IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Modeling the Progression of Speech Deficits in Cerebellar Ataxia using a Mixture Mixed-effect Machine Learning Framework

BIPASHA KASHYAP[1], **(Student Member, IEEE)**, PUBUDU N. PATHIRANA[1], **(Senior Member, IEEE)**, MALCOLM HORNE[2], LAURA POWER[3], DAVID J SZMULEWICZ[234]

[1]Networked and Sensing Control (NSC) lab, School of Engineering, Deakin University, Waurn Ponds, Victoria, AU (e-mail: bkashy@deakin.edu.au, pubudu.pathirana@deakin.edu.au)
[2]Florey Institute of Neuroscience and Mental Health, Parkville, Victoria, AU (e-mail: malcolm.horne@florey.edu.au)
[3]Balance Disorders & Ataxia Service,Royal Victorian Eye and Ear Hospital (RVEEH), St Andrews Place, East Melbourne, Victoria, AU (e-mail: laura_power@live.com.au)
[4]Cerebellar Ataxia Clinic, Alfred Hospital, Prahran, Victoria, AU (e-mail: dsz@me.com)

Corresponding author: Bipasha Kashyap (e-mail: bkashy@deakin.edu.au).

This work is supported by the Florey Institute of Neuroscience and Mental Health, Melbourne, Australia through the National Health and Medical Research Council (NHMRC) GNT1101304 and APP1129595 and CSIRO Data61

ABSTRACT Background: Accurate and reliable prediction of changes in the severity of Cerebellar Ataxia (CA) will be helpful in trials of disease-modifying therapies. This study demonstrated that objective acoustic measures were more sensitive than perceptive analysis through the Scale for the Assessment and Rating of Ataxia (SARA) in assessing CA progression, within a time window of two years (mean). Method: Thirty-seven people with CA were tested at baseline (time point 1, TP1) and two years later (time point 2, TP2). A machine-learning framework with a robust three-step feature selection criterion and a Bayesian data-driven clustering technique based on the multivariate mixture extension of the generalized linear mixed model (GLMM) was used. The outcomes included two (time and cepstral-based) objective speech parameters recorded at TP1 and TP2. For testing subjects, the dynamic prediction was conducted using samples from the posterior distributions of parameter estimates and random effects. This study further employed the penalized expected deviance (PED) criterion for model comparison and the selection of the number of groups in the clustering procedure. Results: First, the selected objective speech metrics in the individual patients showed a significant worsening of the speech impairment ($p<0.001$, Kolmogorov–Smirnov test) from TP1 through TP2. Second, the cluster analysis divided the CA patients into two distinct subgroups showing a strong association between objective speech measures and disease duration, with $\sim$96% of observed values falling within the 95% credible intervals. Third, for the training data, our multivariate model ($PED_{Fea1+Fea2}$=5175; number of groups=2) performed more reliably than the univariate models (PED_{Fea1}=4225, PED_{Fea2}=3850; number of groups=2) in discriminating the CA patients. Fourth, the individual-level predictions of the change in profiles of the objective measures over time were performed for the testing data. Conclusion: Such a framework using objective speech metrics indeed holds promise to predict the rate of clinical progression of Ataxic Speech in individuals with CA.

INDEX TERMS Cerebellar ataxia, cerebellar dysarthria, clustering, mixed-effect modeling, mixture analysis, progression analysis, speech processing.

I. INTRODUCTION

THe cerebellum integrates information from a range of sensory afferent inputs to produce coordinated movement. Cerebellar Ataxia (CA) refers to the uncoordinated movement resulting from dysfunction of the cerebellum caused by many processes, including neurodegeneration, multiple sclerosis, stroke and trauma. Here "CA" will refer specifically to neurodegenerative cerebellar conditions. The

cerebellum regulates many aspects of movements, including movements of the limbs, trunk, balance, gait, eyes and speech, the latter being the specific focus of this study. Cerebellar ataxia of speech is sometimes referred to as *Ataxic Dysarthria* or *Cerebellar Dysarthria* but here is described as 'Ataxic Speech', where the tempo and articulation of speech is affected and the voice is unstable and impaired in quality [1], [2].

Progression of ataxia is reflected in increasing scores on clinical rating scales such as the Scale for the Assessment and Rating of Ataxia (SARA) [3]. SARA is an eight-item clinical scale devised by Schmitz-Hubsch et al. [3] for measuring ataxia severity in different domains, including upright posture, gait, speech, upper and lower limbs. It ranges from a total score of 0 (no ataxia) to 40 (severe ataxia).

Neurodegenerative cerebellar disease generally progresses relatively slowly and consequently measurement instruments must be reasonably sensitive to detect these changes [4]. In multiple system atrophy (MSA) and spinocerebellar ataxia (SCA), the minimum detectable changes is 1-2 points per year [5]–[7] with an average mean deterioration of $1.38/40$ [standardized response mean (SRM) = 0.5] in the total SARA scores [8]. Thus, it is unlikely that the SARA speech scores, which contribute no more than 6 of the total 40 will change significantly in two years. This is not surprising because the SARA was designed to be a composite score for ataxia rather than an independent score of each domain. Furthermore the SARA and other clinical scales are ordinal and the interval between each score most likely does not represent a similar increase in severity: that is they do not linearly correspond to severity. The assumptions underpinning the present study were that the severity of speech ataxia will worsen over time and that the measures to detect this deterioration of speech performance with greater sensitivity and lower variance are the most sensitive measures. This study aimed to design an automated system with objective measures to detect the worsening of speech ataxia with greater sensitivity than clinical scales such as the SARA.

Most recent studies (Table 1) were designed to find objective acoustic features to diagnose Ataxic Speech and measure its severity [12], [13]. The term 'diagnosis' will be used here to mean identification of Ataxic Speech from the speech of non-ataxic subjects. Additionally, most recent studies were cross-sectional [12], [13] with only one longitudinal study [9]. The latter examined the changes over time in perceptual and acoustic features of the speech of individuals with SCA. Previous studies were specifically focused on exploiting time [9], [11]–[15], spectral and cepstral [10] based speech characteristics for the diagnosis of Ataxic Speech. All the previous researchers [9]–[11] used either sustained phonations or connected speech in their studies. However, speech tasks involving syllable repetition have proved to be more useful than sentence utterances [16]–[18] for identifying differences between the speech of ataxic and non-ataxic individuals during perceptual analysis. Variations on the "repeated Consonant-Vowel (C-V) syllable paradigm tasks" have been most commonly used for syllable repetition.

While mixed-effect models have been commonly used to assess longitudinal progression [19]–[21], they are usually only univariate analyses. As the present study assessed different outcomes—that is, continuous objective speech parameters from multiple speech tasks [22], [23], a multivariate model (rather than a univariate model) was warranted. Multivariate generalized linear mixed effect models (GLMMs) can be adapted simultaneously for inference, integrating not only the association of repeated measurements for each outcome within a subject, but also the use of random effects to correlate multiple outcomes. Komarek and Komarkova [24] demonstrated a Bayesian data-driven clustering technique to draw inferences based on a multivariate mixture extension of the classical GLMM [25], [26]. This approach proved to be a computationally efficient and reliable alternative to an existing method [27], and has relevance for predicting deterioration in Ataxic Speech over time. To our knowledge, this is the first study to use optimally-integrated multivariate objective acoustic measures to predict the progression of Ataxic Speech over time.

The main contributions of this study can be summarized as follows:

1) Define a robust three-step objective speech feature selection criterion to select distinctive acoustic features as outcomes from different repeated C-V syllable paradigm speech tasks.
2) Build a multivariate mixture extension of a GLMM for prediction based on the repeated measurements of selected multivariate continuous outcomes.
3) Perform a cluster analysis based on this multivariate model to identify groups of patients with similar characteristics and draw meaningful inferences.
4) Evaluate the model's performance and further compare its performance with univariate analyses.
5) Predict, at a specific timepoint, the probability of a subject belonging to a particular patient group.

The rest of this paper is organized as follows: section II introduces the proposed speech progression assessment framework, data collection strategy and feature selection scheme, and describes the multivariate mixture generalized linear mixed model which will serve as the basis for our clustering procedure; section III describes the results of the research. Section IV discusses the significance of the proposed approach; and section V concludes the paper and explains the future scope.

II. MATERIALS AND METHODS

A. SPEECH PROGRESSION ASSESSMENT FRAMEWORK (SPA_{CA})

The framework of our proposed assessment of Ataxic Speech progression (hereafter, referred to as SPA_{CA}) involved the following steps:

1) *Speech Inputs* generated by instrumental versions of the standard clinical test for assessing ataxic speech.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114328, IEEE Access

TABLE 1. Comparative overview of recent literature in objective assessment of Ataxic Speech

Recent Studies in Ataxic Speech	**Feature Type**	**Brief description**	**Speech Task Considered** *Type* **(Number)**	**Sample Size**	**CA Diagnostic\ CA Severity (Performance)**	**AS Progression\ Prediction Models**
Schalling et al. [9]	Perceptual measures, Objective measures (syllable rate, syllable duration) standard deviation	Time, frequency, amplitude measures	Sustained Phonation */a/* and sentences (2)	3 CA, 6 SCA	No	Yes, but **no prediction modeling**
Jannetts et al. [10]	APQ (%), PPQ(%), (%)RAP, CPP mean, CPP standard deviation	Amplitude, frequency, pitch and voice quality perturbation	Sustained Phonation */a/* and connected speech (2)	10 CA, 43 PD	Yes (Pearson Correlation)	No
Luna Webb [11]	Jitter(%), Shimmer (dB)	Amplitude and frequency perturbation	Sustained Phonation */a/, /i/, /o/* (1)	20 FA, 20 Controls	Yes (TPR for FA\Controls) (1\0.95)	No
Kashyap et al. [12]	RT_Ca, RT_PPa, RT_DR75, RT_RF50 (Hz), RT_Gr (s), RT_Dt50	Time-domain measures (captured using Topographic Prominence based automatic algorithm)	Repeated Consonant-Vowel syllable, *Repeated Ta* */ta/-/ta/-/ta/* (1)	63 CA, 28 Controls	Yes (Acc\TNR\TPR\AUC) 0.84\0.89\0.78\0.91	No
Kashyap et al. [13]	MFCC+MGDCC	Cepstral measures capturing amplitude, phase and spectral fluctuations	Phrase with repeated Consonant-Vowel syllable, *British Constitution* (1)	23 CA, 42 Controls	Yes (Acc\TNR\TPR\AUC) 0.84\0.9\0.75\0.97	No
Current study	Time domain [12], cepstral domain [13]		*Repeated Ta* and *British Constitution* (2)	37 CA	No, covered in our previous studies [12]–[14]	**Yes, with inferences and prediction modeling**

Captions: AS : Ataxic Speech, PD : Parkinson Disease, CA : Cerebellar Ataxia, FA : Friedreich ataxia, SCA : Spinocerebellar ataxia, TNR : True negative rate, TPR : True positive rate, Acc : Accuracy, AUC : Area under the ROC Curve, RT : Reapted Ta, MFCC : Mel-frequency cepstral coefficients, MGDCC : Modified group delay cepstral coefficients, APQ : Amplitude Perturbation Quotient, PPQ : Pitch Period Perturbation Quotien, RAP : Relative Amplitude Perturbation, CPP : Cepstral peak prominence.

2) Speech recordings were *captured* by a condenser microphone clipped at an average distance of 10 cm from the subject's lips, in a quiet room with low ambient noise. Recording was conducted using the BioKinMobi^{TM} [28] application on an Android phone, under the supervision of a trained investigator.
3) Recordings were wirelessly transmitted to a blockchain based distributed *cloud* network [29] where the proposed SPA_{CA} *algorithm* was deployed.
4) Data analysis results were transformed into a *clinically relevant format*.

A pictorial representation of the assessment platform is demonstrated in Figure 1.

B. DATA COLLECTION

Thirty-seven native speakers of Australian English with a bilateral neurodegenerative cerebellar disorder were followed for two years (median: 2, mean: 1.99, standard deviation: 0.02). The demographics and clinical characteristics of participants are summarized in Table 2. Subjects were assessed upon entry to the study and again after two years (median: 2, mean: 1.99, standard deviation: 0.02). None of the participants had received speech therapy prior to the investigation. The speech intelligibility was perceptually scored by an experienced clinician in accordance with SARA, which was on a scale of 0 to 6 as follows: '0', normal speech; '1', disturbed speech; '2', distorted but simple to comprehend speech; '3', words sometimes difficult to understand; '4', several words difficult to understand; '5', only single words are comprehensible; and '6', unintelligible speech.

TABLE 2. Clinical characteristics of the enrolled participants

Characteristics	**TP1**	**TP2**
Age (years)	64.72±9.35	
Disease Duration (years)	10.59±7.04	
TP2-TP1 (years)		1.99±0.02
SARA speech (Item 4)	1.54±1.22	1.89±1.33
Male : Female	20 : 17, n = 37	20 : 17, n = 37
Disease Duration at TP1	>8 years	7
	≤8 years	10
	Unknown	20
***CA Phenotypes**	Pure (central) CA	16
	CABV	9
	CABV+SS	7
	Unknown	5

Captions: TP : timepoint, n = number of patients, CA : Cerebellar Ataxia, CABV : Cerebellar Ataxia with Bilateral Vestibulopathy, SS : Somatosensory impairment. Data presented as Mean ± Standard Deviation (range). *Deep phenotyping has not been undertaken in these subjects.

The assessment consisted of participants performing the following two speech tasks:

1) **Speech Task 1**: Repeat the *syllable /ta/* for five seconds, producing the */ta/-/ta/-/ta/* syllabic train (*Repeated Ta*: hereafter 'RT')).
2) **Speech Task 2**: Utter the phrase *British Constitution* (BC: a classical phrase for eliciting the features of ataxic speech) thrice. Individual's acoustic measures were taken as the average of the three recordings.

These speech tasks resulted in 74 speech recordings (37 from each task) at each timepoint.

This study was approved by the Human Research and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114328, IEEE Access

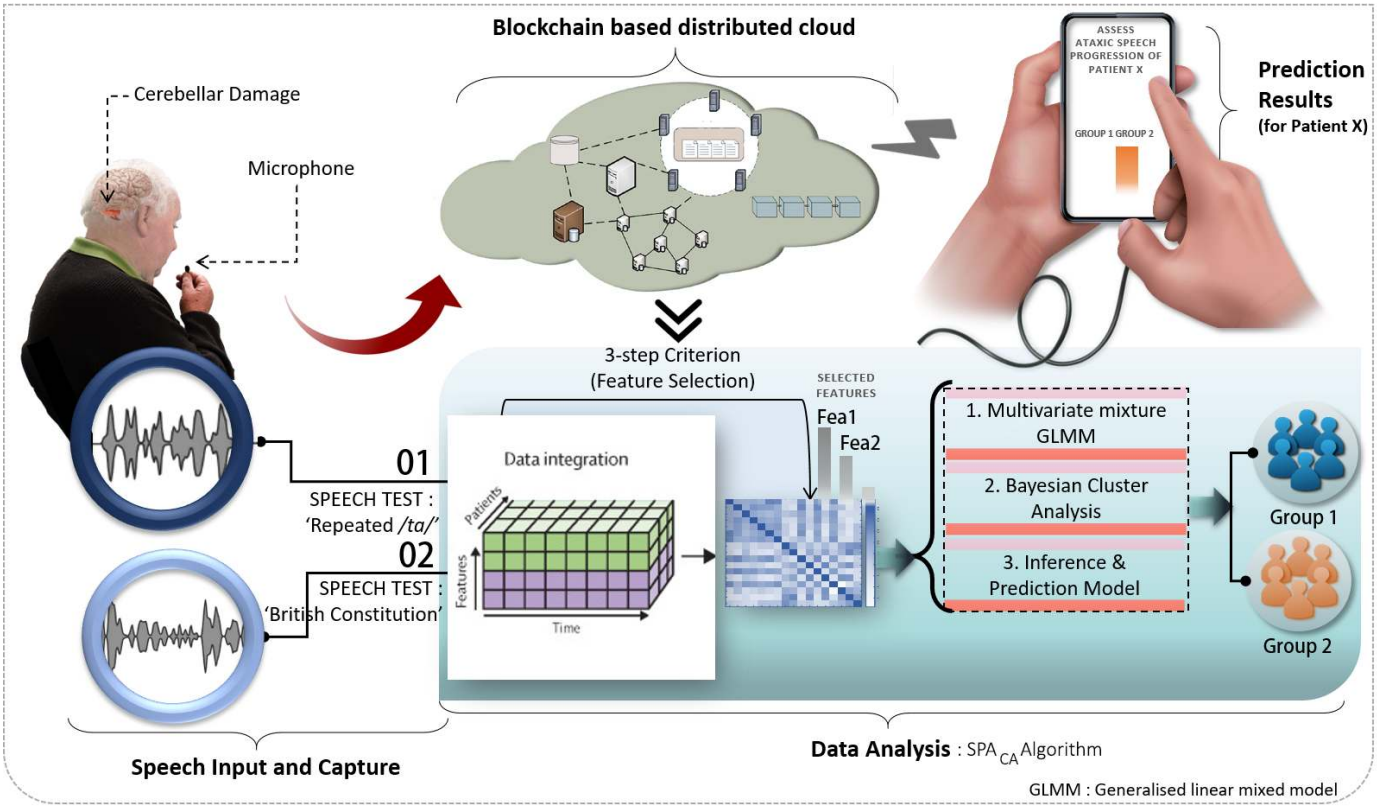

FIGURE 1. Flowchart demonstrating the data flow in the automated SPA_{CA} framework.

Ethics Committee at the Royal Victorian Eye and Ear Hospital, Australia (HREC Reference Number: 11/994H/16) and administered through the Florey Institute of Neuroscience and Mental Health, Melbourne, Australia. It complied with the NHMRC's guidelines for research using human participants. Written consent was sought from all participants prior to their enrolment. The subject in the Figure 1 provided informed consent to publish their image.

C. MACHINE-LEARNING FRAMEWORK

In this study, the data analysis constituted of developing the SPA_{CA} algorithm (Figure 1) to predict the progression in the severity of Ataxic Speech was developed. Its three stages were:

1) Extract and select distinctive acoustic features with a 3-step criterion.
2) Build a multivariate mixture generalized linear mixed model (GLMM) for prediction based on the joint exploitation of the selected features' change over two timepoints.
3) Classify the subjects into two groups based on the mixture extensions of the multivariate model.

D. DATA REPRESENTATION AND OUTCOME DEFINITION

$Data_{tr}$ and $Data_{ts}$ were used to denote a training dataset and a testing dataset with a sample size of N_1 and N_2 respectively. $Data_{tr}$ was used to build the prediction model and $Data_{ts}$ is used to assess the prediction for new subjects. Y_{irj} denoted the objective measure for subject i at time t_{ij}, extracted from a specific speech task. The index $i = 1, 2, ..., N$ represented the subject while the index $j = 1, 2, ..., n$ represented speech measurements from a subject at different timepoints (TP) for a specific speech task. In our designed Ataxic Speech progression study, we used two different speech tasks, and the measurement times followed a protocol with a common set of follow-up times, $t_{ij} = t_j$ where subjects were measured at baseline (that is, $TP1$ or $t_1 = 0$) and after 2 years (that is, $TP2$ or $t_2 = 2$). Exploration of the observed changes in the profiles of objective speech features (Figure 3) suggested that changes in the features for each subject, where the respective values of the features may differ across subjects, could be linearly modeled over time. Being motivated by this inference, we considered Y_{iqj} to be continuous.

E. EXTRACTION AND SELECTION OF SPEECH PARAMETERS

In this study, the time-based features were extracted from speech task 1 and the spectral and cepstral features were extracted from speech task 2. Initially, six topographic prominence–based time features and 23 cepstral features (12 Mel-frequency cepstral coefficients (MFCC) [30] and 11 modified group delay cepstral coefficients (MGDCC) were extracted, applying the same methods that were established in our previous studies for Ataxic Speech diagnosis and severity

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114328, IEEE Access

TABLE 3. Descriptions of objective measures (features)

Speech Task (Parameter Count)	Feature (unit) [Abbreviation]	Brief Description	Domain
Speech Task 1: RT (6)	RT Duration Regularity(s) [RT_Dr50]	Variability in the rhythm of speech production which is an integral measure of timing deficit.	Time-based
	Average RT Peak Prominence [RT_PPr]	Average relative elevation/ peak for a specific /ta/ pulse calculated	
	RT Damping Ratio [RT_DR75]	Average of the /ta/ syllables' damping ratio calculated on the wave data extracted at 75% prominence	
	RT Resonant Frequency (Hz) RT_RF50]	Average of the /ta/ syllables' resonant frequency calculated on the wave data extracted at 50% prominence	
	RT Gap Regularity(s) [RT_Gr]	Variability in the time difference between two consecutive /ta/ syllable peaks	
	Average RT Compensation [RT_Ca]	Average of the differences calculated between the peak and its corresponding prominence for a specific /ta/ syllable pulse	
Speech Task 2: BC (11+12)	Modified Group Delay Cepstral Coefficients [BC_MGDCC1, BC_MGDCC2,...,$BC\_MGDCC11$]	Phased and spectral fluctuations	Cepstral
	Mel Frequency Cepstral Coefficients [BC_MFCC1, BC_MFCC2,...,BC_MFCC12]	Amplitude and spectral fluctuations	Cepstral

prediction [12], [13]. Feature descriptions are presented in Table 3 and the results of the three-step feature selection criterion is described via a pictorial representation in Figure 2.

A feature x was from a specific speech task was selected through the three-step feature selection criterion. In the first step, features (in the training data) that do not change significantly from TP1 to TP2 were eliminated using a mass-univariate approach. The threshold that determined whether changes in the feature observations from TP1 to TP2 were significant was the p-value of 0.001 from a KS test (TH_1). All the significant features with p-value < 0.001 were selected during this step. In the second step, the selected features were pruned using another exclusion mechanism. As the existence of multicollinearity is indicated by an absolute correlation coefficient of > 0.7 between two or more predictors, only features with a spearman correlation < 0.5 (TH_2) were selected. The correlation between features was compared and one of two features with a correlation > 0.5 was removed. In the third step, the selected features were further sorted based on the ANOVA ω^2 effect size (> 0.14) [31], [32], observed power (> 0.85) and F-statistic test ($p = 0.001$). The selected features from the final step were carried forward to design the model. The one-way repeated measures ANOVA [33] statistics of all the selected features from step 2 are tabulated in Table 4.

F. CONSTRUCTING MIXTURE MODEL-BASED CLUSTERING

The clustering procedure used in this study was based on a multivariate mixture GLMM (MMGLMM). The change in the profile of the r^{th} selected feature ($r = 1, 2, ..., R$) belonging to the i^{th} subject ($i = 1, ..., N$) was denoted by the random vector, $Y_{ir} = (Y_{ir1}, ..., Y_{irn_{i,r}})^T$. Furthermore, $Y_i = (Y_{i1}^T, ..., Y_{iR}^T)^T$ represented the random vector of the change in measurements of all selected features at the different timepoints for the i^{th} subject and $Y = (Y_1^T, ..., Y_N^T)^T$ was a random vector representing the available outcomes of all subjects.

This mixed mixture model was designed based on the changes in measurements of the selected speech parameters from TP1 to TP2. Two continuous features were selected through the three-step feature selection criterion, Y_{i1j} and Y_{i2j} with Gaussian distribution. Following the data interpretation in Section II-D, when the subjects $i = 1, 2, ..., N$ were each measured at timepoints $j = 1, 2, ..., n$, the outcome at each timepoint j could be represented with the mean structure of the change in profiles of the two parameters as follows:

$$\begin{aligned} E(Y_{i1j|b_{i1}}) &= b_{i11} + b_{i12}t_{i1j}, and \\ E(Y_{i2j|b_{i2}}) &= b_{i21} + b_{i22}t_{i2j}. \end{aligned} \tag{1}$$

Here, t_{irj} is the time in years and Y_{irj} is the corresponding outcome for $r = 1, 2$, that is the timepoints from the start (TP1) to the follow-up (TP2).

Next, to identify groups of patients with similar characteristics using the change in profiles of the measurements over time, they were classified into two groups (K = 2) in the distribution of the four-dimensional random effect vector, $B_i = (b_{i11}, b_{i12}, b_{i21}, b_{i22},)^T$, where b_{i11}, b_{i21} are random intercepts from each selected feature and b_{i12}, b_{i22} are their respective random slopes.

The estimates of the cluster specific (marginal) mean change in profiles over time is denoted as

$$E(Y_{\cdot,r,\cdot}|u = K) = E_B\left\{E(Y_{\cdot,r,\cdot}|B, u = K)\right\}. \tag{2}$$

TABLE 4. One-way Repeated Measures ANOVA (Within-subjects design) statistics comparison of the acoustic measures (after Step 2) and SARA speech scores.

Speech Task	Feature	Observed Power, $p = 0.001$	Omega squared (ω^2)	F-statistic	Significance (p)
RT	**RT_Dr50**	**0.999**	**0.400**	**50.35**	**0.000**
BC	BC_MFCC3	0.154	0.065	6.166	0.019
BC	BC_MFCC7	0.157	0.066	6.24	0.018
BC	BC_MGDCC2	0.52	0.133	12,39	0.001
BC	**BC_MGCDCC5**	**0.991**	**0.338**	**38.85**	**0.000**
BC	BC_MGDCC6	0.738	0.140	13	0.002
SARA speech scores		0.005	0.838	0.100	0.23

Captions: Omega squared (ω^2) is a measure of effect size and is calculated as, $\omega^2 = \frac{((k-1)(F-1))}{((k-1)(F-1)+nk)}$ where k is the number of levels of the within-subjects factor, F is the value of the F-statistic, and n is the number of participants. ANOVA Omega squared (ω^2) effect values are 0.01 (small), 0.06 (medium) and 0.14 (large).

Here, r = 1, 2 over time and K = 1, 2, representing the two groups.

Following the clustering procedure, the optimal classification of the i^{th} subject to a specific cluster or group was computed based on the subject-level marginalization over the posterior distribution [34] as

$$\pi_{i,K} = \int p_{i,K}(\psi,\theta)p(\psi,\theta|y)d(\psi,\theta) = E\left\{p_{i,K}(\psi,\theta)|y\right\} \approx \frac{1}{M}\sum_{m=1}^{M} p_{i,K}(\psi^{(m)},\theta^{(m)}) = \hat{\pi}_{i,K} \tag{3}$$

A priori, the independence between the mixture related parameters and the GLMM related parameters were assumed as θ and ψ respectively.

According to Bayesian inference,

$$p_{i,K}(\psi,\theta) = P(u_i = K|\psi,\theta,y_i) = \frac{w_K L_{i,K}(\psi,\theta)}{\sum_{l=1}^{2} w_K L_{i,K}(\psi,\theta)}, \tag{4}$$

where $i = 1, 2, ..., N$ and K=1, 2. The i^{th} subject was assigned to group K with the highest value of $\hat{\pi}_{i,K}$ [35], [36].

G. EXTERNAL PREDICTION OF CA SUBJECTS USING THE TESTING DATASET

To classify a test subject into either Group 1 or Group 2 with 95% highest posterior density credible intervals (HPD CI), the patient-specific component probability was computed as:

$$p_{i,K} = p_{i,K}(\psi,\theta). \tag{5}$$

Here, i is a patient and K is the number of groups. Only if the lower limit of the corresponding credible interval approaches a certain threshold, such as 0.5 (considering the classification into K = 2 groups), the patient is categorized into one of the considered groups.

H. MODEL PERFORMANCE STATISTIC

In this study, the penalized expected deviance (PED) was selected as a criterion for choosing the number of clusters/ groups in the multivariate model [37]. Furthermore the PED was also exploited to compare the performance of the multivariate model with univariate models of similar group size. PED has been successfully employed in various applications [38], [39] and is defined as

$$PED = E\left\{D(\psi,\theta)|y\right\} + p_o, \tag{6}$$

Here, $D(\psi,\theta) = 2logL(\psi,\theta)$ is the observed data deviance of the model. The expected deviance, represented as $E\left\{D(\psi,\theta)|y\right\}$ (posterior mean), can be calculated from the Markov chain Monte Carlo (MCMC) sample. The penalty term, optimism, is denoted in Equation 6 by p_o, which can be further computed from importance sampling and two parallel MCMC chains [37].

III. EXPERIMENTAL RESULTS

A. CROSS-VALIDATION AND SELECTED MODEL PARAMETERS

Participants were sorted into a training set (N=32) and a test set (N=5). The time (years) since the baseline visit, and the two objective speech parameters resulting from the three-step parameter selection criterion were the only features included in this study to build the multivariate prediction model. Interestingly, the three-step feature selection criterion resulted in two heterogeneous features, each belonging to a different speech task. This suggested that each provided complementary information to aid in identifying the presence and severity of Ataxic Speech. The two outcomes were:

1) RT parameter (RT_Dr50) (hereafter, referred to as **Fea1**)
2) BC parameter (BC_MGDCC5) (hereafter, referred to as **Fea2**),

as indicated in the Figure 2. The feature descriptions are listed in Table 2. The model was trained by running the prediction algorithm with a burn-in of 1000 iterations [40] and 100 subsequent iterations with 1:10 thinning to provide samples from the joint posterior distribution (two parallel Markov chain Monte Carlo (MCMC) sampled chains with different sets of initial values).

B. COMPARISON OF SPEECH PARAMETERS AT TP1 AND TP2 (APPROXIMATELY 2 YEARS LATER)

The SARA speech scores at TP2 were not significantly different from those at TP1 ($p = 0.1$, KS test). However, there were statistically significant differences ($p \leq 0.001$, KS test) between the two speech features (Fea1 and Fea1, see Section

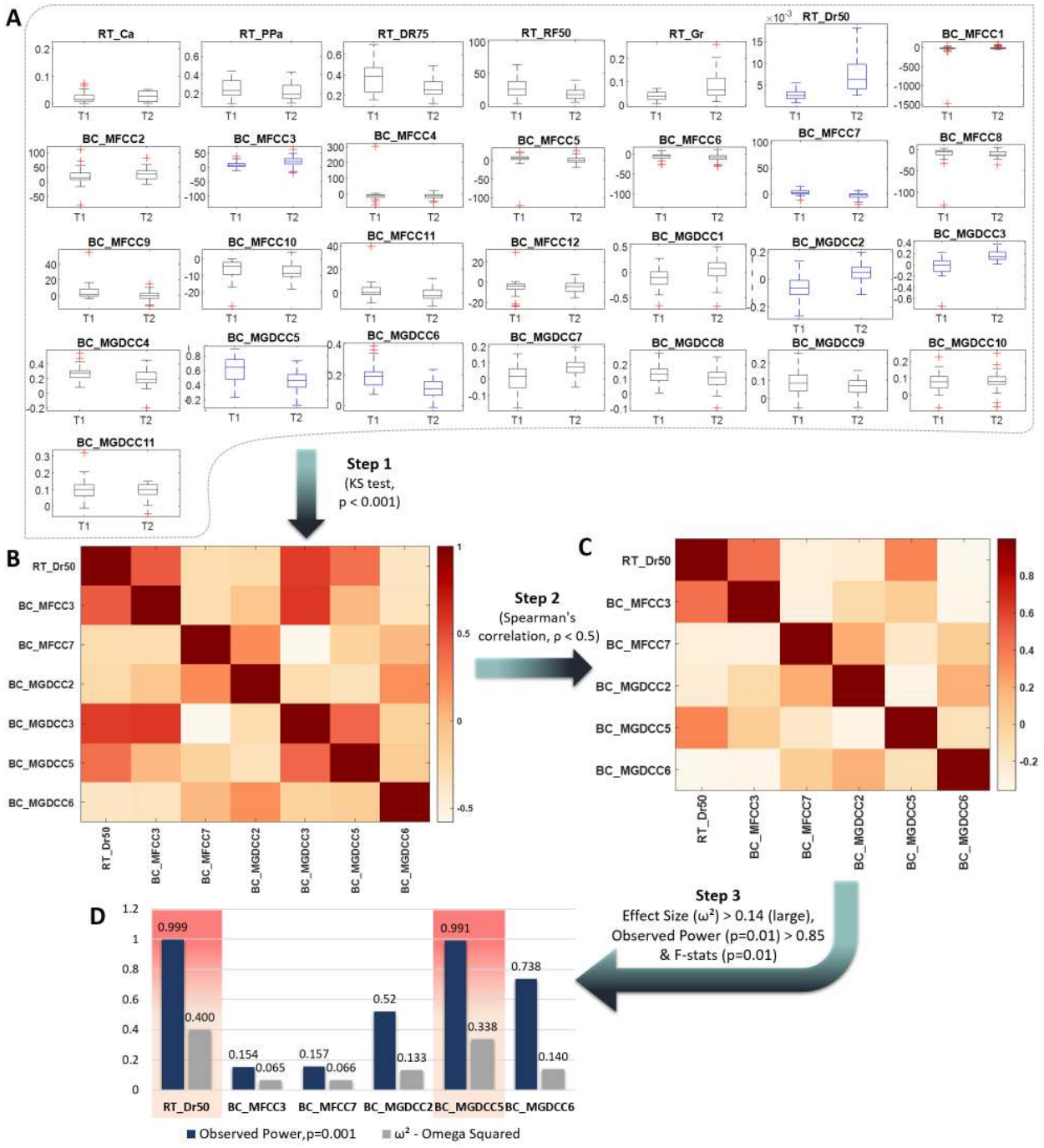

FIGURE 2. Illustration of the 3-step feature selection criterion used in this study. (a) Box plot representation of the initially extracted 29 acoustic measures. Blue boxes indicate the features that are statistically significantly different at p=0.001 and are forwarded to Step 2. (b) Heat map correlation plot of the seven selected features in Step 1. (c) Heat map correlation plot of the six selected features in Step 2 with Spearman's, $\rho < 0.5$. (d) Cluster bar plots indicate the effect size and the observed power of the selected features from Step 2. Red gradient areas highlight the selected features' cluster bar plots having a large effect size ($\omega^2 > 0.14$) and observed power > 0.85.

III-A) extracted from each speech task after the three-step feature selection criterion. Descriptive statistics are recorded in Table 5 and the change in profiles are depicted in Figure 3.

C. GROUP-SPECIFIC MEAN CHANGE FROM TP1 TO TP2

The multivariate mixture GLMM (MMGLMM) for the clustering of ataxic subjects measurements at TP1 was compared with measurements at TP2 for both Fea1 (Y_{i1j}) and Fea2 (Y_{i2j}). Both features were logarithmically transformed be-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114328, IEEE Access

TABLE 5. Descriptive statistics for SARA Speech and Objective Measures (Fea1 and Fea2)

Descriptive Statistics		Objective Measures		SARA Speech
		Fea1	Fea2	
Mean	TP1	0.003	14.176	1.67
	TP2	0.007	3.9	2.10
S.D.	TP1	0.002	9.35	1.3
	TP2	0.004	1.63	1.59
TP1 versus TP2		p<0.001	p<0.001	n.s.

Captions: S.D. : standard deviation, n.s.: not significant, TP1 versus TP2 : comparison between timepoint 1 and timepoint 2/ KS test at p = 0.001.

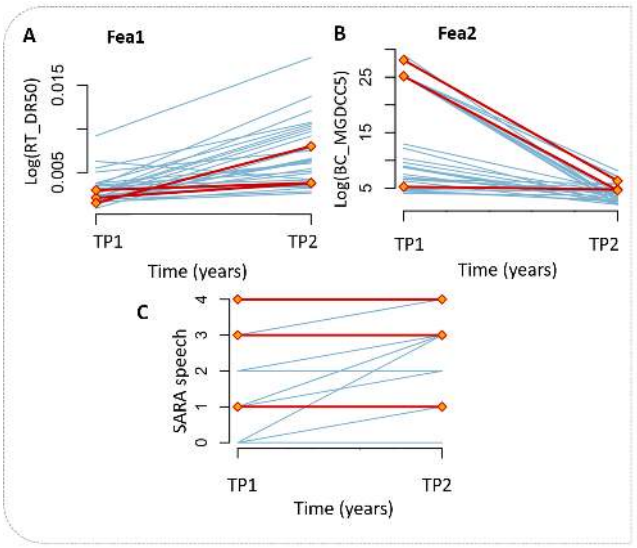

FIGURE 3. Log-transformed change in objective speech parameters (in blue), resulting from the three-step parameter selection criterion **A.** Fea1 and **B.** Fea2 from the two speech tasks and **C.** SARA speech scores, from TP1 to TP2. Red lines indicates the profiles of three selected subjects (ID 5, 10 and 28).

cause they were assumed to have Gaussian distributions. The changes in clusters from TP1 to TP2 for each subject are plotted in Figure 4 and the means are plotted in blue for Group 1 and red for Group 2. Group 1 was thus characterized by a lower Fea1 level at TP1 (baseline). Also, Fea1 increased at a slower rate in Group 1 than in Group 2 (Figure 4(A)). In contrast, Fea2 for Group 1 was remarkably lower at TP1 than for Group 2 and seemed to drop quicker in Group 1 as compared to Group 2(Figure 4(B)).

D. PREDICTION OUTCOMES FOR OUR MODEL, SPA_{CA}

Figure 5 shows the estimates of the posterior distributions of the subject-level component probabilities of the follow-up (TP2) outcomes for two patients in the testing cohort. Patient 49 had a high density of 0.990 (red bar) with a narrow 95% HPD CI (0.970, 1.000) and therefore could be confidently classified into Group 1; conversely, as the posterior probability of Patient 3 belonging to Group 1 was 0 (black vertical line) , with a very narrow 95% HPD CI (1.000, 1.000),they could be classified into Group 2. Majority (∼96%) of the 95% credible intervals included the true observed values in

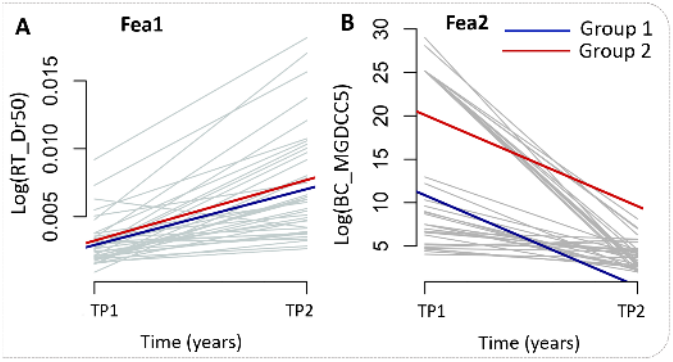

FIGURE 4. Log-transformed change in considered outcomes, **A.** Fea1 and **B.** Fea2 (in grey) along with the estimated cluster specific mean from TP1 to TP2 (Group 1 in blue, Group 2 in red) when classified into $K = 2$ groups.

the training cohort.

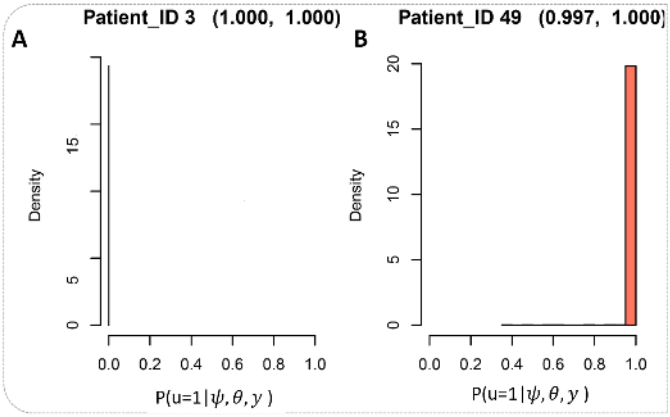

FIGURE 5. Subject-level component probabilities $p_{i,1}(\theta)$ for two subjects,**A.** Patient_ID 3 and **B.** Patient_ID are demonstrated with histograms of the respective sampled values including posterior median 95% HPD CI values.

E. COMPARISON WITH UNIVARIATE MODELS AND MODEL PERFORMANCE

In this study, the optimal model type (univariate\ multivariate) and group size (K=1\ 2\ 3\ 4) were selected using PED [37]. For both the univariate (Figure 6(A), (C)) and multivariate (Figure 6(B)) models, the plots reveal a significant change in the model's deviance when moving from the model with group size K = 1 to the model with group size K = 2.

In the multivariate model (Figure 6(B)), the variability of the posterior distribution of the deviance in a model with group size K = 2 was practically the same as with K = 1. Nonetheless, as opposed to K = 1, the K = 2 deviance posterior distribution was distinctly skewed to the left. When comparing models of group sizes K = 3 and K = 2, as well as K = 4 and K = 3, a similar observation was made.

Under the assumption that there were no associations between the outcomes, Fea1 and Fea2, related methods were used to compare the univariate models (Figure 6(A, B)). The plots indicated that our multivariate model (Figure 6(C)) ($PED_{Fea1+Fea2}$=5175; K=2) performed better than the univariate models (PED_{Fea1}=4225, PED_{Fea2}=3850; K=2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114328, IEEE Access

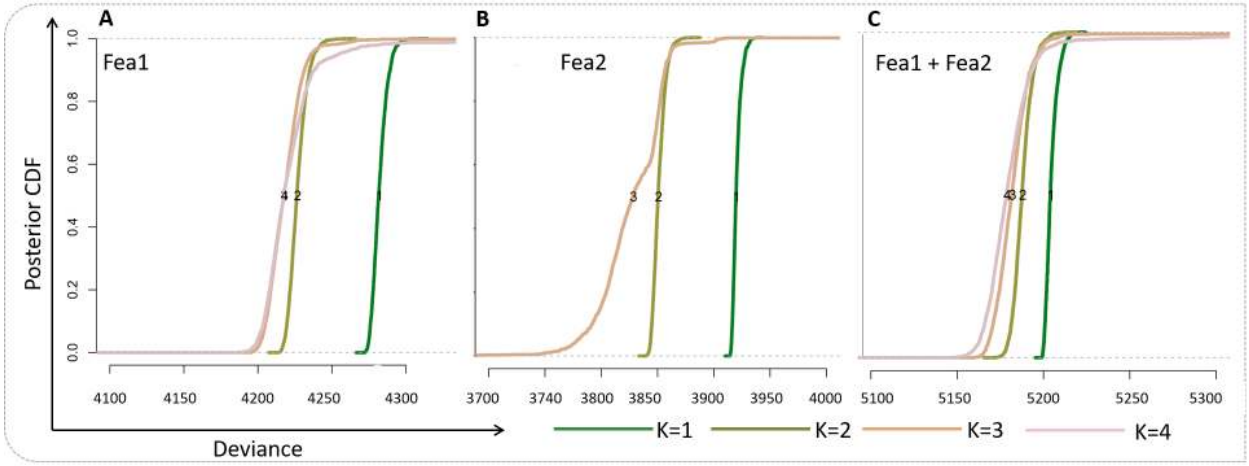

FIGURE 6. Comparisons of the posterior cumulative distribution functions of the observed data deviances for univariate models with parameters **A.** Fea1 and **B.** Fea2 and multivariate models with **C.** multivariate parameters (Fea1 + Fea2), when classified into K = 1; 2; 3; 4. groups.

However, further confirmation would require a larger trial or cohort analysis.

IV. DISCUSSION

In this study, a machine-learning framework based on the multivariate extension of a mixture generalized linear mixed model (GLMM) was introduced for assessing whether an automated analysis of Ataxic Speech parameters could predict deterioration in speech ataxia. Our proposed framework yielded accurate and meaningful results. This led to the identification of a reliable feature profile, specific to speech ataxia obtained from a dataset of 74 speech recordings collected at baseline and approximately 2 years later.

The descriptions of ataxic speech in the literature mostly highlight variability in syllable and pause durations [16], [41], [42] and variations in loudness [43]. We initially selected a feature-set of 29 time and cepstral domain features from our previous objective assessments [24] to examine their contributions to assessing the change in speech ataxia from TP1 to TP2. A robust three-step feature selection criterion was adopted to select the objective measures that best detected the worsening of speech ataxia with greater sensitivity than clinical scales such as the SARA. Interestingly, the three-step feature selection resulted into two heterogeneous objective features, RT_Dr50 (Fea1) and BC_MGDCC5 (Fea2), one from each C-V speech task, indicating that features from different speech tasks capture complementary information on the ataxic symptoms co-existing in different speech dimensions.

In a clinical CA assessment routine, only the most recently available SARA score representing the current patient status is normally used to determine the severity of a patient's speech ataxia. Clearly, such a protocol disregards the existing data on symptoms' evolution over time, which could be more relevant for an accurate classification than simply the last recorded state. To address this shortcoming, the present study proposed a clustering method to jointly consider the entire history of the changes in the profiles of the measurements of all considered parameters. In the framework design, the computational complexity arising from the multivariate extension of the mixture GLMM was handled using a Markov chain Monte Carlo (MCMC) simulation based Bayesian data-driven clustering approach [24]. Interestingly, at a pre-specified time point from the start to follow-up, the changes in profiles of the selected speech parameters over time could be used to classify patients into groups with similar characteristics.

Cluster analysis using the objective measures at the two time points permitted the classification of patients into two groups based on the amplitude of change between the two time points. Group 1 was differentiated from Group 2 by lower values of Fea1 and Fea2 at TP1 and smaller changes in Fea1 and Fea2 between TP1 and TP2 (Figure 4). Most (95%) Group 1 patients had a disease duration of $<$8 years at TP1 whereas in Group 2, the disease duration was $>$8 years. The rate of change (increase\ decrease) in the speech parameters (Fea1\ Fea2) indicated that the speech impairments progressed at a different rate in CA patients who were in their early years of the disease (Group 1) as compared to those in their later years of the disease (Group 2). In contrast, the respective SARA speech scores showed no association with the disease duration, with 11/ 37 CA patients indicating the same SARA speech scores at the two timepoints. This suggested that the proposed method is more sensitive than clinical rating in assessing change in Ataxic Speech over 2 years.

The proposed framework demonstrated promising results with most ($\sim$ 96%) observed values falling within the 95% credible intervals in the training cohort. Multivariate outcomes showed improved performance in prediction compared to univariate models. Notably, to date, the estimation of multivariate outcomes, their changes in trajectory and their joint modeling have not been rigorously studied. Such studies will contribute to a deeper understanding of the heterogeneity of CA and its progression in each manifested domain (for example, speech, as in the present study). Additionally, this

work could cross-examine questions that are clinically and translationally relevant; in particular, designing a patient-specific scheme to predict disease progression using multi-domain (for example, kinematic data extracted from upper-limbs, lower-limbs, gait and balance) longitudinal objective metrics.

V. CONCLUSION

In conclusion, a multivariate mixed-effect model-based framework was applied to predict the progression of Ataxic Speech by jointly incorporating the correlations among multiple outcomes and the correlations among repeated measurements. This is the first study to evaluate the changes in multiple clinically relevant quantitative acoustic parameters over time to predict the progression of Ataxic Speech. The predictive performance of joint multivariate modeling was analysed and compared to univariate modeling using the training cohort. The non-inferiority of the joint multivariate modeling in terms of bias and PED showed the potential of the proposed model for estimating the rate of progression of Ataxic Speech in a clinical environment.

ACKNOWLEDGMENT

The authors would like to thank the Royal Victorian Eye and Ear Hospital (RVEEH), the Florey Institute of Neuroscience and Mental Health, Melbourne, Australia and CSIRO Data61 for their research support.

REFERENCES

[1] W. Ziegler and H. Ackermann, "Subcortical contributions to motor speech: phylogenetic, developmental, clinical," Trends in Neurosciences, vol. 40, no. 8, pp. 458–468, 2017.

[2] H. J. Chenery, J. C. Ingram, and B. E. Murdoch, "Perceptual analysis of the speech in ataxic dysarthria," Australian Journal of Human Communication Disorders, vol. 18, no. 1, pp. 19–28, 1990.

[3] T. Schmitz-Hübsch, S. T. Du Montcel, L. Baliko, J. Berciano, S. Boesch, C. Depondt, P. Giunti, C. Globas, J. Infante, J.-S. Kang et al., "Scale for the assessment and rating of ataxia: development of a new clinical scale," Neurology, vol. 66, no. 11, pp. 1717–1720, 2006.

[4] T. L. Monte, E. da Rosa Reckziegel, M. C. Augustin, L. D. Locks-Coelho, A. S. P. Santos, G. V. Furtado, E. P. de Mattos, J. L. Pedroso, O. P. Barsottini, F. R. Vargas et al., "The progression rate of spinocerebellar ataxia type 2 changes with stage of disease," Orphanet journal of rare diseases, vol. 13, no. 1, pp. 1–8, 2018.

[5] K. Yasui, I. Yabe, K. Yoshida, K. Kanai, K. Arai, M. Ito, O. Onodera, S. Koyano, E. Isozaki, S. Sawai et al., "A 3-year cohort study of the natural history of spinocerebellar ataxia type 6 in japan," Orphanet journal of rare diseases, vol. 9, no. 1, pp. 1–8, 2014.

[6] H. Jacobi, P. Bauer, P. Giunti, R. Labrum, M. Sweeney, P. Charles, A. Dürr, C. Marelli, C. Globas, C. Linnemann et al., "The natural history of spinocerebellar ataxia type 1, 2, 3, and 6: a 2-year follow-up study," Neurology, vol. 77, no. 11, pp. 1035–1041, 2011.

[7] T. Ashizawa, K. P. Figueroa, S. L. Perlman, C. M. Gomez, G. R. Wilmot, J. D. Schmahmann, S. H. Ying, T. A. Zesiewicz, H. L. Paulson, V. G. Shakkottai et al., "Clinical characteristics of patients with spinocerebellar ataxias 1, 2, 3 and 6 in the us; a prospective observational study," Orphanet journal of rare diseases, vol. 8, no. 1, pp. 1–8, 2013.

[8] T. Schmitz-Hübsch, R. Fimmers, M. Rakowicz, R. Rola, E. Zdzienicka, R. Fancellu, C. Mariotti, C. Linnemann, L. Schöls, D. Timmann et al., "Responsiveness of different rating instruments in spinocerebellar ataxia patients," Neurology, vol. 74, no. 8, pp. 678–684, 2010.

[9] E. Schalling, B. Hammarberg, and L. Hartelius, "A longitudinal study of dysarthria in spinocerebellar ataxia (sca): aspects of articulation, prosody, and voice," Journal of Medical Speech-Language Pathology, vol. 16, no. 2, pp. 103–118, 2008.

[10] S. Jannetts and A. Lowit, "Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures," Journal of Voice, vol. 28, no. 6, pp. 673–680, 2014.

[11] S. Luna-Webb, "Comparison of acoustic measures in discriminating between those with friedreich's ataxia and neurologically normal peers," 2015.

[12] B. Kashyap, M. Horne, P. N. Pathirana, L. Power, and D. Szmulewicz, "Automated topographic prominence based quantitative assessment of speech timing in cerebellar ataxia," Biomedical Signal Processing and Control, vol. 57, p. 101759, 2020.

[13] B. Kashyap, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "Quantitative assessment of speech in cerebellar ataxia using magnitude and phase based cepstrum," Annals of biomedical engineering, vol. 48, no. 4, pp. 1322–1336, 2020.

[14] B. Kashyap, D. Szmulewicz, P. N. Pathirana, M. Horne, and L. Power, "Identification of cerebellar dysarthria with siso characterisation," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017, pp. 479–485.

[15] B. Brendel, H. Ackermann, D. Berg, T. Lindig, T. Schölderle, L. Schöls, M. Synofzik, and W. Ziegler, "Friedreich ataxia: dysarthria profile and clinical data," The Cerebellum, vol. 12, no. 4, pp. 475–484, 2013.

[16] H. Ackermann and I. Hertrich, "Dysarthria in friedreich's ataxia: timing of speech segments," Clinical linguistics & phonetics, vol. 7, no. 1, pp. 75–91, 1993.

[17] W. Ziegler and K. Wessel, "Speech timing in ataxic disorders: sentence production and rapid repetitive articulation," Neurology, vol. 47, no. 1, pp. 208–214, 1996.

[18] J. J. Sidtis, J. S. Ahn, C. Gomez, and D. Sidtis, "Speech characteristics associated with three genotypes of ataxia," Journal of communication disorders, vol. 44, no. 4, pp. 478–492, 2011.

[19] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger et al., Analysis of longitudinal data. Oxford University Press, 2002.

[20] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," Biometrics, pp. 963–974, 1982.

[21] J. Ouyang, Q. Zhao, E. V. Sullivan, A. Pfefferbaum, S. F. Taper, E. Adeli, and K. M. Pohl, "Longitudinal pooling & consistency regularization to model disease progression from mris," IEEE Journal of Biomedical and Health Informatics, 2020.

[22] C. McCulloch, S. Searle, and J. Neuhaus, "Generalized, linear and mixed models, 2008. hoboken."

[23] R. A. Jhonson and D. W. Wichern, "Applied multivariate statistical analysis," 2007.

[24] A. Komarek and L. Komárková, "Clustering for multivariate continuous and discrete longitudinal data," The Annals of Applied Statistics, pp. 177–200, 2013.

[25] C. E. McCulloch and S. R. Searle, "Generalized, linear, and mixed models (wiley series in probability and statistics)," 2001.

[26] S.-C. Lin, C.-J. Chen, and T.-J. Lee, "A multi-label classification with hybrid label-based meta-learning method in internet of things," IEEE Access, vol. 8, pp. 42 261–42 269, 2020.

[27] A. Komárek, "A new r package for bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data," Computational Statistics & Data Analysis, vol. 53, no. 12, pp. 3932–3947, 2009.

[28] S. W. Ekanayake, A. J. Morris, M. Forrester, and P. N. Pathirana, "Biokin: an ambulatory platform for gait kinematic and feature assessment," Healthcare technology letters, vol. 2, no. 1, pp. 40–45, 2015.

[29] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "Blockchain for secure ehrs sharing of mobile cloud based e-health systems," IEEE access, vol. 7, pp. 66 792–66 806, 2019.

[30] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, and E. Nöth, "Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases," IEEE journal of biomedical and health informatics, vol. 19, no. 6, pp. 1820–1828, 2015.

[31] T. D. Wickens and G. Keppel, Design and analysis: A researcher's handbook. Pearson Prentice-Hall Upper Saddle River, NJ, 2004.

[32] R. M. Warner, Applied statistics: From bivariate through multivariate techniques. Sage Publications, 2012.

[33] H.-Y. Kim, "Statistical notes for clinical researchers: A one-way repeated measures anova for data with repeated observations," Restorative dentistry & endodontics, vol. 40, no. 1, p. 91, 2015.

[34] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.
[35] D. M. Titterington, U. E. Makov, and A. F. Smith, Statistical analysis of finite mixture distributions. Wiley, 1995.
[36] J. Yang, P. Liu, and S. Li, "A common method for detecting multiple steganographies in low-bit-rate compressed speech based on bayesian inference," IEEE Access, vol. 7, pp. 128 313–128 324, 2019.
[37] M. Plummer, "Penalized loss functions for bayesian model comparison," Biostatistics, vol. 9, no. 3, pp. 523–539, 2008.
[38] P. Y. de la Fé Rodríguez, A. Coddens, E. Del Fava, J. C. Abrahantes, Z. Shkedy, L. O. M. Martin, E. C. Muñoz, L. Duchateau, E. Cox, and B. M. Goddeeris, "High prevalence of f4+ and f18+ escherichia coli in cuban piggeries as determined by serological survey," Tropical animal health and production, vol. 43, no. 5, pp. 937–946, 2011.
[39] C. R. B. Cabral, V. H. Lachos, and M. R. Madruga, "Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population," Journal of Statistical Planning and Inference, vol. 142, no. 1, pp. 181–200, 2012.
[40] A. Johansen, "Markov chain monte carlo," in International Encyclopedia of Education (Third Edition), third edition ed., P. Peterson, E. Baker, and B. McGaw, Eds. Oxford: Elsevier, 2010, pp. 245–252.
[41] B. Brendel, M. Synofzik, H. Ackermann, T. Lindig, T. Schölderle, L. Schöls, and W. Ziegler, "Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with friedreich ataxia," Journal of neurology, vol. 262, no. 1, pp. 21–26, 2015.
[42] H. Ackermann and I. Hertrich, "Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing," Folia phoniatrica et logopaedica, vol. 46, no. 2, pp. 70–78, 1994.
[43] R. D. Kent, J. F. Kent, J. R. Duffy, J. E. Thomas, G. Weismer, and S. Stuntebeck, "Ataxic dysarthria," Journal of Speech, Language, and Hearing Research, vol. 43, no. 5, pp. 1275–1289, 2000.