INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples

Jun Wang[1,2], *Prasanna V. Kothalkar*[1], *Beiming Cao*[1], *Daragh Heitzman*[3]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]MDA/ALS Center, Texas Neurology, Dallas, Texas, United States
{wangjun, prasanna.kothalkar, beiming.cao}@utdallas.edu; dheitzman@texasneurology.com

Abstract

Amyotrophic lateral sclerosis (ALS) is a rapid neurodegenerative disease that affects the speech motor functions of patients, thus causes dysarthria. There is no definite marker for the diagnosis of ALS. Currently, the diagnosis of ALS is primarily based on clinical observations of upper and lower motor neuron damage in the absence of other causes, which is time-consuming, of high cost, and often delayed. Timely diagnosis and assessment for ALS are crucial. Automatic detection of ALS from speech samples would advance the diagnosis of ALS. In this paper, we investigated the automatic detection of ALS from short, pre-symptom speech acoustic and articulatory samples using machine learning approaches (support vector machine and deep neural network). A data set of more than 2,500 speech samples collected from eleven patients with ALS and eleven healthy speakers was used. Leave-subjects-out cross validation experimental results indicate the feasibility of the automatic detection of ALS from speech samples. Adding articulatory motion information (from tongue and lips) further improved the detection performance.

Index Terms: amyotrophic lateral sclerosis, human-computer interaction, computational paralinguistics

1. Introduction

Amyotrophic lateral sclerosis (ALS), also referred to as Lou Gehrig's disease, is the most common motor neuron disease that causes degeneration of both upper and lower motor neurons [1]. There is no cure for ALS. 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 [2]. The diagnosis of ALS is provisional, based primarily on clinical observations of upper and lower motor neuron damage in the absence of other causes [3]. Because the clinicopathologic markers of ALS are poorly defined, patients are often misdiagnosed (up to 45% of the time) or delayed for up to 12 months [4]. One unfortunate consequence of this delay is that by the time of diagnosis, a patient's motor neurons may have been affected. The diagnosis and treatment of ALS will be significantly strengthened when objective, sensitive markers for the disease can be identified [5].

ALS causes dysarthric speech [6]. Speech production decline is among the earliest indicators of bulbar motor involvement due to ALS [7, 8]. Most of the currently used clinical measures are subjective. ALS Functional Rating Scale- Revised (ALSFRS-R) is currently used for monitoring the progression of disability in patients with ALS, which includes self-reported questions on speech, swallowing, feeding and other body motion measures [9]. Speech intelligibility (percentage of words that are understood by listeners who are not familiar with the patients) and speaking rate (number of spoken words per minute, W/M) are other commonly used clinical measures for speech performance [10].

Recent studies have tried to detect bulbar ALS through comprehensive physiological measures, including articulatory, phonatory, respiratory, and laryngeal sub-systems [11, 12]. Although the results are promising, the logistical difficulty of data collection (particularly tongue motion data) prevents these approaches from being practically used. In contrast, speech signals can be collected conveniently in a clinical environment or at home (e.g., through a smart phone). Therefore, speech signals may be a promising source of information for the automatic detection of ALS in practical applications.

The feasibility of the automatic detection of neurological diseases from speech signals have been recently demonstrated, for example, for depression [13, 14], traumatic brain injury [15], and Parkinson's disease detection or severity estimation [16, 17]. Acoustic feature analysis (e.g., formant centralization ratio [18], vowel space area [18], intonation and prosody [19]) showed promising results in the detection of neurological diseases. Hahm and colleagues used quasi-articulatory features that were inversely mapped from acoustic data and showed an improvement for Parkinson's condition estimation [20]. As a motor neuron disease, ALS affects the articulatory patterns, including tongue and lip motion [8]. Thus, articulatory motion information may also provide complementary information that would benefit ALS detection.

To our knowledge, this project is the first that aimed to automatic detection for ALS from speech acoustic and articulatory samples. Speech samples are pseudo words or short phrases that are spoken in daily life (e.g., *how are you?*). Two commonly used machine learning classifiers, support vector machine (SVM) and deep neural network (DNN), were used. Leave-subjects-out cross validation strategy was used in the experimental design, where training data and testing data were from unique talkers.

2. Data Collection

2.1. Participants

Eleven patients with ALS and eleven healthy talkers participated in the data collection. All participants were asked to

http://dx.doi.org/10.21437/Interspeech.2016-1542

Table 1: Patient information.

Subject ID	Gender	Age	Speech Intelligibility (%)	Speaking Rate (W/M)
A01	M	50	100	237
A02	F	62	89	78
A03	F	40	99	135
A04	M	74	97	85
A05	M	69	100	200
A06	M	77	96	177
A07	F	54	96	137
A08	F	72	80	148
A09	M	62	98	172
A10	F	63	94	105
A11	M	44	90	109
Average		60	95	144
SD		12.3	3.9	52.3

repeat a list of sentences (e.g., *how are you doing?*) or eight isolated vowels in /bVb/ form (i.e., /bab/, /bib/, /beb/, /b@b/, /bˆb/, /bcb/, /bob/, /bub/) multiple times. The acoustic output was recorded synchronously.

The speech intelligibility and speaking rate of these patients were evaluated by a certified speech-language pathologist using the Sentence Intelligibility Test (SIT) software [21]. The speech intelligibility scores in Table 1 show that patients were pre-symptom and had normal speech.

2.2. Setup and Procedure

Two electromagnetic articulographs (NDI Wave and Carstens EMA AG500) were used for collecting speech acoustic and articulatory movement data. The two articulographs are based on the same electromagnetic technology by tracking small wired sensors that are attached to the subject's tongue, lips, and head [22]. Thus, we just described the procedure of using Wave in this paper. The two devices have a similar tracking accuracy (0.5 mm) [23, 24].

Four sensors were attached to tongue and lips. The sensors were tongue tip (TT, 5-10 mm to tongue apex), tongue back (TB, 20-30mm back from TT), upper lip (UL, vermilion border of the upper lip at midline), and lower lip (LL, vermilion border of the lower lip at midline). Previous studies indicated that the four-sensor set is optimal for this application (e.g., [25]). Hence, data from these sensors were used in analysis. Meanwhile, an additional sensor was attached to the middle point of the forehead. The head sensor data were used for calculating head-independent data of other sensors. The positions of the five sensors attached to a participant's head, tongue and lips were illustrated in Figure 1.

Invalid samples were rare and excluded from the analysis. A total of 2,567 valid samples were collected with each sample containing both acoustic and articulatory information. 1,832 of the samples were from healthy speakers; 735 samples were from patients with ALS. ALS patients produced less number of samples than healthy talkers, because some patients with ALS did not complete the whole task.

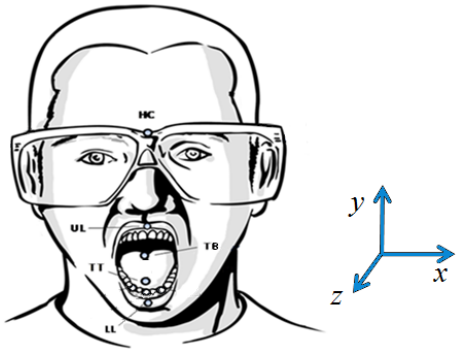

Figure 1: *Sensor locations in data collection. Labels are described in text.*

3. Method

The major design of ALS detection involved two major steps: feature preparation and classification. Feature preparation was to extract and select a set of content-independent acoustic and articulatory features and speaker characteristics from speech and articulatory samples. Classification was to distinguish if a sample is from a healthy or ALS speaker.

3.1. Feature extraction

The script provided in [26] was modified (70 ms window size and 35 ms frame shift) and used to extract acoustic and articulatory features from acoustic and articulatory motion data respectively. The script extracted 6,373 pre-defined acoustic features including these with jitter, shimmer, and MFCC. However, low frequency articulatory movement data do not contain all these features. Thus, we disabled the following feature groups when using the tool to extract articulatory features:

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

In each feature group, individual features were calculated, for example, mean, flatness, posamean, rqmean, range, maxPos, minPos, centroid, stddev, skewness, kurtosis, etc. Please refer to [26] for a detailed description 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 to test our ALS detection approaches. In addition, the articulatory features from tongue and lips were used individually to advance the understanding of individual contribution of each articulator to distinguishing ALS from healthy speakers.

3.2. Feature selection using randomized logistic regression

Feature selection is crucial for high dimensional classification tasks, like that in this project. We used Randomized Logistic Regression (RLR) as the feature selection procedure [27]. RLR is a stability selection technique where any specified selection algorithm can be applied along with subsampling [27]. RLR uses logistic regression as the selection algorithm. Logistic regression classifier assumes a parametric form for the distribution $P(Y|X)$ and the model can be defined as:

$$P(Y=1|X) = \frac{1}{1+\exp(w_o + \sum_{i=1}^{n} w_i X_i)} \quad (1)$$

where Y is the boolean class vector, $X =< X_1, X_2, ...X_n >$ is any discrete or continuous valued data vector and $W =< w_0, w_1, ...w_n >$ is the weight vector to be learned from the training data. The implementation of RLR in [28] was used in this project.

Examples of the selected acoustic features are the first quartile of MFCC, the arithmetic means of the F0 contour, and the first quartile of the F0 contour. The selected articulatory features included the first quartile of FFT magnitude and the minimum range relative to the segment length of FFT magnitude.

3.3. Speaker normalization using i-Vectors

I-vectors are a widely used technique for speaker verification, which maps speaker-related information to low-dimensional fixed-length vectors [29]. In this paper, i-vectors were used to reduce the undesired speaker variability that may lower the performance.

I-vector algorithm assumes a linear dependence between the speaker-adapted information and the speaker-independent information, which can be modeled as equation:

$$s = m + Tw \tag{2}$$

where m is the mean supervector of Gaussian mixture model (GMM) representing universal background model (UBM). T is also referred to as the i-vector extractor that is a low-rank matrix representing subspace containing important variability in the mean supervector space [29], and w is a standard normal distributed vector. The UBM used in this stage was represented as a diagonal covariance Gaussian mixture model (GMM), which was trained using EM algorithm based on data from other speakers [29].

The concatenation of i-vectors and the selected RLR features was used for classification.

3.4. Support Vector Machine

Support Vector Machine (SVM) is a classification technique that produces a separating line with the maximum margin from the nearest data points belonging to either class [30]. SVM solves a quadratic optimization problem that maximizes the distance between the separating hyperplanes passing through points belonging to each class and the data points on the boundary and, in the meanwhile, satisfying the class membership requirement of the points [31]. A kernel function is used to describe the distance between two samples (i.e., r and s in Equation 3). 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) [22]:

$$K_{RBF}(r, s) = exp(1 - \gamma||r - s||). \tag{3}$$

Please refer to [31] for more details about the implementation of the SVM. All feature values were normalized by groups (ALS and Healthy) using z-score before they were fed into SVM.

3.5. Deep Neural Network

Deep neural networks (DNN) have recently been used in pattern recognition and speech recognition successfully [32]. DNN is a multiple-layer neural network with each layer having multiple nodes connected to the nodes of the next layer, from input layer to output layers. The DNN training approach based on restricted Boltzmann machines (RBMs), which are subsequently fine-tuned using backpropagation algorithm. The weights for nodes in hidden layers at iteration $(t + 1)$ are updated based on iteration (t) using stochastic gradient descent using the following equation:

$$w_{ij}(t + 1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{4}$$

where w_{ij} is the weight between nodes i and j in neighboring layers, η is the learning rate, and C is the cost function.

A detailed explanation and further discussion of the DNN can be found in [32, 33]. Considering the trade-off between performance and time cost based on our preliminary analysis, we specified the DNN with four layers with each layer having 256 nodes. As required by the DNN implementation [34], all feature values were normalized between 0 and 1 by groups (ALS and healthy).

3.6. Experimental Design

To understand the performance using acoustic signals only and if adding articulatory information can benefit the classification, three sub-configurations were used in the experiment, where data were separated into three groups, acoustic, acoustic + lip data, acoustic + lip data + tongue data.

Leave-one-subject-pair-out cross validation was used to test the performance of SVM. In each execution, data samples from one ALS patient and one healthy speaker were used for testing, and the rest for training. DNN takes longer time for training, thus, in this stage, 4-fold cross validation strategy was used. DNN classification requires a separate validation set [34]. Therefore, in each execution using DNN, data samples from three ALS patients and three healthy speakers were used for testing; samples from another three ALS patients and three healthy speakers were used for validation; the rest samples were used for training. The averaged performance of all the execution was considered the overall performance.

Accuracy, sensitivity, and specificity were used as the major performance measures. Accuracy is the number of true positives plus true negatives divided by the number of all testing samples. Sensitivity is the number of true positives divided by the sum of numbers of true positives and false negatives, which means the probability that a patient is classified as positive who actually has the disease. Specificity is the number of true negatives divided by the number of true negatives and false positives, which means the probability that a subject is classified as negative who is healthy.

4. Results and Discussion

Figure 2 (left part) gives the results using SVM. When only acoustic data were used, the overall accuracy, specificity and sensitivity were all above guess level (50%). Adding lip data (from both upper lip and lower lip) significantly increased the accuracy, specificity, and sensitivity. In addition, adding both lip data and tongue data (from both tongue tip and tongue body back) further increased the accuracy to 80.91%, the specificity to 80.51%, and sensitivity to 81.90%. Table 2 gives the summed classification matrix of the cross validations using SVM with all acoustic, lip, and tongue features.

Figure 2 (right part) gives the results using DNN. A promising result was obtained even when only acoustic data were used. Adding lip data (from both upper lip and lower lip) increased the accuracy, specificity, and sensitivity. In addition, adding both lip data and tongue data (from both tongue tip and tongue

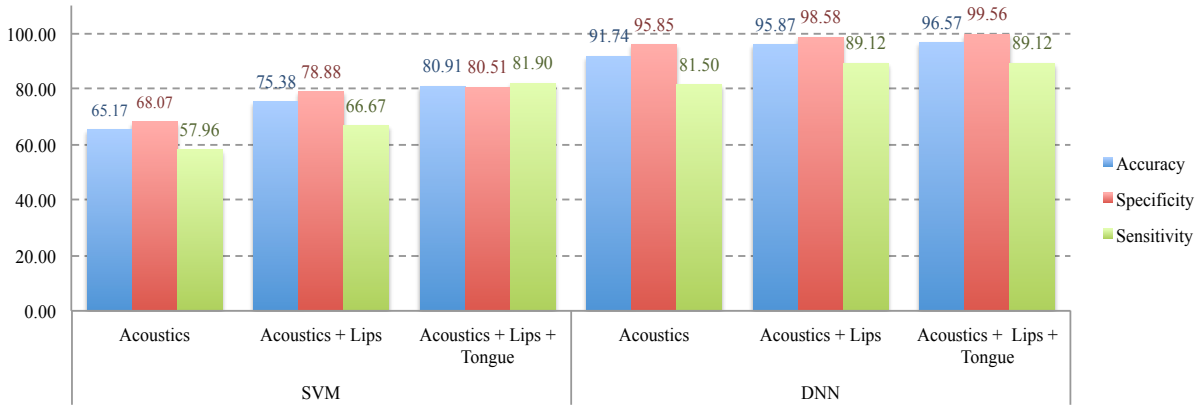

Figure 2: *Performances of ALS detection from combined types of data (acoustic, lip, and tongue data) using SVM or DNN, respectively.*

Table 2: Classification matrix using SVM with acoustic, lip, and tongue features. The overall accuracy is 80.91%.

	Healthy	**ALS**	
Healthy	1475	357	80.51% (Specificity)
ALS	133	602	81.90% (Sensitivity)

Table 3: Classification matrix using DNN with acoustic, lip, and tongue features. The overall accuracy is 96.57%.

	Healthy	**ALS**	
Healthy	1824	8	99.56% (Specificity)
ALS	80	655	89.12% (Sensitivity)

body back) increased the accuracy to 96.57%, the specificity to 99.56%, and sensitivity to 89.12%, which were the best performance in this experiment. Table 3 gives the classification matrix using DNN with all data (acoustic + lip + tongue data).

The experimental results demonstrated the feasibility of automatic detection of ALS from short speech samples. In addition, adding articulatory information from lips significantly improved the performance. Adding tongue data can further improve the overall accuracy. The sensitivity for both SVM and DNN was improved as lip data were added. When tongue data was added on top of acoustic and lip data, SVM obtained an even higher sensitivity.

Using articulatory movement data in practice has a logistical obstacle, cause articulatory movement data are relatively difficult to collect [22] (compared with acoustic data). However, lip and tongue data can be converted using inverse (acoustic-to-articulatory) mapping [20].

DNN outperformed SVM in all configurations, as shown in Figure 2. This is consistent with our recent work on Parkinson's condition estimation from speech samples [20]. Although DNN training cost is about 3-8 minutes for one cross validation (on the selected features), the testing time for each sample is comparable with SVM (in milliseconds for one sample). The experiment was executed on a PC with Intel i7 CPU 2.4GHz with 8 GB RAM running Ubuntu Linux.

Analysis of sensitivity and specificity (e.g., receiver operating characteristic, ROC), would help to tell the tradeoff between Type-I and Type-II errors with different parameters in the classifiers. Next step of this work would include a ROC analysis.

We think the performance could be even better if better selected data stimuli were used (e.g., using samples of only one short phrase). We used all samples in the data set to keep the maximum number of samples. The data set contained a rich set of stimuli (20 short phrases, and eight CVCs). Although we extracted acoustic and articulatory features and speaker characteristics from these samples, there might be still a level of content variation in the data. Most of work for detection of neurological disease from speech in literature used a single or only a few stimuli (e.g., sustained vowels [17] or a short syllable [35]).

Limitations.The healthy talkers and the ALS patients are not exactly age- or gender-matched (some are younger than these patients). This unbalanced subject group may cause some unexpected bias in terms of group distinctiveness. As we actively collect data from more subjects, a larger data set with age-matched subjects will be used to verify if the performance level in the paper can be generalized to a larger population.

5. Conclusions and Future Work

This paper demonstrated the feasibility of the automatic detection of ALS from pre-symptom, intelligible speech samples. Experiments using a data set collected from eleven patients with ALS and eleven healthy talkers showed promising results. The experiments also demonstrated that adding articulatory information could improve the detection performance. Particularly, even adding lip information on top of acoustic data could significantly improve the performance. In the future, a larger data set will be used to verify our approach for ALS detection from speech samples. The data set will include a balanced, age- and gender-matched ALS patients and healthy talkers.

6. Acknowledgements

This work was supported by the National Institutes of Health through grants R03 DC013990 and R01 DC013547, and the American Speech-Language-Hearing Foundation through a New Century Scholar grant. We would like to thank Dr. Jordan R. Green, Dr. Thomas F. Campbell, Dr. Yana Yunusova, Dr. Seongjun Hahm, Dr. Myungjong Kim, Dr. Panying Rong, Dr. Anusha Thomas, Jennifer McGlothlin, Denisse Amaranta Soler Morales, Jana Mueller, Victoria Juarez, Saara Raja, Soujanya Koduri, Kumail Haider 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] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, pp. 293–299, 2000.

[4] Y. Iwasaki, K. Ikeda, and M. Kinoshita, "The diagnostic pathway in amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 2, pp. 123–126, 2001.

[5] M. Cudkowic, M. Qureshi, and J. Shefner, "Measures and markers in amyotrophic lateral sclerosis," *Journal of Pharmacology and Experimental Therapeutics*, vol. 1, pp. 273–283, 2004.

[6] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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] J. V. squez Correa, J. Orozco-Arroyave, J. Arias-Londono, J. F. Vargas-Bonilla, and E. N. th, "New computer aided device for real time analysis of speech of people with Parkinsons disease," *Fac. Ing. Univ. Antioquia*, vol. 72, pp. 87–103, 2014.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for ag500, electromagnetic articulograph," *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 547–555, 2009.

[25] 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.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Procj. of the ACM International Conference on Multimedia*, 2013, pp. 835–838.

[27] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] P. Matjka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. ernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4828–4831.

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

[31] 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.

[32] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[33] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[34] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010.

[35] 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.