SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

Predicting Intelligible Speaking Rate in Individuals with Amyotrophic Lateral Sclerosis from a Small Number of Speech Acoustic and Articulatory Samples

Jun Wang[1,2], *Prasanna V. Kothalkar*[1], *Myungjong Kim*[1], *Yana Yunusova*[3]
Thomas F. Campbell[2], *Daragh Heitzman*[4], *Jordan R. Green*[5]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]Department of Speech-Language Pathology
University of Toronto, Toronto, Canada
[4]MDA/ALS Center, Texas Neurology, Dallas, Texas, United States
[5]Department of Communication Sciences and Disorders
MGH Institute of Health Professions, Boston, MA, United States
wangjun@utdallas.edu

Abstract

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurological disease that affects the speech motor functions, resulting in dysarthria, a motor speech disorder. Speech and articulation deterioration is an indicator of the disease progression of ALS; timely monitoring of the disease progression is critical for clinical management of these patients. This paper investigated machine prediction of intelligible speaking rate of nine individuals with ALS based on a small number of speech acoustic and articulatory samples. Two feature selection techniques - decision tree and gradient boosting - were used with support vector regression for predicting the intelligible speaking rate. Experimental results demonstrated the feasibility of predicting intelligible speaking rate from only a small number of speech samples. Furthermore, adding articulatory features to acoustic features improved prediction performance, when decision tree was used as the feature selection technique.

Index Terms: amyotrophic lateral sclerosis, intelligible speaking rate, support vector regression

1. Introduction

Amyotrophic lateral sclerosis (ALS), also referred to as Lou Gehrig's disease, is a fast progressive neurological disease that causes degeneration of both upper and lower motor neurons and affects various motor functions, including speech production [1, 2]. The typical survival time is 2-5 years from the onset time [2]. ALS affects between 1.2 and 1.8 /100,000 individuals and the incidence is increasing at a rate that cannot be accounted for by population aging alone [3]. Approximately 30% of patients present with significant speech abnormalities at disease onset; of the remaining patients, nearly all will develop speech deterioration as the disease progresses [4, 5]. Technology for objective, accurate monitoring of speech decline is critical for providing timely management of speech deterioration in ALS and for extending their functional speech communication. Currently, ALS Functional Rating Scale-Revised (ALSFRS- R) - a self-report evaluation - is used for monitoring the progression of changes across motor function [6]. ALS-FRS-R includes 3 questions pertaining to speech, swallowing, and salivation. Commonly used clinical measures for communication efficiency include speech intelligibility (percentage of words that are understood by listeners) and speaking rate (number of spoken words per minute, WPM), which are not closely correlated. Intelligible speaking rate (also called the communication efficiency index) combines intelligibility and rate in a form of speech intelligibility $\times$ speaking rate, providing an index of intelligible spoken words per minute (WPM) [7, 8, 9].

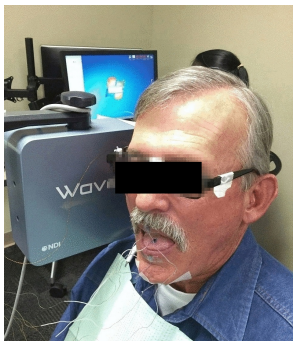
Recent studies have tried to predict the rate of speech intelligibility decline of ALS using an interpretable model based on a comprehensive data set with measures from articulatory, respiratory, resonatory, and phonatory subsystems [10, 11, 12]. Although this approach is promising for understanding the mechanisms of speech decline in ALS, it may not be suitable for clinical environment, given the skill level and the significant time demands required for the data collection and analysis. Novel turnkey and automated speech assessment approaches are, therefore, needed to facilitate clinical diagnosis and management.

Speech signals can be collected using any audio collection devices such as a smart phone and thus can be a great source of information for dysarthria severity estimation. The feasibility of using speech signals revealed promising results in a number of recent studies for disease detection and severity estimation in depression [13, 14], traumatic brain injury [15], and Parkinson's disease detection or severity estimation [16, 17, 18, 19, 20, 21]. Our recent work also showed the feasibility of detection of ALS from speech samples [22]. Estimating the progression of ALS from speech samples using data-driven approaches, however, has rarely been attempted.

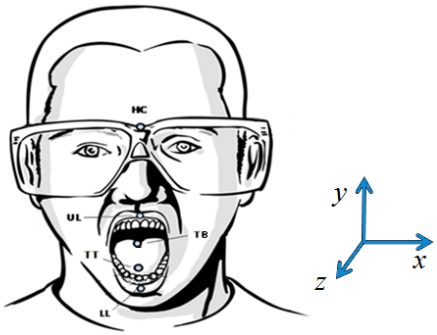
Automatic speech recognition (ASR) systems are a promising but relatively unexplored solution [23, 24]. One significant limitation of ASR for this application is, however, that most approaches require a prohibitively large number of speech samples, since the approach is based on counting the percentage of correctly recognized words. This might be impractical for persons with motor speech disorders due to patient fatigue or variable responses. Yet another challenge of ASR approach is the potential performance variability caused by different speech recognition systems, which is currently understudied.

This project investigated the estimation of speech deteriora-

10.21437/SLPAT.2016-16



(a) *Wave System*



(b) *Sensor Locations. Labels are described in text.*

Figure 1: *Data collection setup.*

tion due to ALS from a number of short speech samples. Data-driven approaches were used to predict intelligible speaking rates of individuals with ALS. As ALS is a motor neuron disease, it affects the articulatory movements including tongue and lip motion patterns [9]. Thus we also tested if the inclusion of articulatory movement data on top of acoustic data can benefit the prediction. Previous studies by Hahm and colleagues used quasi-articulatory features that were inversely mapped from acoustic data, which resulted in improvement for detections of Parkinson's condition estimation [25]. We hypothesized that adding articulatory information to the acoustic data might also benefit the speech performance prediction in ALS.

To our knowledge, this project is the first that aims to predict communication efficiency (intelligible speaking rate or intellgible rate) in ALS from a small number of speech (acoustic and articulatory) samples using data-driven approaches. Speech samples are short phrases that are spoken in daily life (e.g., *How are you doing?*). A pre-defined set of speech features was extracted from acoustic and articulatory samples to represent various characteristics of the speech. Two feature selection techniques were used together with support vector regression (SVR) to predict the intelligible speaking rate. We chose to predict intelligible speaking rate (rather than speech intelligibility and speaking rate) at this stage, because intelligible speaking rate is the measure that better represents the communication efficiency level of ALS patients [8]. To understand if articulatory movement data can improve the prediction, three combinations of features (acoustic, acoustic + lip data, acoustic + lip + tongue data) were tested.

Table 1: Speech intelligibility, speaking rate, and intelligible speaking rate in each recorded session.

Session ID	Speech Intelligibility (%)	Speaking Rate (WPM)	Intelligible Rate (WPM)
S01	95.45	136.60	130.38
S02	80.00	147.98	118.38
S03	100.00	182.33	182.33
S04	98.18	172.54	169.40
S05	79.09	121.10	95.78
S06	99.00	164.189	162.54
S07	98.18	110.47	108.46
S08	0.00	41.05	0.00
S09	94.55	111.11	105.05
S10	80.91	108.20	87.54
S11	23.64	80.29	18.98
S12	99.00	108.73	107.64
S13	96.36	33.33	32.12
S14	79.09	71.88	56.85
Average	80.25	113.56	92.59
SD	29.37	40.04	53.51

2. Data Collection

2.1. Participants

Nine patients (five females) with ALS participated in 14 sessions of data collection. The average age at their first visit was 61 years (SD = 11). Table 1 gives the speech intelligibility, speaking rate, and intelligible speaking rate values for each recorded session. Three of the participants contributed data more than once. S04-S05 were from the same participant but with a year gap. S06-08 were from another patient, with five months and nine months intervals between each two consecutive visits. S09-11 were from another patient with four months and eight months gaps between each two consecutive visits.

2.2. Setup and Procedure

An electromagnetic articulograph (Wave system, NDI Inc., Waterloo, Canada) was used for collecting speech acoustic and articulatory data synchronously. Wave is one of the two commonly used electromagnetic motion tracking technologies by tracking small wired sensors that are attached to the subject's tongue, lips, and head [26]. Figure 1a pictures the device and the patient setup. The spatial accuracy of motion tracking using Wave is 0.5 mm when sensors are in the central space of the magnetic field [27].

After a participant was seated next to the Wave magnetic field generator, sensors were attached to the participant's forehead, tongue, and lips. The head sensor was used to track head movement for head-correction of other sensor's data. The four-sensor set - tongue tip (TT, 5-10 mm to tongue apex), tongue back (TB, 20-30 mm back from TT), upper lip (UL), and lower lip (LL) - was used for our experiments as previous studies indicated that the set is optimal for this application [28, 29, 30]. The positions of five sensors attached to a participant's head, tongue and lips were shown in Figure 1b.

All participants were asked to repeat a list of pre-defined phrases multiple times. The phrases were selected based on lists of phrases that are commonly spoken by AAC (alternative and augmentative communication) users in their daily life [31, 32]. The acoustic and articulatory data were recorded synchronously.

Speech intelligibility and speaking rate were obtained by a certified speech-language pathologist with the assistance of SIT software [33]. Intelligible rate was the multiplication of speech intelligibility and speaking rate. The range of intelligible rate in this data set was between 0-182 words per minute (WPM).

2.3. Data Processing

While raw acoustic data (sampling rate 16Khz, 16 Bit resolution) were used directly for feature extraction, a processing procedure was performed on the articulatory data prior to analysis. The two steps of articulatory data processing included head correction and low pass filtering. The head translations and rotations were subtracted from the tongue and lip data to obtain head-independent tongue and lip movements. The orientation of the derived 3D Cartesian coordinates system is displayed in Figure 1b, in which x is left-right, y is vertical, and z is front-back directions. A low pass filter (i.e., 20 Hz) was applied to remove noise [26].

Invalid samples were rare and were excluded from the analysis. A valid sample contained both valid acoustic and articulatory data. A total of 944 valid samples were recorded. The range of number of samples from individual patients was from 39 to 80.

3. Method

The method of intelligible speaking rate prediction in this project involved two major steps: feature preparation and regression, where feature preparation included feature extraction and selection. The goal of feature extraction was to obtain content-independent acoustic and articulatory features from the data samples. Feature selection was to reduce the data size by choosing the best features for regression. Regression aimed to predict a target score (intelligible speaking rate) from features that are extracted from a data sample.

3.1. Feature Extraction

The script provided in [22] was used for extracting acoustic and articulatory features from acoustic and articulatory motion data, respectively. The script was modified based on that provided in [34]. The window size was 70 ms and the frame shift was 35 ms. The script extracted up to 6,373 pre-defined acoustic features that were categorized in groups such as jitter, shimmer, MFCC, and spectral features. However, low frequency articulatory data do not contain these information. The following feature groups were disabled for articulatory feature extraction [22] :

Jitter, Shimmer, logHNR, Rfilt, Rasta, MFCC, Harmonicity, and Spectral Rolloff.

For each feature group, the following features were calculated and fed into the final feature set before fed into a feature selection technique: mean, flatness, posamean (position of the algorithmic mean), range, maxPos, minPos, centroid, stddev, skewness (a measure of the asymmetry of the spectral distribution around its centroid), kurtosis (an indicator for the peakedness of the spectrum), etc. Please refer to [34, 35] for details of these features.

Therefore, for each dimension (x, y, or z) of a sensor, 1,200 features were extracted. In total, 20,733 features (6,373 acoustic feature + 3,600 articulatory features $\times$ 4 sensors (Tongue Tip, Tongue Body Back, Upper Lip, and Lower Lip) were used in the regression test.

3.2. Feature Selection

Feature selection [36] was performed to reduce the data to the most significant features. We used decision tree regression and gradient boosting as the feature selection procedures.

3.2.1. Decision Tree

Decision trees are rule-based, non-linear classification/regression models that perform recursive partitioning on the data by separating the data into disjoint branches (thus forming a tree structure) for classification or regression [37]. There are a number of ways to measure the quality of a split or branching. We used MSE (mean squared error) as the measure in this project, which is equal to variance reduction as feature selection criterion.

Decision tree-based regression fits the best least squared error criterion to the data. The expected value at each leaf node that minimizes this least squared error is the average of the target values within each leaf l.

$$v_l = \frac{1}{|D_l|} \sum_{D_l} y_i \tag{1}$$

where D_l is the set of samples that are partitioned to leaf l and y_i is the target value of sample i in set D_l.

The splitting criterion is to minimize the fitting error of the resultant tree. The fitting error was defined as the average of the squared differences between the target values Y_l at a leaf node l and the mean value v_l. Error of a tree was defined as the weighted average of the error in its leaves and the error of a split is the weighted average of the error of its resulting sub-nodes.

3.2.2. Gradient Boosting

Gradient boosting [38] applies boosting to regression models by selecting simpler base learners to current pseudo residuals by minimizing least squares loss at each iteration. The pseudo residuals are the gradient of the loss functional that is to be minimized, with respect to model values at each training data point, evaluated at the current step. Given training samples $x_i \in R^d, i = 1, \ldots, n$, and a regression vector $\mathbf{y} \in R^n$ such that $y_i \in R$ we want to find a function $F^{(*)}(\mathbf{x})$ that maps $\mathbf{x}$ to y, to minimize the expected value of some specified loss function $\Delta(y, F(\mathbf{x}))$ over the joint distribution of all ($\mathbf{x}$,y) values. Boosting approximates $F^{(*)}(\mathbf{x})$ by a stage-wise summation of the form

$$F(\mathbf{x}) = \sum_{\mathbf{i=0}}^{\mathbf{N}} \gamma_\mathbf{i} \mathbf{g_i}(\mathbf{x}; \mathbf{a_i}) \tag{2}$$

where the functions $g_i(\mathbf{x}; \mathbf{a_i})$ are chosen as base classifiers of $\mathbf{x}$ in stage i where $\mathbf{a_i}$ is set of parameters. γ_i is the expansion coefficient for stage i.

Gradient boosting solves for arbitrary loss functions for each stage in two steps. First, it fits the function $g_i(\mathbf{x}; \mathbf{a_i})$ to current pseudo residuals by minimizing the least squares loss. Second, the optimal value of the expansion coefficient γ_i was found by single parameter optimization based on a general loss criterion. We selected gradient boosting in this experiment because the model generally works well with small datasets [38].

3.3. Selected Features

The partial lists of features that were selected by decision tree and gradient boosting are given below. Decision tree selected 57 features in total; while gradient boosting selected 517 features. The features selected from articulatory data are indicated in parenthesis; otherwise, the features are selected from acoustic data. Below are the top 10 selected features by decision tree:

1. audspec_lengthL1norm_sma_lpgain
2. shimmerLocal_sma_de_iqr1-3
3. logHNR_sma_percentile99.0
4. F0final_sma_quartile3
5. mfcc_sma[7]_quartile1
6. pcm_fftMag_spectralFlux_sma_quartile1
7. audSpec_Rfilt_sma_de[2]_quartile1
8. pcm_fftMag_fband3-8_sma_de_stddevRisingSlope (TTz)
9. F0final_sma_stddev
10. pcm_fftMag_spectralKurtosis_sma_peakMeanAbs (TTy)

where mfcc stands for mel-frequency cepstral coefficients 1-12; fft denotes fast Fourier transform; pcm means pulse-code modulation, the standard digital representation of analog signals; quartile1 denotes the first quartile (the 25% percentile); quartile 2 denotes the second quartile (the 50% percentile); quartile 3 denotes the third quartile (the 75% percentile); iqr1-3 means the inter-quartile range: quartile3-quartile1; Mag means magnitude; Rfilt means Relative Spectral Transform (RASTA)-style filtered; F0final means the smoothed fundamental frequency (pitch) contour; stddev denotes the standard deviation of the values of the contour; kurtosis is an indicator for the peakedness of the spectrum; sma means smoothing by moving average; de means delta; stddevRisingSlope is the standard deviation of rising slopes, i.e. the slopes connecting a valley with the following peak. The suffix sma appended to the names of the low-level descriptors indicates that they were smoothed by a moving average filter with window length 3 [35]. Spectral flux (F_S^t) for N FFT bins at time frame t is computed as

$$F_S^t = \sqrt{\frac{1}{n}\sum_{f=1}^{N}(\frac{X^t(f)}{E^t} - \frac{X^{t-1}(f)}{E^{t-1}})^2} \tag{3}$$

where E^t is energy at time frame t; $X^t(f)$ is the FFT bin f based on data X at time t. Further, audspec stands for auditory spectrum; shimmerLocal is the local (frame-to-frame) Shimmer (amplitude deviations between pitch periods); lpgain implies the linear predictive coding gain; lengthL1norm is the magnitude of the L1 norm; percentile99.0 is the outlier-robust maximum value of the contour, represented by the 99% percentile and logHNR is the log of the ratio of the energy of harmonic signal components to the energy of noise like signal components. A more descriptive explanation, for example for mfcc_sma[7]_quartitle1, is the 25% percentile of the 7^{th} MFCC that was smoothed using an averaging filter with window length 3.

Below are the top 10 selected features by gradient boosting:

1. audspec_lengthL1norm_sma_lpgain
2. pcm_fftMag_fband1000-4000_sma_percentile1.0
3. F0final_sma_linregc2
4. logHNR_sma_percentile99.0
5. mfcc_sma[6]_quartile2
6. pcm_fftMag_fband3-8_sma_de_lpgain(TBx)
7. audspecRasta_lengthL1norm_sma_peakDistStddev
8. pcm_fftMag_spectralFlux_sma_stddevRisingSlope
9. F0final_sma_percentile99.0
10. audSpec_Rfilt_sma[19]_iqr1-3

where percentile1.0 is the outlier-robust minimum value of the contour, represented by the 1% percentile; linregc2 is the offset (c from y = mx+c) of a linear approximation of the contour; fband denotes frequency band; audspecRasta is the Relative Spectral Transform applied to Auditory Spectrum.

The features were selected based on a feature importance score, which is based on the (normalized) total reduction of the variance brought by that feature [37]. These features with highest importance scores were selected.

3.4. Support Vector Regression

Support vector regression is a regression technique that is based on support vector machine [39], was used as the regression model in this project. SVR is a soft-margin regression technique that depends only on a subset of the training data, because the cost function for building the model does not care about training points that are beyond the margin [40], which is similar with SVM. Details on the introduction of SVR can be found in [41]. We used LIBSVM to implement the experiment [42]. After a preliminary test, ν-SVR [43] outperformed or was comparable to others, thus was selected for regression in this experiment. ν-SVR is a variation of standard SVR, which uses ν to control the ϵ. Given training vector $x_i \in R^d, i = 1, \ldots, n$, and a regression vector $\mathbf{y} \in R^n$ such that $y_i \in R$, the SVR optimization problem is

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b},\xi,\epsilon\in\mathcal{R}^{\mathbf{n}}} \quad & \frac{1}{n}\sum_{i=1}^{n}(\xi_i) - \nu\epsilon + \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^\intercal\phi(\mathbf{x_i}) + \mathbf{b}) \geq \epsilon - \xi_\mathbf{i}, \\ & \xi_i \geq 0, i = 1, \ldots, n, \qquad \epsilon \geq 0 \end{aligned} \tag{4}$$

A kernel function is used to describe the distance between two samples (i.e., r and s in Equation 5). The following radial basis function (RBF) was used as the kernel function K_{RBF} in this study, where γ is an empirical parameter ($\gamma = 1/n$, by default, where n is the number of features) [26]:

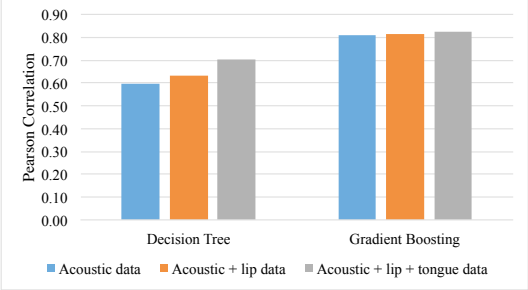
$$K_{RBF}(r, s) = exp(1 - \gamma||r - s||). \tag{5}$$

Please refer to [42] for more details about the implementation of the SVR. All feature values were normalized using z-score before they were fed into SVR.

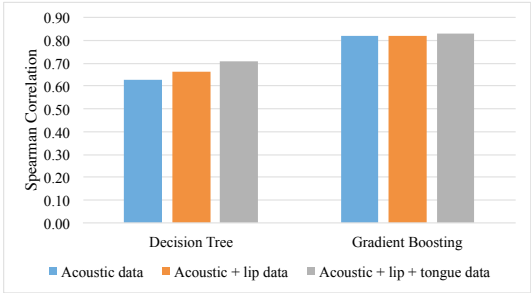
3.5. Experimental Design

As mentioned previously, we tested the prediction on three configurations of data to understand the performance using acoustic signals only and if adding articulatory information is beneficial for the regression. The three configurations of data were acoustic data only, acoustic + lip data, and acoustic + lip data + tongue data.

Three-fold cross validation strategy was used, where all 14 sessions of data were divided into three groups with a balanced distribution of intelligible rates. Initially all the 14 data collections were arranged in ascending order by intelligible rate



(a) *Results measured by Pearson correlation.*



(b) *Results measured by Spearman correlation.*

Figure 2: *Experimental results based on acoustic data only, acoustic + lip data, and acoustic + lip + tongue data using support vector regression and two feature selection techniques.*

(labelled from 1 to 14). Then a jack-knife strategy was used to choose the groups as testing data and the rest as training data. The three folds are sessions (1, 4, 7, 10, 13), (2, 5, 8, 11, 14), (3, 6, 9, 12) (in Table 1). The last validation had four sessions for testing. The data size for testing was about 120 - 360 samples (and the rest for training) in each validation.

Two correlations, Pearson and Spearman, were used to evaluate the performance of the regression. We used both correlations just in case they provide complementay information, becuase of their different characteristics. Pearson correlation is more sensitive than Spearman correlation for outliers [34]; Pearson is typically applied for normally distributed data. The data size is relatively small and the distribution was unknown in this project. Thus, using both correlations (rather than just one of them) may provide more detailed information for interpreting the experimental results. A higher correlation between the estimated rate and the actual rate indicates a better performance.

4. Results and Discussion

Figure 2 gives the results of the regression experiments using SVR and two feature selection techniques, decision tree and gradient boosting, based on acoustic data only, acoustic + lip data, and acoustic + lip + tongue data. The results were measured by Pearson correlation (Figure 2a) and Spearman correlation (Figure 2b). As shown in Figure 2a, the three data configurations obtained Pearson correlations, 0.60, 0.63, and 0.70, respectively when using decision tree, and 0.81, 0.82, and 0.83 when using gradient boosting. The three data configurations obtained Spearman correlations, 0.62, 0.66, and 0.71 respectively when using decision tree, and 0.82, 0.82, and 0.83 when using gradient boosting. There was no difference among the values measured by Pearson or Spearman correlation.

The experimental results indicated the feasibility of predicting intelligible speaking rate from a small number of speech acoustic (and articulatory) samples.

In addition, the results demonstrated that adding articulatory data could improve the performance when using decision tree as the feature selection but not when using gradient boosting. When lip data were added to the acoustic data, the prediction performance was improved when decision tree was used for feature selection. Adding both lip and tongue data obtained the best performance. These findings are consistent with the literature that speech motor function decline (particularly in the articulatory subsystem) are early indicators of the bulbar deterioration in ALS [7]. The added benefit of articulatory data was not obtained when using gradient boosting possibly because this approach was more effective in selecting acoustic features than using the decision tree approach, which required the added articulatory features.

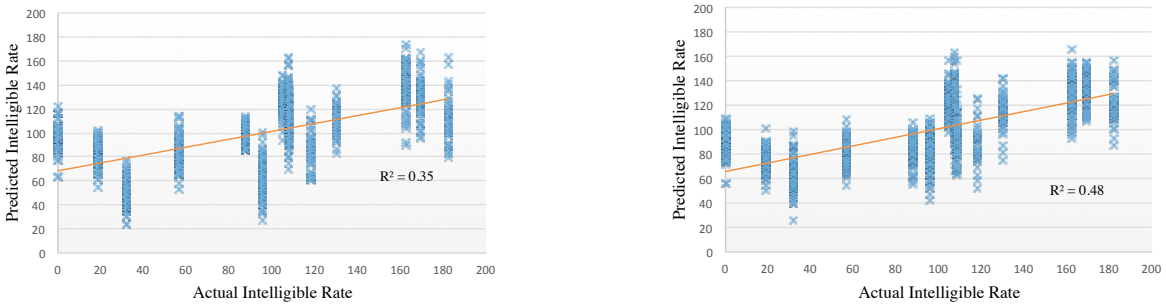
These finding suggested the possibility, in the future, of developing mobile technologies that can collect speech acoustic and lip (via a webcam) as a practical tool for monitoring the ALS speech performance decline as an indicator of disease progression. There are currently logistical obstacle for acquiring tongue data [26] (compared with acoustic data). However, with the availability of portable devices such as portable ultrasound, we anticipated that tongue data will be more accessible in the near future. An alternative solution for tongue data collection is acoustic-to-articulatory inverse mapping [25].

Although comparison of the feature selection techniques was not a focus in this paper, the experimental results indicated that gradient boosting outperformed decision tree. Gradient boosting was so powerful such that adding articulatory information did not show benefit. This finding suggested that feature selection is critical. More feature selection techniques will be explored in the next step of this study.

Figure 3 gives the scatter plots of the measured intelligible rate and predicted intelligible rates using SVR + decision tree on acoustic features only (Figure 3a) and using both acoustic and articulatory features (Figure 3b). Each marker (cross) in the figure represents the measured and predicted intelligible rates on one data sample (a short phrase produced by a patient). As described earlier, each patient produced multiple samples in one session.

A linear regression was applied on both Figure 3a and 3b. The R-squared values illustrated how close the data are to the fitted regression line. A larger value is better. As illustrated in Figure 3, adding articulatory data on top of acoustic data obtained a larger R^2 value, which indicated articulatory features (tongue + lips) improved the prediction on top of acoustic features (using decision tree as the feature selection technique). Specifically, adding articulatory data significantly improved the prediction for some sessions, for example, S13 (with intelligible rate 32.12 WPM) and S05 (with intelligible rate 95.78 WPM). A further analysis is needed to discover how articulatory data affect the prediction performance for these sessions (or patients).

Limitation. The current approach was purely data-driven and used a large number of low-level acoustic and articulatory features. Inclusion of high-level, interpretable features would help the understanding of how these individual features could contribute to the speech decline. Examples of interpretable features include formant centralization ratio [19], intonation [20], and prosody [44], which have already been used for other diseases (e.g., Parkinson's disease).



(a) *Acoustic data only.*

(b) *Acoustic + lip + tongue data.*

Figure 3: *Scatter plots of actual intelligible speaking rate (words per minute) and the predicted values using SVR + decision tree for two data configurations: (a) acoustic data only, and (b) acoustic + lip + tongue data.*

5. Conclusions and Future Work

This paper investigated the automatic assessment of speech performance in ALS from a relatively small number of speech acoustic and articulatory samples. Support vector regression with two feature selection techniques (decision tree and gradient boosting) were used to predict intelligible speaking rate from speech acoustic and articulatory samples. Experimental results showed the feasibility of intelligible speaking rate prediction from acoustic samples only. Adding articulatory data further improved the performance when decision tree was used as the feature selection technique. Particularly, even only lip information was added, the prediction performance was significantly improved. The best results were obtained when both lip and tongue data were added.

The next step of this research would further verify this finding using a larger data set and other feature selection and regression techniques (e.g., deep neural network [25]).

6. Acknowledgements

This work was in parted supported by the National Institutes of Health through grants R01 DC013547 and R03 DC013990, and the American Speech-Language-Hearing Foundation through a New Century Scholar grant. We would like to thank Dr. Panying Rong, Dr. Anusha Thomas, Jennifer McGlothlin, Jana Mueller, Victoria Juarez, Saara Raja, Soujanya Koduri, Kumail Haider, Beiming Cao and the volunteering participants.

7. References

[1] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, pp. 942–955, 2011.

[2] M. Strong and J. Rosenfeld, "Amyotrophic lateral sclerosis: A review of current concepts," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 4, pp. 136–143, 2003.

[3] E. Beghi, G. Logroscino, A. Chi, O. Hardiman, D. Mitchell, R. Swingler, and B. J. Traynor, "The epidemiology of ALS and the role of population-based registries," *Biochimica et Biophysica Acta*, vol. 1762, pp. 1150–1157, 2011.

[4] R. D. Kent, R. L. Sufit, J. C. Rosenbek, J. F. Kent, G. Weismer, R. E. Martin, and B. Brooks, "Speech deterioration in amyotrophic lateral sclerosis: a case study," *Journal of Speech, Language and Hearing Research*, vol. 34, pp. 1269–1275, 1991.

[5] S. E. Langmore and M. Lehman, "The orofacial deficit and dysarthria in ALS," *Journal of Speech and Hearing Research*, vol. 37, pp. 28–37, 1994.

[6] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, and BDNF_ALS_Study_Group, " The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the Neurological Sciences*, vol. 169, pp. 13–21, 1999.

[7] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, pp. 494–500, 2013.

[8] 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, pp. 296–301, 1981.

[9] Y. Yunusova, J. R. Green, L. Greenwoode, J. Wang, G. Pattee, and L. Zinman, "Tongue movements and their acoustic consequences in ALS," *Folia Phoniatrica et Logopaedica*, vol. 64, pp. 94–102, 2012.

[10] Y. Yunusova, J. S. Rosenthal, J. R. Green, P. Rong, J. Wang, and L. Zinman, " Detection of bulbar ALS using a comprehensive speech assessment battery," in *Proc. of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2013, pp. 217–220.

[11] P. Rong, Y. Yunusova, J. Wang, and J. R. Green, " Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioral Neurology*, no. 183027, pp. 1–11, 2015.

[12] P. Rong, Y. Yunusova, J. Wang, L. Zinman, G. L. Pattee, J. D. Berry, B. Perry, and J. R. Green, "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," *PLoS ONE*, vol. 11, no. 5, p. e0154971, 2016.

[13] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[14] T. F. Quatieri and N. Malyska, " Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. of INTERSPEECH*, 2012, pp. 1059–1062.

[15] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, " Using isolated vowel sounds for classification of mild traumatic brain injury," in *Proc. of ICASSP*, 2012, pp. 7577–7581.

[16] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, pp. 842–855, 2011.

[17] M. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1015–1022, 2009.

[18] A. Tsanas, M. Little, P. McSharry, J. Spielman, and L. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 2012.

[19] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant Centralization Ratio (FCR): A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 114–125, 2010.

[20] S. Skodda, W. Grnheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199 – e205, 2011.

[21] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, and E. Noth, "New computer aided device for real time analysis of speech of people with parkinson's disease," *Revista Facultad de Ingenieria Universidad de Antioquia*, no. 72, pp. 87–103, 2014.

[22] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples," in *Proc. of INTERSPEECH*, 2016.

[23] T. Haderlein, C. Moers, B. Mobius, F. Rosanowski, and E. Noth, "Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation," *Proceedings of Text, Speech and Dialogue (TSD), ser. Lecture Notes in Artificial Intelligence*, vol. 6836, pp. 195–202, 2011.

[24] R. Vich, J. Nouza, and M. Vondra, "Automatic speech recognition used for intelligibility assessment of text-to-speech systems," *In Verbal and Nonverbal Features of Human-Human and Human-Machine Interactions, Lecture Notes in Computer Science*, vol. 5042, pp. 136–148, 2008.

[25] S. Hahm and J. Wang, "Parkinsons condition estimation using speech acoustic and inversely mapped articulatory data," in *Proc. of INTERSPEECH*, 2015, pp. 513–517.

[26] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[27] J. Berry, "Accuracy of the NDI wave speech research system." *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–301, 2011.

[28] J. Wang, J. Green, and A. Samal, "Individual articulator's contribution to phoneme production," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7785–7789.

[29] J. Wang, S. Hahm, and T. Mau, "Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition," in *Proc. of ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 79–85.

[30] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, pp. 15–26, 2016.

[31] D. R. Beukelman, K. M., Yorkston, M. Poblete, and C. Naranjo, "Analysis of communication samples produced by adult communication aid users," *Journal of Speech and Hearing Disorders*, vol. 49, pp. 360–367, 1984.

[32] J. Wang, A. Samal, J. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 4985–4988.

[33] D. R. Beukelman, K. M. Yorkston, M. Hakel, and M. Dorsey, "Speech Intelligibility Test (SIT) [Computer Software]," 2007.

[34] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. H. nig, J. R. Orozco-Arroyave, E. Noth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinsons & Eating Condition," in *Proc. of INTERSPEECH*, 2015, pp. 478–482.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Resarch*, vol. 3, pp. 1157–1182, 2003.

[37] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[38] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[40] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, *Support Vector Regression Machines*. MIT Press, 1997, vol. 9.

[41] A. J. Smola and B. Schlkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[43] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[44] S. Skodda, H. Rinsche, and U. Schlegel, "Progression of dysprosody in Parkinson's disease over time - A longitudinal study," *Movement Disorders*, vol. 24, no. 5, pp. 716–722, 2009.