HHS Public Access

Author manuscript

Int J Speech Lang Pathol. Author manuscript; available in PMC 2019 November 08.

Published in final edited form as:

Int J Speech Lang Pathol. 2018 November ; 20(6): 669–679. doi:10.1080/17549507.2018.1508499.

Automatic Prediction of Intelligible Speaking Rate for Individuals with ALS from Speech Acoustic and Articulatory Samples

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

[1]Speech Disorders & Technology Lab, Department of Bioengineering, University of Texas at Dallas, Richardson, Texas, United States

[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

Abstract

Purpose: This research aimed to automatically predict intelligible speaking rate for individuals with Amyotrophic Lateral Sclerosis (ALS) based on speech acoustic and articulatory samples.

Method: Twelve participants with ALS and two normal subjects produced a total of 1,831 phrases. NDI Wave system was used to collect tongue and lip movement and acoustic data synchronously. A machine learning algorithm (i.e. support vector machine) was used to predict intelligible speaking rate (speech intelligibility × speaking rate) from acoustic and articulatory features of the recorded samples.

Result: Acoustic, lip movement, and tongue movement information separately, yielded a R^2 of 0.652, 0.660, and 0.678 and a Root Mean Squared Error (RMSE) of 41.096, 41.166, and 39.855 words per minute (WPM) between the predicted and actual values, respectively. Combining acoustic, lip and tongue information we obtained the highest R^2 (0.712) and the lowest RMSE (37.562 WPM).

Conclusion: The results revealed that our proposed analyses predicted the intelligible speaking rate of the participant with reasonably high accuracy by extracting the acoustic and/or articulatory features from one short speech sample. With further development, the analyses may be well-suited for clinical applications that require automatic speech severity prediction.

Name: Jun Wang, Address: BSB 13.302, 800 W Campbell Rd. Richardson, TX 75080, Country: United States, Phone: (972)-883-6821, wangjun@utdallas.edu

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Keywords

amyotrophic lateral sclerosis; dysarthria; speech kinematics; intelligible speaking rate; machine learning; support vector machine

Introduction

Amyotrophic lateral sclerosis (ALS), also referred to as Lou Gehrig's disease, is a rapidly progressive, neurodegenerative disease that causes degeneration of both upper and lower motor neurons and affects various motor functions, including speech production. The typical survival time of individuals affected by ALS is 2 to 5 years from the symptom onset (Strong et al., 2003). ALS affects between 1.7 and 2.3 /100,000 individuals in the world, but the incidence is increasing at a rate that cannot be accounted for by population aging alone (Beghi et al., 2006). Although speech intelligibility decline may not be the initial symptom at disease onset, nearly all patients with ALS will develop speech (aka bulbar) impairment as the disease progresses (Beukelman, Fager, & Nordness, 2011).

The long-term goal of this research is to develop automated assessments of speech impairments in neurodegenerative diseases, specifically in ALS. This research is motivated by the need for objective, reliable, and accurate diagnostic tools for identifying symptom onset and for monitoring the progression of bulbar dysfunction in ALS (Green et al., 2013). Recent findings suggested that current best practices, which rely primarily on patient symptom reports or clinical ratings, are inadequate for the early detection and clinic monitoring of bulbar motor involvement (Allison, Yunusova, Campbell, Wang, Berry, & Green, 2017; Green et al., 2013). Current best practice for speech assessment typically include a clinician estimates of speech severity, speech intelligibility, and speaking rate (Green et al., 2013; Kent et al., 1991; Yunusova, Green, Greenwood, Wang, Pattee, & Zinman, 2012). Another motivating factor for the automated approach is to minimise patient and clinician burden. Bulbar motor assessments, such as the standard oral motor examination, can be time intensive to administer and fatiguing to patients.

One promising approach to automatic speech assessment is machine learning classification – a technique that forms the basis of automatic speech recognition and is now being used to detect abnormal speech patterns (Mitchell, 1997; Kim, Wang, & Kim, 2016). Widely used machine learning algorithms for speech analysis include support vector machine (SVM) (Cortes & Vapnik, 1995), artificial neural network (ANN) or simply neural network (Bishop, 1995), and hidden Markov models (Rabiner, 1990). The basic concept of machine learning classification is to train a model using a subset of speech data (training data set) and then test the predictive accuracy of the model on a different subset of speech samples (testing data set). The model does not have prior knowledge associated with the prediction problem. Speech data are usually in the form of features that are extracted from speech recordings. A number of open-source algorithms, such as openSMILE (Schuller et al., 2015; Eyben, Wöllmer & Schuller, 2010), are now publicly available for extracting a large number of features from speech audio recordings. OpenSMILE extracts up to 6,373 acoustic features from the speech recordings. Prediction accuracy is, therefore, not only dependent on the

power of the classifier but on how well the selected features – either collectively or individually - represent speech patterns. Researchers are actively searching for the best acoustic speech features for detecting a variety of neurologic conditions including depression (Cummins, Scherer, Krajewski, Schnieder, Epps, & Quatieri, 2015; Quatieri & Malyska, 2012), traumatic brain injury (Falcone, Yadav, Poellabauer, & Flynn, 2013), and Parkinson's disease (Hahm & Wang, 2015; Tsanas, Little, McSharry, & Ramig, 2011; Little, McSharry, Hunter, Spielman & Ramig, 2009; Sapir, Ramig, Spielman, & Fox, 2010; Skodda, Grönheit & Schlegel, 2011; Vásquez Correa, Orozco Arroyave, Arias-Londoño, Vargas Bonilla, & Noth, 2014).

Two long-standing challenges for leveraging the power of acoustic features for speech diagnostics have been (1) difficulties extracting them reliably from disordered speech (Kim, Wang, & Kim, 2016; Kim, Kim, Yoo, Wang, & Kim, 2017) and (2) the large number of speech samples often required for model building, which may be challenging for patients with ALS who experience fatigue while speaking.

To address these limitations, investigators have begun to explore the (1) added value of articulatory features that are extracted directly from recordings of speech movements (Wang, et al., 2016b) and (2) methods of machine classification that require only a small number of training samples. Although only a few studies have been conducted on the diagnostic efficacy of oral-articulatory kinematic features (Green et al., 2013; Rong et al., 2016), our preliminary study on nine patients suggested that the most robust detection of abnormal speech due to ALS is obtained when the machine learning regression model is provided speech acoustic and oral-articulatory kinematic data (Wang, et al., 2016b).

The current study extends this previous, preliminary work in several significant ways. First, the data set is twice as large (e.g. compare 9 to 14 subjects and 944 to 1,831 phrase samples) and the subjects range substantially in the degree of their speech intelligibility impairment. Second, the number of acoustic and articulatory feature groups used in the model is much more exhaustive than the previous. Wang et al., 2016b only compared three groups of data, acoustics, acoustics + lip, and acoustics + lip + tongue; while this study compare up to seven groups, acoustic, lip, tongue, lip + tongue, acoustics + lip, acoustics + tongue, acoustics + lip + tongue. This exhaustive comparison provided a powerful experimental design to evaluate the effectiveness of both acoustics and articulatory features in the automatic prediction of abnormal speech decline due to ALS.

In this paper, we used acoustic and articulatory features as inputs to a machine learning algorithm for automated prediction of intelligible speaking rate based on a single, short speech sample. Intelligible speaking rate, also called communication efficiency, is the multiplication of speech intelligibility score and the speaking rate (Yorkston & Beukelman, 1981). Speech acoustic and articulatory samples were collected when a subject produced short common speech phrases (e.g. *how are you doing?*). A machine learning algorithm (i.e. SVM regression) was used to predict/estimate an individual's intelligible speaking rate. The prediction performance was measured by the coefficient of determination (R^2) and the difference (root mean squared error, RMSE) between the actual and predicted intelligible speaking rates. To determine if the added value of articulatory information contributed to

speech impairment prediction, experiments were conducted on exhaustive combinations of subset of data (e.g. acoustic, lip movement, tongue movement, and their combinations).

Method

Participants

Twelve individuals with ALS (6 females) participated in the data collection. The subjects were selected with a distribution of intelligible speaking rate from zero to more than 200. The average age of these participants was 58 (Standard Deviation = 10), with ranges from 44 to 72 years. To provide a better distribution of intelligible speaking rate scores across all data sessions, two healthy controls (females, N01 and N02) were included (ages 65 and 63, respectively). Data collected from a total of 25 sessions were used.

Table I lists all sessions with speaking rate, speech intelligibility, and intelligible speaking rate scores by participant and session. Subject IDs starting with letter “A” denote patients with ALS; subject IDs starting with letter “N” denote those who are healthy. Seven patients provided data in multiple sessions. Five patients and the two healthy subjects had a single recording session. The speech intelligibility scores ranged from 0 to 100%; the speaking rate ranged from 33.3 to 235 words per minute (WPM); the intelligible speaking rate ranged from 0 to 235 WPM (Table I).

Articulatory motion tracking device

An electromagnetic articulograph (Wave Speech Research System, Northern Digital Inc., Waterloo, Canada; see Figure 1) was used for the collection of articulatory and acoustic data. The voltage induced in the sensor coils by alternating magnetic field is recorded and translated into position and orientation data. The spatial accuracy of motion tracking using Wave is 0.5 mm when sensors are in the central space of the magnetic field (Berry, 2011). Sampling rate was set at 100 Hz for articulatory recording and 22 kHz for acoustic recording. The acoustic data was synchronously recorded with the articulatory data using a microphone. Figure 1(b) illustrates the positions of the five sensors attached to a participant’s head, tongue, and lips. The head centre (HC) sensor was on the bridge of the glasses. We used glasses, rather than taping the sensor to the skin directly, to avoid skin movement artefact during speaking. The movement data of HC were used to calculate the head-independent data of other sensors. Tongue tip (TT) and tongue body back (TB) sensors were attached at the mid-line of the tongue (Wang, Green, Samal, & Yunusova, 2013). TT was 5–10 mm from the tongue apex. TB was as far back as possible and about 30 to 40 mm from TT (Wang et al., 2013). Lip sensors were attached to the vermilion borders of the upper lip (UL) and lower lip (LL) at mid-line. The four-sensor set was found optimal for this application (Wang, Green, & Samal, 2013; Wang, Hahm, & Mau, 2015; Wang, Samal, Rong, & Green, 2016c). Here, optimal means the set has the minimum number of sensors but contains no less information than other sets with more sensors. Figure 1(b) also shows the 3D Cartesian coordinate system derived for our articulatory data movement. Here, x is left-right, y is vertical, and z is front-back.

Procedure

After signing the consent forms, participants were seated with their head within a calibrated magnetic field. Five sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape, following the layout of Figure 1(b). A three-minute training session helped the participants to adapt to the wired sensors before the formal data collection.

All participants were asked to read a list of 20 common phrases, often used in alternative and augmentative communication (AAC) technologies. Example phrases are "*how are you doing?*", "*good afternoon*", and "*I need to make an appointment*". The phrases were repeated sequentially multiple times by all subjects.

Additionally, speech intelligibility and speaking rate were measured by a certified Speech-Language Pathologist using the software Sentence Intelligibility Test (SIT) (Yorkston, Beukelman, Hakel, & Dorsey, 2007). In each session, the SIT software generated a random list of eleven sentences with increasing length from five to fifteen words. Participants were asked to read sentences once, and the audio was recorded. A certified speech-language pathologist who was unfamiliar to the speakers transcribed the words by typing what she heard in the SIT software after the session. The SIT software then calculated speech intelligibility (percentage of correctly perceived words) by comparing how many words were understood correctly and speaking rate by how many (correct or incorrect) words were produced per minute. Finally, the intelligible speaking rate was calculated by multiplying the speech intelligibility score by the speaking rate. Intelligible speaking rate is the percentage of correctly perceived words per minute, which was used as the measure for speech severity of individuals with ALS in this project.

Data processing

Prior to analysis, the quality of each kinematic recording was visually inspected. Sixty-eight invalid samples (e.g. dropped recording frames in the articulatory movement recordings, incorrect pronunciations, sensor falling off) were disregarded. Only recordings with both valid acoustic and articulatory data were considered for this experiment. Each continuous recording of 20 phrases was parsed into 20 individual data files containing both the acoustic and articulatory information. A total of 1,832 valid phrase samples were collected.

A preprocessing procedure was applied on articulatory movement data before data analysis, which included head movement correction and low pass filtering (i.e. with a cut-off frequency 20 Hz). The head translations and rotations were subtracted from tongue and lip sensor trajectories to obtain head-invariant tongue and lip movements. This correction was performed automatically by the NDI Wave system. Low pass filtering was done off-line by the software SMASH (Green, Wang, & Wilson, 2013).

Data analysis

The automatic prediction of intelligible speaking rate from acoustic and articulatory data involved three steps: (1) feature extraction, (2) feature selection, and (3) regression (i.e. prediction of the intelligible speaking rate from the selected features). The goal of feature

extraction was to obtain statistical acoustic and articulatory features from the data samples (i.e. acoustic and articulatory signals obtained from each phrase). Feature selection was employed to reduce the dimensionality of the feature set by choosing the best feature set for regression analysis. Regression aimed to predict a target score (intelligible speaking rate) from features that were extracted and selected from a single phrase. Figure 2 gives a schematic description of the data analysis flow, where acoustic and articulatory signals relative to a single phrase were used as input for feature extraction. The best features were then selected and fed into a machine learning algorithm (SVM regression) to predict the intelligible speaking rate.

Feature extraction.—Acoustic and articulatory features were extracted from the acoustic and articulatory signals of each phrase, respectively, by using the publicly available tool openSMILE (Eyben, Wöllmer & Schuller, 2010; Schuller et al., 2015).

The acoustic feature set was composed of 6,373 pre-defined statistical measures (e.g. mean, standard deviation) of acoustic parameters such as Jitter, Shimmer, MFCC, logHNR estimated within small temporal windows (e.g. 35 ms length). For an exhaustive description of the acoustic feature set, please refer to Eyben Wöllmer & Schuller, 2010 and Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013.

Likewise, openSMILE was used for the extraction of the articulatory features from sensor trajectories. However, given the small frequency range of articulatory movements (10's Hz) with respect to the acoustic signal (10000's Hz), the estimation of several parameters typical of the speech signal (*Jitter, Shimmer, logHNR, Rfilt, Rasta, MFCC, Harmonicity, and Spectral Rolloff*) was disabled. A set of 1,200 measures for each coordinate (x, y, and z) of each sensor (TT, TB, UL, and LL) was extracted, for a total of 14,400 articulatory features.

Thus, 20,773 features (6,373 acoustic + 14,400 articulatory features) were extracted for a phrase sample.

Feature selection using gradient boosting.—The goal of feature selection is to select the most important features for the regression analysis (prediction). This step is essential because it helps avoiding overfitting, removing redundant and irrelevant information (features). The gradient boosting algorithm (Friedman, 2002) was used to reduce the dimension of the feature set. Our preliminary work suggested that gradient boosting works well for this task when compared to other approaches, such as standard decision trees (Wang et al., 2016b).

Gradient boosting is an ensemble machine learning algorithm used for regression, classification, and feature selection composed of several decision trees as a base learner (i.e. non-linear classification/regression models that perform recursive partitioning on the data by separating them into disjoint branches) added sequentially. Specifically, each decision tree is fitted on the residuals of the previous one, rather than on the variable to predict (in our case, intelligible speaking rate). The residuals give a measure of how the previous tree correctly performed the prediction: the higher the residuals, the worse was the prediction. This procedure allows the new added trees to focus more on those instances that previous models

found difficult to predict, in order to reduce the variance and improve the stability of the whole model in the prediction.

Despite its wide use in classification / regression problems (Breiman et al., 1984), gradient boosting can also be used for feature selection, choosing those features that were more useful in the construction of trees within the whole model. In this work, features were selected based on a feature importance score (in percentage) that depends on the total reduction of the variance brought by that feature (Breiman et al., 1984). The features with the highest importance score whose sum was 100% were selected.

Intelligible Speaking Rate Prediction using SVM regression.—SVM is a machine learning algorithm extensively used for classification and regression (the latter is also called support vector regression - SVR) (Drucker et al., 1996). We used SVM because SVM is a widely used machine learning classifier that can be rapidly implemented and showed great performance in our previous studies (e.g. Wang, Green, Samal, & Yunusova, 2013; Wang et al., 2015, 2016a, 2016b), where we used SVM, neural networks, decision tree, Gaussian mixture model (GMM), hidden Markov model (HMM), and other classifiers. The goal of SVM for classification is to find the best hyperplane in the feature space that allows the maximum margin separation between data samples of two classes. If data are not linearly separable in the original feature space, they are mapped (using a so-called “kernel function”) into a higher-dimensional space where the linear separation can be performed. The concept of SVR is similar: data samples are mapped into a higher-dimensional feature space where a linear model can be fitted in order to describe the data accurately.

In this study, we used a variation of the standard SVR called ν-SVR, where an extra parameter ν is used to control the maximum deviation from the actual intelligible speaking rate. This choice was based on preliminary tests where ν-SVR outperformed or was comparable to other SVR variants, such as ε-SVR (unpublished).

A radial basis function (RBF) was used as the kernel function for the algorithm. Consistently with our previous studies (Wang, Green, Samal, & Yunusova, 2013; Wang, Samal, Rong, & Green, 2016), preliminary tests confirmed that RBF outperformed other kernels including linear and polynomial functions. The implementation of SVR in the open access machine learning tool Weka was used (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009).

Regression (Prediction) experimental setup.—To understand how the prediction performance changed using different types of information (acoustic and/or articulatory information), seven combinations of the three groups of selected features were tested: (1) acoustic, (2) lips, (3) tongue, (4) acoustic + lips, (5) acoustic + tongue, (6) lips + tongue, and (7) acoustic + lips + tongue.

Three-fold cross validation strategy was used for testing our regression model performance. The dataset was divided into three parts: two-third considered for training and the remaining one-third for testing. The procedure was repeated three times, so that each third of the

dataset was considered as test set (once) and training set (twice). In each execution, every single phrase sample was used as the input for prediction.

For measuring the performance of our regression model, we used the coefficient of determination (R^2 value) between the actual and predicted (intelligible speaking rate) values. In addition, Root Mean Squared Error (RMSE) values, sample standard deviation of the differences between predicted values and observed values, were used to quantify the difference between the predicted scores and the true values. Lower RMSE indicated a better performance. We did not use false negative and positives as the additional measures, because the goal in the project is to predict continuous values, rather a binary outcome. Although we could possibly use a ROC curve analysis to set a criterion value (cut point for prediction errors) to determine false positives and negatives, it is beyond the scope of the current analyses with a page limitation. Thus, we will leave it as a future direction for exploration.

The overall prediction performance was obtained by averaging the values of R^2 and RMSE obtained during the three executions of the cross-validation.

Result

Selected features

A total of 499 acoustic and articulatory features were selected by the gradient boosting algorithm. Two hundred and sixty-six of them were selected from acoustic data; 117 were selected from lip movement data; and 116 were selected from tongue movement data. Table II gives three examples (the best three) of the selected features with explanations (Eyben et al., 2010) and selection weights. Those features were selected from the combined acoustic and articulatory data. The features selected from articulatory data are accompanied by the sensor name and in dimension in parentheses; otherwise, the features are selected from acoustic signals. Selection weight (the third column) gives how much information the individual feature account for (in percentage) among all the selected features.

More descriptive explanations of these selected feature are given under the name in the first column; The second column gives detailed explanation about these terminologies used in the feature name. For example, shimmerLocal_sma_de_quartile1 is the 25% percentile of the delta value for the local pitch period deviations that was smoothed using an averaging filter with window length 3, where quartile1 means the 25% percentile, *de* denotes delta value, *sma* denotes an averaging filter with window length 3, and *shimmerLocal* means the local pitch period deviations. A ranked list of the top 40 selected features and detailed explanation are provided in the online supplementary material.

Intelligible Speaking Rate Prediction (Regression)

Figure 3 gives the results of the regression experiments using the selected 390 acoustic + lip + tongue movement features. The R^2 was 0.712 ($p < 0.0001$). The 390 features were further selected from the initially selected 499 features (Gradient Boosting was applied again on the 499 features, which returned 390 features). The R^2 was 0.680 if using all 499 features.

Detailed results of the regression experiments using different combinations of data sources (acoustic, lip, and tongue motion data) are reported in the online supplemental materials.

Figure 4 summarises the all the results (R^2 values in Figures 3 - 9) of the regression experiments using individual or combined acoustic, lip movement, and tongue movement features. A higher R^2 value indicated a higher correlation between the predicted intelligible speaking rate and the actual intelligible speaking rate. When using only individual source of information (acoustic, lip, or tongue), tongue information yielded the highest R^2 value (0.678); acoustic information only yielded the lowest R^2 value (0.652), which was lower than that obtained using lip information only (0.660). When two sources of information were combined, lip and tongue together obtained the highest R^2 value (0.702). The best performance ($R^2 = 0.712$) was obtained when all the three sources of information (i.e. acoustic, lip, and tongue movement information) were used together.

Figure 5 summarises the all the RMSE values (in WPM) in the regression experiments using individual or combined acoustic, lip movement, and tongue movement features. A smaller RMSE value indicated the predicted intelligible speaking rate is closer to the actual intelligible speaking rate. When using only individual source of information (acoustic, lip, or tongue), tongue movement information yielded the lowest (best) RMSE value (39.86); lip movement information yielded the highest (worst) value RMSE (41.17), which was higher than that obtained using acoustic information only (40.10). When two sources of information were combined, acoustic and tongue together obtained the lowest (best) RMSE value (39.49). The best performance (RMSE = 37.51) was obtained when all the three sources of information (i.e. acoustic, lip, and tongue movement information) were used together.

Discussion

This study used a machine learning-based approach (feature selection + SVM regression) to predict intelligible speaking rate of patients with ALS based on a single speech acoustic and articulatory sample. Acoustic and articulatory data were collected from twelve participants with ALS and two healthy controls. Results revealed that by extracting the acoustic and/or articulatory features from one short speech sample, our approach predicted the intelligible speaking rate of the participant with a reasonable high accuracy (a high R^2 and a low RMSE value).

Selected features

In this project, we used data features extracted by openSMILE, a widely used data-driven approach without a priori assumption about their feature connection to physiological functions. The ranked, top 40 selected features are given as Supplementary materials. The findings about the relative importance of these features were somewhat consistent with the literature on voice changes due to ALS (Tomik, Tomik, Wiatr, Składzień, Stręk, & Szczudlik, 2015; Wang, Kothalkar, Cao, & Heitzman, 2016; Wang, Kothalkar, Kim, Yunusova, Campbell, Heitzman, & Green, 2016) and other neurological disorders including Parkinson's disease (Cummins, Scherer, Krajewski, Schnieder, Epps, & Quatieri, 2015; Quatieri & Malyska, 2012; Vásquez Correa, Orozco Arroyave, Arias-Londoño, Vargas

Bonilla, & Noth, 2014; Tsanas, Little, McSharry, & Ramig, 2011). For example, the above literature found that hoarseness, voice range, amplitude of vibration showed significant abnormalities with repeated examination of patients with ALS. In addition, Tomik and colleagues (2015) found jitter and NHR (noise-to-harmonic ratio) in women with ALS were increased in all longitudinal examinations. Our study cohort was comprised of more than half female participants.

Almost half of the selected feature were associated with the articulatory subsystem (tongue or lips), which is consistent with prior findings demonstrating that, among the four speech subsystems (i.e. articulatory, phonatory, resonatory, respiratory), impairment to the articulatory subsystem has the greatest impact on speech intelligibility (Rong et al., 2016). As indicated in the list of top ranked features, one feature from lower lip (LLy) was ranked up as high as third, which suggested lip movement features provided complementary information that acoustic data did not contain. More interestingly, multiple features (33rd, 38th and 39th) in the list were from the *x* dimension of upper lip (UL*x*), lower lip (LL*x*) and tongue back (TB*x*) respectively. Although *x* (left-right) movement is not a major component of speech movements in healthy talkers (Wang, Green, Samal, & Yunusova, 2013; Westbury, 1994), the current findings suggest that movement in this dimension may become more prominent in persons with ALS.

Intelligible Speaking Rate Prediction

The experimental results demonstrated the effectiveness of automatic estimation of intelligible speaking rate from single speech acoustic and articulatory samples. The results demonstrated that adding articulatory (lip and/or tongue movement) information could significantly improve the prediction performance. These findings are consistent with the literature that speech motor function decline (particularly in the articulatory subsystem) may be an early indicator of the bulbar deterioration in ALS (Allison et al., 2017; Green et al., 2013; Rong et al., 2015) and adding articulatory information significantly improved the performance of automatic detection of ALS through speech samples (Wang et al., 2016a).

When compared the prediction performance using acoustic, lip, and tongue motion data separately, tongue data yielded the highest R^2 value (0.678) and lowest (best) RMSE value (39.86). The finding that tongue movement information yielded the best performance was somewhat incongruous with the fact that the top three selected features (see details in the previous sub-section) were not tongue features (but rather, two acoustic features and one lip feature from LL_y). However, we think these low-level features (on small segments) may not be statistically comparable with the descriptive, high-level articulators.

The RMSE was as low as 37 words per minute when using both acoustic and articulatory information in prediction. This level of performance is encouraging considering the large distribution of the intelligible speaking rate (0 to more than 200) included in our cohort. Further research is required to identify features and classifiers that will lower the RMSE values. We did not directly compare the RMSE number with that in our prior study (Wang et al., 2016b), because the prior study was preliminary. However, findings in this work were consistent with these in Wang et al., 2016b. For example, data from more articulatory flesh points yielded better performance (lower RMSE).

This approach is theoretically (phrase) content independent because the features were low-level and are not relevant with the content. This means the technique is not dependent on the list of phrases. In other words, if a smaller set of unique phrases (i.e. one unique phrase) was used, the prediction performance could be even better. However, the predictability of the phrases (different intelligibility levels on the same phrases) may affect the prediction performance in practice. Further work is needed to confirm if the predictability variation would cause the prediction measure variation.

The current approach was purely data-driven and used a large number of low-level acoustic and articulatory features. Additional insights into the physiologic mechanism that drive speech decline will require the extraction of more interpretable speech features such as formant centralisation ratio (Fletcher, McAuliffe, Lansford, & Liss, 2017; Sapir et al., 2010), articulation entropy (Jiao et al., 2017), intonation (Skodda et al., 2011), prosody (Skodda, Rinsche & Schlegel, 2009), formants (Horwitz-Martin et al, 2016), and speech pauses (Rong, Yunusova, Wang, Zinman, Pattee, Berry, & Green, 2016).

Future work will explore other powerful feature selection techniques and machine learning classifiers. Better feature selection will provide more information for the purpose of prediction. A more powerful machine learning classifier (e.g. deep neural network) may be better able to capture the abnormal speech patterns that is more related to speech decline due to ALS.

Limitations and potential clinical application

Although the present study yielded promising results with a novel technology, the data set contains a relatively small number of patients. A future study with a larger number of patients will further refine and validate this approach.

Another limitation is that we used the manual marking from only one SLP as the reference/gold standard, which may have a bias when compared to the machine prediction results. Measures from multiple SLPs will be added in the future direction of this work.

In this study, we used 3-fold cross validations. A larger number of folds (e.g. 6 folds) may be helpful to reduce the bias. A further analysis on the variance of these individual validations is needed to explore the effects by the number of folds, although we believe the number of folds may not play a critical role because the final, reported performance was the average of all validations.

Although recording articulatory data in clinical setting is currently logistically difficult, new low-cost motion capture systems are rapidly proliferating. For example, many of the next generation of smart phones will be equipped with 3D depth sensing technology, which can be used to record lip and jaw motion data. With further development, the proposed analyses will be well-suited for clinical applications that require automatic speech severity prediction. For example, acoustic and lip motion information will be collected easily using a portable device such as mobile phones, which will be a fit for future wide-scale clinical implementation. The mobile app will also serve as a convenient tool for visual display of the prediction results.

Conclusion

This paper investigated the automatic estimation/prediction of speech severity (measured by intelligible speaking rate) in ALS from single speech acoustic and articulatory samples. Gradient boosting was used as the feature selection technique. A machine learning algorithm (support vector machine) was used to predict intelligible speaking rate from speech acoustic and articulatory (tongue and lip movement) samples. Experimental results showed that reasonably accurate estimates of intelligible speaking rate can be generated from acoustic samples only. Tongue and lip motion information yielded a higher accuracy than using acoustic information only. Furthermore, combining acoustic and articulatory (tongue and lip movement) information obtained the best prediction performance. These findings provided preliminary support for employing machine learning models for predicting speech performance (measured by intelligible speaking rate) in individuals with ALS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was supported by the National Institutes of Health through grants R01DC013547, R03DC013990, K24DC016312, and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant. We would like to thank Dr. Anusha Thomas, Jennifer McGlothlin, Brian Richburg, Kristin Teplansky, Jana Mueller, Saara Raja, Heather Xiao, and the volunteering participants.

References

Allison K, Yunusova Y, Campbell T, Wang J, Berry J, & Green J (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. DOI: 10.1080/21678421.2017.1303515 [PubMed: 28355886]

Beukelman D, Fager S, & Nordness A (2011). Communication support for people with ALS, Neurology Research International, no. 714693, 6 pages.

Beghi E, Logroscino G, Chiò A, Hardiman O, Mitchell D, Swingler R, & EURALS Consortium. (2006). The epidemiology of ALS and the role of population-based registries. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1762(11), 1150–1157. [PubMed: 17071060]

Berry JJ (2011). Accuracy of the NDI wave speech research system. Journal of Speech, Language, and Hearing Research, 54(5), 1295–1301.

Bishop CM (1995). Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA.

Breiman L, Friedman J, Stone CJ, & Olshen RA (1984). Classification and regression trees. CRC press.

Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, & 1A complete listing of the BDNF Study Group. (1999). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. Journal of the Neurological Sciences, 169(1), 13–21. [PubMed: 10540002]

Cortes C, & Vapnik V (1995). Support-vector networks. Machine learning, 20(3), 273–297.

Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, & Quatieri TF (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49.

Drucker H, Burges CJ, Kaufman L, Smola A, & Vapnik V (1997). Support vector regression machines. Advances in Neural Information Processing Systems, 9, 155–161.

Eyben F, Wöllmer M, & Schuller B (2010, 10). Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia (pp. 1459–1462). ACM.

Falcone M, Yadav N, Poellabauer C, & Flynn P (2013). Using isolated vowel sounds for classification of mild traumatic brain injury. *In* Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7577–7581). IEEE.

Friedman JH (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378.

Green JR, Wang J, & Wilson DL (2013, 8). SMASH: a tool for articulatory data processing and analysis. *In* Interspeech (pp. 1331–1335).

Green JR, Yunusova Y, Kuruvilla MS, Wang J, Pattee GL, Synhorst L, & Berry JD (2013). Bulbar and speech motor assessment in ALS: Challenges and future directions. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 14(7–8), 494–500. [PubMed: 23898888]

Hahm S, & Wang J (2015). Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data. Proceedings of Interspeech (pp. 513–517). International Speech and Communication Association.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, & Witten IH (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18.

Horwitz-Martin RL, Quatieri TF, Lammert AC, Williamson JR, Yunusova Y, Godoy E, & Green JR (2016). Relation of Automatically Extracted Formant Trajectories with Intelligibility Loss and Speaking Rate Decline in Amyotrophic Lateral Sclerosis. Proceedings of Interspeech (pp. 1205–1209).

Jiao Y, Berisha V, Liss J, Hsu S-C, Levy E, & McAuliffe M (2017). Articulation entropy: An unsupervised measure of articulatory precision, IEEE Signal Processing Letters, 24(4): 485–489.

Kent RD, Sufit RL, Rosenbek JC, Kent JF, Weismer G, Martin RE, & Brooks BR (1991). Speech Deterioration in Amyotrophic Lateral Sclerosis. A Case Study. Journal of Speech, Language, and Hearing Research, 34(6), 1269–1275.

Kiernan MC, Vucic S, Cheah BC, Turner MR, Eisen A, Hardiman O, & Zoing MC (2011). Amyotrophic Lateral Sclerosis. The Lancet, 377(9769), 942–955.

Kim M, Kim Y, Yoo J, Wang J, & Kim H (2017). Regularized speaker adaptation of KL-HMM for dysarthric speech recognition, IEEE Transactions on Neural Systems & Rehabilitation Engineering, 25(9): 1581–1591. [PubMed: 28320669]

Kim M, Wang J, & Kim H (2016). Dysarthric speech recognition using Kullback-Leibler divergence-based hidden markov model. Proceedings of Interspeech (pp. 2671–2671).

Langmore SE, & Lehman ME (1994). Physiologic deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis. Journal of Speech, Language, and Hearing Research, 37(1), 28–37.

Little MA, McSharry PE, Hunter EJ, Spielman J, & Ramig LO (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022. [PubMed: 21399744]

Mitchell T (1997). Machine Learning, McGraw Hill, 414 pages.

Quatieri TF, & Malyska N (2012). Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity. Proceedings of Interspeech (pp. 1059–1062).

Rabiner LR (1990). A tutorial on hidden Markov models and selected applications in speech recognition In Readings in Speech Recognition, Waibel Alex and Lee Kai-Fu (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 267–296.

Rong P, Yunusova Y, Wang J, & Green JR (2015). Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. Behavioural Neurology, 1–11.

Rong P, Yunusova Y, Wang J, Zinman L, Pattee GL, Berry JD, & Green JR (2016). Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. PloS ONE, 11(5), e0154971, 1–19. [PubMed: 27148967]

Sapir S, Ramig LO, Spielman JL, & Fox C (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. Journal of Speech, Language, and Hearing Research, 53(1), 114–125.

Schölkopf B, Smola AJ, Williamson RC, & Bartlett PL (2000). New support vector algorithms. Neural Computation, 12(5), 1207–1245. [PubMed: 10905814]

Schuller BW, Steidl S, Batliner A, Hantke S, Hönig F, Orozco-Arroyave JR, & Weninger F (2015). The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition. Proceedings of Interspeech (pp. 478–482).

Skodda S, Rinsche H, & Schlegel U (2009). Progression of dysprosody in Parkinson's disease over time—a longitudinal study. Movement Disorders, 24(5), 716–722. [PubMed: 19117364]

Skodda S, Grönheit W, & Schlegel U (2011). Intonation and speech rate in Parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission. Journal of Voice, 25(4), e199–e205. [PubMed: 21051196]

Smola AJ, & Schölkopf B (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199–222.

Strong M, & Rosenfeld J (2003). Amyotrophic lateral sclerosis: a review of current concepts. Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders, 4(3), 136–143. [PubMed: 13129799]

Tomik J, Tomik B, Wiatr M, Składzień J, Stręk P, & Szczudlik A (2015). The evaluation of abnormal voice qualities in patients with amyotrophic lateral sclerosis. Neurodegenerative Diseases, 15(4), 225–232. [PubMed: 25967115]

Tsanas A, Little MA, McSharry PE, & Ramig LO (2011). 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, 8(59), 842–855.

Vásquez Correa JC, Orozco Arroyave JR, Arias-Londoño JD, Vargas Bonilla JF, & Noth E (2014). New computer aided device for real time analysis of speech of people with Parkinson's disease. Revista Facultad de Ingeniería Universidad de Antioquia, (72), 87–103.

Wang J, Green JR, & Samal A (2013, 5). Individual articulator's contribution to phoneme production. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 7785–7789). IEEE.

Wang J, Green JR, Samal A, & Yunusova Y (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. Journal of Speech, Language, and Hearing Research, 56(5), 1539–1551.

Wang J, Hahm S, & Mau T (2015, September). Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition. Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (pp. 79–85).

Wang J, Kothalkar PV, Cao B, & Heitzman D (2016a). Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. Proceedings of Interspeech (pp. 1195–1199).

Wang J, Kothalkar PV, Kim M, Yunusova Y, Campbell TF, Heitzman D, & Green JR (2016b). Predicting Intelligible Speaking Rate in Individuals with Amyotrophic Lateral Sclerosis from a Small Number of Speech Acoustic and Articulatory Samples. Proceedings of the ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies (pp. 91–97).

Wang J, Samal A, Rong P, & Green JR (2016c). An optimal set of flesh points on tongue and lips for speech-movement classification. Journal of Speech, Language, and Hearing Research, 59(1), 15–26.

Weninger F, Eyben F, Schuller BW, Mortillaro M, & Scherer KR (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. Frontiers in Psychology, 4:292. [PubMed: 23750144]

Westbury J (1994). X-ray microbeam speech production database user's handbook. Unpublished manuscript, University of Wisconsin–Madison

Yorkston KM & Beukelman DR (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate," Journal of Speech and Hearing Disorders, 46, 296–301. [PubMed: 7278175]

Yorkston K, Beukelman D, Hakel M, Dorsey M. Sentence Intelligibility Test, Speech Intelligibility Test. Lincoln, Neb, USA: Madonna Rehabilitation Hospital; 2007.

Yunusova Y, Green JR, Greenwood L, Wang J, Pattee GL, & Zinman L (2012). Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. Folia Phoniatrica et Logopaedica, 64(2), 94–102. [PubMed: 22555651]

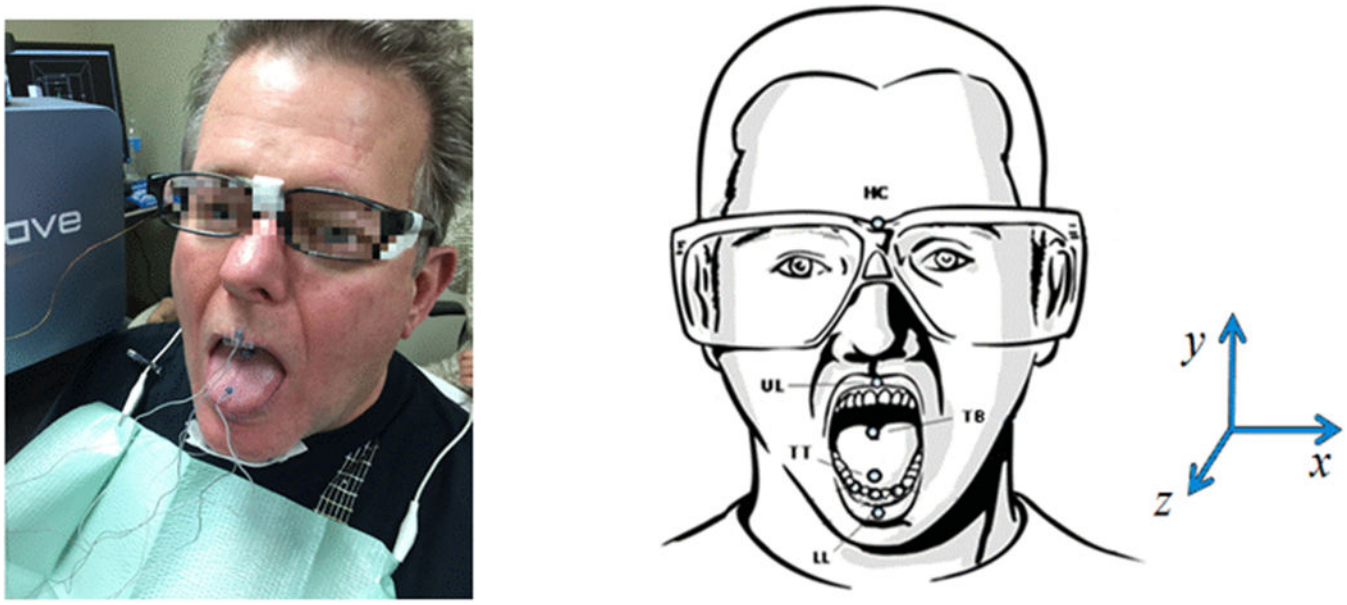

Figure 1.
Data collection setup. The left picture shows the Wave Speech Research System. The right picture illustrates sensor locations. Sensor labels are described in text.

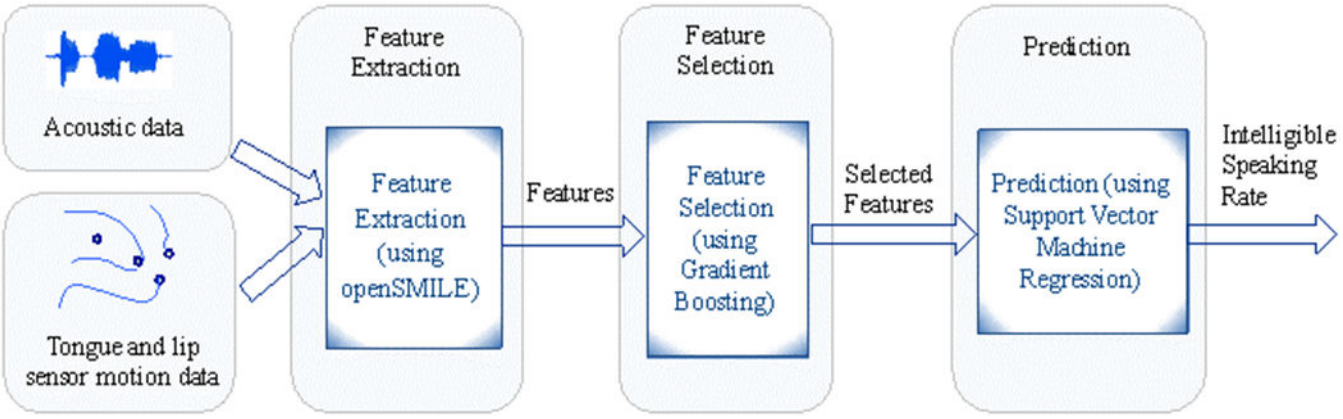

Figure 2.
Schematic description of the data analysis flow.

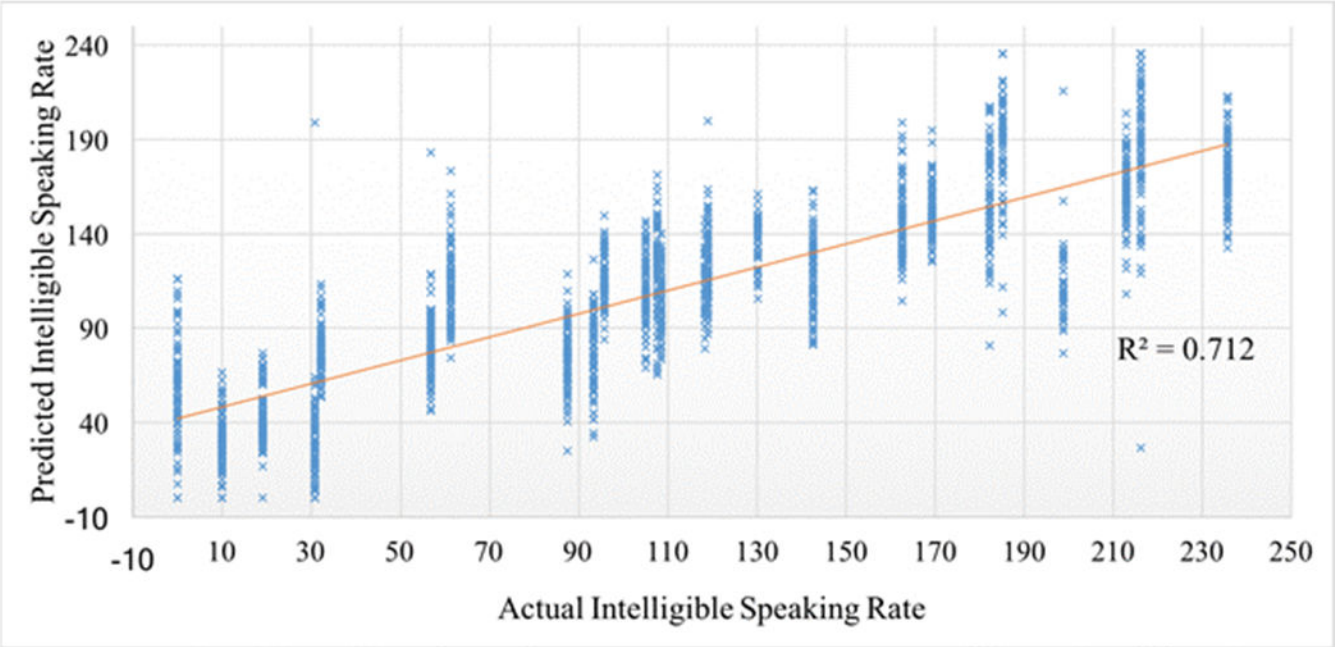

Figure 3.

Scatter plots of actual intelligible speaking rate (words per minute) and the predicted values using 390 *acoustic* + *lip* + *tongue movement features*. The 390 features were further selected from the initially selected 499 features.

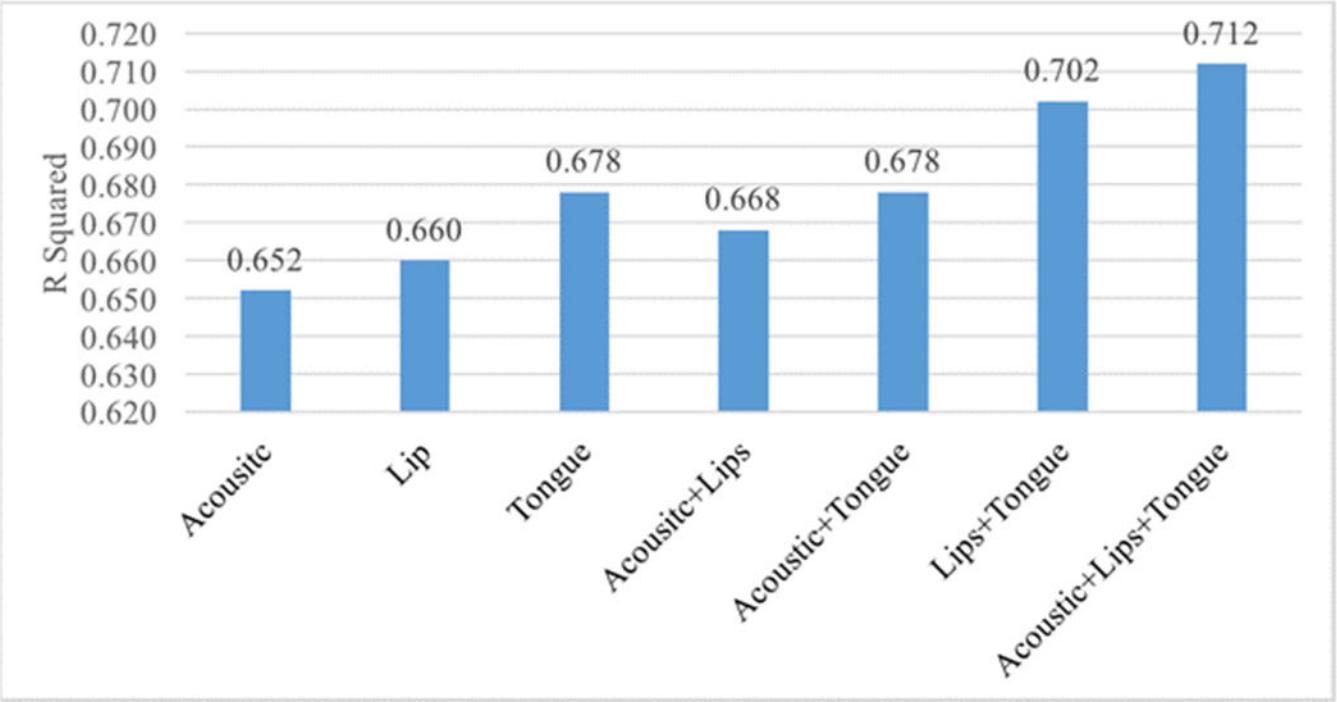

Figure 4.

R^2 values in the intelligible speaking rate prediction using individual or combined values of *acoustic* + *lip* + *tongue movement features*.

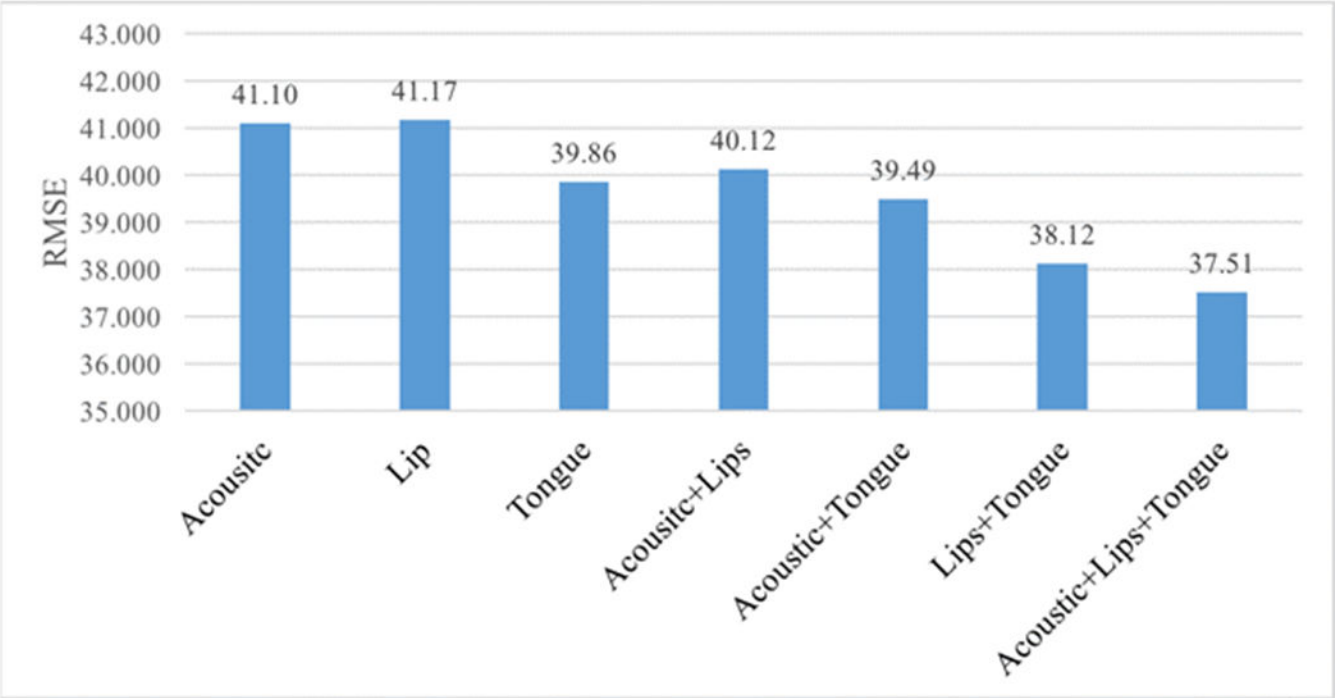

Figure 5.
RMSE values in the intelligible speaking rate prediction using individual or combined values of *acoustic* + *lip* + *tongue movement features*.

Table I:

Speech Intelligibility, Speaking Rate, and Intelligible Speaking Rate of the subjects in each recorded session

Subject ID	Session ID	Speech Intelligibility (%)	Speaking Rate (WPM)	Intelligible Rate (WPM)
A01	S01	95.45	136.36	130.16
	S02	96.36	123.60	119.10
A02	S03	80.00	147.98	118.38
A03	S04	100.00	182.33	182.33
	S05	96.36	218.54	210.59
	S06	100.00	235.71	235.71
A04	S07	98.18	172.54	169.40
	S08	97.27	146.67	142.67
	S09	79.09	121.10	95.78
A05	S10	99.09	164.18	162.69
	S11	98.18	110.47	108.46
	S12	0.00	41.05	0.00
A06	S13	94.55	111.11	105.05
	S14	80.91	108.20	87.54
	S15	23.64	80.29	18.98
A07	S16	99.00	108.73	107.64
	S17	92.73	100.61	93.3
A08	S18	96.36	33.33	32.12
A09	S19	79.09	71.88	56.85
A10	S20	100.00	216.16	216.16
A11	S21	13.00	76.31	9.92
	S22	40.00	76.92	30.77
A12	S23	100.00	212.90	212.90
N01	S24	100.00	194.12	185.29
N02	S25	99.09	200.61	198.78

Table II:

Example selected features

Feature (Explanation in parenthesis)	Detailed Explanation	Selection Weight (%)
shimmerLocal_sma_de_quartile1 *(25% percentile of the delta value for the local pitch period deviations that was smoothed using an averaging filter with window length 3)*	*shimmerLocal*: the local (frame-to-frame) Shimmer (pitch period amplitude deviations) 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. *de*: delta *quartile1*: the first quartile (the 25% percentile) *quartile2*: the second quartile (the 50% percentile) *quartile3*: denotes the third quartile (the 75% percentile)	1.47083
pcm_fftMag_fband1000-4000_sma_percentile1.0 *(the outlier-robust minimum value of contour, represented by the 1% percentile of the pulse-code modulation magnitude after fast Fourier transform using frequency band between 1000 and 4000 hz)*	*pcm*: pulse-code modulation, the standard digital representation of analog signals *fft*: fast Fourier transform *Mag*: magnitude *fband*: frequency band *percentile1.0*: the outlier-robust minimum value of the contour, represented by the 1% percentile *percentile99.0*: the outlier-robust maximum value of the contour, represented by the 99% percentile	1.39430
pcm_fftMag_spectralFlux_sma_lpgain ***(LLy)*** *(spectralFlux of the pulse-code modulation amplitude after fast Fourier transform and smoothed by moving average on linear predictive coding energy)*	*lpgain* implies the linear predictive coding gain. *Gain* means the energy of the frame. Explanation and calculation of *spectral flux* is given in the Appendix. This feature was calculated from LLy data (y coordinate of Lower Lip)	1.32485