INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic assessments of dysarthric speech: the usability of acoustic-phonetic features

Loes van Bemmel[1], *Chiara Pesenti*[2], *Xue Wei*[3], *Helmer Strik*[1,3,4]

[1]Centre for Language and Speech Technology, Radboud University, the Netherlands
[2]Department of Humanities, Università degli studi di Torino, Italy
[3]Centre for Language Studies, Radboud University, the Netherlands
[4]Donders Institute for Brain, Cognition and Behaviour, the Netherlands

{loes.vanbemmel, xue.wei, helmer.strik}@ru.nl, chiara.pesenti@edu.unito.it

Abstract

Individuals with dysarthria suffer from difficulties in speech production and consequent reductions in speech intelligibility, which is an important concept for diagnosing and assessing effectiveness of speech therapy. In the current study, we investigate which acoustic-phonetic features are most relevant and important in automatically assessing intelligibility and in classifying speech as healthy or dysarthric. After feature selection, we applied a stepwise linear regression to predict intelligibility ratings and a Linear Discriminant Analysis to classify healthy and dysarthric speech. We observed a very strong correlation between actual and predicted intelligibility ratings in the regression analysis. We also observed a high classification accuracy of 98.06% by using 17 features and a comparable, high accuracy of 96.11% with only two features. These results indicate the usefulness of the acoustic-phonetic features in automatic assessments of dysarthric speech.

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

1. Introduction

Dysarthria comprises a set of motor speech disorders caused by a neurological injury such as Parkinson's disease or stroke. Individuals with dysarthria experience difficulties in speech production which can lead to a reduction in speech intelligibility. As a consequence, they may gradually lose contact with friends and family and eventually become isolated from social life and society. Such outcomes can severely affect their quality of life. In order to alleviate such repercussions and improve speech intelligibility, speech therapy has been found to be effective.

Speech intelligibility is an important concept in speech-language pathology that has been used to diagnose and assess the effectiveness of speech therapy. In the clinical practice of speech therapy, a common definition of intelligibility proposed by Hustad is that "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" [1]. In line with this definition, intelligibility has been measured through various methods, such as by having human listeners transcribe what they hear [2] [3] or by ratings on a visual analogue scale (VAS) [4] [5]. In fact, ratings through VAS have been shown to be a reliable and valid intelligibility measure [4] [5] and have been widely used in research and clinical practice.

However, such procedures that rely on human listeners are quite time-consuming and labor-intensive. Although it may be feasible to apply these rating procedures in research, easy-to-use tools are highly desirable in clinical practice. For these reasons, it is necessary to explore procedures that employ automation to assess speech intelligibility and to identify dysarthric speakers. Many researchers have employed Automatic Speech Recognition (ASR) or more sophisticated machine learning (ML) algorithms to obtain embeddings of dysarthric speech [6] [7]. However, it is hard to interpret how these ML-based embeddings are related to speech intelligibility and to properties of dysarthric speech that can be addressed in speech therapy.

Procedures using acoustic-phonetic analysis can provide detailed diagnostic information about deviations in dysarthric speech and have been shown to be promising for assessing dysarthric speech [8] [9]. Based on automatic identifications of the most relevant features of dysarthric speech, it is possible to establish reliable, accurate and non-invasive assessment tools to assess the intelligibility and to distinguish dysarthric speech from healthy speech, as well as diagnose the type and extent of speech disorders. In addition, previous studies [8] [9] [10] have shown that therapeutical treatment through a serious game can lead to improvements in speech intelligibility and increases in loudness and intensity although such improvements and increases were speaker-dependent. Anyhow, these studies further indicate the possibility of using acoustic-phonetic features to assess speech intelligibility and to classify dysarthric speakers from healthy speakers.

However, so far, it is still unclear what acoustic-phonetic features are the most relevant and important in the assessments of intelligibility and classification of speech as healthy or dysarthric. Thus, in the current paper, we analyze a set of acoustic-phonetic features and focus on addressing two research questions:
(1) To what extent can intelligibility ratings be predicted by objective, acoustic-phonetic features?
(2) To what extent can acoustic correlates of intelligibility classify healthy and dysarthric speech?

Specifically, we start with a feature selection and a stepwise linear regression analysis to identify the most relevant features for predicting intelligibility ratings. Then, we explore different combinations of the features selected by the regression analysis to identify the most relevant features for classifying healthy and dysarthric speech. Based on the analyses, we intend to provide insights into automatic assessments of dysarthric speech.

2. Materials and methods

2.1. Speakers and speech materials

To cover various types of speech materials, we used two three-word lists, two semantically unpredictable sentences (SUS), and four meaningful sentences selected from the commonly-used, phonetically-balanced text "Papa and Marloes" [11] in Dutch. The speakers involved in this study were those who participated in a project which was aimed at developing a serious game for conducting research on speech disorder treatment through ASR-

10.21437/Interspeech.2023-2017

based technology[1] [10]. All speakers were native speakers of Dutch, with five being healthy (4 male and 1 female) and thirteen having dysarthria (10 male and 3 female). The healthy speakers were aged between 61 and 69 (M = 65.0, SD = 3.4). The dysarthric speakers were aged between 53 and 75 (M = 64.2, SD = 6.4), ten of them had Parkinson's and three had had a Cerebral Vascular Accident. All eighteen speakers provided recordings of the selected speech materials without experiencing the game. In addition, in order to have a larger dataset, we also included the recordings that were collected from eight of the dysarthric speakers (5 male and 3 female) immediately after the game. In total, there are 208 utterance items (from this point called 'items'). As we did not collect the data, we do not have permission to publicly release it.

2.2. Intelligibility ratings and acoustic-phonetic features

For each of the 208 items, eleven intelligibility rating scores were collected from eleven speech therapists (all female) as listeners through VAS ranging from 0 (not intelligible) to 100 (intelligible) as described in [12]. The inter-listener reliability for the VAS scores through the Intraclass Correlation Coefficients was >0.9, which signifies an excellent inter-listener reliability [12]. We calculated the mean values of the VAS scores for each item before conducting the statistical analyses in Section 2.3. Moreover, for each item, we extracted a total of 103 acoustic-phonetic features including 88 eGeMAPS features [13] [14] and 15 Praat features related to duration, formant frequency, pitch, intensity and gravity center [15].

2.3. Statistical analyses

To avoid an overfitting problem, we first applied a feature selection by employing a LASSO analysis, with an alpha of 1 and 3 folds. Based on the selected features of LASSO, we further applied regression and classification analyses to answer our two research questions, respectively.

Specifically, to address the first research question about predicting intelligibility ratings, we applied a both-direction stepwise linear regression with VAS scores as the dependent variable in a generalized linear model (GLM). The Akaike Information Criterion was used in the stepwise approach to decide which acoustic-phonetic features to include in the final GLM model. Both the LASSO analysis and the stepwise linear regression approach were implemented by using the *GLMnet* package [16] in R (version 4.2.2).

To address the second research question about classifying healthy and dysarthric speech, we used the features selected by the final GLM model and applied a Linear Discriminant Analysis (LDA) classification. In detail, based on the significance of the features selected by the final GLM model, we first applied LDA on multiple subsets of these features. That is, we started with the most significant feature selected by the final GLM model and then gradually add another feature that was the most significant in the rest of the features according to their significance levels reported by the final GLM model. In addition, we exhaustively applied LDA on each pair of features selected by the final GLM model. These analyses aimed to study which combination of features performs the best in classifying healthy and dysarthric speech. Note that for classification, we only involved the items for the five healthy speakers and for the eight dysarthric speakers to have a rather balanced dataset, leading to a total of 104 items. The LDA analyses were implemented by using the standard parameters of the LDA in the *scikit-learn* package in python [17] with a Leave-One-Speaker-Out (LOSO) cross-validation scheme [18].

[1] According to the Ethics Committee at Radboud University, Ethics approval was not required.

	Feature	Signf.	Sign
1	mean of F2 bandwidth	***	-
2	mean of harmonic difference H1-A3	***	+
3	50th percentile of pitch in semitone ($F0 > 27.5Hz$)	***	+
4	mean of MFCC 4 over voiced regions	***	-
5	mean of harmonic difference H1-H2	***	-
6	mean of spectral slope 500-1500 Hz over voiced regions	***	+
7	coefficient of variation of MFCC 3 over voiced regions	***	+
8	coefficient over variation of spectral flux	***	+
9	maximal intensity	***	+
10	coefficient of variation of HNR	***	+
11	coefficient of variation of F2	***	+
12	coefficient of variation of the F3 bandwidth	***	-
13	center of gravity	**	+
14	standard deviation of the loudness slope of rising signal parts	**	-
15	the number of loudness peaks per second	**	+
16	coefficient of variation of pitch in semitone ($F0 > 27.5Hz$)	**	+
17	pitch variability	*	-
18	standard deviation of intensity	*	+
19	coefficient of variation of harmonic difference H1-H2	*	+
20	mean of F3 bandwidth	*	+
21	mean of MFCC 2	*	-
22	coefficient of variation of MFCC 1		-
23	standard deviation of pitch in semitone ($F0 > 27.5Hz$)		-
24	coefficient of variation of MFCC 4 over voiced regions		-

Table 1: *The features selected by the final GLM model in the stepwise analysis in order of their significance. 0 '***' 0.001 '**' 0.01 '*' >0.01 ' '., where the top feature has the smallest p value. The sign column indicates whether the GLM coefficient is positive (+) or negative (-).*

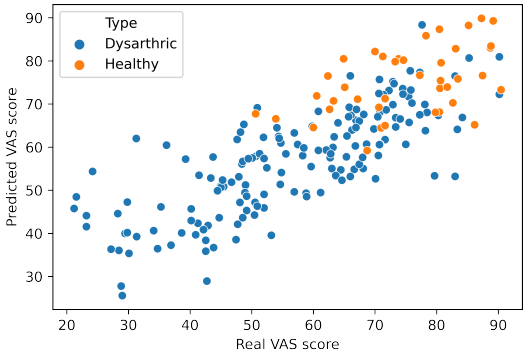

Figure 1: *The scattergram for the actual and predicted VAS scores per item by the final GLM model. The correlation is 0.80 ($p < .0001$).*

3. Results

3.1. Regression for predicting intelligibility ratings

The LASSO analysis selected 53 out of 103 acoustic-phonetic features. Based on these 53 features, the stepwise GLM analysis further selected and reported 24 features in the final model. These 24 features were ranked according to their significance levels in the final GLM model and are shown in Table 1 together with their signs of the GLM coefficient. The final GLM model had a R^2 of 0.65 and an RMSE of 9.89. The correlation between the actual VAS scores (M=60.38, SD=16.65) and

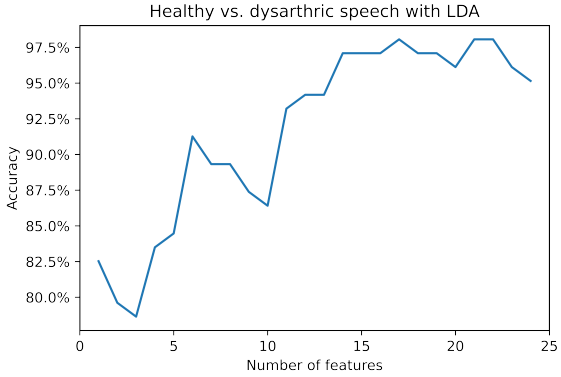

Figure 2: *Accuracy results for classifying healthy and dysarthric speech with different subsets of features based on their significance in Table 1.*

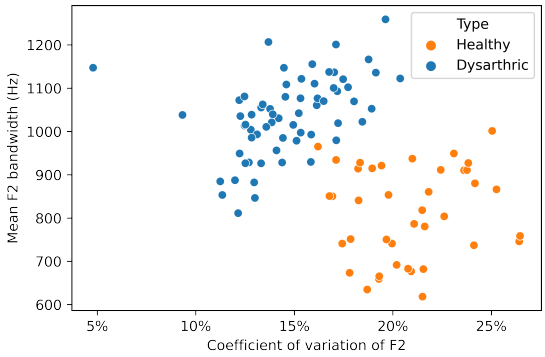

Figure 3: *Scattergram of coefficient of variation of F2 (%) vs. mean of F2 bandwidth (Hz) for healthy (orange) and dysarthric (blue) speech.*

the predicted scores was 0.80 ($p < .0001$). Figure 1 presents a scattergram between the actual and predicted scores.

3.2. Classification for distinguishing healthy and dysarhtic speech

As mentioned in Section 2.3, we applied LDA on 24 subsets of the 24 features to classify healthy and dysarthric speech. In detail, we first applied LDA on the feature subset of top 1 (i.e., the first feature in Table 1). Then, we applied LDA on the subset of top 2 (i.e., the first and second features in Table 1) and repeated this procedure until all 24 features selected by the final GLM model were involved (i.e., the subset of top 24). This led to accuracy results of classification with the number of features ranging from 1 to 24, as shown in Figure 2. We observed the highest accuracy value of 98.06% when using the subsets of top 17, top 21, and top 22. We also examined the accuracy results per item to study whether speech materials (i.e., word lists, SUS, and meaningful sentences) play a role in the classification. We did not observe any high error values for specific speech materials nor for specific items.

In addition to applying LDA on subsets of features, we exhaustively applied LDA on each pair of 24 features selected by the final GLM model. We observed that with the mean of the second formant frequency (F2) bandwidth (i.e., F2frequency_sma3nz_stddevNorm in eGeMAPS feature set script) and the coefficient of variation of F2 (i.e., F2bandwidth_sma3nz_amean), we can already reach a comparable, high accuracy value of 96.11%. The scattergram for these two features is presented in Figure 3 with different colours for healthy and dysarthric speech.

4. Discussion

In this study, we explored the usefulness of acoustic-phonetic features in automatic assessments of dysarthric speech. In particular, we examined to what extent the acoustic-phonetic features can be used to predict intelligibility ratings and to classify healthy and dysarthric speech. Briefly, to a certain extent, the acoustic-phonetic features can be used to predict intelligibility ratings, and to a great extent, the acoustic-phonetic features can be used to classify healthy and dysarthric speech. Below we discuss our results in more detail.

4.1. Regression for predicting intelligibility ratings

Regarding the regression analysis between acoustic-phonetic features and the intelligibility ratings, it can be clearly seen in Table 1 that most of the selected features showed to be significant in explaining the variance in the intelligibility ratings. On the other hand, it is difficult to interpret how every single feature contributes to predicting intelligibility ratings because they are selected through a stepwise approach and these features may work in an interactive manner in the prediction.

In addition, the 0.65 proportion of the variance in intelligibility ratings can be explained by the 24 features in Table1 which were selected by a stepwise linear regression analysis. Also, the correlation between the actual and predicted VAS scores in the final regression model is very strong [19] as shown in Figure 1. These findings seem to be in line with that by Xue et al. [15]. The authors reported a moderate correlation between intelligibility ratings obtained through VAS and an acoustic-probability index calculated based on 15 acoustic-phonetic features, which were related to pitch, intensity, and formant frequency, extracted from both healthy and dysarthric speech. Also, many researchers examined intelligibility measures [10] and acoustic features at different granularity levels [9] [8] before and after speech therapy and found significant differences although the differences were speaker-dependent. These findings together with the results of this study further indicate the usefulness of acoustic-phonetic features in automatic assessments of dysarthric speech.

On the other hand, although the regression model led to a very strong correlation between the actual and predicted VAS scores in Figure 1, some spots presenting dysarthric speech at the bottom-left part in Figure 1 were not well predicted and had higher predicted scores compared to the actual VAS. A possible explanation is that the amount of items with relatively low VAS scores is limited, thus leading to worse predictions. We also observed several spots presenting healthy speech had lower predicted scores. This may be due to the overlap of VAS scores between healthy and dysarthric speech. However, having overlaps of intelligibility measures between healthy and dysarthric speech is inevitable due to the nature of how intelligibility is measured involving human listeners. Thus, these findings should be further elaborated in a larger dataset.

4.2. Classification for distinguishing healthy and dysarthric speech

Regarding the classification of speech as healthy or dysarthric, the usefulness of the acoustic-phonetic features is supported by the following findings. First, the mean of F2 bandwidth, which was the most significant feature selected through the stepwise regression analysis, resulted in a high classification accuracy value of 82.50%. Second, when having this feature and the coefficient of variation of F2 as independent variables, the accuracy increased to 96.11%. Also, Figure 3 shows a clear distinction between healthy and dysarthric speech. Third, although the classification accuracy decreased when having the three most significant features, it remains above 78% as shown in Figure 2. Fourth, the accuracy values gradually increased when involving more features as independent variables. These results suggest that stepwise regression analysis can be used as a method of feature selection, and the selected features can result in good classification results.

Furthermore, it is expected that the accuracy increases when the number of features in classification increase because more features can help classify healthy and dysarthric speech. However, having more features may increase the possibility of having an overfitting issue. Therefore, as explained in Section 2.3, we applied both LASSO and stepwise regression analyses for feature selection. As the consequence, it is much less likely to have the overfitting issue since the number of features has been reduced from 103 to 24 and is much smaller compared to the number of items (i.e., 104). Also, we applied LOSO cross-validation and used a simple linear model for classification.

Note that although the classification accuracy gradually increased when involving more features, we observed decreases in the curve in Figure 3. It might be that some features selected by the stepwise regression analysis with high significance may not contribute to better classification results. Also, although we exhaustively examined each pair of the 24 features for classification and found the most successful pair of features (i.e., the mean of F2 bandwidth and the coefficient of variation of F2), this exhaustive-examination approach may not be feasible when using larger datasets. On the other hand, simply involving more features can lead to very high accuracy values that were comparable to or higher than that of the most successful pair of features, and thus, it is more feasible than exhaustively examining all pairs of features, especially on larger datasets.

Moreover, one of the benefits of using acoustic-phonetic features instead of embeddings based on sophisticated ML algorithms is that we can better interpret the most important features. Also, these features are more useful in helping speech-language therapists for diagnosis. For example, we found the most successful pair of features (i.e., the mean of F2 bandwidth and the coefficient of variation of F2) in distinguishing healthy and dysarthric speakers as shown in Figure 3. Taking together the signs of the coefficients in the GLM model presented in Table 1, it seems that F2 frequencies of healthy speech are rather spread according to a higher standard deviation (sign = '+'), and the F2 bandwidth is smaller according to a lower mean (sign = '-') compared to dysarthric speech. Since F2 contributes to distinguishing different vowels [20], these results seem to indicate that healthy speakers can better control their vowel articulation as they can place their F2 widely (high standard deviation) and precisely (small bandwidth).

4.3. Limitations

One limitation of this study may be that we exhaustively examined every pair of features for classification to find the most relevant ones. This may not be feasible when larger datasets or more features are involved. Also, some features, such as features about Mel-Frequency Cepstral Coefficients (MFCC) as reported in Table 1, are more difficult to interpret than F2 frequency. Another limitation is that the speech examined in this study is read speech, whereas other speech types, such as spontaneous speech, may have different acoustic characteristics, leading to different results. Also, we used the data that was collected by others. Due to the General Data Protection Regulation, we cannot make the data public. Future research may elaborate our findings on a public dataset.

4.4. Future research

Future research may further elaborate our findings on a larger dataset that has more speakers. Using a larger dataset allows for more sophisticated statistical analyses and reduces the risk of having overfitting problems. It may also be interesting to repeat our analyses for different measures of intelligibility obtained at different levels of granularity or for a different language. Future research could also extend the exploration of this study to features based on different syntactic-semantic structures. They may lead to different results for regression or classification examinations.

5. Conclusions

This study explored the usefulness of acoustic-phonetic features for predicting intelligibility ratings measured through VAS and for classifying healthy and dysarthric speech. By using LASSO and the stepwise linear regression analyses, we succeeded in reducing the number of features from 103 to 24. By using these 24 features, an R-squared score of 0.65 was found. These 24 features also contributed to the classification of speech as healthy or dysarthric by the LDA model, with the highest accuracy of 98.06% using the first 17 features based on the significance reported by the stepwise linear regression model. By exhaustively examining every pair of the 24 features, we found the most successful pair of features, i.e., the mean of F2 bandwidth (i.e., F2bandwidth_sma3nz_amean) and the coefficient of variation of F2 (i.e., F2frequency_sma3nz_stddevNorm). These two features alone were able to obtain a similar, high accuracy value of 96.11%. These results seem to indicate that it is possible to use acoustic-phonetic features for automatic assessments of dysarthric speech.

6. Acknowledgements

We thank Fleur Boogmans for collecting the VAS scores and Mario Ganzeboom for collecting the recordings.

7. References

[1] K. C. Hustad, "The relationship between listener comprehension and intelligibility scores for speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 562–573, 2008.

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

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

[4] D. Abur, N. M. Enos, and C. E. Stepp, "Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in parkinson's disease with variable listener exposure," *American Journal of Speech-Language Pathology*, vol. 28, no. 3, pp. 1222–1232, 2019.

[5] K. L. Stipancic, K. Tjaden, and G. Wilding, "Comparison of intelligibility measures for adults with parkinson's disease, adults with multiple sclerosis, and healthy controls," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 2, pp. 230–238, 2016.

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

[7] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 2825–2828.

[8] C. Pesenti, L. Van Bemmel, R. van Hout, and H. Strik, "The effect of ehealth training on dysarthric speech," in *Proceedings of the RaPID Workshop-Resources and ProcessIng of Linguistic, Para-linguistic and Extra-linguistic Data from People With Various Forms of Cognitive/Psychiatric/Developmental Impairments-Within the 13th Language Resources and Evaluation Conference*, 2022, pp. 1–8.

[9] L. van Bemmel, C. Cucchiarini, and H. Strik, "Using feature selection to evaluate pathological speech after training with a serious game," *International Conference of Experimental Linguistics 2021*, pp. 245–248, 2021.

[10] M. Ganzeboom, M. Bakker, L. Beijer, H. Strik, and T. Rietveld, "A serious game for speech training in dysarthric speakers with parkinson's disease: Exploring therapeutic efficacy and patient satisfaction," *International Journal of Language & Communication Disorders*, vol. 57, no. 4, pp. 808–821, 2022.

[11] J. Van DeWeijer and I. Slis, "Nasaliteitsmeting met de nasometer [measuring nasality using the nasometer]," *Logopedie Foniatrie*, vol. 63, p. 97–101, 1991.

[12] F. Boogmans, "Beoordeling spraakverstaanbaarheid van dysartrische en referentie spraak middels een luisterexperiment [assessment of intelligiblity of dysarthric and reference speech by means of a listening experiment]," *Master thesis*, 2020.

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[14] W. Xue, C. Cucchiarini, R. van Hout, and H. Strik, "Acoustic correlates of speech intelligibility: The usability of the egemaps feature set for atypical speech," in *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*. [Sl: sn], 2019, pp. 48–52.

[15] W. Xue, R. van Hout, F. Boogmans, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Speech intelligibility of dysarthric speech. human scores and acoustic-phonetic features," *Proceedings of Interspeech 2021*, pp. 2911–2915, 2021.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–20, 2010.

[17] 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.

[18] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[19] J. D. Evans, *Straightforward Statistics for the Behavioural Sciences*. Brooks/Cole Publishing Company, 1996.

[20] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," *Journal of Communication Disorders*, vol. 74, pp. 74–97, 2018.