Using feature selection to evaluate pathological speech after training with a serious game

Loes van Bemmel[1,2], Catia Cucchiarini[3], Helmer Strik[3,4,5]
[1]Department of Artificial Intelligence, Radboud University, The Netherlands
[2]Institute for Computing and Information Sciences, Radboud University, The Netherlands
[3]Centre for Language and Speech Technology, Radboud University, The Netherlands
[4]Centre for Language Studies, Radboud University, The Netherlands
[5]Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands
https://doi.org/10.36505/ExLing-2021/12/0062/000535

Abstract

To evaluate the effectiveness of speech therapy, speech features before and after treatment can be compared, focussing on those features that changed most during treatment. In the current study acoustic features were automatically extracted from speech of patients affected by Parkinson's Disease who had received speech treatment. Praat and openSMILE were used for feature extraction. Through feature selection, the top ten most characterizing features for pre vs. post-treatment were found. Further analysis of these features confirmed that after treatment the speakers spoke louder with lower pitch, which were the goals of the treatment.

Keywords: Parkinson's disease, pathological speech, feature selection

Introduction

Evaluating the effectiveness of speech therapy is a complex issue. First, because there are many measures and procedures to obtain these measures from speech data. Second, because the evaluations are generally based on human ratings, which are time-consuming, error-prone and may contain an element of subjectivity. Objective metrics derived from acoustic measurements would seem to be an interesting alternative, but high quality objective speech measures are difficult to obtain, non-trivial to interpret, and the differences between before and after treatment might be small and non-significant. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2015) is a standardized set of features that were chosen for their demonstrated theoretical relevance, potential to distinguish important aspects of speech production, and ease in automatic computation. In addition, extra features were calculated with a Praat script. In the current study we investigate the usability of these acoustic features to evaluate the effectiveness of speech therapy that was provided through a serious game. We address the following research questions:

1) Can we employ automatic feature selection to find the most characterizing features for further analysis out of a large number of measures?
2) Do these features indicate the effectiveness of the therapy by showing a decrease in pitch and an increase in loudness?

Methodology

Speech data

The speech data were collected from eight native Dutch adults suffering from Parkinson's Disease (PD) who read seven Dutch sentence prompts. The prompts are taken from the story "Papa en Marloes" (Van de Weijer and Slis, 1991) and from apple pie recipes also used by Ganzeboom et al. (2018). There were three recording times: T1 four weeks pre-experiment, T2 pre-treatment, and T3 post-treatment after four weeks of training. The treatment in this experiment is the second version of the serious speech training game "Treasure Hunters" (https://waag.org/project/chasing and Ganzeboom et al, 2018). The main goal of the game is to improve intelligibility of speech through Pitch Limiting Voice Treatment, thus increasing loudness but reducing pitch.

Acoustic features and feature selection

Praat (Boersma and Weenik, 2020) and openSMILE (Eyben et al., 2010) were used to automatically extract 103 features per recording. The 15 Praat features are duration, four formants, pitch variance, gravity center, and the mean, minimum, maximum and standard deviation of pitch and intensity. OpenSMILE was used to extract the 88 eGeMAPS feature set (Eyben et al., 2015). The 103 features were extracted on three different levels: full, word and phoneme. Segmentations on these levels were obtained by a forced aligner. (https://webservices.cls.ru.nl).

Classification with Support Vector Machine and feature selection with Recursive Feature Elimination (RFE; Guyon et al., 2002) were used to determine which of the features changed most between pre and post-treatment, and thus showed the highest scores for pre vs. post-treatment classification. The top ten features were analysed in more detail using classification and statistical analysis Precautions were taken to reduce the risk of overfitting on the small dataset: using a simple model (linear SVM), speaker-based normalization and Leave One Subject Out cross validation (Sakar et al., 2013), and Matthew's Correlation Coefficient (MCC) instead of accuracy.

Results

Using the feature ranking gained by RFE for the pre vs. post-treatment contrast, the MCC scores per time contrast and segmentation level for linear SVM classification led to the results shown in Figure 1. Six out of the top ten T2 vs. T3 features were found to have a significant time effect with $p<0.05$ when tested with a GLM analysis, shown in Table 1.

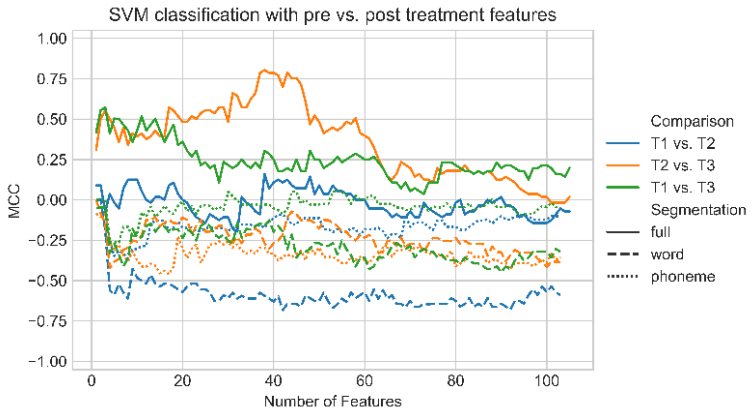

Figure 1. Classification scores of the pre-experiment (T1) vs. pre-treatment (T2) vs. post-treatment (T3) features on three linguistic segmentation levels using SVM classification with a linear kernel and the T2 vs. T3 ranked features.

Table 1. GLM p-values for T2 vs. T3 time effects and mean values for the top ten full-segmentation T2 vs. T3 features.

Rank	Feature	Mean T2	Mean T3	p
1	F3amplitudeLogRelF0_sma3nz_stddevNorm	-0.672	-0.708	**
2	slopeV0-500_sma3nz_amean	0.028	0.030	**
3	F0semitoneFrom27.5Hz_sma3nz_percentile50	29.503	29.140	**
4	HNRdBACF_sma3nz_amean	7.011	6.980	
5	mfcc1_sma3_amean	15.799	17.089	**
6	hammarbergIndexUV_sma3nz_amean	8.822	9.041	
7	mfcc2_sma3_stddevNorm	2.006	-2.635	
8	mfcc2_sma3_amean	4.645	3.989	**
9	mfcc3_sma3_stddevNorm	1.177	1.155	
10	mfcc3V_sma3nz_stddevNorm	0.902	0.957	*

Note: ** $p < 0.01$, * $p < 0.05$

Discussion

The ranked pre vs. post-treatment features reach the highest classification scores around 40 features for the full-segmentation level and for the T2 vs. T3 contrast, while other segmentations and time contrasts result in lower scores. Subsequent analysis of the top ten features, both statistically and manually, confirms that training with "Treasure Hunters" significantly changes the loudness and pitch for speakers in ways that were expected. Aside from the final feature evaluation the entire process is automatic. Future research could include more data, more features, and more classification methods.

Conclusions

We succeeded in obtaining a feature ranking by using a Recursive Feature Elimination method based on Support Vector Machine classification, which answers our first research question. Statistically significant differences were found for the top ten measures. After treatment, the eGeMAPS features HammarbergIndex and slopeV0-500 had higher values, indicating that loudness had increased, and F0semitone-Median was lower, indicating that pitch was lower. These results provide a positive answer to our second research question.

References

Boersma, P. and Weenik, D.. 2020. Praat: doing phonetics by computer (version 6.1.22). http://www.praat.org

Eyben, F., Wöllmer, M., Schuller, B.. 2010. Opensmile: the munich versatile and fast open audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. 1459-1462.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S.. 2015. The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing. 7(2), 190-202.

Ganzeboom, M., Bakker, M., Beijer, L., Rietveld, T., Strik, H.. 2018. Speech training for neurological patients using a serious game. British Journal of Educational Technology. 49(4), 761-774.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V.. 2002. Gene selection for cancer classification using support vector machines. Machine learning, 4(6), 389-422.

Sakar, B.E., Isenkul, M.E., Sakar, C.O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O.. 2013. Collection and analysis of parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics. 17(4), 828-834.

Van de Weijer, J., Slis, I.. 1991. Nasaliteitsmeting met de nasometer. Logopedie en Foniatrie 63(97), 101.