INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Detecting Manifest Huntington's Disease Using Vocal Data

*Vinod Subramanian*[1], *Namhee Kwon*[1], *Raymond Brueckner*[1], *Nate Blaylock*[1], *Henry O'Connell*[1], *Luis Sierra*[2], *Clementina Ullman*[2], *Karen Hildebrand*[2], *Simon Laganiere*[2]

[1]Canary Speech, USA
[2]Beth Israel Medical Center, USA

{vinod, namhee, ray, nate, henry}@canaryspeech.com
{lsierra1, cullman, khildebr, slaganie}@bidmc.harvard.edu

## Abstract

Huntington's disease (HD) is an autosomal-dominant neurodegenerative disorder that leads to the devastating loss of motor control - including severe speech impairment. Current models are insufficient to predict the onset or progression of manifest symptoms and early signs of the disease remain challenging to detect and monitor. Therefore, we propose a purely speech-based, non-invasive approach to discriminate Huntington's Disease patients who are exhibiting early signs of disease from those who are not. We study various features derived from speech and machine learning models to classify HD patients. Our results show that Random Forest classifiers leveraging language features perform very well with an unweighted accuracy of 0.95. In addition, we analyze the statistical significance of features, the importance of different questions asked to the patients, and other classification problems in Huntington's disease to provide a strong foundation for this field of research.

**Index Terms**: Huntington's disease, dysarthria, healthcare, biomarker

## 1. Introduction

Huntington's Disease (HD) is defined by Walker [1] as "an autosomal-dominant, progressive neurodegenerative disorder with a distinct phenotype, including chorea and dystonia, incoordination, cognitive decline, and behavioral difficulties." HD can be diagnosed by a genetic test; however, there is no cure for HD. Clinical trials that target the earliest stage of the disease are currently underway; however the ability to properly time these disease-modifying interventions is critically dependent on the sensitivity of early biomarkers of the disease. Finding the earliest and most reliable biomarkers, including changes in speech, remains a priority. Currently, speech changes in Huntington's disease are recorded by an examiner using a relatively crude scale of clinical dysarthria.

The aim of this paper is to solve the problem of relying on clinical assessments of speech by automating the process using machine learning models that can predict whether an HD patient is manifesting subtle speech abnormalities, based on vocal input. Progressive dysarthria results from an inability to fully control the articulatory-phonatory muscles and is one of the most common symptoms in HD patients [2].

Using audio to detect symptoms of Huntington's Disease is a newly emerging field. Before detailing prior work, it is important to understand the labels used in HD data. Patients with genetic confirmation of Huntington's Disease can be divided into two groups: a) premanifest (Pre) group, which includes patients who are not yet exhibiting signs and symptoms of HD, and b) manifest (MF), which includes patients who are exhibiting these signs. HD studies also often include a healthy control group (CT) for comparisons. These three groups allow for different contrasts when approaching this problem as a classification task.

Perez et al. [3] attempted to discriminate the CT from the HD group on a dataset of people reading passages and extracted pause, filler, and goodness of pronunciation features. Using a *k*-nearest neighbor algorithm, they achieved a classification accuracy of 0.87. Rusz et al. [4] adopted the same dataset, but used reading passages and monologues. They extracted features related to vowel articulation, pitch, speech rate, and pauses and trained a Support Vector Machine (SVM) classifier on all combinations of features in pairs.

In a later study, Rusz et al. [5] classified CT vs. Pre on a dataset of sustained vowel sounds, extracting features that identify vocal breaks, frequency perturbations, as well as MFCCs and delta-MFCCs. They trained an SVM classifier and reported a sensitivity of 0.91 and a specificity of 0.79. A 3-class classification between CT, Pre, and MF groups was performed by Riad et al. [6], where the authors also used a dataset of sustained vowels. They used the same features as [5], along with features that capture the modulation of the pitch amplitude, and trained a Logistic Regression model achieving a classification accuracy of 0.56. Finally, Romana et al. [7] tackled the problem of discriminating the Pre and MF groups, which is closely related to the classification task in this paper. Their dataset consisted of sustained vowel sounds and a passage reading. They developed a feature set to capture distortions in the vowels and trained a logistic regression model achieving an accuracy of 0.80.

Our contribution addresses three fundamental shortcomings identified in prior work:

1. We compare various machine learning models and features to classify HD patients as either manifest or premanifest.
2. We compare and contrast the performance of various models based on the type of vocal training data, specifically free speech, passage reading, and the Stroop test [8].
3. We build a model based on the clinical dysarthria scale to assess the relative efficacy of identifying early speech changes.

## 2. Dataset

The data collection was performed at Beth Israel Deaconess Medical Center in Boston and the speech audio was recorded in the approved research locations within the institution. 76 subjects were collected in total, including 44 patients with Huntington's disease at various stages of the disease and 32 healthy persons in the control group (CT). The HD and CT groups were balanced w. r. t. sex, age, and education. All HD patients were assessed by a neurologist using the 124-point Unified Hunt-


10.21437/Interspeech.2023-1096

ington's Disease Rating Scale - Total Motor Score (UHDRS-TMS) [9, 10]. The mean was 15.20 and the standard deviation was 15.55. A subject was categorized as manifest if UHDRS-TMS $> 8$, otherwise as premanifest. Using the very low cut-off score, we categorized the patients at the early manifestation stage as manifest. Dysarthria represented one item in UHDRS-TMS ranging between 0 (no dysarthria) and 4 (mute). HD patients in this cohort had scores between 0 and 2, where a score of 0 was defined to be healthy, and the rest as having dysarthria. The number of subjects per category is shown in Table 1.

| Category | | No. of Subjects |
|---|---|---|
| Control | | 32 |
| Huntington's Disease | Premanifest | 21 |
| | Manifest w/o Dysarthria | 8 |
| | Manifest w/ Dysarthria | 15 |

Table 1: *Data distribution over categories*

We developed a mobile application that prompted multiple questions and the audio recordings were captured using a tablet under the supervision of a clinician. To analyze the voice characteristics in various types of speech, we designed three different kinds of questions including the Stroop Color-Word Test (SCWT) [11], reading passages, and spontaneous speech. The SCWT is a cognitive control test composed of three tasks: 1) word reading asking to read the name of the color printed in black ink, 2) color naming asking to name different color patches, and 3) color-word where the participant was required to answer the color that the word was printed in. For example, if the word "RED" was printed in "BLUE", the person had to say "BLUE" while ignoring the word "RED". For the reading test we asked participants to read a caterpillar passage [12] known to be phonetically balanced. For free speech, three questions are asked: "How are you?", "How to make a peanut butter and jelly sandwich?", and a picture description (cookie theft). Table 2 shows the audio durations per question type.

| Prompts | [Mean±Std] | [Min, Max] |
|---|---|---|
| Stroop: Word-Reading | $47.8 \pm 1.8$ | [44.2, 54.5] |
| Stroop: Color-Naming | $48.2 \pm 2.2$ | [45.0, 60.9] |
| Stroop: Color-Word | $48.4 \pm 1.7$ | [45.1, 53.7] |
| Read | $74.5 \pm 18.7$ | [1.4, 133.0] |
| How-Are-You | $24.7 \pm 21.7$ | [2.3, 166.4] |
| Narrative (How-to) | $31.4 \pm 11.3$ | [1.8, 98.9] |
| Picture Description | $41.9 \pm 13.2$ | [13.9, 103.0] |

Table 2: *Audio duration [seconds]*

# 3. Methods

## 3.1. Features

We generated various fixed-length representations of each audio sample as an input representation for the subsequent classification and prediction steps, resulting in three different feature categories: acoustic-prosodic features, language features, and speaker embedding features. For each question prompt a feature vector was extracted and the feature vectors belonging to a speaker were concatenated and fed into the classification model.

Acoustic-prosodic features are commonly adopted in audio classification tasks and have proven effective in detecting several types of diseases [13, 14]. **Acoustic** features represent statistics of frame-based signal descriptors including Mel-Frequency Cepstral Coefficients (MFCC) [15], Perceptual Linear Prediction (PLP) coefficients [16], and pitch-related information [17]. **Prosodic** features instead capture supra-segmental aspects of the modulations of human articulatory organs during speech, such as normalized fundamental frequency ($f0$) and energy patterns [13]. We generated both acoustic and prosodic features using the Kaldi toolkit [18].

**Language** features on the other side capture lexical information. Such features were previously shown to be correlated to the detection of Huntington's disease [3, 19]. We generated the language features from the transcription result of an automatic speech recognition (ASR) system. We computed syllable duration, filler (*"ah", "hmm", "eh", "uh"*, etc.) ratio, spoken words ratio, word repetition ratio, pause duration, etc. Additionally, we derived question-specific features for read speech and spontaneous free speech. For read speech, we expanded upon features used in previous work [3] by including errors such as insertions, deletions, and substitutions to evaluate how each word is articulated. For spontaneous free speech including *How-Are-You*, *How-to*, and *picture description*, semantic features are also added such as lexical complexity (n-gram probability based on a language model generated using the Librispeech dataset [20]), word usage (SMOG reliability) score [21], word ambiguity, and familiarity score based on linguistic research [22]. The feature vector dimensions differ across question types and range from 20 to 53.

Speaker embedding features include **i-vectors** [23] and **x-vectors** [24]. They were originally proposed for robust speaker recognition [25]. I-vectors are based on factor analysis using Gaussian Mixture Models and x-vectors are built using deep neural networks. Given the success of x-vectors as an audio representation for depression analysis [26, 27], we explored them as a feature to detect Huntington's disease.

## 3.2. Machine Learning Models

Using the various features described in Section 3.1, we built and compared three different model architectures: Random Forests (RF), Support Vector Machine (SVM), and Logistic Regression (LR). In order to compare our results to prior work on predicting Huntington's disease from audio samples [5, 7], we trained SVMs with a radial basis function (RBF) kernel and LRs with $L_2$ penalty as baseline references. We additionally investigated RF classifier models since they had been reported to outperform other model types, in particular logistic regression, on a number of (typically small datasets) in the research area of bioinformatics [28]. The Random Forest classifier was constructed using the Gini impurity and the maximum number of features was defined as the square root of the feature dimensions.

# 4. Experiments and Results

We investigated multiple tasks: Huntington's disease vs. control group, manifest vs. premanifest, and dysarthria vs. no-dysarthria. The three classification models (cf. Section 3.2) were trained using the five sets of features (cf. Section 3.1). We performed the 10-fold cross-validation where no speaker was duplicated between folds and the label distribution in each fold reflected the overall data distribution.

Our evaluation metrics included *sensitivity* (Sen) and *specificity* (Spe) in addition to *accuracy* (Acc). Sensitivity is defined as the positive class recall and specificity as the negative class

recall. We also reported the *unweighted accuracy* (UA), i. e. the average recall, as long as the *Area Under the Curve* (AUC).

### 4.1. Manifest vs. Premanifest Classification

We examined the 15 feature-model combinations on the task of classifying premanifest and manifest patients. The results are shown in Table 3.

| Models | Features | Acc | UA | Sen | Spe | AUC |
|---|---|---|---|---|---|---|
| **RF** | Acoustic | 0.86 | 0.86 | 0.91 | 0.81 | 0.92 |
| | Prosodic | 0.75 | 0.75 | 0.74 | 0.76 | 0.83 |
| | Language | **0.95** | **0.95** | **0.96** | **0.95** | **0.97** |
| | i-vector | 0.73 | 0.72 | 0.83 | 0.62 | 0.77 |
| | x-vector | 0.75 | 0.74 | 0.87 | 0.62 | 0.84 |
| **LR** | Acoustic | 0.68 | 0.68 | 0.69 | 0.66 | 0.84 |
| | Prosodic | 0.59 | 0.59 | 0.60 | 0.57 | 0.67 |
| | Language | 0.75 | 0.75 | 0.65 | 0.86 | 0.85 |
| | i-vector | 0.72 | 0.72 | 0.82 | 0.62 | 0.79 |
| | x-vector | 0.84 | 0.84 | 0.91 | 0.76 | 0.87 |
| **SVM** | Acoustic | 0.52 | 0.51 | 0.83 | 0.19 | 0.51 |
| | Prosodic | 0.47 | 0.46 | 0.78 | 0.14 | 0.46 |
| | Language | 0.77 | 0.77 | 0.78 | 0.76 | 0.77 |
| | i-vector | 0.68 | 0.67 | 0.87 | 0.48 | 0.67 |
| | x-vector | 0.68 | 0.67 | 0.78 | 0.57 | 0.67 |

Table 3: *Classification results for manifest vs. premanifest*

The language features consistently performed well in every architecture and the Random Forest (RF) classifier outperformed the other models across all metrics. The combination of language features and RF classifier resulted in an absolute UA margin of 9% over the 2nd-best model. To verify that the RF model's high performance was not just by chance, e. g. by fortunate initialization of parameters, 5 different versions of the RF model were trained and evaluated; the standard deviation for UA is less than 0.01.

### 4.2. Dysarthria Classification

In this experiment, we trained classifiers to differentiate between dysarthria and non-dysarthria in the Huntington's disease group. Dysarthria is often observed in many HD patients [19] and in our dataset the dysarthria class is a complete subset of the manifest class (cf. Table 1). Following the approach with the manifest classification task in Section 4.1, all feature-model combinations were trained and evaluated. Again the Random Forest Classifier performed best; hence, for brevity, Table 4 only reports the results of the RF model.

| Features | Acc | UA | Sen | Spe | AUC |
|---|---|---|---|---|---|
| Acoustic | 0.70 | 0.58 | 0.20 | 0.96 | 0.67 |
| Prosodic | 0.64 | 0.53 | 0.20 | 0.86 | 0.67 |
| Language | **0.82** | **0.78** | **0.67** | **0.89** | **0.86** |
| i-vector | 0.66 | 0.53 | 0.13 | 0.93 | 0.71 |
| x-vector | 0.68 | 0.60 | 0.33 | 0.86 | 0.76 |

Table 4: *RF classification results for Dysarthria*

Similar to the results of discriminating between manifest and premanifest (cf. Section 4.1), the Random Forest classifier using the language features outperformed all other model-feature combinations. Interestingly, the dysarthria model's performance was lower than the manifest model, even though dysarthria is defined as a speech difficulty and is a complete subset of the manifest HD group in our data. A deeper analysis will be presented in Section 5.3.

### 4.3. Huntington's Disease Classification

There are many sub-problems in Huntington's disease research as seen from prior work (cf. Section 1). To efficiently compare the Random Forest classifier using language features with prior research, we trained our model to perform 4 different classification tasks. First, we examined the discrimination between all HD patients (Pre + MF) and the (healthy) control group (CT). Second, we performed the 3-class classification manifest (MF) vs. premanifest (Pre) vs. control group (CT). As shown in Table 5, the performance metrics were lower than the previous results for MF vs. Pre. This result is not surprising because the premanifest patients do not exhibit any evident symptoms, and actually many premanifest patients' UHDRS-TMS was 0 in our dataset. This intuition is confirmed in the third classification experiment between Pre vs. CT. Therefore, discriminating the combination of the premanifest and control group vs. the manifest group leads to results on par with the results from Section 4.1. A detailed analysis supporting these findings is presented in Section 5.1.

| Classification | Acc | UA | Sen | Spe | AUC |
|---|---|---|---|---|---|
| (Pre + MF) vs. CT | 0.73 | 0.72 | 0.76 | 0.69 | 0.72 |
| MF vs. Pre vs. CT | 0.60 | 0.59 | N/A | N/A | N/A |
| Pre vs. CT | 0.65 | 0.62 | 0.43 | 0.81 | 0.63 |
| MF vs. (Pre + CT) | **0.96** | **0.95** | **0.91** | **0.98** | **0.97** |

Table 5: *RF performance for different classification tasks*

## 5. Discussion

### 5.1. Feature Correlation Analysis

As the results presented in Section 4 showed that the language features (when used with RF models) outperform the other feature types on several tasks, we conducted a statistical feature analysis among control (CT), premanifest HD (Pre), and manifest HD (MF) groups. From a total of 241 language features across all question prompts, the three features *speech rate*, *silence ratio*, and *average pause length* were chosen and analyzed for the passage reading, the "How are you?" question (HowAY), and the Stroop-Word Reading (SWR) test. Not only do these feature analyses help to understand the dataset and tasks at hand. But these specific features are also comparable to the ones reported in [19] so it supports the comparison of the two datasets. First, we performed the median version of Levene's test across the three groups where the null hypothesis is that each group's value distributions have equal variance. After confirming that most features had an equal variance in the distributions among the groups, a one-way ANOVA test was performed between the three combinations of the groups: control vs. premanifest, control vs. manifest, and premanifest vs. manifest. The mean and standard deviation of the feature values are reported in Table 6.

The ANOVA test showed that all three features were significantly different in distributions between control vs. manifest and premanifest vs. manifest except for the HowAY question. To investigate more on this, we performed question group analysis in Section 5.2.

| Features | Group | Mean (Standard Deviation) | | | Levene | ANOVA p-value | | |
|---|---|---|---|---|---|---|---|---|
| | | Control | Premanifest HD | Manifest HD | | CT vs Pre | CT vs MF | Pre vs MF |
| **speech rate** | SWR | 2.16 (0.38) | 1.98 (0.38) | 1.38 (0.34) | 0.834 | 0.120 | <0.001 | <0.001 |
| **%silence** | SWR | 0.95 (0.05) | 0.92 (0.08) | 0.84 (0.09) | 0.009 | 0.185 | <0.001 | 0.005 |
| **Pause** | SWR | 26.76 (26.84) | 48.00 (73.52) | 129.55 (90.77) | <0.001 | 0.150 | <0.001 | 0.003 |
| **speech rate** | Read | 3.14 (0.37) | 3.01 (0.40) | 2.17 (0.45) | 0.788 | 0.238 | <0.001 | <0.001 |
| **%silence** | Read | 0.93 (0.04) | 0.93 (0.03) | 0.86 (0.12) | 0.015 | 0.822 | 0.003 | 0.017 |
| **Pause** | Read | 22.01 (14.98) | 23.67 (12.36) | 92.41 (169.0) | 0.072 | 0.681 | 0.025 | 0.076 |
| **speech rate** | HowAY | 2.64 (0.69) | 2.36 (0.68) | 2.09 (0.39) | 0.044 | 0.149 | 0.001 | 0.130 |
| **%silence** | HowAY | 0.90 (0.13) | 0.86 (0.10) | 0.83 (0.14) | 0.426 | 0.239 | 0.054 | 0.376 |
| **Pause** | HowAY | 65.54 (110.18) | 106.98 (85.04) | 112.29 (101.90) | 0.630 | 0.157 | 0.122 | 0.856 |

Table 6: *Statistical analysis of language features*

None of the features reported showed a significant difference between the control and premanifest group, which explains the result of Section 4.3. Huntington's disease is a genetic disorder that manifests usually between the age of 30 and 50, so it is important to detect the manifest state in a timely manner because the current treatments for Huntington's disease focus on symptom control and occupational therapy. Although our model does not detect Huntington's disease at the premanifest state consistently, our highly accurate manifest classification model should prove to be very useful in the field.

### 5.2. Question Group Analysis

Table 6 shows that each feature had different $p$ values based on the question. So, we built classification models between manifest and premanifest using features extracted per question group separately. The goal of this experiment was to compare each question group and find stronger question types if they exist.

| Models | Features | Acc | UA | Sen | Spe | AUC |
|---|---|---|---|---|---|---|
| **RF** | Stroop | 0.88 | 0.88 | 0.91 | 0.86 | 0.95 |
| | read | **0.93** | **0.93** | **0.96** | 0.90 | **0.98** |
| | free | 0.73 | 0.73 | 0.74 | 0.71 | 0.77 |
| **LR** | Stroop | 0.91 | 0.91 | 0.91 | 0.90 | 0.91 |
| | read | 0.88 | 0.89 | 0.83 | **0.95** | 0.92 |
| | free | 0.84 | 0.84 | 0.87 | 0.81 | 0.90 |
| **SVM** | Stroop | 0.87 | 0.86 | 0.92 | 0.81 | 0.86 |
| | read | 0.71 | 0.72 | 0.58 | 0.86 | 0.72 |
| | free | 0.64 | 0.63 | 0.65 | 0.62 | 0.64 |

Table 7: *Results of per-question-group models using the language features*

The classification performance for each question group using the language features is shown in Table 7. No model outperformed the best RF model using all question groups, yet the models using Stroop test or read speech performed well compared to the model using free speech. This result suggests that the type of speech data captured has an impact on the ability to classify between manifest and premanifest. Reading and the Stroop test are structured because there is a script or an expected answer, whereas free speech questions such as "How are you" can be answered in a wide variety of ways and are potentially impacted by other factors such as personality, culture, and native language. We conjecture that our dataset is not big enough to capture the variations in free speech patterns. which could potentially also be affected by the interview environment. This is supported by the fact that the durations of free speech answers varied the most with respect to their averages (cf. Table 2).

### 5.3. Dysarthria Model Error Analysis

Our dataset offers the unique opportunity to analyze dysarthria alongside Huntington's disease. To our knowledge, this is the first study comparing dysarthria and Huntington's disease performance on the same dataset. In contrast to the intuition that the voice analysis model is good at detecting dysarthria, it actually turned out that the dysarthria model's prediction accuracy was lower than the accuracy of the manifest classification model (cf. Table 3 and Table 4).

We performed an error analysis on the false positive prediction of the dysarthria Random Forest model. All false positives, i. e. speakers misclassified as having dysarthria, were in the manifest group without dysarthria. This phenomenon was observed in all 5 versions of RF training experiments both using language features and acoustic features. As shown in Table 1, dysarthria is a subset of manifest and our model was confused about the positive manifest label with a positive dysarthria label consistently. We conjecture that in those cases there is some evidence of dysarthria contained in the features that, however, is not reflected in the dysarthria labels, against which the models are evaluated. This situation could arise in the early stages of dysarthria and was also observed in an earlier study [19].

## 6. Conclusion

We developed several different classification models to distinguish between manifest vs. premanifest for Huntington's disease. We extracted various features from three types of questions: the Stroop test, reading, and spontaneous free speech. The Random Forest classification model using language features outperformed other models with an unweighted accuracy (UA) of 0.95 and AUC of 0.97. Then, we explored using dysarthria labels since it is a vocal disorder and a symptom of manifest HD. The best dysarthria classification model reported a UA of 0.78, where we observed that the model was mistaken between manifest and dysarthria, but never between premanifest and dysarthria. Finally, to standardize Huntington's disease research, we perform a statistical analysis of features for the different questions asked and show that there is statistical significance between premanifest and manifest groups. Then models are trained per question group and show that the Stroop test and read speech are better for detecting manifest in Huntington's disease compared to free speech. Encouraged by our success in accurately classifying manifest Huntington's disease, we plan to extend our research to other diseases that correlate with motor symptoms such as Parkinson's disease.

# 7. References


[1] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, Jan. 2007.

[2] L. Hartelius, A. Carlstedt, M. Ytterberg, M. Lillvik, and K. Laakso, "Speech disorders in mild and moderate huntington disease: Results of dysarthria assessments of 19 individuals," *Journal of Medical Speech-Language Pathology*, vol. 11, no. 1, pp. 1–15, 2003.

[3] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A. Roberts, and E. Mower Provost, "Classification of Huntington Disease Using Acoustic and Lexical Features," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1898–1902.

[4] J. Rusz, J. Klempíř, T. Tykalova, E. Baborová, R. Čmejla, E. Růžička, and J. Roth, "Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication," *Journal of Neural Transmission*, vol. 121, pp. 1529–1539, 2014.

[5] J. Rusz, C. Saft, U. Schlegel, R. Hoffman, and S. Skodda, "Phonatory Dysfunction as a Preclinical Symptom of Huntington Disease," *PLoS ONE*, vol. 9, no. 11, Nov. 2014.

[6] R. Riad, H. Titeux, L. Lemoine, J. Montillot, J. H. Bagnou, X.-N. Cao, E. Dupoux, and A.-C. Bachoud-Lévi, "Vocal Markers from Sustained Phonation in Huntington's Disease," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1893–1897.

[7] A. Romana, J. Bandon, N. Carlozzi, A. Roberts, and E. M. Provost, "Classification of Manifest Huntington Disease Using Vowel Distortion Measures," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 4966–4970.

[8] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology*, vol. 18, no. 6, pp. 643–662, Dec. 1935.

[9] H. S. Group, "Unified Huntington's disease rating scale: Reliability and consistency," *Movement Disorders*, vol. 11, no. 2, pp. 136–142, Mar. 1996.

[10] T. A. Mestre, M. J. Forjaz, P. Mahlknecht, F. Cardoso, J. J. Ferreira, R. Reilmann, C. Sampaio, C. G. Goetz, E. Cubo, P. Martinez-Martin, G. T. Stebbins, and Members of the Movement Disorder Society Committee on Rating Scales Development, "Rating scales for motor symptoms and signs in huntington's disease: Critique and recommendations," *Mov. Disord. Clin. Pract.*, vol. 5, no. 2, pp. 111–117, Mar. 2018.

[11] F. Scarpina and S. Tagini, "The stroop color and word test," *Front. Psychol.*, vol. 8, p. 557, Apr. 2017.

[12] R. Patel, K. Connaghan, D. Franco, E. Edsall, D. Forgit, L. Olsen, L. Ramage, E. Tyler, and S. Russell, ""The Caterpillar": A Novel Reading Passage for Assessment of Motor Speech Disorders," *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, vol. 22, no. 1, pp. 1–9, Feb. 2013.

[13] N. Kwon and S. Kim, "Depression severity detection using read speech with a divide-and-conquer approach," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021, pp. 633–637.

[14] S. Kim, N. Kwon, H. O'Connell, N. Fisk, S. Ferguson, and M. Bartlett, ""how are you?" estimation of anxiety, sleep quality, and mood using computational voice analysis," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020, pp. 5369–5373.

[15] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2494–2498.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, Dec. 2011.

[19] J. C. S. Chan, J. C. Stout, C. A. Shirbin, and A. P. Vogel, "Listener Detection of Objectively Validated Acoustic Features of Speech in Huntington's Disease," *Journal of Huntington's disease*, vol. 11, no. 1, pp. 71–79, 2022.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[21] K. L. Grabeel, J. Russomanno, S. Oelschlegel, E. Tester, and R. E. Heidel, "Computerized versus hand-scored health literacy tools: a comparison of simple measure of gobbledygook (smog) and flesch-kincaid in printed patient education materials," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, p. 38, 2018.

[22] S. Cho, N. Nevler, S. Ash, S. Shellikeri, D. J. Irwin, L. Massimo, K. Rascovsky, C. Olm, M. Grossman, and M. Liberman, "Automated analysis of lexical features in frontotemporal degeneration," *Cortex*, vol. 137, pp. 215–231, 2021.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[25] A. Khosravani, C. Glackin, N. Dugan, G. Chollet, and N. Cannings, "The Intelligent Voice 2016 Speaker Recognition System," *CoRR*, vol. abs/1611.00514, 2016. [Online]. Available: http://arxiv.org/abs/1611.00514

[26] J. V. Egas-Lopez, G. Kiss, D. Sztaho, and G. Gosztolya, "Automatic Assessment of the Degree of Clinical Depression from Speech Using X-Vectors," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8502–8506.

[27] N. Kwon, S. Hossain, N. Blaylock, H. O'Connell, N. Hachen, and J. Gwin, "Detecting Anxiety and Depression from Phone Conversations using x-vectors," in *Proc. Workshop on Speech, Music and Mind*, 2022, pp. 1–5.

[28] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC bioinformatics*, vol. 19, pp. 1–14, 2018.