



A simple method for predicting Clinical Scores in Huntington's Disease by leveraging ASR's uncertainty on spontaneous speech

Hadrien Titeux^{1,*}, Quang Tuan Rémy Nguyen^{1,3,*}, Andres Gil-Salcedo^{1,3}, Anne-Catherine Bachoud-Levi^{1,3}, Emmanuel Dupoux²

¹NPI/ENS/INSERM/UPEC/PSL Research University, France

²CoML/ENS/CNRS/EHESS/PSL Research University, France

³Huntington's Disease National Reference Center, Henri-Mondor Hospital, APHP, Créteil, France

hadrien.titeux@ens.fr, quangtuan.nguyen@aphp.fr

Abstract

Recent automatic speech recognition (ASR) models struggle to correctly transcribe pathological speech but implicitly capture phonetic, syntactic, and semantic properties. Unlike state-of-the-art methods that rely on speech features based on time-consuming handcrafted speech transcription, we propose a simple and fully automated approach using ASR log-probabilities to quantify intelligibility in spontaneous speech of patients with Huntington's disease. By linking this measure with clinical scores, we explore its potential as a scalable and lightweight biomarker for disease progression. Our findings suggest that ASR-derived uncertainty offers a novel, efficient, and non-invasive alternative for clinical assessment.

Index Terms: Huntington's disease, speech biomarkers, intelligibility, automatic speech recognition, pathological speech, Whisper ASR, clinical score prediction

1. Introduction

Huntington's disease (HD) is an autosomal dominant neurodegenerative disease due to CAG repeats in the mutant Htt gene [1]. This complex disease is characterized by a triad of motor, cognitive, and psychiatric symptoms leading to progressive disability [2]. With the increasing development of innovative disease-modifying therapies [3], there remains a critical need for biomarkers that are not only sensitive to disease progression and cost-effective, but also need to be ecologically valid, i.e., easily generalizable to a patient's actual functioning in their daily environment. Current HD assessment methods are based heavily on neurological examinations, time-consuming cognitive batteries [4], brain imaging [5] or invasive analysis of cerebrospinal fluid [6].

Recent studies have explored speech as a potential biomarker in HD, demonstrating the efficacy of distinguishing HD patients from controls [7], and identifying prodromal stages [8]. These approaches have shown results in the prediction of clinical outcomes, including motor, functional, and cognitive severity [9, 10, 11, 12]. [9] notably achieved good performances using a simple counting task, using mostly acoustic features, but also collateral tracks, sequence errors, and perseverations that relied on speech therapist transcriptions. Although these findings are promising, the reliance on time-consuming human transcriptions presents a significant limitation for the future automation of speech biomarkers. Automatic speech recognition (ASR) faces challenges in neurologic disorders due to increased errors, especially in dysarthric speech [13]. Although not strictly evaluated in HD, ASR can perform even worse in HD, which combines both dysarthria and linguistic impair-

ments [14, 15, 16, 17]. However, these transcription errors, while problematic in linguistic studies when a precise transcription is needed, may offer valuable insights when analyzed from a different perspective.

Several studies have used ASR-derived features to predict intelligibility [18] using the estimated probabilities of specific phonemes occurring at a given time in speech (called posterior probabilities). This methodology has been used in clinical applications, for example for the classification of neck cancer [19] or dysphonia [20]. [21] demonstrated that uncertainty in vowel predictions by an ASR was associated with clinical scores (Spearman correlations up to 0.51) and longitudinal progression of cerebellar ataxia. Different combinations of acoustic and ASR-derived features (phoneme or syllable probability) allowed a regression of the disease score (with Spearman correlations of up to 0.56) in participants with Parkinson's disease [22].

Beyond phoneme or vowel probabilities, recent ASRs such as Whisper[23] use actionable log-probabilities to indicate the probability that each token will appear in the output sequence given the context of the previous tokens. [24] correlated intelligibility with ASR uncertainty, using an approach based on decoded tokens log-probabilities. Such log-probabilities models take into account not only phonetics, but also syntax and lexico-semantic plausibility [25]. This approach could be particularly relevant for HD where speech is impaired at different levels of language (phonetics, phonology, morphology, syntax, and lexico-semantics) [26]. Our study aims to use Whisper-produced log-probabilities from spontaneous speech to predict the clinical aspects of HD, potentially offering a novel automated approach to speech-based biomarkers in this complex neurodegenerative disease.

2. Clinical database

2.1. Medical Assessment

Carriers of at least 36 CAG repeats in the Htt gene were recruited at the National Reference Center for Huntington's Disease at the Henri-Mondor Hospital in Créteil, France, in two prospective cohorts: BIOHD (NCT01412125) and CAPIT-HD beta from Repair-HD (NCT03119246). All participants signed an informed consent. The studies were conducted in accordance with the Declaration of Helsinki, current guidelines for good clinical practice, and local laws and regulations.

Certified neurologists assessed participants using the Unified Huntington's Disease Rating Scale (UHDRS [27]), which includes the Total Motor Score (TMS). We calculated the composite UHDRS (cUHDRS) as described by [28]. Functional decline was assessed using the Total Functional Capacity score (TFC) [29]. The participants' clinical characteristics are sum-

*These authors contributed equally.

marized in Table 1.

Table 1: *Participants demographics and clinical scores. Values are mean \pm standard deviation*

Number of participants	82
Sex	F44/M38
Age	52.97 \pm 10.53
CAG Repeats	43.56 \pm 2.85
cUHDRS	10.33 \pm 3.94
TMS	28.63 \pm 16.68
TFC	10.77 \pm 2.20

2.2. Voice Database

All participants completed a standardized speech battery that included (1) constrained tasks (such as counting backwards and forwards) and (2) spontaneous speech tasks, where participants were asked to narrate a sad, angry, joyful stories, the red-riding hood story, their most recent 24 hours and to describe a picture. The entire battery took less than 15 minutes and was recorded with a Zoom H4nPro (sampled at 44.1 kHz with 16-bit resolution). Using the Praat software [30] and the Seshat platform [31], speech pathologists blindly provided (1) diarization (if necessary) and (2) annotations of the linguistic content and possible linguistic anomalies (collateral tracks).

3. Features

Two sets of features were extracted from our audio corpus, **Counting** and **LogprobASR**.

3.1. Counting features

The **Counting** features set was computed from the "counting" task of our audio corpus. We used the features defined in [9]. We only used features computed from the backward counting part of the task, as they hold most of the predictive power [9]. These state-of-the-art features required careful annotations by speech-language pathologists, which allowed us to extract precise lexical and phonetic features from the participants' speech, as well as the collateral tracks (such as fillers or abnormal vocalizations).

3.2. Intelligibility features

The **LogprobASR** features set was computed from the "last 24h" spontaneous speech task, which was the most emotionally neutral task and easy to use in future clinical applications. We used the annotations from the speech pathologists to reliably obtain audio segments free of any non-participant speech. The amount of audio per participant was 73(\pm 38) seconds. We then fed those audio segments to the Whisper[23] * ASR model. This model works in two steps: it first uses its internal Voice Activity Detection (VAD) filter to re-segment the input audio into smaller segments; then, it produces a transcription for each of these smaller segments. For each transcription of segment s , Whisper also outputs the average log-probability of the transcriptions. We set the temperature parameter of the decoding to 0.0 to ensure a deterministic behavior. The formula for that log-probability is:

*Using `large-v3`, on whisper package version 20231117

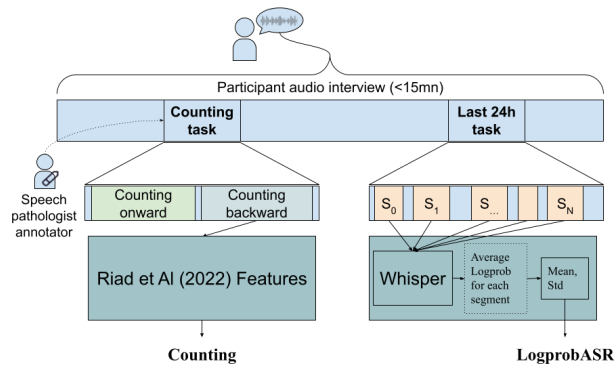


Figure 1: *Diagram of Counting and LogprobASR feature sets extraction from a single participant's interview.*

$$AvgLogProb(s) = \frac{1}{T} \sum_{i=0}^T \log p_{\theta}(x_i | x_{<i}) \quad (1)$$

Where $s = (x_0, \dots, x_T)$ is the sequence of decoded tokens. $\log p_{\theta}(x_i | x_{<i})$ is the log-likelihood of the i -th decoded token, conditioned on the previously decoded tokens ($x_{<i}$).

As "last 24h" was a spontaneous speech, there is a variable amount of decoded segments per participant speaker. We call $P = (AvgLogprob(s_0), \dots, AvgLogProb(s_n))$ the array of all these average log-probabilities for a given participant. To obtain a fixed-length array of features for each participant, we pooled the values of P by computing their mean and standard deviation. In short, for a participant P :

$$LogprobASR(P) = (mean(P), std(P)) \quad (2)$$

4. Methods

4.1. Statistical analysis

The patient sample for analysis consisted of 82 participants who met two criteria: (1) completion of both the "last 24h" task and the "counting" task, and (2) availability of annotated data from the "counting" task required for the **Counting** feature extraction. To investigate the relationship between the extracted audio features and clinical scores, we assessed the correlation between the **LogprobASR**-mean feature and the clinical outcomes (cUHDRS, TMS, and TFC). We employed both Pearson and Spearman correlation coefficients to capture different aspects of these relationships. Pearson's correlation was used to measure linear associations, assuming normally distributed data, while Spearman's rank correlation provided a non-parametric assessment, capturing potential monotonic but non-linear relationships.

4.2. Feature Sets

We considered three distinct feature sets to predict clinical scores: **LogprobASR**, **Counting**, and **Demog**. The **Demog** feature set included three demographic variables: age, sex, and number of CAG triplets (which is obtained when HD is confirmed with the Htt mutation). In addition to evaluating each feature set individually, we explored various combinations of feature sets to assess whether their integration could improve or degrade predictive performance. By comparing

Figure 2: Chart of MAE prediction results for cUHDRS, TMS and TFC. ”+” indicates a combination of two feature sets. All predictions are done using linear regression, except for feature sets containing **Counting**. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$ after Bonferroni correction and is measured versus MAE values from **Demog**. Bars around mean values indicate the 95% confidence interval.

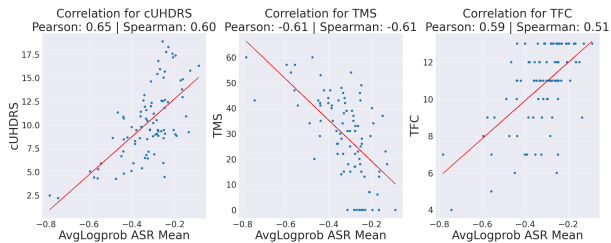
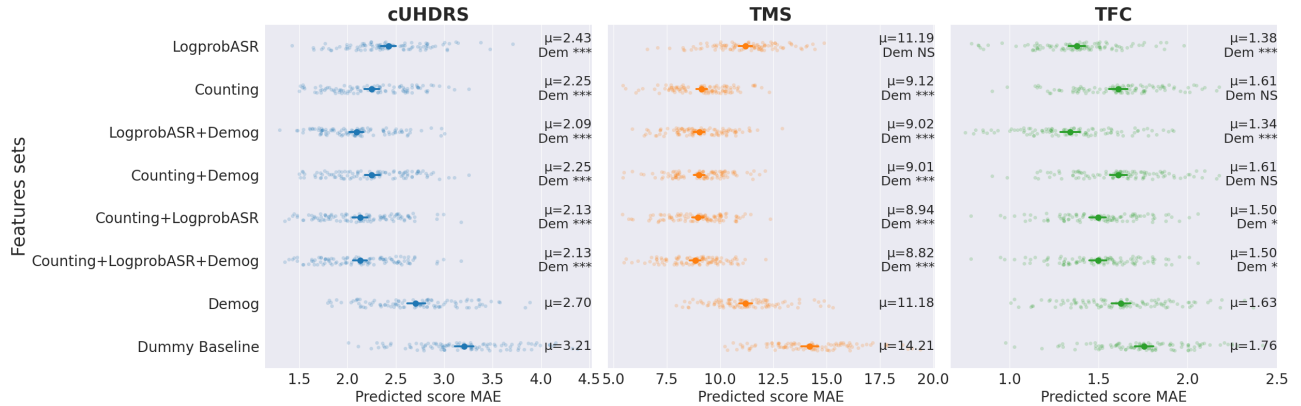


Figure 3: Chart for correlations between **LogprobASR-mean** and clinical scores. The red line is a linear regression.

models trained on single feature sets versus combinations (e.g., **LogprobASR+Demog**, **Counting+Demog**, or all three combined), we aimed to identify potential synergies or redundancies between features. This approach allows us to determine whether complementary information from different sources enhances prediction accuracy.

To establish a ”Dummy Baseline” for model performance, we also used a dummy model that predicts the mean of the target variable, allowing us to compare the predictive value of each feature set against a simple, non-informative benchmark.

4.3. Models

We employed different machine learning models based on the nature of each feature set. For **LogprobASR** and **Demog**, we used linear regression due to its simplicity and efficiency, particularly effective when dealing with low-dimensional data. In the case of **LogprobASR**, which contains only two features, linear regression minimizes the risk of overfitting without the need for complex regularization techniques.

For the **Counting** feature set, we opted for ElasticNet regression to remain consistent with the methodology of [10, 9]. ElasticNet, which combines L1 and L2 regularization, is better suited for handling **Counting**’s larger set of features.

We used scikit-learn’s [32] implementations of both models for training and testing. To ensure robustness against potential outliers and differences in feature scales, we applied the `RobustScaler` class from scikit-learn to all individual features.

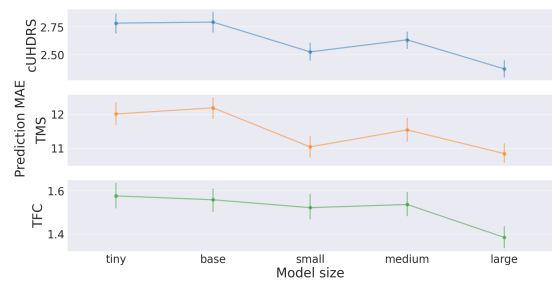


Figure 4: MAE of clinical score predictions using the **LogprobASR** feature set across different Whisper model sizes, over 100 folds with 80/20 splits with Linear Regression. Error bars represent standard deviations across folds.

4.4. Prediction Assessment

Model performance was evaluated using Mean Absolute Error (MAE) and the coefficient of determination (R^2) for each clinical score. We conducted a 100-fold cross-validation procedure with an 80/20 train/test split in each fold. We used a t-test to compare the MAE of the **Demog** set to the six other feature sets. P-values were corrected for multiple comparisons for the six feature set and three clinical scores using bonferroni correction. For each combination of feature sets and models, we report the mean and standard deviation of MAE and R^2 across the 100 folds in Table 2.

4.5. Post-hoc analysis

In a post-hoc analysis, we aimed to evaluate the variations in clinical score prediction performance across different Whisper model sizes. To achieve this, we repeated our previous machine learning protocol using **LogprobASR** for each clinical score and each model size. This allowed us to assess the impact of model complexity on prediction accuracy and determine whether larger models provided any advantage over smaller ones. We report the MAE variations in Figure 4.

5. Results & Discussion

ASR log-probabilities (Figure 3) strongly correlated to cUHDRS (Pearson 0.65, Spearman 0.60), TMS (Pearson -0.61, Spearman -0.61), and TFC (Pearson 0.59, Spearman 0.51).

Table 2: MAE and R2 of predictions over 100 folds. Values are the mean over 100 folds \pm standard deviation. Lowest MAE and highest R2 are bolded. The **Counting** feature set baseline was the previous state-of-the-art.

Features sets	cUHDRS MAE \downarrow	R2 \uparrow	TMS MAE \downarrow	R2 \uparrow	TFC MAE \downarrow	R2 \uparrow
Dummy Baseline	3.21 \pm 0.52	-0.10 \pm 0.13	14.21 \pm 2.03	-0.09 \pm 0.14	1.76 \pm 0.27	-0.13 \pm 0.23
Demog	2.70 \pm 0.47	0.16 \pm 0.21	11.18 \pm 1.58	0.29 \pm 0.18	1.63 \pm 0.29	-0.02 \pm 0.35
Counting[9]	2.25 \pm 0.41	0.46 \pm 0.14	9.12 \pm 1.41	0.52 \pm 0.15	1.61 \pm 0.26	0.04 \pm 0.20
Counting+Demog	2.25 \pm 0.41	0.46 \pm 0.14	9.01 \pm 1.37	0.54 \pm 0.14	1.61 \pm 0.26	0.04 \pm 0.20
LogprobASR	2.43 \pm 0.41	0.35 \pm 0.19	11.19 \pm 1.61	0.28 \pm 0.19	1.38 \pm 0.24	0.24 \pm 0.27
LogprobASR+Demog	2.09 \pm 0.36	0.49 \pm 0.20	9.02 \pm 1.34	0.53 \pm 0.16	1.34 \pm 0.27	0.25 \pm 0.32
Counting+LogprobASR	2.13 \pm 0.40	0.51 \pm 0.13	8.94 \pm 1.36	0.54 \pm 0.14	1.50 \pm 0.24	0.16 \pm 0.19
Counting+LogprobASR+Demog	2.13 \pm 0.40	0.51 \pm 0.13	8.82 \pm 1.33	0.56 \pm 0.14	1.50 \pm 0.24	0.16 \pm 0.19

We replicated the MAE reported by [9] using the same **Counting** feature set for each of the clinical scores, with all the clinical predictions superior to **Demog**.

The **LogprobASR** feature set demonstrated comparable cUHDRS and TFC predictions to the **Counting** feature set (Figure 2), proving the potential of Whisper log-probabilities in HD assessment. This approach aligns with several studies that have employed similar methods based on intelligibility for clinical predictions [18, 19, 20, 22, 21]. Our methodology can be conceptualized within the framework of human psycholinguistics as a hierarchical predictive model, as proposed by [33]. In this model, language comprehension is based on context encoded at various levels of internal representation, ranging from phonology to semantics. Whisper’s ability to indicate the likelihood of tokens (not just acoustic phonemes or syllables) potentially offers a wider range of internal representation levels, thus capturing a broader syntax and lexico-semantic context [25]. These log-probabilities could be particularly valuable in complex neurological diseases like HD, where intelligibility is affected not only by motor but also by cognitive or linguistic impairment.

In detail (Table 2), we observed a superior performance of the combination of **LogprobASR** with simple **Demog** features, particularly for cUHDRS and TFC prediction where **LogprobASR+Demog** outperformed **Counting+Demog** and **LogprobASR+Counting**. This synergy suggests a potential influence of participant characteristics such as sex, age, and CAG repeat length on **LogprobASR** performance, consistent with previous HD studies [34] showing associations between age, CAG repeats, and disease severity. Moreover, age and sex may influence voice characteristics, and thus ASR predictions. Thus, incorporating demographic information could enhance the predictive power of **LogprobASR**, possibly by adjusting for individual variations in HD-related speech patterns. These findings underscore the importance of including demographic features in future automatic speech models for HD.

Interestingly, the **Counting** feature set provided better TMS prediction than **LogprobASR**, with performance further enhanced by combining **Counting** and **LogprobASR** features. This discrepancy likely stems from the more acoustic nature of **Counting**, which may be more closely related to dysarthria (mainly a speech motor disorder) and therefore more related to the motor aspects assessed by TMS. **Counting** [9] should still be used preferentially when the evaluation concerns motor aspects. In contrast, none of the combinations using **Counting** features performed better than **LogprobASR+Demog** to predict cUHDRS and TFC. This suggests that cUHDRS, a composite general score, and TFC, which assesses daily functional autonomy, might encompass a broader range of linguistic or

cognitive functions beyond motor skills. These findings indicate that some general HD scores could be influenced by a more complex interplay of neurological factors. In such cases **Counting** from [9] should be avoided in favor of Whisper log-probabilities calculated from a spontaneous speech task.

The log-probabilities are inherently dependent on both (1) the size of Whisper model and (2) its built-in Voice Activity Detection (VAD). Regarding (1), smaller model size could not achieve comparable predictive performance across all the three clinical scores (Figure 4). For (2), we replicated our analysis using our human-annotated VAD boundaries, yielding to only marginal predictive performance gains (which are not reported in this paper). This suggested that the predictions are primarily driven by the log-probabilities themselves, and not critically biased by the choice of Whisper’s VAD boundaries.

6. Conclusion & Future Work

Our study introduced a novel approach to predicting clinical scores in Huntington’s disease (HD) using Automatic Speech Recognition (ASR) log-probabilities derived from the Whisper model. This method, which estimates speech intelligibility, has shown promising results in predicting key clinical measures including cUHDRS, TMS, and TFC, particularly when combined with demographic characteristics such as age, sex, and CAG repeat count. The use of Whisper log-probabilities offers several advantages for clinical assessment in HD as it provides (1) an off-the-shelf solution for automatic analysis of spontaneous speech, and (2) captures a broader range of features, including syntax and lexico-semantic context, which may be particularly relevant in complex neurodegenerative disorders like HD.

Whisper log-probabilities are an off-the-shelf solution and our approach demonstrates promising results for automatic spontaneous speech analyses; however, its applicability remains limited to this specific ASR. In future work, we plan to evaluate whether similar predictive performance can be achieved with alternative ASR models. This will determine the generalizability and robustness of logprob-derived features for clinical score prediction in Huntington’s disease.

7. References

- [1] M. Macdonald, “A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes,” *Cell*, vol. 72, no. 6, pp. 971–983, Mar. 1993.
- [2] M. J. U. Novak and S. J. Tabrizi, “Huntington’s disease,” *BMJ*, vol. 340, no. jun30 4, pp. c3109–c3109, Jun. 2010.
- [3] S. J. Tabrizi, C. Estevez-Fraga, W. M. C. Van Roon-Mom, M. D. Flower, R. I. Scahill, E. J. Wild, I. Muñoz-Sanjuán, C. Sampaio,

- A. E. Rosser, and B. R. Leavitt, "Potential disease-modifying therapies for Huntington's disease: lessons learned and future opportunities," *The Lancet Neurology*, vol. 21, no. 7, pp. 645–658, Jul. 2022.
- [4] J. C. Stout, S. Queller, K. N. Baker, S. Cowlshaw, C. Sampaio, C. Fitzer-Attas, B. Borowsky, and the HD-CAB Investigators, "HD-CAB: A cognitive assessment battery for clinical trials in Huntington's disease^{1,2,3}," *Movement Disorders*, vol. 29, no. 10, pp. 1281–1288, Sep. 2014.
- [5] S. L. Mason, R. E. Daws, E. Soreq, E. B. Johnson, R. I. Scahill, S. J. Tabrizi, R. A. Barker, and A. Hampshire, "Predicting clinical diagnosis in Huntington's disease: An imaging polymarker," *Annals of Neurology*, vol. 83, no. 3, pp. 532–543, Mar. 2018.
- [6] R. I. Scahill, P. Zeun, K. Osborne-Crowley, E. B. Johnson, S. Gregory, C. Parker, J. Lowe, and et al., "Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease Young Adult Study (HD-YAS): a cross-sectional analysis," *The Lancet Neurology*, vol. 19, no. 6, pp. 502–512, Jun. 2020.
- [7] 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.
- [8] T. Kouba, W. Frank, T. Tykalova, A. Mühlbäck, J. Klempf, K. S. Lindenberg, G. B. Landwehrmeyer, and J. Rusz, "Speech biomarkers in Huntington's disease: A cross-sectional study in pre-symptomatic, prodromal and early manifest stages," *Euro J of Neurology*, vol. 30, no. 5, pp. 1262–1271, May 2023.
- [9] R. Riad, M. Lunven, H. Titeux, X.-N. Cao, J. Hamet Bagnou, L. Lemoine, J. Montillot, A. Sliwinski, K. Youssov, L. Cleret De Langavant, E. Dupoux, and A.-C. Bachoud-Lévi, "Predicting clinical scores in Huntington's disease: a lightweight speech test," *J Neurol*, vol. 269, no. 9, pp. 5008–5021, Sep. 2022.
- [10] 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.
- [11] M. Perez, A. Romana, A. Roberts, N. Carlozzi, J. A. Miner, P. Dayalu, and E. M. Provost, "Articulatory Coordination for Speech Motor Tracking in Huntington Disease," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 1409–1413.
- [12] A. S. Nunes, M. Pawlik, R. K. Mishra, E. Waddell, M. Coffey, C. G. Tarolli, R. B. Schneider, E. R. Dorsey, A. Vaziri, and J. L. Adams, "Digital assessment of speech in Huntington disease," *Front. Neurol.*, vol. 15, p. 1310548, Jan. 2024.
- [13] B. G. Schultz, V. S. A. Tarigoppula, G. Noffs, S. Rojas, A. Van Der Walt, D. B. Grayden, and A. P. Vogel, "Automatic speech recognition in neurodegenerative disease," *Int J Speech Technol*, vol. 24, no. 3, pp. 771–779, Sep. 2021.
- [14] 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, p. e113412, Nov. 2014.
- [15] A. P. Vogel, C. Shirbin, A. J. Churchyard, and J. C. Stout, "Speech acoustic markers of early stage and prodromal Huntington's disease: A marker of disease onset?" *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, Dec. 2012.
- [16] W. Hinzen, J. Rosselló, C. Morey, E. Camara, C. Garcia-Gorro, R. Salvador, and R. De Diego-Balaguer, "A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage Huntington's disease," *Cortex*, vol. 100, pp. 71–83, Mar. 2018.
- [17] M. Teichmann, E. Dupoux, P. Cesaro, and A.-C. Bachoud-Lévi, "The role of the striatum in sentence processing: Evidence from a priming study in early stages of Huntington's disease," *Neuropsychologia*, vol. 46, no. 1, pp. 174–185, Jan. 2008.
- [18] M. Karbasi and D. Kolossa, "ASR-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, p. 108606, Dec. 2022.
- [19] C. Fredouille, A. Ghio, I. Laaridh, M. Lalain, and V. Woisard, "Acoustic-phonetic decoding for speech intelligibility evaluation in the context of Head and Neck Cancers."
- [20] M. G. Tulics, G. Szaszak, K. Meszaros, and K. Vicsi, "Using ASR Posterior Probability and Acoustic Features for Voice Disorder Classification," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. Mariehamn, Finland: IEEE, Sep. 2020, pp. 000 155–000 160.
- [21] D. Y. Isaev, R. M. Vlasova, J. M. Di Martino, C. D. Stephen, J. D. Schmahmann, G. Sapiro, and A. S. Gupta, "Uncertainty of Vowel Predictions as a Digital Biomarker for Ataxic Dysarthria," *Cerebellum*, vol. 23, no. 2, pp. 459–470, Apr. 2023.
- [22] A. Zlotnik, J. M. Montero, R. San-Segundo, and A. Gallardo-Antolín, "Random forest-based prediction of Parkinson's disease progression using acoustic, ASR and intelligibility features," in *Interspeech 2015*. ISCA, Sep. 2015, pp. 503–507.
- [23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *International Conference on Machine Learning*, 2022.
- [24] Z. Tu, N. Ma, and J. Barker, "Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction," *INTERSPEECH*, 2022.
- [25] N. Ballier, A. Méli, M. Amand, and J.-B. Yunès, "Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English."
- [26] C. Jacquemot and A.-C. Bachoud-Lévi, "Striatum and language processing: Where do we stand?" *Cognition*, vol. 213, p. 104785, Aug. 2021.
- [27] "Unified Huntington's disease rating scale: Reliability and consistency," *Movement Disorders*, vol. 11, no. 2, pp. 136–142, Mar. 1996.
- [28] S. A. Schobel, G. Palermo, P. Auinger, J. D. Long, S. Ma, O. S. Khwaja, D. Trundell, M. Cudkowicz, S. Hersch, C. Sampaio, E. R. Dorsey, B. R. Leavitt, K. D. Kiebertz, J. J. Seigny, D. R. Langbehn, S. J. Tabrizi, and For the TRACK-HD, COHORT, CARE-HD, and 2CARE Huntington Study Group Investigators, "Motor, cognitive, and functional declines contribute to a single progressive factor in early HD," *Neurology*, vol. 89, no. 24, pp. 2495–2502, Dec. 2017.
- [29] I. Shoulson, "Huntington disease: Functional Capacities in patients treated with neuroleptic and antidepressant drugs," *Neurology*, vol. 31, no. 10, pp. 1333–1333, Oct. 1981.
- [30] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT," vol. 5, no. 9, 2001.
- [31] H. Titeux, R. Riad, X.-N. Cao, N. Hamilakis, K. Madden, A. Cristia, A.-C. Bachoud-Levi, and E. Dupoux, "Seshat: A tool for managing and verifying annotation campaigns of audio data."
- [32] 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.
- [33] G. R. Kuperberg and T. F. Jaeger, "What do we mean by prediction in language comprehension?" *Lang. Cogn. Neurosci.*, vol. 31, no. 1, pp. 32–59, 2016.
- [34] S. J. Tabrizi, R. I. Scahill, G. Owen, A. Durr, B. R. Leavitt, R. A. Roos, B. Borowsky, B. Landwehrmeyer, C. Frost, H. Johnson, D. Craufurd, R. Reilmann, J. C. Stout, D. R. Langbehn, and TRACK-HD Investigators, "Predictors of phenotypic progression and disease onset in premanifest and early-stage huntington's disease in the TRACK-HD study: analysis of 36-month observational data," *Lancet Neurol.*, vol. 12, no. 7, pp. 637–649, Jul. 2013.