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

Neurophysiological Vocal Source Modeling for Biomarkers of Disease

Gregory Ciccarelli[1], *Thomas F. Quatieri*[1], *Satrajit S. Ghosh*[2,3]

[1]MIT Lincoln Laboratory, Lexington, Massachusetts, USA
[2]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[3]Department of Otolaryngology, Harvard Medical School, Boston, Massachusetts, USA
gregory.ciccarelli@ll.mit.edu, quatieri@ll.mit.edu, satra@mit.edu

Abstract

Speech is potentially a rich source of biomarkers for detecting and monitoring neuropsychological disorders. Current biomarkers typically comprise acoustic descriptors extracted from behavioral measures of source, filter, prosodic and linguistic cues. In contrast, in this paper, we extract vocal features based on a neurocomputational model of speech production, reflecting latent or internal motor control parameters that may be more sensitive to individual variation under neuropsychological disease. These features, which are constrained by neurophysiology, may be resilient to artifacts and provide an articulatory complement to acoustic features. Our features represent a mapping from a low-dimensional acoustics-based feature space to a high-dimensional space that captures the underlying neural process including articulatory commands and auditory and somatosensory feedback errors. In particular, we demonstrate a neurophysiological vocal source model that generates biomarkers of disease by modeling vocal source control. By using the fundamental frequency contour and a biophysical representation of the vocal source, we infer two neuromuscular time series whose coordination provides vocal features that are applied to depression and Parkinson's disease as examples. These vocal source coordination features alone, on a single held vowel, outperform or are comparable to other features sets and reflect a significant compression of the feature space.

Index Terms: depression, Parkinson's, neural computational modeling, vocal biomarkers

1. Introduction

Traditional feature engineering of speech or other modalities, best embodied by deep neural networks, often take a discriminative approach to machine learning problems. Features may or may not be interpretable and algorithms are more tuned to optimize a performance metric than to yield a model that mechanistically explains the data. Consequently, large, representative data sets may be needed, and the final result, while potentially accurate, may yield little insight into the underlying disorder.

In this work, we develop a complementary approach to traditional feature engineering schemes. Instead of relying solely on acoustic properties or more interpretable phonological characterizations of speech such as jitter, shimmer, pitch, and formants [1, 2, 3], we present a neurocomputational framework that aims to model the speech production process of an individual with a particular disorder. Within a computational biophysical framework depicted in Figure 1, features may reflect internal or latent model parameters. This approach is a step towards a brain basis for the success of previous biomarkers, towards novel biomarkers, and towards increased clinical acceptance of final assessment systems. As an example application, we develop a neurophysiological vocal source model with two muscle parameters and their coordination from which we derive features. We apply these features to depression in the Audio Visual Emotion Challenge database [4] and to Parkinson's in the mPower database [5].

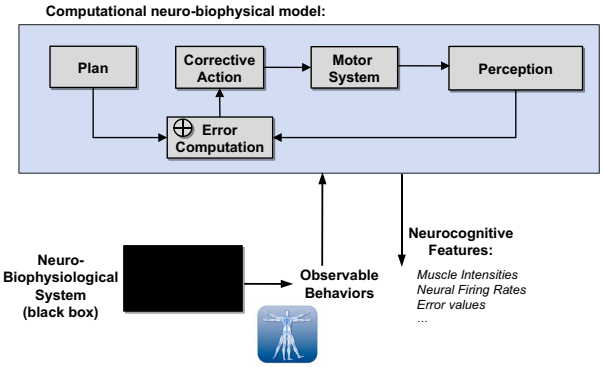

Figure 1: *Neurocomputational framework for biomarkers.*

2. Perception-Action (PA) Framework

We view the perception-action cycle as the governing paradigm for neuromotor control [6]. As illustrated in Figure 1, the cycle contains several elements: a plan, an error computation, a corrective action, a motor system, and perception. The plan consists of the goals as well as a preliminary notion of how to manipulate the motor system to achieve those goals. The motor system is the effectors such as the vocal folds or articulators in speech, or legs in walking, or ocular muscles in eye tracking. The perceptual step observes the outcome of the motor system which can be the auditory and also the associated tactile and proprioceptive signals. The perceptual step and the plan dynamically interact to create corrective actions that are sent to the motor system. This is a general framework that applies not only to speech motor control but to many forms of motor control including gait analysis and eye tracking as examples.

Computational modeling provides a constrained brain and motor feature space for exploration. Potential features that may

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force contract #FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Satrajit Ghosh was partly supported by a joint collaboration between the McGovern Institute for Brain Research and MIT Lincoln Laboratory and by NIH 1R01EB020740.

http://dx.doi.org/10.21437/Interspeech.2016-292

be extracted include the inferred neural firing rates of hypothesized areas of brain function for the different components, muscle intensities of the motor system, and the internal error signals and updates. Under neurological insults, each component of the perception-action cycle may be differentially affected which results in potentially distinct signatures that are a combination of each of the latent features. This provides a step towards characterizing the disorder that is informed by neuroscience and may provide testable hypotheses on what speech features may be expected under different insults to the brain. While we have focused on depression and Parkinson's disease in this article, our computational framework can generalize to many neurological disorders such as traumatic brain injury, autism, multiple sclerosis, schizophrenia, and Alzheimer's to name a few.

Williamson *et al.* [7] investigated a Spanish Parkinson's disorder corpus introducing a new feature that was biophysically motivated. Specifically, the formant trajectories extracted from the speech waveform were inputs to the Directions Into Velocities of Articulators (DIVA) neurocomputational model of speech motor control [8]. By running DIVA, they estimated latent articulation parameters that generated the target formant tracks. Using coordination of these tracks as features provided a performance gain.

In this paper, we build upon this approach by adopting the DIVA control architecture and a biophysically motivated model of the vocal source, with coordination of muscles driving the vocal folds as measured through a multi-scale cross correlation technique [9]. We focus initially only on the motor muscle components; yet to be explored is a more extensive characterization of the underlying complex neural motor system.

3. Databases

The depression corpus is derived from the 2013 AVEC set [4]. Depression severity was quantified with the Beck Depression Inventory II [10] scale, and the range of scores was between 0 and 45 in this corpus. Subjects participated in a variety of speech tasks, but we only analyze the held /a/ vowel which was produced at a comfortable pitch, a high pitch, a loud intensity, and a soft or low intensity. The vowels were manually segmented from the recording. The recording setup was in a quiet, laboratory environment. A total of 55 subjects pooled from the train and development subsets had /a/ vowels for analysis.

The Parkinson's corpus was available through the recently published mPower Parkinson's study [5]. The held /a/ vowel recordings were self-made on a variety of iPhone devices (4S, 5, 5c, 5S, 6, 6Plus) and iPod touch (5G) in a multitude of environments with varying levels of background noise. Both subjects who have been diagnosed with Parkinson's disorder and subjects without a Parkinson's diagnosis self-enrolled. Due to the recency of the dataset's publication, Parkinson's severity scores were not available for regression, so we classified samples as Parkinson's (PD) or healthy control (HC). We processed a subset of the full mPower dataset (Figure 2) to enforce a level of homogeneity within our PD group and our HC group. When noticed, files with artifacts (e.g., dog barks) were removed from the matched set, but exhaustive manual inspection was not possible.

4. Vocal Source Model

A novel contribution of this research is a feature set derived from a biophysically inspired model of the vocal source and computationally plausible neural control mechanism. We

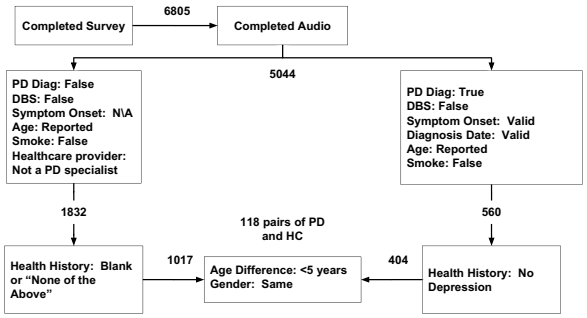

Figure 2: *Number of mPower subjects available after each selection stage.*

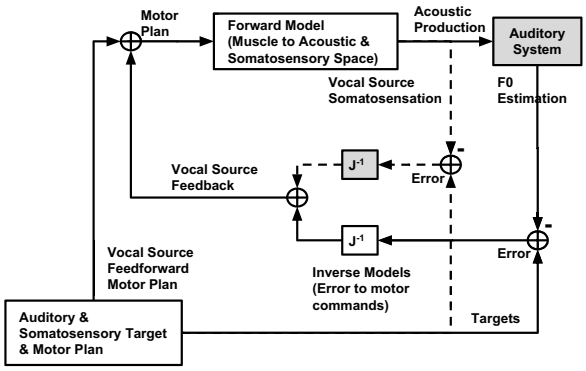

Figure 3: *Neurocomputational control framework for the vocal source. The biophysical source model enters in the forward model and auditory inversion blocks. Dotted lines and gray modules are not used in the results.*

sought a model of the vocal folds and their control mechanism in order to derive unobserved but existent muscle activations in the larynx. This approach is similar to, but substantially more developed than, Williamson *et al.* [7] in that we use the same control system paradigm, but differs in that we focus on the vocal source rather than vocal tract, and we introduce a biophysically inspired model of the vocal source. We defer a discussion of the broader implications of this approach until later, and here focus on implementation.

4.1. Control Framework

We adopted the neurocomputational control scheme hypothesized by Guenther *et al.* [8] in the Directions into Velocities of Articulators model. The model, adapted for vocal source control, is shown in Figure 3. In this scheme, there is an auditory target, which we set as the extracted fundamental frequency time series of the input speech. The forward model transforms the unobserved motor space parameters to the observed auditory space. A second component, termed the inverse model, transforms the error between the observed auditory component and the produced auditory signal into a feedback motor update. In the system used for this work, we have omitted the auditory estimation step and the somatosensory feedback as unnecessary for illustrating the central theme of using a neurocomputational model to extract features.

4.2. Implementation

The system is trained in an iterative process to determine the latent parameters necessary to reproduce the auditory target. When there is a sufficient match between the auditory target and the model production, the latent parameters are assumed to be representative of the true latent parameters of the system. In our implementation, we have a one dimensional auditory target, the fundamental frequency, and we have a two dimensional latent space that nominally represents the neural activation to the cricothyroid (CT) and thyroarytenoid (TA) muscles of the larynx.

The CT and TA muscles along with subglottal pressure and other intrinsic laryngeal muscles influence fundamental frequency [11], but we focus on the CT and TA muscles to capture their dominant influence while maintaining tractability. The CT and TA muscles are innervated by the superior laryngeal nerve and recurrent laryngeal nerve respectively, and both nerves are branches of the tenth cranial nerve whose nucleus is in the brainstem [12]. In the context of the perception-action model, we understand the neural signals to the CT and TA muscles as the net contribution of a planned fundamental frequency trajectory determined by prefrontal cortex, limbic system, and basal ganglia integration, and corrective actions based on auditory and somatosensory error signals. The muscles, acting as the motor system and our forward model, translate the neural commands by their interaction with the airstream into a new glottal flow waveform. The new muscle state and the acoustic consequences are perceived and compared to the plan, and the neural activations are updated as needed.

Our forward model is inspired by the Titze and Story biophysical model [13], and we use its computational implementation and extension by Zañartu [14]. We created a mapping of the CT and TA values to the fundamental frequency estimated by a peak picking algorithm of a generated glottal flow waveform for a generic laryngeal system. We approximated the mapping with a quadratic polynomial to quickly evaluate produced fundamental frequency for given CT and TA values because solving the differential equation governing glottal flow is computationally infeasible for seconds worth of speech. Furthermore, with a closed form, differentiable forward model, we were able to quickly compute an inverse of the forward model. We took partial derivatives of the fundamental frequency with respect to the two muscle activations to create a Jacobian matrix, and used the Moore-Penrose inverse of a matrix ($\boldsymbol{A}^{\dagger} \equiv (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$) in MATLAB (Natick, MA) to create a pseudo-inverse. The pseudo-inverse of the Jacobian converts error signals in sensory space to corresponding motor changes.

We used Praat to extract the fundamental frequency (f_0) trajectory from the vowel waveform [15], shifted it to our model's f_0 range, and input the trajectory to our vocal source model to infer the hidden CT and TA activation time series. Then we perform multiscale correlation of the CT and TA time series jointly and individually after Williamson *et al.* [9] to extract the eigenvalues that are our neurocomputational source (NCS) features. The joint analysis (Figure 5, NCS) yields 84 features, and each individual analysis (NCSct and NCSta) yields 42 eigenvalues.

4.3. Muscle Activation

An example of the fundamental frequency extracted from the waveform (True), the model-generated fundamental frequency (Model), and the inferred CT and TA muscle activations are given in Figure 4. We acknowledge the strong correlation between the inferred CT value and the fundamental frequency and

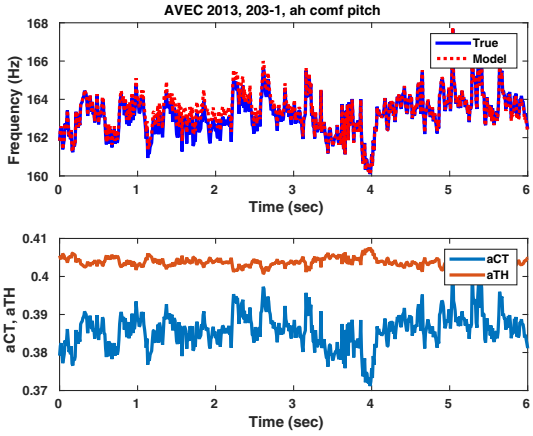

Figure 4: *True and model-generated fundamental frequency (top) and inferred CT and TA muscle activations (bottom).*

argue that this is reasonable. Mathematically, the correlation occurs because the gradient of the fundamental frequency surface is strongly aligned with CT, so small changes in muscle activation will impact fundamental frequency most when the CT muscle is changed. The gradient's alignment is consistent with computational simulations of the two muscles in [11] and is consistent with vocal source physiology.

The CT, when tensed, pulls the cricoid and thyroid cartilage to directly tense the vocal cords and therefore increase fundamental frequency. The CT has a larger influence than the TA on changes in fundamental frequency because the CT directly regulates vocal fold tension whereas the TA has an indirect effect. To understand the difference in magnitude of influence of the two muscles, we appeal to the body cover model of the vocal folds that describes the vocal folds as a muscular body loosely connected to a covering with different mechanical properties than the body. The TA may differentially slacken the cover component of the model while increasing the tension of the body. Depending on the net tension increase or decrease of the body cover system, the fundamental frequency may increase or decrease [16, 17]. *Ex vivo* laryngeal stimulation experiments confirm the dominant role of the CT muscle and the nuanced role of the TA muscle [18].

5. Machine Learning

5.1. Depression Prediction

To predict depression severity, we use the extremely randomized trees (ERT) regression algorithm implemented in scikit-learn [19]. ERT is an ensemble learning method in which multiple decision trees are constructed and their performance combined. ERTs make no statistical assumptions about feature dependence, can be used for classification and regression, and empirically have been successful in a range of applications [20]. We randomly split our dataset into a train set containing approximately 80 percent of the subjects and a test set containing the remaining subjects. We build the regression model on the train set and evaluate on the test set. We average across all vowels and all sessions to achieve a final predicted subject score and we compare against the average session score for the subject. We repeat the cross validation 100 times.

We compare our NCS features against the 6553 features

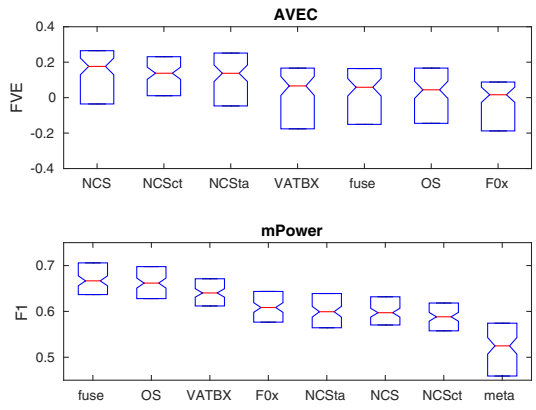

Figure 5: *The NCS features have the highest median performance in depression and have a median above demographic (meta) data in Parkinson's.*

from Opensmile (OS) [21] and against the 339 Voice Analysis Toolbox features (VATBX) from Tsanas *et al.* [22, 23, 24]. Opensmile and VATBX have respectively been used in depression and Parkinon's assessment. The OS features consist of a core set of descriptors such as spectral band statistics, fundamental frequency, and mel frequency cepstral coefficients and their derivatives that are augmented by numerous summary statistics including the min, max, mean, percentiles, and ranges. The VATBX features emphasize statistical characterization of the source with jitter, shimmer, Teager energy operator, glottal quotient, and entropy measures, but also includes mel frequency cepstral coefficients. We also compare to a multiscale correlation analysis of the fundamental frequency that is input to the neurocomputational model (F0x, 42 features). The "fuse" feature set is the concatenation of all the individual feature sets.

5.2. Parkinson's Detection

We follow a nearly identical approach for Parkinson's as with depression, but we use the ERT algorithm for classification. Each vowel from each subject contributes one sample to the performance calculation per each of the 100 test iterations. We also include age and gender information (meta) in each speech feature set in addition to testing the meta information alone. Age and gender were not released with the AVEC database.

6. Results

For AVEC, we report the fraction of variance explained (FVE). FVE is a normalized mean square error ($\text{FVE} \equiv 1 - \text{MSE}/\sigma^2_{truth}$) that is upper bounded by 1 (best) but not lower bounded. A FVE of zero is obtained by predicting the mean of the test set, but a FVE less than zero can be obtained if predictions are worse than if the mean had been used. For mPower, we report the F_1 score ($\text{F}_1 \equiv 2 \cdot \text{PR} \cdot \text{RC}/(\text{PR} + \text{RC})$). Precision, PR, is the probability of a declared Parkinson's sample belonging to a Parkinson's subject, and recall, RC, is the probability of a Parkinson's sample being declared as a Parkinson's sample. The F_1 score varies between 0 and 1 (best).

Figure 5 contains the median (center line) and 25^{th} and 75^{th} percentiles (box edges) of the test iterations. For AVEC, a FVE of zero is a baseline that corresponds to predicting the average depression score. For mPower, we use the age and gender attribute as the baseline feature set because, to our knowledge, no previous speech baseline exists.

7. Discussion

In depression, the neurocomputational source features (NCS) appear to contribute predictive power, and in the mPower set, NCS, VATBX, and OS have medians above a baseline system that uses only age and gender information. Both results imply speech has some discriminatory power, and this is consistent with previous studies. The neurocomputational source features appear to provide greater relative advantage on the depression set than the Parkinson's set whereas the VATBX features seem to be most useful for Parkinson's. OS's poor performance in depression may reflect overfitting to the training set with its large number of features.

The NCS features have a median performance above VATBX in depression. The VATBX features are source features, but they are statistical quantities that do not appear to generalize across disorders. We hypothesize that the NCS model is detecting a fundamental neural dysfunction rather than a statistical aberration. Also, NCS has a higher median than F0x, NCSta, and NCSct in depression, so we hypothesize that a biophysical model that considers coordination across components may allow a more complete description of the disorder than individual neural signals.

Potential avenues of expansion for the lower level neural physiology include additional models for the other intrinsic laryngeal muscles, respiratory muscle control of the subglottal pressure, and excitation-contraction muscle dynamics. Models for higher level control include prosodic planning for affective vs linguistic prosody, limbic system influence on plans under depression, and the changes due to the basal ganglia's degradation under Parkinson's.

8. Conclusions

Our investigation demonstrated a unique approach to identifying and utilizing vocal features in held /a/ vowels for assessing depression and Parkinson's. We presented a neurocomputational framework for biomarker discovery in which neural physiology and biomechanical principles are used to map low dimensional feature data into a higher dimensional but constrained space. While our examples were limited to the speech source for depression and Parkinson's, this general paradigm has broad applicability to neuroscience and physiological status monitoring applications. The framework naturally admits multi-modal integration, so such diverse metrics as functional MRI times series, gait accelerometry data, and speech can be unified. By combining medical knowledge with observations, we can bring forth new sets of features via the computational models that may be more discriminative and sensitive than traditional approaches as well as provide directions for treatment and a better understanding of mental or physiological disorders.

9. Acknowledgements

We thank Matías Zañartu for his body cover model, and Adam Lammert, Daryush Mehta, and Chris Smalt for helpful discussions. PD data was contributed by users of the Parkinson's mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse doi:10.7303/syn4993293.

10. References

[1] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[2] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[3] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 65–72.

[4] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[5] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey *et al.*, "The mpower study, parkinson disease mobile data collected using researchkit," *Scientific Data*, vol. 3, 2016.

[6] V. Cutsuridis, A. Hussain, and J. G. Taylor, *Perception-Action Cycle: Models, Architectures, and Hardware*. Springer Science & Business Media, 2011.

[7] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. Ciccarelli, and D. D. Mehta, "Segment-dependent dynamics in predicting parkinsons disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and language*, vol. 96, no. 3, pp. 280–301, 2006.

[9] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using eeg spatiotemporal correlation structure," *Epilepsy & Behavior*, vol. 25, no. 2, pp. 230–238, 2012.

[10] A. T. Beck, R. A. Steer, and G. K. Brown, *Manual for the Beck depression inventory-II*. San Antonio, TX: The Psychological Corporation, 1996.

[11] I. R. Titze and B. H. Story, "Rules for controlling low-dimensional vocal fold models with muscle activation," *The Journal of the Acoustical Society of America*, vol. 112, no. 3, pp. 1064–1076, 2002.

[12] A. Yau and S. Verma, "Laryngeal nerve anatomy," http://emedicine.medscape.com/article/1923100-overview, 2013, accessed: 2016-03-28.

[13] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1249–1260, 1995.

[14] M. Zañartu Salas, "Influence of acoustic loading on the flow-induced oscillations of single mass models of the human larynx," Ph.D. dissertation, Purdue University West Lafayette, 2006.

[15] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[16] I. R. Titze, E. S. Luschei, and M. Hirano, "Role of the thyroarytenoid muscle in regulation of fundamental frequency," *Journal of Voice*, vol. 3, no. 3, pp. 213–224, 1989.

[17] C. R. Larson, G. B. Kempster, and M. K. Kistler, "Changes in voice fundamental frequency following discharge of single motor units in cricothyroid and thyroarytenoid muscles," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 4, pp. 552–558, 1987.

[18] D. K. Chhetri, J. Neubauer, and D. A. Berry, "Neuromuscular control of fundamental frequency and glottal posture at phonation onset," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1401–1412, 2012.

[19] 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.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[22] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[23] A. Tsanas, "Accurate telemonitoring of parkinsons disease symptom severity using nonlinear speech signal processing and statistical machine learning," Ph.D. dissertation, University of Oxford, 2012.

[24] ——, "Automatic objective biomarkers of neurodegenerative disorders using nonlinear speech signal processing tools," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2013, pp. 37–40.