ordinal ratings are often used in clinical assessments not because of an inherent preference for discrete categories, but because they facilitate more reliable subjective ratings. In these cases the coarse nature of the ordinal scale is undesirable as it inhibits the ability to detect subtle changes in a patient's status. As a pertinent example of this, we consider the challenge of measuring the decline in speech motor control experienced by individuals with amyotrophic lateral sclerosis (ALS). ALS is a progressive neurological disease that inhibits the ability of the brain to control muscle movements. The ALS functional rating scale-revised (ALSFRS-R) is a clinical measure of the decline in motor control experienced in this population (Cedarbaum et al. 1999). The ALSFRS-R measures motor control across 12 common motor tasks along a 5-point ordinal scale (0 - 4). While the ordinal scale provides clearly defined benchmarks for assessment that ease administration and in-

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

terpretation of the measure, the coarse measurement scale limits its ability to track small changes along specific dimensions of motor control. As a result, the components of the ALSFRS-R may lack the sensitivity of instrumentation-based measures of motor function (Allison et al. 2017; Andres et al. 2017). Using a general regression model that treats the ordinal measures of the ALSFRS-R as if they were continuous can avoid binding the predicted speech scores to the same coarse scale of the original ALSFRS-R measure. Although this approach has been employed in some prior studies (Wisler et al. 2019), the degree to which the enhanced resolution offered by these estimates is actually useful remains to be studied.

In this paper we focused on the speech component of the ALSFRS-R that is designed to measure patients' speech production. We approached this problem with two main goals. First, we examined the degree to which the residuals in ALSFRS-R speech score predictions contain salient information about the patients' speech production that eludes the coarse measure of the ALSFRS-R. To accomplish this goal, we used participants' intelligible speaking rate (ISR, understandable words per minute) as a high-resolution measure of speech production and measure the relationship between the predicted speech scores and ISR when controlling for the ground truth ALSFRS-R speech scores. Note that our goal here was not actually to predict ISR, but to evaluate whether the predicted speech score measures capture information related to ISR beyond what is captured by the initial labels used to train the model. Therefore, our goal in this analysis was not simply to evaluate the supervised learning models' ability to reproduce the valuable information captured by the ALSFRS-R, but its ability to improve upon it by representing the same information along a more granular scale. Our second goal was to examine the role of model complexity in the models' ability capture this latent information. To accomplish this goal, we repeated the initial analysis for a range of different models, and examined the degree to which different model estimates are predictive of both patient's speech score and their ISR.

2 Data Collection

2.1 Participants

The data used for this project, a subset of a larger set that is being collected, contains forty-five speakers diagnosed with ALS at early-onset. Participants were asked to attend four data collection sessions at four to six month intervals. At each session, participants or caregivers completed the ALSFRS-R. Although the ALSFRS-R is completed in its entirety, this paper primarily focuses on the component of the ALSFRS-R measuring speech production which assigns patients a speech score of 0-4 according to the following criteria:

- **4**: Normal Speech
- **3**: Detectable speech disturbance
- **2**: Intelligible with repeating
- **1**: Speech combined with nonvocal communication
- **0**: Loss of useful speech

Table 1: Summary of the key diagnostic measures for the participants.

	Male	Female
Number of Participants	26	19
Age	57 $\pm$ 1.77	60.5 $\pm$ 2.15
ALSFRS-R	36.8 $\pm$ 1.20	36.6 $\pm$ 1.95
Speech Score	3.3 $\pm$ 0.16	3 $\pm$ 0.24

Speech intelligibility (percentage of understandable words, judged by listeners) and speaking rate (words produced per minute) were assessed by a speech-language pathologist using the Sentence Intelligibility Test (SIT) software (Dorsey et al. 2007). ISR, a measure of communication efficiency, was also calculated (speech intelligibility $\times$ speaking rate) (Yorkston and Beukelman 1981). The range of ISR in our data is from 0 to 235.71 words per minute (wpm).

The set of participants included in this analysis contained 19 female speakers and 26 male speakers. Participants' ages ranged from 39 to 81 years, with an average age of 58.56 (excluding two participants who did not disclose their date of birth). A detailed breakdown of the participants' diagnostic information as measured at the beginning of their participation in the study is displayed in Table 1.

2.2 Stimuli and Procedure

The participants were asked to produce 20 sentences in a fixed order, such as *I need some assistance* and *call me back when you can*. A complete list of the stimuli used for data collection can be found in (Wisler et al. 2019). The sentences were selected because they are commonly used in augmentative and alternative communication (AAC) devices (Beukelman et al. 1984). All speech stimuli were presented on a TV screen in front of the participants. The stimuli were repeated for a total of four recordings at the participants habitual speaking rate among other speech tasks. In some cases, participants were unable to complete the entire recording process, and only a subset of the regular 80 recordings could be included in the analysis. In total, we used 5,288 recordings were collected across the 45 participants at 75 different recording sessions. The audio signals used in this analysis were collected using a Shure Microflex microphone with a sampling rate of 22kHz was positioned approximately 15 cm from each speaker's mouth.

3 Methods

The first step in the proposed system was to extract features from the phrase-level acoustic speech signal. The features used in this paper were based on the Mel-frequency cepstral coefficients (MFCCs). Although MFCCs do not provide comprehensive representation of the effects of ALS on the acoustic signal, as they ignore characteristics of pitch, they have been shown to be an effective tool for characterizing motor-speech disorders (Benba, Jilbab, and Hammouch 2015; Williamson et al. 2015; Tu, Berisha, and Liss 2017) and provide a suitable test case for this study. We extracted the first fourteen MFCCs, along with their first and second derivatives, thus totaling 42 low-level descriptors at the frame level. These frame-level features are then aggregated using five different statistical functionals: mean, median, standard deviation, skewness and pairwise variability, leading to a total of 210 acoustic features.

Once the features were extracted from each participants frame-level speech, we were left with a 5288x210 feature matrix that was used to both train and evaluate our regression model. To partition the data in such a way that maximizes the amount of training data while still effectively evaluating out-of-sample performance, we used leave-one-participant-out cross-validation (CV). Thus at each step of the CV the model was trained on 42 participants, while the single remaining participant was held out for evaluation.

As a baseline, we considered a simple artificial neural network with one hidden layer containing eight artificial neurons. Each artificial neuron is fully connected to the 210 input features, and passes the linear combination of inputs through a hyperbolic tangent sigmoid activation function. This ANN was trained using the scaled conjugate gradient backpropagation to generate continuous predictions of the ALSFRS-R speech scores based on the previously outlined feature set and cross-validation procedure. As the ALSFRS-R speech scores being predicted exist along a five-level scale, this task is not well suited to evaluation by Pearson or Spearman correlation coefficients. Instead, we evaluated our predictions using Goodman and Kruskal's gamma (Goodman and Kruskal 1979). This measure is calculated by looking at each pair of samples in which both scores (the original speech score and the predicted speech score) differ across samples and counting both the number of times that the two measures agree in the ranking of the two samples (N_s) and the number of times that the two measures disagree in the ranking of the two samples (N_d). From these counts the Goodman and Kruskal's gamma is calculated by

$$G = \frac{N_s - N_d}{N_s + N_d}. \tag{1}$$

Following this initial assessment, where the predictions were evaluated based on a comparison to the original speech score values, we conducted a follow-up analysis where they were compared with participants' intelligible speaking rate. The goal of this analysis was to assess the degree to which predictions generated by this model contain useful information about participants' speech severity beyond what is contained in the original ALSFRS-R speech score values. For this assessment we performed a regression analysis where the dependent variable is intelligible speaking rate and the independent variables are the individual's ALSFRS-R speech score (y) and the model residuals. We calculate the model residuals as

$$r = y_p - y. \tag{2}$$

where y_p represents the predicted speech score. Typically we think of residuals as errors in model predictions; however, this thinking is overly simplistic in cases where the variable being predicted is an imperfect representation of the underlying quantity it tries to measure. Speech motor control is too complex to be completely characterized by only five categories. It is possible that some part of the residuals in our predictions represents latent information about speech motor function that is not captured by the ordinal rankings of the ALSFRS-R. As there is no perfect measure for overall speech motor control, ISR serves as a reasonable proxy.

To evaluate how the performance varies with model complexity, we conducted a second experiment where we tested eight single-layer ANNs and eight two-layers ANNs with the number of neurons per layer selected according to the following exponential pattern: $N_{ANN} = [2, 4, 8, 16, 32, 64, 128, 256]$. To overcome random variations in each model's performance, due to the non-deterministic training process, we averaged the results across a 10-trial Monte Carlo simulation. The performance of each model is evaluated based on the strength of the relationship between the model's predictions and the two clinical quantities of interest (speech score and ISR).

4 Results & Discussion

4.1 Experiment 1: Baseline model performance

Based on the initial analysis of the single-layer 8 neuron ANN, we found that the sample-level predictions are moderately correlated, $G = 0.466$, with the true speech score measures. When predictions are aggregated to the session-level (averaged across the $\approx$ 80 phrase-level predictions), this correlation goes up to $G = 0.594$. Functionally, this means that if one participant is assigned a higher speech score than another, they are also assigned a higher speech score by the proposed model 79.7% of the time. While this may seem low, it is similar to the strength of correlation exhibited between different clinical measures of speech severity. As an example, using the same correlation analysis to compare participant's speech score with their intelligible speaking rate, we found that the strength of this correlation, $G = 0.591$, was not much different. Furthermore, the Pearson correlation between the estimated scores and ISR ($\rho = 0.758$) was higher than that of the original speech scores and ISR ($\rho = 0.662$). A visual depiction of the aggregated predicted speech scores relative to the true speech scores is displayed in Figure 1.

Having validated the general efficacy of the model in predicting the participants' speech scores, our main interest in this analysis is to examine whether the residuals provide useful information for predicting ISR beyond what was offered by the original speech score values. To accomplish this, we construct a regression model which attempts to fit the residuals and true speech score values to ISR according to the following equation

$$ISR = I + b \cdot r + C_y \cdot y \tag{3}$$

where I represents the intercept, b represents the coefficient of effect for the residual and C_y represents a variable coefficient of effect that is assigned uniquely to each level of participants' speech score (y). The results of this regression model are summarized in Table 2.

The results of this analysis provided strong evidence of a relationship between the prediction residual and ISR, since the effect of residuals in the regression model was significantly below the standard $p = 0.05$ threshold. While the

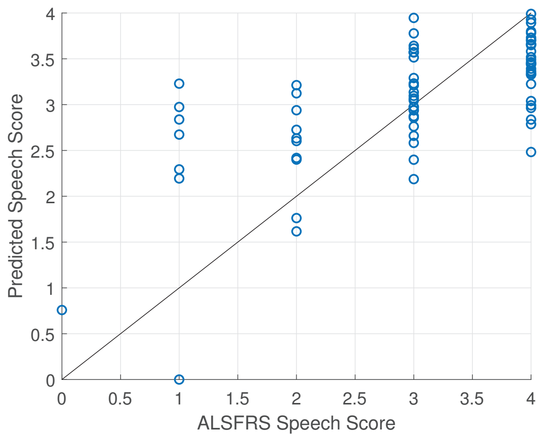

Figure 1: Scatter plot illustrating the relationship between the predicted and true speech score values.

Table 2: Summary of regression results predicting ISR based on participant speech score values and model residuals from the 8 neuron ANN. Note that a speech score of zero is the considered the baseline for this model. Therefore the intercept represents the effect of a zero speech score, and the effects of other speech scores are estimated relative to this baseline.

	Estimate	SE	t-stat	p-value
Intercept	-35.55	35.72	-1.00	0.32
Residual	47.81	8.28	5.78	$< \mathbf{0.001}$
Speech Score 1	35.45	37.87	0.94	0.35
Speech Score 2	73.21	36.92	1.98	0.05
Speech Score 3	159.68	36.21	4.41	$< \mathbf{0.001}$
Speech Score 4	219.50	37.36	5.88	$< \mathbf{0.001}$

effects of speech score in the model were strongly significant for scores of 3 & 4, speech scores of 1 & 2 don't show significant effects. This was likely due to both them being lower levels (the observed effect is relative to the baseline score of zero), and fewer numbers of samples available at these levels (7 and 10 respectively). In total, this regression model was able to explain 64.3% of the variance in the ISR values. This constitutes a significant improvement over the 43.8% of the variance that the speech scores were able to explain in isolation. To further illustrate the significance of this regression model, a visual depiction of the relationship between the ANN prediction residuals and ISR is displayed in Figure 2.

These results provide strong evidence that the ANN prediction residuals have a direct relationship with participants' intelligible speaking rate. Based on the results summarized in Table 2, a 1-point increase in prediction residuals corresponds to a 47.81 words per minute increase in ISR. This finding has significant clinical implications, as it suggests that data-driven estimates of diagnostic measures may be more sensitive than the original measures themselves.

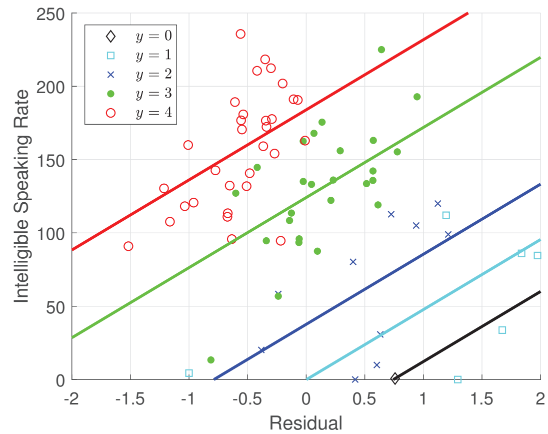

Figure 2: Scatter plot illustrating the relationship between the residuals in predictions of ALSFRS-R speech score and participant's intelligible speaking rate at each speech score level (y) for the initial 8 neuron ANN.

4.2 Experiment 2: Model comparison

In this section we will look at the results, when the previous experiment is expanded to consider a range of different ANN architectures, including ANNs with one and two hidden layers each layer containing N_{ANN} neurons ($N_{ANN} = [2, 4, 8, 16, 32, 64, 128, 256]$). For this experiment we summarize the effectiveness of each model according to two metrics. The first metric is the Goodman & Kruskal's gamma summarizing the strength of correlation between the model's speech score predictions and the true speech score. The second metric is that amount of variance of ISR that can be explained by a linear regression model (like the one summarized in Table 2) of the residuals and true speech score values. The average performance across each metric, along with their standard errors, is displayed as a function of the number of model parameters in Figure 3. As the ANN's are fully connected, the number of parameters referred to on the x-axis is equal to $210N_{ANN} + N_{ANN}$ for the weights and biases corresponding to connections between the features and the first hidden layer, $N_{ANN} + 1$ for the weights and biases corresponding to connections between the last hidden layer and the output layer, and $N_{ANN}^2 + N_{ANN}$ for the weights and biases corresponding to connections between hidden layers in the two-layer case.

There are several interesting observations that can be made from these results. In both measures, we found that the performance peaks at a certain complexity level, after which point increasing the size of the model only degrades the performance. While this finding is not surprising, what is interesting is the stark difference in where this peak is located across the two measures. When performance was measured by comparing our predictions to the ground truth ALSFRS-R speech score, both the 1 & 2-layer ANNs achieved optimal performance for a layer width of 64 neurons. However when measured based on their ability to predict ISR, we found better performance from the lower-complexity mod-

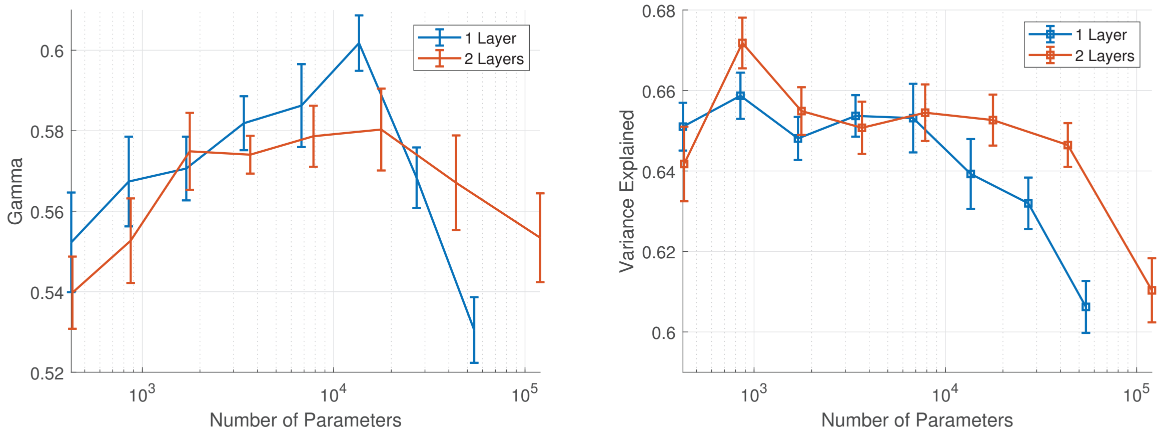

Figure 3: Comparison of results based on model complexity: (Left) Rank correlation (Gamma) between predicted and true speech ALSFRS-R score values (Right) Variance of ISR explained by prediction residuals.

els, with peaks occurring at a layer width of 4 neurons. Although the performance deviations illustrated in these plots were relatively small, and further research is needed to evaluate the robustness of these findings, this provides evidence for a hidden cost associated with increasing model complexity that is not captured by a traditional evaluation of the regression models (at least in the presence of imperfect labels). One possible explanation is that additional complexity of the wider models is predominately used to learn characteristics of the speech score that are artifacts of the ordinal scale that it is measured on, and thus do not contain useful information in predicting speech severity along other metrics.

Another interesting finding is that while the single layer ANN generally outperformed the 2-layer ANN in the speech score evaluation, the inverse is true for the ISR evaluation. Looking at this result in combination with our previous observation, it seems that modeling decisions that make the ANN less prone to over-fitting / memorization (i.e. more layers, with fewer parameters per-layer) tend to lead to predictions which are more strongly correlated with the ISR relative to the ALSFRS-R speech score. Although there has been research on the prediction of other diagnostic measures for individuals with ALS (such as the ALSFRS-R Bulbar Score (Wisler et al. 2019) or ISR (Wang et al. 2016)), this paper represents the first reported work on the prediction of the ALSFRS-R speech scores. Because of this there are no suitable published baselines for direct comparison of the presented results, and it is highly likely the accuracy of the speech score predictions reported here could be improved with additional refinement of the acoustic features and supervised learning strategies that are used. Future work will consider a broader category of regression models to compare with ANNs.

5 Conclusion

In this paper we examined the ability of artificial neural networks to learn clinically-relevant information about patients with ALS beyond what is present in the ordinal labels that the model is trained on. We found that even when trained on a coarse ordinal measure for speech motor control, the continuous ratings generated by the ANN accurately rank-ordered participant's by ISR in 70.5% of cases where there was no difference in the baseline ordinal rating. Additionally, when comparing models of varying complexity, we found that models with fewer parameters were generally capable of characterizing these subtle deviations in motor control even when exhibiting inferior performance in the original prediction task.

6 Acknowledgments

This work was supported by the National Institutes of Health (NIH) under award numbers R01DC013547 and R01DC016621. We also thank Jennifer McGlothlin, Brian D. Richburg, Beiming Cao, Carolina Uzquiano, Alyssa Shrode, Brittany Shrode, Bjorn Bleta, and the volunteering participants.

References

Allison, K. M.; Yunusova, Y.; Campbell, T. F.; Wang, J.; Berry, J. D.; and Green, J. R. 2017. The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to als. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 18(5-6):358–366.

An, K.; Kim, M.; Teplansky, K.; Green, J. R.; Campbell, T. F.; Yunusova, Y.; Heitzman, D.; and Wang, J. 2018. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. *Proc. Interspeech 2018* 1913–1917.

Andres, P. L.; Allred, M.; Stephens, H. E.; Proffitt Bunnell, M.; Siener, C.; Macklin, E. A.; Haines, T.; English, R. A.; Fetterman, K. A.; Kasarskis, E. J.; et al. 2017. Fixed dynamometry is more sensitive than vital capacity or als rating scale. *Muscle & nerve* 56(4):710–715.

Benba, A.; Jilbab, A.; and Hammouch, A. 2015. Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines. *International Journal on Electrical Engineering and Informatics* 7(2):297.

Berisha, V.; Utianski, R.; and Liss, J. 2013. Towards a clinical tool for automatic intelligibility assessment. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2825–2828. IEEE.

Beukelman, D. R.; Yorkston, K. M.; Poblete, M.; and Naranjo, C. 1984. Frequency of word occurbence in communication samples produced by adult communication aid users. *Journal of Speech and Hearing Disorders* 49(4):360–367.

Borrie, S. A.; McAuliffe, M. J.; and Liss, J. M. 2012. Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech, Language, and Hearing Research*.

Cedarbaum, J. M.; Stambler, N.; Malta, E.; Fuller, C.; Hilt, D.; Thurmond, B.; Nakanishi, A.; Group, B. A. S.; complete listing of the BDNF Study Group, A.; et al. 1999. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the neurological sciences* 169(1-2):13–21.

Cheng, J.; Wang, Z.; and Pollastri, G. 2008. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1279–1284. IEEE.

Chu, W., and Keerthi, S. S. 2007. Support vector ordinal regression. *Neural computation* 19(3):792–815.

Dorsey, M.; Yorkston, K.; Beukelman, D.; and Hakel, M. 2007. Speech intelligibility test for windows.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2002–2011.

Goodman, L. A., and Kruskal, W. H. 1979. Measures of association for cross classifications. In *Measures of association for cross classifications*. Springer. 2–34.

Hsu, S.-C.; Jiao, Y.; McAuliffe, M. J.; Berisha, V.; Wu, R.-M.; and Levy, E. S. 2017. Acoustic and perceptual speech characteristics of native mandarin speakers with parkinson's disease. *The Journal of the Acoustical Society of America* 141(3):EL293–EL299.

Liss, J. M.; Spitzer, S. M.; Caviness, J. N.; and Adler, C. 2002. The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America* 112(6):3022–3030.

Norel, R.; Pietrowicz, M.; Agurto, C.; Rishoni, S.; and Cecchi, G. 2018. Detection of amyotrophic lateral sclerosis (als) via acoustic analysis. *bioRxiv* 383414.

Orozco-Arroyave, J.; Hönig, F.; Arias-Londoño, J.; Vargas-Bonilla, J.; Daqrouq, K.; Skodda, S.; Rusz, J.; and Nöth, E. 2016. Automatic detection of parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America* 139(1):481–500.

Tu, M.; Berisha, V.; and Liss, J. 2017. Objective assessment of pathological speech using distribution regression. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5050–5054. IEEE.

Wang, J.; Kothalkar, P. V.; Kim, M.; Yunusova, Y.; Campbell, T. F.; Heitzman, D.; and Green, J. R. 2016. Predicting intelligible speaking rate in individuals with amyotrophic lateral sclerosis from a small number of speech acoustic and articulatory samples. In *Workshop on Speech and Language Processing for Assistive Technologies*, volume 2016, 91. NIH Public Access.

Wang, J.; Kothalkar, P. V.; Kim, M.; Bandini, A.; Cao, B.; Yunusova, Y.; Campbell, T. F.; Heitzman, D.; and Green, J. R. 2018. Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples. *International journal of speech-language pathology* 20(6):669–679.

Williamson, J. R.; Quatieri, T. F.; Helfer, B. S.; Perricone, J.; Ghosh, S. S.; Ciccarelli, G.; and Mehta, D. D. 2015. Segment-dependent dynamics in predicting parkinson's disease. In *Sixteenth annual conference of the international speech communication association*.

Wisler, A.; Teplansky, K.; Green, J. R.; Yunusova, Y.; Campbell, T.; Heitzman, D.; and Wang, J. 2019. Speech-based estimation of bulbar regression in amyotrophic lateral sclerosis. In *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 24–31.

Yorkston, K. M., and 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.