HHS Public Access

Author manuscript

IEEE Access. Author manuscript; available in PMC 2024 November 27.

Published in final edited form as:

IEEE Access. 2024 ; 12: 62328–62340. doi:10.1109/access.2024.3393243.

Sensitive Quantification of Cerebellar Speech Abnormalities Using Deep Learning Models

KYRIAKOS VATTIS[1,2], **BRANDON OUBRE**[1,2], **ANNA C. LUDDY**[1], **JESSEY S. OUILLON**[1], **NICOLE M. EKLUND**[1], **CHRISTOPHER D. STEPHEN**[1,2,3], **JEREMY D. SCHMAHMANN**[1,2,3], **ADONAY S. NUNES**[1,2], **ANOOPUM S. GUPTA**[1,2,3]

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA

[2]Harvard Medical School, Boston, MA 02115, USA

[3]Department of Neurology, Ataxia Center, Massachusetts General Hospital, Boston, MA 02114, USA

Abstract

Objective, sensitive, and meaningful disease assessments are critical to support clinical trials and clinical care. Speech changes are one of the earliest and most evident manifestations of cerebellar ataxias. This work aims to develop models that can accurately identify and quantify clinical signs of ataxic speech. We use convolutional neural networks to capture the motor speech phenotype of cerebellar ataxia based on time and frequency partial derivatives of log-mel spectrogram representations of speech. We train classification models to distinguish patients with ataxia from healthy controls as well as regression models to estimate disease severity. Classification models were able to accurately distinguish healthy controls from individuals with ataxia, including ataxia participants who clinicians rated as having no detectable clinical deficits in speech. Regression models produced accurate estimates of disease severity, were able to measure subclinical signs of ataxia, and captured disease progression over time. Convolutional networks trained on time and frequency partial derivatives of the speech signal can detect sub-clinical speech changes in ataxias and sensitively measure disease change over time. Learned speech analysis models have the potential to aid early detection of disease signs in ataxias and provide sensitive, low-burden assessment tools in support of clinical trials and neurological care.

Corresponding author: Anoopum S. Gupta (agupta@mgh.harvard.edu).

AUTHOR CONTRIBUTION

Kyriakos Vattis and Anoopum S. Gupta contributed to the conception and design of the study; Anna C. Luddy, Jessey S. Ouillon, Nicole M. Eklund, Christopher D. Stephen, Jeremy D. Schmahmann, and Anoopum S. Gupta contributed to data acquisition; Kyriakos Vattis, Adonay S. Nunes, and Anoopum S. Gupta contributed to data analysis; Brandon Oubre reviewed analyses, corrected errors, and anonymized data for public release; Kyriakos Vattis, Brandon Oubre, and Anoopum S. Gupta contributed to drafting the text and figures; and all authors revised the manuscript for intellectual content.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at Massachusetts General Hospital under Protocol No. 2016P001048 (7/11/2016), No. 2019P002752 (1/2/2020), and No. 2019P003458 (4/2/2020).

COMPETING INTERESTS

The authors declare no competing interests.

Keywords

Ataxia; deep learning; speech

I. INTRODUCTION

Cerebellar ataxias (CA) are disorders associated with impaired function of the cerebellum. People with CA often experience clumsiness, unsteady gait, and slurred speech. The etiology of CA is heterogeneous, though many associated diseases are neurodegenetive. Progression can vary by both etiology and individual. CA can be hereditary—such with ataxia-telangiectasia (AT), spinocerebellar ataxias (SCAs), and Friedreich's ataxia (FRDA)—or acquired—such as with idiopathic late-onset cerebellar ataxia (ILOCA) and multiple system atrophy (MSA). Dominant hereditary CAs have an average prevalance of $2.7/10^5$, with SCA type 3 (SCA-3) being the most common dominant ataxia. Autosomal recessive CAs have an average prevelance of $3.3/10^5$, with FRDA being the most frequent followed by AT [1].

Initial diagnosis and monitoring of disease progression currently rely on neurologist-performed assessments of motor and cognitive behavior. Clinical rating scales—such as the Scale for the Assessment and Rating of Ataxia (SARA) [2], the International Cooperative Ataxia Rating Scale (ICARS) [3], and the Brief Ataxia Rating Scale (BARS) [4]—are structured, semi-quantitative scales used to track disease progression in natural history studies and clinical trials to determine the efficacy of new therapies. Unfortunately, these assessments have several limitations: they are subjective, depend on the rater's clinical experience, and are limited by human perception. These scales are necessarily discrete and coarse to achieve adequate intra- and inter-rater reliability. Furthermore, as they depend on a trained clinician and are typically performed in-person, these assessments have limited accessibility and cannot be performed frequently to account for short-term fluctuations in symptoms [5], [6]. These factors contribute to the challenge of measuring disease change, and thereby to long and large clinical trials, which increases costs and places high burden on patients. Finally, these tools are inadequate for assessing early stages of disease, which may be the most impactful time to intervene in order to prevent neurodegeneration, but is also when signs are least detectable.

Quantitative phenotyping attempts to address this gap through the use of technology to quantify important aspects of motor and cognitive behavior [7]. There has been increasing interest in the use of inertial measurement units (IMUs) that record accelerometer and gyroscope data to quantify gait or limb movement [8], [9], [10], [11], [12], [13] features, computer mouse tasks that assess arm motor control [14], and eye tracking devices to quantify eye movement abnormalities [15], [16]. Such data can be used to generate granular descriptions of disease phenotypes and severity, which can be extended to the home setting for accessible and frequent assessments.

Speech is a promising source for quantitative behavioral biomarkers in neurological conditions. Analysis of speech has been widely used to detect and quantify the severity of Parkinson's Disease using machine learning techniques such as support vector machines

and random forests [17], [18], [19], k-nearest neighbors [20], parallel neural networks [21], and gradient boosting classifiers [22] applied to acoustic speech features. Deep learning approaches, such as bidirectional long-short term memory models [23], have also been proposed for analysis of parkinsonian speech. Related work in multiple system atrophy [24] and multiple sclerosis [25], [26] identified quantifiable characteristics of speech digression and their potential usage for severity estimation. Mixed-effects models have also been proposed to analyze speech in the context of amyotrophic lateral sclerosis (ALS) [27].

Speech changes have been extensively characterized in CA [28], [29], [30], [31] and are an early disease feature of SCAs [32]. Cerebellar dysfunction results in many changes in acoustic properties of speech, including long-term variability of the fundamental frequency, shimmer, peak amplitude variation, jitter, slowed speech rates, increased utterance duration, and slowed and irregular alternating syllable rates [33], [34]. Computational analysis of speech in ataxias has been relatively limited. Kashyap, et al. published a series of works with promising results based on conventional machine-learned modeling of acoustic features of the tongue-twister phrase “British Constitution” [35] and the “ta-ta-ta” syllable repetition task [36]. Furthermore, using the same tasks, a multivariate mixture extension of the generalized linear mixed model and cluster analysis was used to model the progression of speech deficits [37]. Song et al. were also able to distinguish between hypokinetic dysarthria, CA, and healthy controls using convolutional networks applied to audio waveforms of passage reading and counting tasks [38].

Given the acoustic and temporal signatures of ataxic speech, we hypothesized that convolutional models trained on the time and frequency gradients of log-mel spectrograms would learn useful representations for detecting ataxia and estimating its severity, including in individuals without clinically apparent deficits in their speech. This manuscript is structured as follows: Sec. II describes the study population, the collected data, data pre-processing, and model training and validation. Sec. III presents the performance of models trained to distinguish healthy and ataxic speech and to estimate disease severity. We also attempt to provide insight into how the model works using integrated gradients. Finally, in Sec. IV, we discuss our results in relation to other works and propose future research directions.

II. METHODS

A. DATA COLLECTION AND PARTICIPANTS

Both in-person and at-home data were collected from 228 unique participants across 463 sessions. A total of 203 of the sessions corresponded to at-home data collection, where data were sampled from the built-in microphone of a laptop provided to the participant or from the participant’s home computer or mobile device. The remaining 260 sessions were collected in-person, where data collection was supervised by study personnel and data were sampled from a lapel-attached microphone connected to an iPad. A total of 157 out of 228 participants had a CA diagnosis; the remaining 71 participants were neurologically healthy controls. Ataxia diagnoses included spinocerebellar ataxia (5 SCA-1, 3 SCA-2, 21 SCA-3, 7 SCA-6, 10 other SCAs), ataxia-telangiectasia (45), multiple system atrophy (9), autosomal recessive cerebellar ataxia (3), Friedreich’s Ataxia (4) and a heterogeneous distribution of

other types of ataxia (50). Table 1 summarizes the demographics of the study population. Participants performed a variety of speech tasks with audio recorded at 44.1 kHz. For this analysis, only audio from repetitive utterances of the syllables “la-la-la”, “go-go-go”, and “me-me-me” were used. These tasks have previously been used for clinical evaluation of ataxia [2], [3], [4] and were chosen to isolate key phenotypic features, including speech rate, articulatory variability, naturalness, and intelligibility [26]. All participants provided written assent and/or consent, and the study was approved by the Institutional Review Board at Massachusetts General Hospital with protocol numbers of 2016P001048 (7/11/2016), 2019P002752 (1/2/2020) and 2019P003458 (4/2/2020).

B. PRE-PROCESSING OF SPEECH DATA

Audio recordings were first downsampled to a rate of 8 kHz. A noise reduction algorithm based on spectral gating [39], [40] was then used to estimate a noise threshold for each frequency band and then filter detected noise from the signal. Mel-frequency spectrograms, which represent the spectrum of signal frequencies as a function of time, were then computed using Librosa [41]. Spectrograms were constructed by segmenting the audio signal into 128 ms overlapping windows with a step of 20 ms and applying a short-time Fourier transform to each window. The resulting spectrograms were passed through a mel-space frequency transformation resulting in 128 frequency bins. Bin magnitudes were log-transformed. Fig. 1 illustrates an example spectrogram.

Each spectrogram was then split into non-overlapping 1 s frames (i.e., each frame was a 51 × 128 matrix), with the last frame zero-padded in the time dimension. The WebRTC Voice Activity Detector was then used to assign a score in the range of 0–1 to each frame, defined as the fraction of instances (i.e., time bins) in which voice was detected. Frames with a voice activity score of less than 0.6 were regarded as noise and discarded from further analysis. Finally, all frames were globally re-scaled to a range of 0–1 and partial derivatives were calculated along the time and frequency dimensions.

C. ATAXIA CLASSIFICATION

One objective of this work was to create a model to distinguish between participants with ataxia and controls. The repetitive speech tasks used in this study were expected to create distinct patterns in the frequency-time space, as seen in Fig. 1. These patterns can occur in multiple places within the sample and patterns that are signs of CA should be present across samples from different individuals. We chose to employ convolutional networks as translational invariance could take advantage of the repetitive signal.

We trained classification models using two different architectures: 1) *ResNet* 18 [42], and 2) a simple, three-layer convolutional network. *ResNet* 18 is a Deep Residual Network utilizing two-dimensional (2D) convolutions. The simple convolutional network also used 2D convolutions, but did not include residual connections. Both models treated spectrogram or single partial derivative inputs analogously to monochrome images. When models were configured to use both time and frequency partial derivatives as inputs, the derivatives were concatenated along the channel dimension. (I.e., such inputs were treated analagously to two-color images, with each color corresponding to a different partial derivative.) Therefore,

the 2D convolutions used in both models were applied to the frequency-time space and could learn features that combined information across both dimensions. Both models were trained using *Adam* [43] to optimize the Cross Entropy Loss. Model outputs were the probability that the input frame originated from a participant with ataxia, P(Ataxia). Model checkpoints were saved throughout the training process, and models with parameters that minimized the loss across one epoch on the validation dataset were chosen for performance estimations.

D. SEVERITY ESTIMATION

A second objective of this work was to create a model to estimate ataxia severity, as measured by the BARS total score ($BARS_{total}$), which ranges from 0–30, and the speech component of BARS ($BARS_{speech}$), which ranges from 0–4. A modified version of BARS incorporating half-points [44] was used and higher numbers represent more severe clinical signs. $BARS_{total}$ depends on assessments of several motor domains, including eye movements (e.g., saccades and smooth pursuit), limb reaching movements, and gait. The total score was used as a training label because aggregate scores tend to be more robust to errors and provide greater resolution. In contrast, $BARS_{speech}$ more closely reflects performance on the considered tasks. Individuals with natural speech and rapid consonant production are given a $BARS_{speech}$ score of 0. Individuals with consonant production irregularities are scored as 0.5. Mildly slurred speech with all words being intelligible corresponds to a score of 1. $BARS_{speech}$ scores then increase as speech becomes more slurred and less intelligible, with a score of 4 corresponding to absent or unintelligible speech.

Similarly to the models trained for classification, a *ResNet* 18 model with a single output was trained to estimate ataxia severity. The model was trained to optimize the Mean Squared Error loss using *Adam* [43]. The model with parameters that minimized the loss across one epoch on the validation dataset was chosen for performance estimations.

E. DATA AUGMENTATION

A combination of data augmentation techniques were applied to the training data for the classification task. These augmentations included 1) under sampling the majority class (ataxia participants) to balance the data set, 2) linear mix-up [45], and 3) cropping along the time dimension to give the illusion of slower speech. Only cropping was used when training the regression model. Finally, all inputs—including the validation samples that did not undergo the aforementioned augmentations—were resized to 100 × 100 matrices to accommodate the large size of *ResNet* 18. For training, each frame was considered as an independent sample. For all validation purposes, the outputs corresponding to frames originating from the same participant during the same session were aggregated such that their median value was considered as the single model output.

F. CROSS VALIDATION

Classification model performance was validated using a 5-fold participant-based (as opposed to frame-based) cross validation procedure. In each fold of cross validation, the data were randomly split into a training data set containing 80% of the ataxia participants and 80% of the controls, while the rest were reserved as a testing set. By repeating this procedure five

times, all participants appeared in the testing set during one of the five folds. Splitting the data based on the participant instead of the frame ensured that frames originating from the same individual could not be included in both the training and testing set for any given fold. In other words, this form of cross validation simulated the application of trained models to previously unobserved subjects, thereby ensuring that the reported performance should generalize to similar populations. The downside of this approach was that the training data set size was not exactly the same across all folds since each participant's task duration varied. Reported performance metrics were averaged over each fold and plots display the pooled data from all folds. The performance of regression models were similarly evaluated using 10-fold participant-based cross validation to ensure a representative distribution of BARS scores within the training data.

G. PERFORMANCE EVALUATION

Classification model performance was assessed using the area under the ROC curve (AUC) and the class-weighted F1 score. Regression model performance was assessed using the mean absolute error (MAE) and coefficient of determination (R^2). Additional analyses (described below) were performed for the best-performing *ResNet18* model for each task.

Additional analyses for the best-performing classification model included a Mann-Whitney U-test on $P(Ataxia)$ to assess the ability of the classification model to distinguish between controls and ataxia participants without clinically observed speech deficits (i.e., $BARS_{speech}^{clin} = 0$). Mann-Whitney U-tests were also performed separately for controls and ataxia participants to compare model performance between sexes. Spearman's correlation (ρ) was used to evaluate the relationship between $P(Ataxia)$ and participant age for participants with ataxia. To determine the classification's dependence on ataxia severity, Mann-Whitney U-tests were used to compare $P(Ataxia)$ between ataxia participants with $BARS_{total} < 15$ and those with $BARS_{total} >= 15$. A similar comparison was done for $BARS_{speech} < 1.5$ versus $BARS_{speech} >= 1.5$.

As with the classification model, additional regression analyses included a Mann-Whitney U-test on $BARS_{speech}^{pred}$ to assess the speech score regression model's ability to distinguish between controls and ataxia participants without clinically observed speech deficits. Another Mann-Whitney U-test on $BARS_{speech}^{pred}$ was used to determine the model's sensitivity to mild speech severity by comparing ataxia participants with $BARS_{speech}^{clin} = 0$ and $BARS_{speech}^{clin} = 0.5$. Mann-Whitney U-tests were also performed separately for controls and ataxia participants on the absolute errors of $BARS_{speech}^{pred}$ and $BARS_{total}^{pred}$ to compare model performance between sexes. Spearman's correlation (ρ) was used to evaluate the relationship between the absolute error of model outputs and participant age. Finally, one-sample *t*-tests between subsequent visits were used to evaluate the regression models' sensitivity to disease progression for ataxia participants who participated in the study multiple times. Progression was further assessed using Spearman's correlation (ρ) between the changes in model estimates versus elapsed time.

Effects sizes for Mann-Whitney U-tests are reported in terms of the common language effect size (f), which is equivalent to AUC. Rank-biserial correlation (r) can be derived from the

common language effect size using the formula $r = 2f - 1$. Effect sizes for one-sample t-tests are reported using Cohen's d.

H. MODEL INTERPRETABILITY

Integrated Gradients (IG) [46] were used to better understand the speech information used by the models to make predictions. IG were calculated for correctly classified tests for the classification model. One set of IG were obtained for each input component (i.e., time derivatives and frequency derivatives) and were studied separately. The absolute values of the IGs were aggregated once along the frequency dimension and once along the time dimension to obtain a salience score for each bin in time and frequency, respectively. To visualize the IGs with respect to a speech sample, Principal Component Analysis (PCA) was performed on each mel spectrogram. The first principal component, which captured the oscillatory behavior of the repeated syllable speech task, was used to set boundaries for single syllables (see Fig. 7).

An analysis was then performed to determine if certain syllable components (in time and frequency) were preferentially informative in generating model predictions. Starting with the frequency-aggregated IGs, the temporal location of the three instances with the three highest salience scores were identified and their temporal position was 0–1 normalized with respect to the syllable endpoints. A similar procedure was applied to the time-aggregated IGs, although the locations of the three highest peaks were identified in frequency space, with range of 1–100.

III. RESULTS

Table 2 summarizes model performance across a variety of tasks (classification of control vs. ataxia participants, $\text{BARS}_{\text{total}}$ regression, and $\text{BARS}_{\text{speech}}$ regression), model architectures, and model inputs.

A. ATAXIA CLASSIFICATION

Of the experiments in Table 2, a model with both frequency and time partial derivatives of the mel spectrogram as model inputs performed the best at classification of controls vs. ataxia participants. Results reported in this subsection are based on this model. The AUC was 0.89±0.03 across all cross validation folds (Fig. 2). Fig. 3 illustrates the relationship of $\text{BARS}_{\text{total}}$ and $\text{BARS}_{\text{speech}}$ with $P(Ataxia)$. Each marker represents whether the participant was a control (blue) or had ataxia (red), was male (dot) or female (cross), and the participant's age (marker size). The dotted line demonstrates that an operating threshold of P(Ataxia) > 0.6 well-separated both groups. Table 3 shows the normalized confusion matrix corresponding to this threshold. The associated sensitivity, specificity, and class-weighted F1 scores were 0.88 ± 0.08, 0.87 ± 0.02, and 0.86 ± 0.02 across all validation folds, respectively.

A Mann-Whitney U-test between controls and ataxia participants with $BARS_{speech} = 0$ was significant ($f = 0.80$, $p = 2 \times 10^{-7}$), providing strong evidence that the model captured ataxic speech features that were not detected by the neurologists' clinical assessments. Furthermore, the median $P(Ataxia)$ was 0.12 for controls and 0.94 for ataxia participants,

as shown in Fig. 3. The F1 score for this subset of the cohort (i.e., controls and ataxia participants with $BARS_{speech} = 0$) was 0.74.

The Mann-Whitney U-tests comparing model performance between sexes were not significant for participants with ataxia ($f = 0.56$, $p = 0.07$), but for controls indicated that the model predicted higher $P(Ataxia)$ for males ($f = 0.62$, $p = 0.02$). This bias may partly be due to the fact that the ataxia cohort contained more males, which could lead to some male characteristics being considered as ataxic by the model. The Spearman correlation between participant age and $P(Ataxia)$ among participants with ataxia was not significant ($\rho = 0.08$, $p = 0.15$), which is further supported by the age-stratified metrics presented in Table 4. The Mann-Whitney U-tests between groups of ataxia participants with different severities revealed that the more severe group had significantly higher $P(Ataxia)$ when comparing both $BARS_{total} < 15$ vs. $BARS_{total} >= 15$ ($f = 0.59$, $p = 7 \times 10^{-3}$) and $BARS_{speech} < 1.5$ vs. $BARS_{speech} >= 1.5$. ($f = 0.64$, $p = 7 \times 10^{-5}$).

B. SEVERITY ESTIMATION

Of the experiments listed in Table 2, a *ResNet* 18 model with time partial derivatives (excluding frequency partials) of the mel spectrogram as model inputs performed the best at severity estimation. Two separate models with the same inputs and archiecture were trained to estimate $\text{BARS}_{\text{total}}$ and $\text{BARS}_{\text{speech}}$. Results reported in this subsection are based on these models. Fig. 4 illustrates the relationship between $\text{BARS}^{\text{clin}}$ and $\text{BARS}^{\text{pred}}$. (Pooled validation data from all ten cross-validation folds are plotted together.) Strong agreement was observed for both speech score estimations (MAE = 0.33 and $R^2 = 0.73$) and total BARS estimations (MAE = 3.5 and $R^2 = 0.61$). This performance is comparable to approaches using wearable inertial measurement unit data [8]. However, the models tended to underestimate scores for individuals with higher severity, especially for $\text{BARS}^{\text{pred}}_{\text{speech}}$, which may due to the relative lack of training data in the high $\text{BARS}_{\text{speech}}$ range.

The Mann-Whitney U-test between ataxia participants with very mild speech severity and those without clinically noted deficits (i.e., $\text{BARS}^{\text{clin}}_{\text{speech}} = 0.5$ vs. $\text{BARS}^{\text{clin}}_{\text{speech}} = 0$) indicated a significant difference in the estimated speech scores ($f = 0.81$, $p = 2 \times 10^{-4}$). Thus, the models could distinguish between small variations in severity, even in very mild individuals. Though the comparison between ataxia participants with $\text{BARS}^{\text{clin}}_{\text{speech}} = 0$ and healthy controls did not reach significance ($f = 0.58$, $p = 0.09$), it is worth noting that the *ResNet* 18 model using both time and frequency derivative inputs did show significant differences ($f = 0.65$, $p = 0.005$) between median $\text{BARS}^{\text{pred}}_{\text{speech}} = 0.017$ for controls and median $\text{BARS}^{\text{pred}}_{\text{speech}} = 0.067$ for ataxia participants. Fig. 5 illustrates these results for both models.

The Mann-Whitney U-test comparing absolute errors between males and females indicated no dependence between participant sex and $\text{BARS}^{\text{pred}}_{\text{speech}}$ for controls ($f = 0.56$, $p = 0.25$) and participants with ataxia ($f = 0.51$, $p = 0.84$). There was, however, a significant difference between participant sex and the absolute error of $\text{BARS}^{\text{pred}}_{\text{total}}$ for both controls ($f = 0.64$, $p = 4 \times 10^{-3}$) and participants with ataxia ($f = 0.58$, $p = 0.047$). This difference may be explained by both 1) there being more male participants in the data set and 2) that male participants in the ataxia cohort had higher average $BARS^{clin}_{total}$ (9.7 points) than females (8.0 points). No

correlation was observed between absolute error and age for $BARS_{speech}$ ($\rho = 0.10$, $p = 0.07$), though there was a weak correlation for $BARS_{total}$ ($\rho = 0.22$, $p = 3 \times 10^{-5}$). These results suggest that the severity estimation models performed similarly well across age groups.

One-sample *t*-tests comparing model estimates between subsequent participant visits indicated significant differences for both $\text{BARS}_{\text{speech}}$ ($d = 0.50$, $p = 8 \times 10^{-3}$) and $\text{BARS}_{\text{total}}$ ($d = 0.43$, $p = 0.02$), which demonstrates that the regression models were able to detect disease progression. Moderate Spearman correlation between changes in estimated scores and the elapsed time between visits was also observed for $\text{BARS}_{\text{total}}$ ($\rho = 0.45$, $p = 9 \times 10^{-3}$) and $\text{BARS}_{\text{speech}}$ ($\rho = 0.46$, $p = 8 \times 10^{-3}$). This result indicates that the models captured larger speech differences over larger time intervals, as would be expected in neurodegenerative diseases. Fig. 6 illustrates changes in estimated scores for repeat participants.

C. MODEL INTERPRETABILITY

Fig. 7 illustrates the time derivative-based IGs and frequency derivative-based IGs for a single speech sample. The frequency-aggregated IGs for each type of IG are overlaid in red alongside the first principal component in white to visualize the part of the speech input in time contributed to correct model predictions. Fig. 8 shows the distribution of peak saliency locations across all samples, which reveals a structured pattern. The top left panel of Fig. 8 shows that time partial derivatives are most important at the beginning of the syllable and, to a lesser extent, at the end. Syllable endpoints have rapid temporal change and may be especially informative in ataxia, where there is known to be a slow alternating motion rate and rhythmic irregularities. The bottom left panel of Fig. 8 shows that the temporal importance distribution of the frequency partial derivatives has a peak in the middle of the syllable. A potential explanation of this finding is that the model used frequency-based acoustic properties of syllable midpoints to make more accurate predictions. Finally, the right column of Fig. 8 shows that low frequencies are predominantly the most important part of the spectrum, although the distribution of time partial derivatives over frequencies has a heavier tail across higher frequencies and potentially a second peak (Fig. 8, top right panel).

IV. DISCUSSION

Speech can be a powerful signal for diagnosis, severity estimation, and progression measurement in CAs. Alternating syllables tasks in speech (and alternating motions in other motor domains) have been widely used by neurologists to aid their assessment of ataxia patients [2], [3], [4]. This work therefore leveraged the translational invariance of convolutional neural networks to capture disease-relevant information from repetitive speech tasks. The results show that convolutional networks trained using the time and frequency partial derivatives of the log-mel spectograms of speech recordings can accurately distinguish between healthy individuals and individuals with CAs, including those without clinically-observed speech impairment. Furthermore, we found that similarly trained regression models could accurately and sensitively estimate both speech and total clinical severity, capture disease progression over time, and were sensitive to subtle speech differences in very mild individuals. These results support that our approach for identifying and measuring speech changes has the potential to both support early detection of ataxias

(e.g., as a component of a screening tool) and to produce sensitive speech measures for clinical trials.

The proposed models performed similarly to both feature-based and deep learning-based models for analysis of dysarthric speech in recent literature. Kashyap, et al. proposed feature-based models based on tongue-twisters [35], a repetitive “ta-ta-ta” task [36], and repeated utterance of the phrase “British Constitution” [47] that had healthy vs. ataxia AUCs of 0.87, 0.91, and 0.97 respectively. (Though the “British Constitution” model reported very high AUC, the results were validated on a small test set of thirteen participants.) Furthermore, Kashyap, et al. showed that models trained on the “ta-ta-ta” and “British Constitution” tasks were able to capture disease progression over a two-year period [37]. Though some works reported high accuracy using CNN-based and gated recurrent unit-based deep learning models applied to mel spectrograms [48], [49], such results are likely optimistic as they were achieved using methodologies that did not ensure participant independence between training and testing sets [50]. A recent work by Song, et al., however, provides a comparable benchmark for the proposed models’ performance [38]. In that work, CNN-based models trained on the audio waveforms of participants reading a prepared passage achieved an AUC of 0.92. The same work further proposed a patch-wise wave splitting algorithm that improved performance to an AUC of 0.96, though such an algorithm may not be applicable to the relatively short repetitive tasks used herein. In contrast, key contributions of the current work include that the proposed classification model learned sub-clinical information from a modestly-sized data set and that a similarly-structured regression model accurately estimated dysarthria severity and disease progression (despite being trained in a cross-sectional manner).

In addition to speech, a variety of other techniques and technologies have been used to detect ataxia and assess disease severity, including inertial measurement units (IMUs) that quantify gait or limb movement [8], [9], computer mouse tasks that assess arm motor control [14], and eye tracking devices that quantify eye movement abnormalities [15]. Our approach complements these prior works by using convolutional neural networks based on mel spectrogram time and frequency gradient inputs to probe the effects of ataxia on vocal motor impairments in an unbiased fashion. In the future, we will seek to combine this assessment technique with quantitative assessment tools in other behavioral domains to improve the ability to identify early disease signs and quantify disease changes over time.

One limitation of deep learning models, especially in health care settings, is a lack of understanding of the information leveraged by the model. As a step toward addressing this limitation, we utilized integrated gradients to calculate saliency scores for the model inputs. For the classification model, we found that time partial derivatives were most important at the beginning and end of syllables, while frequency partial derivatives were most important in the middle. This result matches intuition that the most temporally dynamic parts of the syllable may carry information about speech frequency changes in time, and that less dynamic parts encapsulate frequency-based acoustic properties of the syllables, all of which are informative in detecting the presence of ataxia.

Another key feature of the proposed models is their scalability. This study included data from multiple contexts, including data collected in-person or at home, with a computer or mobile device, and using a built-in or lavalier microphone. Thus, this assessment technique only requires performance of a repetitive speech task for less than a minute and may be performed at home using everyday technologies. The ability to obtain at-home assessments facilitates both participation independent of geographical location and more frequent and longitudinal data collection. The scalability of the proposed model also has the potential to be further improved. For example, transfer learning could be used to fine-tune the trained model to new devices or to specific populations.

There were some limitations to this study. Most importantly, our classification performance may be inflated compared to real-world performance as a dedicated validation set (i.e., separate from the test set) was not used to select the best-performing models during the training process. Additionally, the distribution of ataxia and control subjects in our dataset does not reflect the population distribution. (In our dataset, controls were the minority group, while in reality, the opposite is true.) Though we were able to account for this issue during model training, our test set still suffered from this imbalance. Additional strategies to address this imbalance, such as class-weighted loss functions, may yield improved results. Though our dataset is relatively large for studies in the field, our models would benefit from even larger training datasets. In particular, the *ResNet18* classification model did not noticeably outperform the comparison simple CNN. If supported by sufficient data, even larger models (e.g., *ResNet* 101) could potentially demonstrate superior performance. The heterogeneity of CA and potential for different manifestations of speech deficiencies in each CA type could also affect our models and should be investigated. Finally, we were unable to account for cultural and geographical biases (e.g., language and accent), although we expect repetitive single syllable tasks to be less sensitive to these factors.

V. CONCLUSION

Speech changes are one of the earliest and most evident manifestations of cerebellar ataxia. We showed that convolutional neural networks, with time and frequency partial derivatives of the speech signal as input, are able to accurately separate healthy controls from patients with ataxia, including ataxia participants with no detectable clinical deficits in speech. Similar regression models also provided accurate and sensitive estimates of speech severity. We showed that integrated gradients could be used to understand how these networks use information about different parts of the syllable to generate predictions. These speech analysis tools have the potential to assist with early detection of ataxia, provide low-burden monitoring tools for neurological care, and provide speech outcome measures for use in natural history studies and interventional trials. Further work is needed to fine-tune and validate these models on larger and more balanced datasets that better reflect the true distribution of the population.

ACKNOWLEDGMENT

The authors thank Michael Brenner for useful conversations. This project utilized resources at the Martinos-MLSC cluster.

This project was funded by NIH R01 NS117826, the University of Pennsylvania Orphan Disease Center in partnership with the Ataxia-Telangiectasia Children's Project, and Biogen Inc.

Biographies

KYRIAKOS VATTIS received the Ph.D. degree in physics from Brown University. He worked in the field of theoretical astrophysics and cosmology, focusing on a variety of problems, including dark matter and primordial black holes. He used data analysis methods, such as Markov Chain Monte Carlo (MCMC) and machine learning to model dark matter properties and how they affect the cosmological evolution. During his postdoctoral appointment with the Department of Neurology, Massachusetts General Hospital, he was involved in analyzing audio, video, wearable, and physiological data recorded during cognitive and motor assessments, using machine learning classification and prediction, to study neurodegenerative phenotypes and their evolution.

BRANDON OUBRE received the Ph.D. degree in computer science from the Manning College of Information and Computer Sciences, University of Massachusetts Amherst. He is currently a Research Fellow in neurology with Massachusetts General Hospital and Harvard Medical School. His research interests include understanding, monitoring, and improving human health, using insights derived from time-series sensor data. He is particularly interested both in work supporting unobtrusive assessment of disease signs using wearable and ubiquitous technologies and in analyzing rich, multi-modal data to more sensitively measure disease progression. He received the Outstanding Dissertation Award from the University of Massachusetts Amherst.

ANNA C. LUDDY received the Bachelor of Science degree in neuroscience from Trinity College, in 2020. She hopes to pursue a career in neuropsychology, and is especially interested in health equity, health literacy, and working with pediatric populations.

JESSEY S. OUILLON received the Bachelor of Arts degree in psychology from Carnegie Mellon University, in 2016. She enjoys working with adults and children, who have neurological disorders and hopes to contribute to clinical research that can improve healthcare for these individuals.

NICOLE M. EKLUND received the Bachelor of Arts degree in psychology and neuroscience from Syracuse University, in 2019. She hopes to attend graduate school and pursue a career researching neurodegenerative diseases, focusing on preventatives and treatments.

CHRISTOPHER D. STEPHEN is a Neurologist, a Clinician-Scientist, and an Epidemiologist with the Movement Disorders Unit, Ataxia Center; Dystonia Clinic, Department of Neurology, Massachusetts General Hospital (MGH); and the Performing Arts Clinic, Department of Neurology, Brigham and Women's Hospital. He is an Instructor in neurology with the Harvard Medical School. His clinical interests include rare movement disorders (particularly ataxia and dystonia and their overlap), motor functional movement disorders (FND), and neurological problems in musicians. Closely associated with his clinical expertise, his research interests include the quantitative assessment of movement disorders, with a view to their use as potential motor biomarkers of disease severity and progression. Utilizing movement analysis techniques, he has quantitatively studied eye movements, static posture, and motion sensor-based analysis in movement disorders, as part of his highly collaborative research. He is a member of the MGH Center for Rare Neurological Diseases and has also been involved in imaging and other clinical research in the rare ataxias, late-onset GM2 gangliosidosis; and within the Ataxia Center, on MRI imaging biomarkers in the sporadic ataxia multiple system atrophy of the cerebellar type (MSA-C). He is also a member of the MGH Collaborative Center for X-Linked Dystonia Parkinsonism (XDP) Clinical Research Team, where he is also a Co-Lead Investigator of the Coordinating Center for the XDP-TRACK Natural History of XDP. He is a key member of the MGH FND Research Program and has led or contributed to multiple publications

with the goal of advancing FND patient care. As an epidemiologist, he studies movement disorders and FND, utilizing large longitudinal cohort studies, large databases, and health services research. He is a member of the International Parkinson and Movement Disorders Society Rare Movement Disorders, Functional Movement Disorders, and Ataxia Study Groups.

JEREMY D. SCHMAHMANN received the Medical degree (Hons.) from the University of Cape Town. He completed residency from the Neurological Unit, Boston City Hospital, and postdoctoral fellowship with Prof. Deepak Pandya with the Department of Anatomy and Neurobiology, Boston University School of Medicine. He is a Professor of neurology with the Harvard Medical School, and a Senior Clinical Neurologist with Massachusetts General Hospital, where he is the Founding Director of the Ataxia Center, the Director of the Laboratory for Neuroanatomy and Cerebellar Neurobiology, and a Founding Member of the Cognitive Behavioral Neurology Unit. His clinical and research efforts focus on the anatomical substrates of intellect and emotion, anatomy, clinical neurology, and basic science of the ataxias and other cerebellar disorders. He is a fellow of the American Academy of Neurology, the American Neurological Association, and the American Neuropsychiatric Association; a member of the Medical and Scientific Research Advisory Board of the National Ataxia Foundation and the Clinical Research Consortium for the Study of Cerebellar Ataxias; an Executive of the Society for Research on the Cerebellum and Ataxias; and a past President of the American Neuropsychiatric Association. In 2000, he received the Norman Geschwind Prize from the American Academy of Neurology and the Behavioral Neurology Society for his description of the cerebellar cognitive affective syndrome and its neurobiological and theoretical underpinnings.

ADONAY S. NUNES received the Ph.D. degree in biomedical physiology from Simon Fraser University, Vancouver. During his degree, he used data analysis methods involving parametric statistics and machine learning to model the population and predict their symptom severity. His research focused on understanding the underlying oscillatory activity and synchrony of neural ensembles that characterize individuals with Autism and in children born very preterm. Currently, he is interested in extracting neurodegenerative phenotypes for machine learning classification and prediction by integrating video, wearable, and physiological data recorded during cognitive and motor assessments.

ANOOPUM S. GUPTA received the bachelor's degree in electrical engineering from Georgia Tech, the Medical degree from the School of Medicine, University of Pittsburgh, and the Ph.D. degree in robotics from the School of Computer Science, Carnegie Mellon University. He completed a residency in neurology from the Mass General Brigham-Harvard Neurology Residency Program (Mass General and Brigham and Women's Hospital). His clinical interests include neurogenetic and neurodegenerative diseases, including cerebellar ataxias, Parkinson's disease, atypical parkinsonian disorders, and Huntington's disease. His research program involves the use of accessible technologies combined with signal processing, computer vision, and machine learning methods to understand how neurodegenerative diseases impact motor and cognitive behavior. Broadly, he is involved in research at the intersection of computational science and medicine.

CODE AVAILABILITY

Code relevant to the models and analysis can be found at https://github.com/neuropheno-org/ieee_access_2023_ataxic_speech.

DATA AVAILABILITY

Raw data cannot be publicly released as the recordings and spectrograms could potentially be used to identify study participants. However, de-identified features computed using the OpenSMILE toolkit [51] have been published on IEEE DataPort (https://dx.doi.org/10.21227/brvz-0g97). These data include the Prosody, IS10, eGeMAPS, emobase, and ComParE features computed using the default configuration files provided by OpenSMILE version 3.0.1. Provided labels include the BARS speech subscore, BARS total score, gender, decade of life, whether the subject had an ataxia-related diagnosis, and the elapsed number of days between BARS administration and data collection. Many labels, such as collection dates and specific diagnoses, have been obfuscated or redacted to maintain privacy given the rarity of the diseases in the data.

REFERENCES

[1]. Ruano L, Melo C, Silva MC, and Coutinho P, "The global epidemiology of hereditary ataxia and spastic paraplegia: A systematic review of prevalence studies," Neuroepidemiology, vol. 42, no. 3, pp. 174–183, 2014. [PubMed: 24603320]

[2]. Schmitz-Hübsch T et al. , "Scale for the assessment and rating of ataxia: Development of a new clinical scale," Neurology, vol. 66, no. 11, pp. 1717–1720, Jun. 2006. [PubMed: 16769946]

[3]. Trouillas P, Takayanagi T, Hallett M, Currier RD, Subramony SH, Wessel K, Bryer A, Diener HC, Massaquoi S, Gomez CM, Coutinho P, Hamida MB, Campanella G, Filla A, Schut L, Timann D, Honnorat J, Nighoghossian N, and Manyam B, "International cooperative ataxia rating scale for pharmacological assessment of the cerebellar syndrome," J. Neurological Sci, vol. 145, no. 2, pp. 205–211, Feb. 1997.

[4]. Schmahmann JD, Gardner R, Macmore J, and Vangel MG, “Development of a brief ataxia rating scale (BARS) based on a modified form of the ICARS,” Movement Disorders, vol. 24, no. 12, pp. 1820–1828, Sep. 2009. [PubMed: 19562773]

[5]. Schuller KA, Vaughan B, and Wright I, “Models of care delivery for patients with Parkinson disease living in rural areas,” Family Community Health, vol. 40, no. 4, pp. 324–330, 2017. [PubMed: 28820786]

[6]. Saadi A, Himmelstein DU, Woolhandler S, and Mejia NI, “Racial disparities in neurologic health care access and utilization in the United States,” Neurology, vol. 88, no. 24, pp. 2268–2275, Jun. 2017. [PubMed: 28515272]

[7]. Gupta AS, “Digital phenotyping in clinical neurology,” in Seminars Neurology. New York, NY, USA: Thieme Medical Publishers, Inc., 2022.

[8]. Knudson KC and Gupta AS, “Assessing cerebellar disorders with wearable inertial sensor data using time-frequency and autoregressive hidden Markov model approaches,” 2021, arXiv:2108.08975.

[9]. Ilg W, Seemann J, Giese M, Traschütz A, Schöls L, Timmann D, and Synofzik M, “Real-life gait assessment in degenerative cerebellar ataxia,” Neurology, vol. 95, no. 9, pp. e1199–e1210, Sep. 2020. [PubMed: 32611635]

[10]. Gupta AS, Luddy AC, Khan NC, Reiling S, and Thornton JK, “Real-life wrist movement patterns capture motor impairment in individuals with ataxia-telangiectasia,” Cerebellum, vol. 22, no. 2, pp. 261–271, Mar. 2022. [PubMed: 35294727]

[11]. Lee J, Oubre B, Daneault J-F, Stephen CD, Schmahmann JD, Gupta AS, and Lee SI, “Analysis of gait sub-movements to estimate ataxia severity using ankle inertial data,” IEEE Trans. Biomed. Eng, vol. 69, no. 7, pp. 2314–2323, Jul. 2022. [PubMed: 35025733]

[12]. Khan NC, Pandey V, Gajos KZ, and Gupta AS, “Free-living motor activity monitoring in ataxia-telangiectasia,” Cerebellum, vol. 21, no. 3, pp. 368–379, Jun. 2022. [PubMed: 34302287]

[13]. Oubre B, Daneault J-F, Whritenour K, Khan NC, Stephen CD, Schmahmann JD, Lee SI, and Gupta AS, “Decomposition of reaching movements enables detection and measurement of ataxia,” Cerebellum, vol. 20, no. 6, pp. 811–822, Dec. 2021. [PubMed: 33651372]

[14]. Gajos KZ, Reinecke K, Donovan M, Stephen CD, Hung AY, Schmahmann JD, and Gupta AS, “Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection,” Movement Disorders, vol. 35, no. 2, pp. 354–358, Feb. 2020. [PubMed: 31769069]

[15]. Velázquez-Pérez L, Seifried C, Abele M, Wirjatijasa F, Rodríguez-Labrada R, Santos-Falcón N, Sánchez-Cruz G, Almaguer-Mederos L, Tejeda R, Canales-Ochoa N, Fetter M, Ziemann U, Klockgether T, Medrano-Montero J, Rodríguez-Díaz J, Laffita-Mesa JM, and Auburger G, “Saccade velocity is reduced in presymptomatic spinocerebellar ataxia type 2,” Clin. Neurophysiol, vol. 120, no. 3, pp. 632–635, Mar. 2009. [PubMed: 19201647]

[16]. Chang Z, Chen Z, Stephen CD, Schmahmann JD, Wu H-T, Sapiro G, and Gupta AS, “Accurate detection of cerebellar smooth pursuit eye movement abnormalities via mobile phone video and machine learning,” Sci. Rep, vol. 10, no. 1, pp. 1–10, Oct. 2020. [PubMed: 31913322]

[17]. Asgari M and Shafran I, “Extracting cues from speech for predicting severity of Parkinson’S disease,” in Proc. IEEE Int. Workshop Mach. Learn. Signal Process, Aug. 2010, pp. 462–467.

[18]. Hazan H, Hilu D, Manevitz L, Ramig LO, and Sapir S, “Early diagnosis of Parkinson’s disease via machine learning on speech data,” in Proc. IEEE 27th Conv. Electr. Electron. Engineers Isr., Nov. 2012, pp. 1–4.

[19]. Tsanas A, Little MA, McSharry PE, Spielman J, and Ramig LO, “Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease,” IEEE Trans. Biomed. Eng, vol. 59, no. 5, pp. 1264–1271, May 2012. [PubMed: 22249592]

[20]. Chen HL, Huang CC, Yu XG, Xu X, Sun X, Wang G, and Wang SJ, “An efficient diagnosis system for detection of Parkinson’s disease using fuzzy k-nearest neighbor approach,” Expert Syst. Appl, vol. 40, no. 1, pp. 263–271, 2013.

[21]. Åström F and Koker R, “A parallel neural network approach to prediction of Parkinson’s disease,” Expert Syst. Appl, vol. 38, no. 10, pp. 12470–12474, Sep. 2011.

[22]. Jain D, Mishra AK, and Das SK, “Machine learning based automatic prediction of Parkinson’s disease using speech features,” in Proc. Int. Conf. Artif. Intell. Appl., 2021, pp. 351–362.

[23]. Quan C, Ren K, and Luo Z, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," IEEE Access, vol. 9, pp. 10239–10252, 2021.

[24]. Rusz J, Tykalová T, Salerno G, Bancone S, Scarpelli J, and Pellecchia MT, "Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy," J. Neurol, vol. 266, no. 6, pp. 1394–1404, Jun. 2019. [PubMed: 30859316]

[25]. Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J, Uher T, Vaneckova M, Andelova M, Novotna K, Kadrnozkova L, and Horakova D, "Characteristics of motor speech phenotypes in multiple sclerosis," Multiple Sclerosis Rel. Disorders, vol. 19, pp. 62–69, Jan. 2018.

[26]. Noffs G, Boonstra FMC, Perera T, Kolbe SC, Stankovich J, Butzkueven H, Evans A, Vogel AP, and van der Walt A, "Acoustic speech analytics are predictive of cerebellar dysfunction in multiple sclerosis," Cerebellum, vol. 19, no. 5, pp. 691–700, Oct. 2020. [PubMed: 32556973]

[27]. Stegmann GM, Hahn S, Liss J, Shefner J, Rutkove S, Shelton K, Duncan CJ, and Berisha V, "Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis," Npj Digit. Med, vol. 3, no. 1, pp. 1–5, Oct. 2020. [PubMed: 31934645]

[28]. Schalling E and Hartelius L, "Speech in spinocerebellar ataxia," Brain Lang., vol. 127, no. 3, pp. 317–322, Dec. 2013. [PubMed: 24182841]

[29]. Schalling E, Hammarberg B, and Hartelius L, "Perceptual and acoustic analysis of speech in individuals with spinocerebellar ataxia (SCA)," Logopedics Phoniatrics Vocology, vol. 32, no. 1, pp. 31–46, Jan. 2007. [PubMed: 17454658]

[30]. Schirinzi T, Sancesario A, Bertini E, Castelli E, and Vasco G, "Speech and language disorders in friedreich ataxia: Highlights on phenomenology, assessment, and therapy," Cerebellum, vol. 19, no. 1, pp. 126–130, Feb. 2020. [PubMed: 31701351]

[31]. Brendel B, Synofzik M, Ackermann H, Lindig T, Schölderle T, Schöls L, and Ziegler W, "Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with friedreich ataxia," J. Neurol, vol. 262, no. 1, pp. 21–26, Jan. 2015. [PubMed: 25267338]

[32]. Vogel AP, Magee M, Torres-Vega R, Medrano-Montero J, Cyngler MP, Kruse M, Rojas S, Cubillos SC, Canento T, Maldonado F, Vazquez-Mojena Y, Ilg W, Rodríguez-Labrada R, Velaáquez-Pérez L, and Synofzik M, "Features of speech and swallowing dysfunction in pre-ataxic spinocerebellar ataxia type 2," Neurology, vol. 95, no. 2, pp. e194–e205, Jul. 2020. [PubMed: 32527970]

[33]. Kent RD, Kent JF, Duffy JR, Thomas JE, Weismer G, and Stuntebeck S, "Ataxic dysarthria," J. Speech, Lang., Hearing Res, vol. 43, no. 5, pp. 1275–1289, Oct. 2000, doi: 10.1044/jslhr.4305.1275.

[34]. Rosen KM, Folker JE, Vogel AP, Corben LA, Murdoch BE, and Delatycki MB, "Longitudinal change in dysarthria associated with friedreich ataxia: A potential clinical endpoint," J. Neurol, vol. 259, no. 11, pp. 2471–2477, Nov. 2012. [PubMed: 22669353]

[35]. Kashyap B, Pathirana PN, Horne M, Power L, and Szmulewicz D, "Automated tongue-twister phrase-based screening for cerebellar ataxia using vocal tract biomarkers " in Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2019, pp. 7173–7176.

[36]. Kashyap B, Horne M, Pathirana PN, Power L, and Szmulewicz D, "Automated topographic prominence based quantitative assessment of speech timing in cerebellar ataxia," Biomed. Signal Process. Control, vol. 57, Mar. 2020, Art. no. 101759.

[37]. Kashyap B, Pathirana PN, Horne M, Power L, and Szmulewicz DJ, "Modeling the progression of speech deficits in cerebellar ataxia using a mixture mixed-effect machine learning framework," IEEE Access, vol. 9, pp. 135343–135353, 2021.

[38]. Song J, Lee JH, Choi J, Suh MK, Chung MJ, Kim YH, Park J, Choo SH, Son JH, Lee DY, Ahn JH, Youn J, Kim K-S, and Cho JW, "Detection and differentiation of ataxic and hypokinetic dysarthria in cerebellar ataxia and parkinsonian disorders via wave splitting and integrating neural networks," PLoS ONE, vol. 17, no. 6, Jun. 2022, Art. no. e0268337. [PubMed: 35658000]

[39]. Sainburg T, "Timsainb/noisereduce: V1.0," Zenodo, Version db94fe2, Jun. 2019. [Online]. Available: https://github.com/timsainb/noisereduce, doi: 10.5281/zenodo.3243139.

[40]. Sainburg T, Thielk M, and Gentner TQ, “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires,” PLOS Comput. Biol, vol. 16, no. 10, Oct. 2020, Art. no. e1008228. [PubMed: 33057332]

[41]. McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E, and Nieto O, “Librosa: Audio and music signal analysis in Python,” in Proc. 14th Python Sci. Conf., 2015, pp. 18–25.

[42]. He K, Zhang X, Ren S, and Sun J, “Deep residual learning for image recognition,” 2015, arXiv:1512.03385.

[43]. Kingma DP and Ba J, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[44]. Zhou H, Nguyen H, Enriquez A, Morsy L, Curtis M, Piser T, Kenney C, Stephen CD, Gupta AS, Schmahmann JD, and Vaziri A, “Assessment of gait and balance impairment in people with spinocerebellar ataxia using wearable sensors,” Neurological Sci., vol. 43, no. 4, pp. 2589–2599, Apr. 2022.

[45]. Zhang H, Cisse M, Dauphin YN, and Lopez-Paz D, “Mixup: Beyond empirical risk minimization,” 2017, arXiv:1710.09412.

[46]. Sundararajan M, Taly A, and Yan Q, “Axiomatic attribution for deep networks,” 2017, arXiv:1703.01365

[47]. Kashyap B, Pathirana PN, Horne M, Power L, and Szmulewicz D, “Quantitative assessment of speech in cerebellar ataxia using magnitude and phase based cepstrum,” Ann. Biomed. Eng, vol. 48, no. 4, pp. 1322–1336, Apr. 2020. [PubMed: 31965359]

[48]. Dumane P, Hungund B, and Chavan S, “Dysarthria detection using convolutional neural network,” in Proc. 3rd Int. Conf. Adv. Technol. Societal Appl., vol. 1, 2021, pp. 449–457.

[49]. Shih D-H, Liao C-H, Wu T-W, Xu X-Y, and Shih M-H, “Dysarthria speech detection using convolutional neural networks with gated recurrent unit,” Healthcare, vol. 10, no. 10, p. 1956, Oct. 2022. [PubMed: 36292403]

[50]. Gholamiangonabadi D, Kiselov N, and Grolinger K, “Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection,” IEEE Access, vol. 8, pp. 133982–133994, 2020.

[51]. Eyben F, Wollmer M, and Schuller B, “Opensmile: The Munich versatile and fast open-source audio feature extractor,” in Proc. 18th ACM Int. Conf. Multimedia, Oct. 2010, pp. 1459–1462.

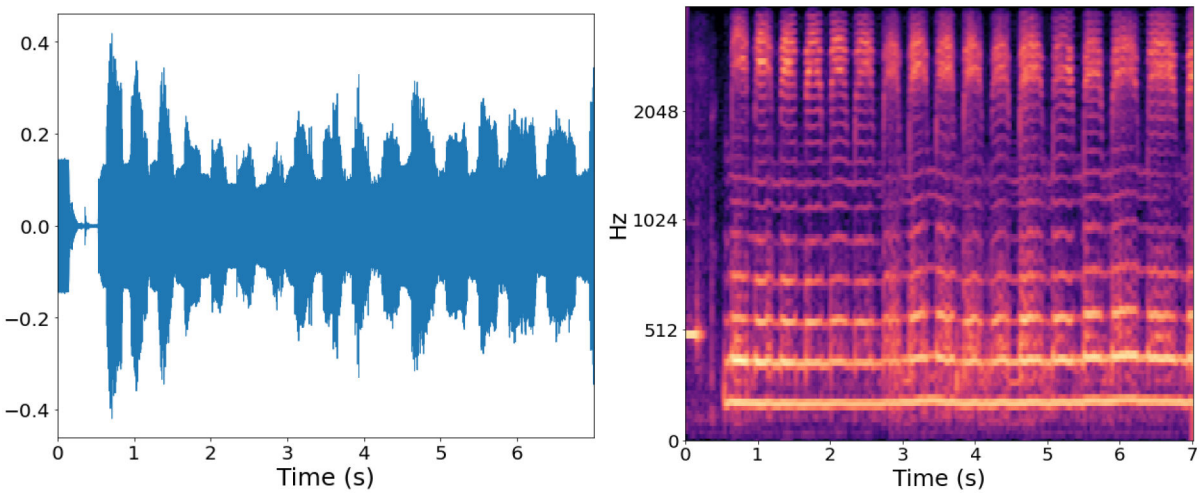

FIGURE 1.

Audio sample from the “me-me-me” syllable repetition task. The left panel shows the audio waveform as a function of time. The right panel shows the log spectrogram (brighter colors represent greater signal power).

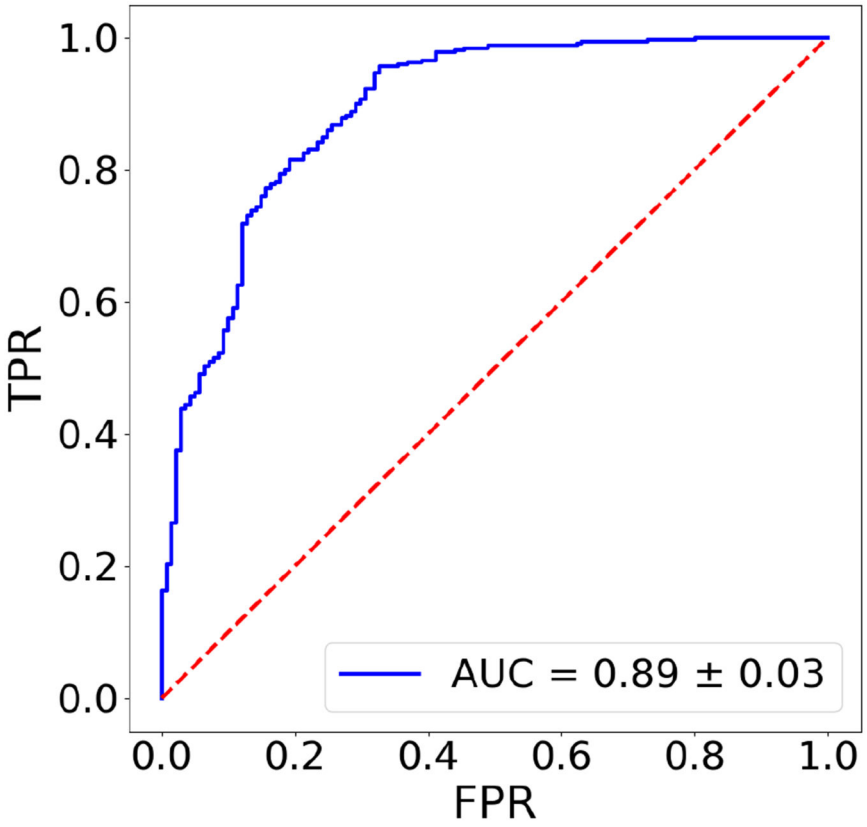

FIGURE 2.
ROC curve for the classification model.

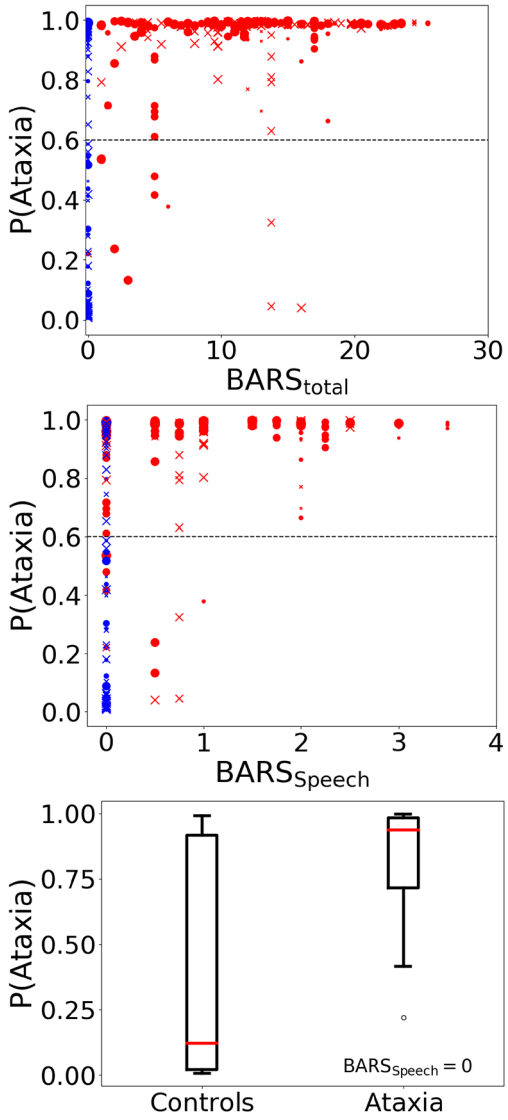

FIGURE 3.

Probability of a sample originating from an individual with ataxia plotted against $\mathrm{BARS}_{\mathrm{total}}$ (top) and $\mathrm{BARS}_{\mathrm{Speech}}$ (middle) using the *ResNet* 18 model. The dashed line was the decision threshold chosen for reported metrics. Blue denotes controls and red denotes participants with ataxia. Dots denote males and crosses denote females. Marker size increases with age. The bottom panel shows the distribution of P(Ataxia) for controls versus participants with ataxia and $\mathrm{BARS}_{\mathrm{Speech}} = 0$.

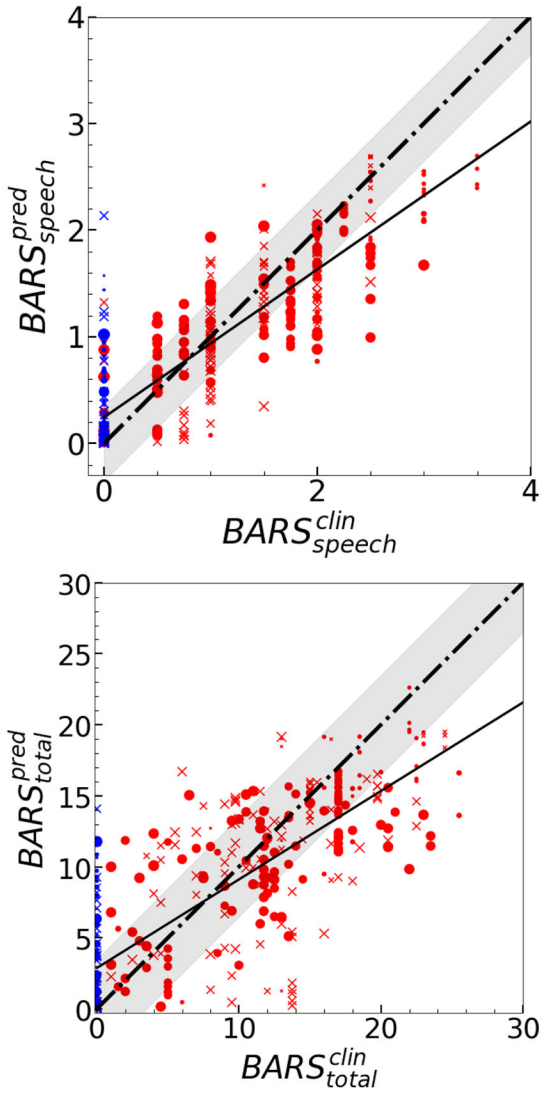

FIGURE 4.

Estimations of $BARS_{total}$ (top) and $BARS_{Speech}$ (bottom). The dot-dashed line represents perfect agreement with clinical scores. The solid line denotes the best fit and the grey band shows the MAE interval. Blue denotes controls and red denotes participants with ataxia. Dots denote males and crosses denote females. Marker size increases with age.

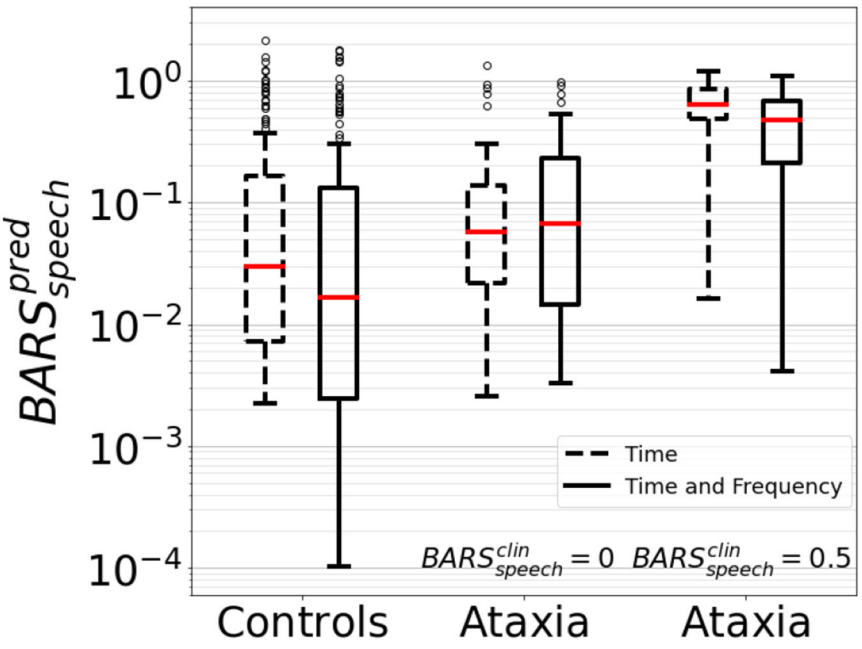

FIGURE 5.

The distribution of the samples for controls on the left, ataxia participants with $\mathrm{BARS}^{\mathrm{clin}}_{\mathrm{speech}} = 0$ in the center, and $\mathrm{BARS}^{\mathrm{clin}}_{\mathrm{speech}} = 0.5$ on the right. Dashed lined boxes correspond to using only time partial derivatives as model inputs. Solid lined boxes correspond to using both time and frequency partial derivatives as model inputs.

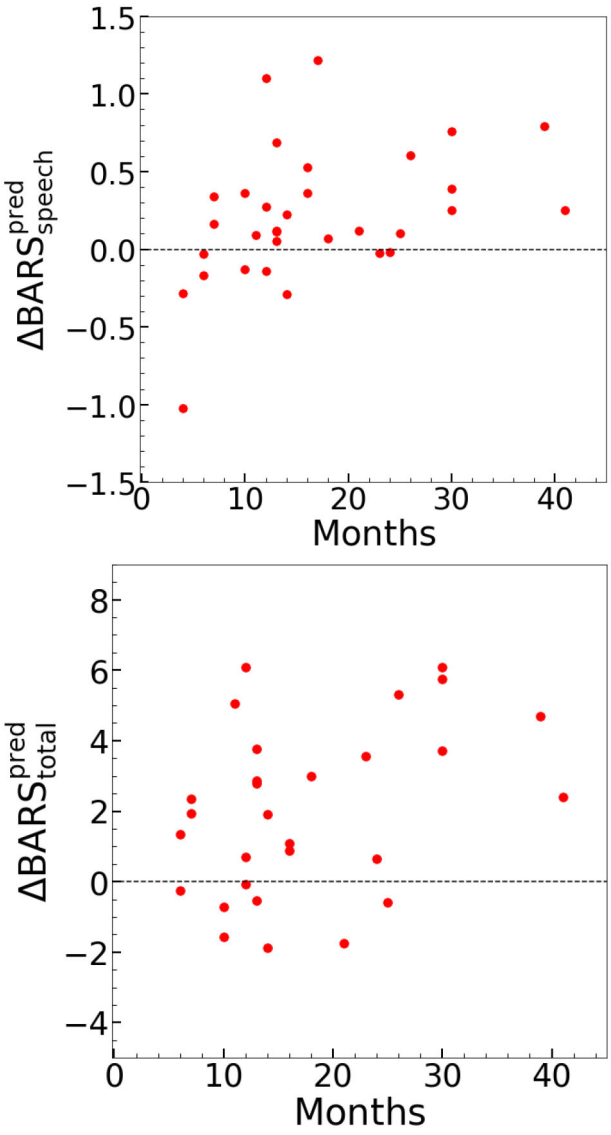

FIGURE 6.

Change in estimated $BARS_{speech}$ and $BARS_{total}$ for participants with multiple sessions at least a month apart. Changes were calculated using the first and last session on record.

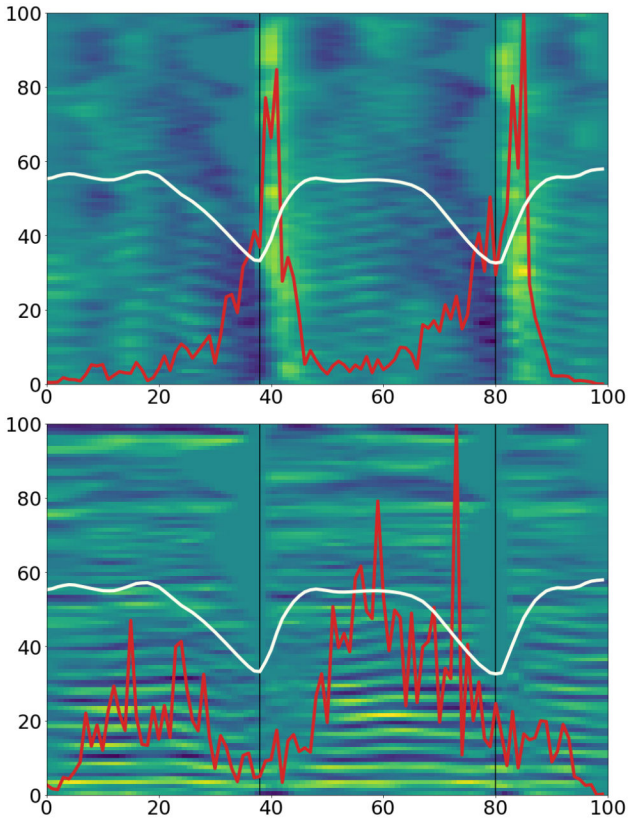

FIGURE 7.

Time (top panel) and frequency (bottom panel) gradients of a sample mel-spectrogram. Positive values are green and negative values are blue. The first principal component of the spectrogram is overlaid as an ivory line. The frequency aggregated salience score is overlaid in red. Vertical black lines indicate the beginning and end of a syllable.

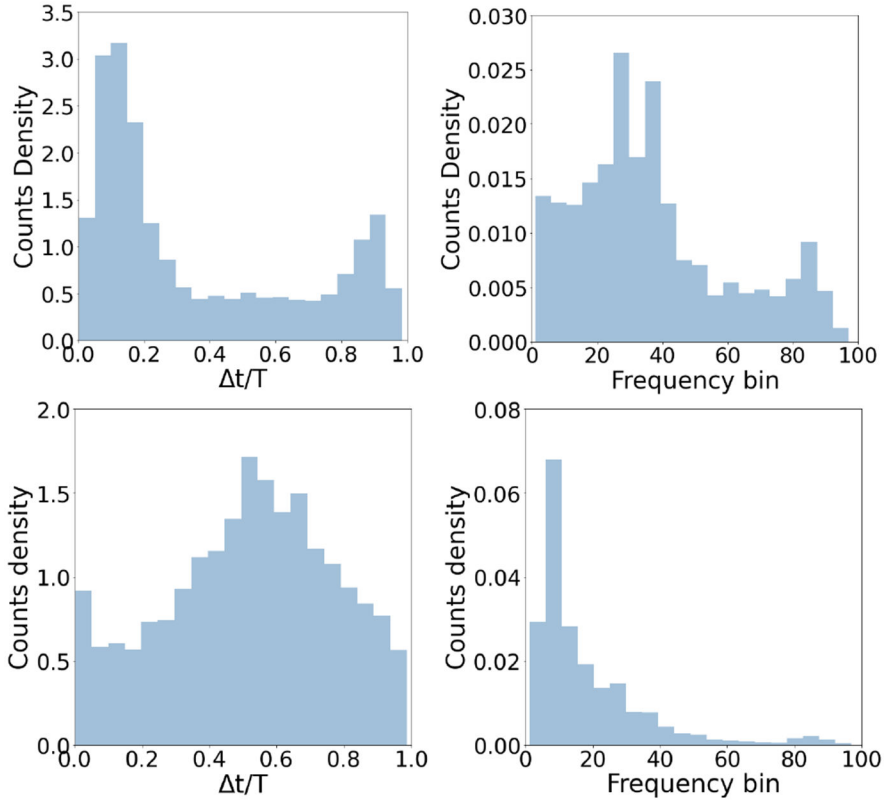

FIGURE 8.

Left column: The distribution of the location of instances with maximum frequency-aggregated salience score with respect to their closest minimum of the first principal component on the left. Right column: The distribution of the location of instances with maximum time-aggregated salience score per frequency bin. The top row uses time gradients while the bottom row uses the frequency gradients.

TABLE 1.

Participant demographics. M: Mean; SD: Standard deviation; SCA: Spinocerebellar ataxia; AT: Ataxia-Telangiectasia; MSA: Multiple system atrophy; FRDA: Friedreich's Ataxia; ARCA: Autosomal recessive cerebellar ataxia.

Diagnosis	No. Participant	No. Male	Age at First Session M±SD (Range)	BARS at First Session M±SD (Range)
Healthy Control	**71**	**35**	**26.5 ± 21.3 (2–86)**	—
Total Ataxia	**157**	**91**	**44.3 ± 25.0 (2–82)**	**11.2 ± 5.9 (0–25.5)**
AT	45	25	10.2 ± 6.4 (2–29)	15.3 ± 6.1 (3.5–22.5)
SCA-1	5	1	52.8 ± 16.9 (30–69)	7.7 ± 2.6 (4.5–10.5)
SCA-2	3	1	61.0 ± 6.6 (54–67)	12.0 ± 1.3 (10.5–13.0)
SCA-3	21	10	54.8 ± 12.1 (32–80)	11.8 ± 4.9 (4.0–21.0)
SCA-6	7	6	68.0 ± 6.2 (61–75)	11.9 ± 4.9 (2.0–17.0)
other SCA	10	7	56.2 ± 16.5 (23–77)	11.2 ± 5.0 (3.5–20.0)
MSA	9	6	64.7 ± 9.1 (54–77)	13.4 ± 5.5 (6.0–23.0)
FRDA	4	3	44.2 ± 13.7 (33–61)	16.5 ± 2.8 (13.5–19.0)
ARCA	3	3	49.0 ± 13.5 (36–63)	15.2 ± 7.3 (10.0–23.5)
Other Ataxias	50	29	59.2 ± 15.0 (22–82)	8.8 ± 5.8 (0.0–22.5)

TABLE 2.

Model performance summary. Bold entries denote the best *ResNet18* task performance.

	Classification				$BARS_{total}$		$BARS_{speech}$	
	ResNet18		Simple CNN		ResNet18			
	AUC	F1	AUC	F1	MAE	R^2	MAE	R^2
log-mel spectrogram	0.86 ± 0.01	0.80 ± 0.02	—	—	3.6	0.58	0.39	0.66
Time derivative‡	0.86 ± 0.06	0.82 ± 0.05	—	—	**3.5**	**0.61**	**0.33**	**0.73**
Time and frequency derivative (all)†	**0.89 ± 0.03**	**0.86 ± 0.02**	0.90 ± 0.03	0.84 ± 0.06	3.5	0.58	0.35	0.69
Time and frequency derivative ("go-go")	0.85 ± 0.05	0.82 ± 0.04	—	—	—	—	—	—
Time and frequency derivative ("me-me")	0.89 ± 0.04	0.83 ± 0.03	—	—	—	—	—	—
Time and frequency derivative ("la-la")	0.86 ± 0.04	0.82 ± 0.02	—	—	—	—	—	—

†Features used for classification results in Section II-C.

‡Features used for regression results in Section II-D.

TABLE 3.

Normalized confusion matrix for the classification model with a decision threshold of P(Ataxia) > 0.6.

	Predicted Control	Predicted Ataxia
Control	0.67	0.33
Ataxia	0.04	0.96

TABLE 4.

Classification performance of *ResNet*18 stratified by participant age.

Age	No.	F1	AUC
0–20	59	0.88	0.94
21–40	37	0.85	0.97
41–60	59	0.87	0.88
> 60	61	0.94	0.93