UNSUPERVISED CLUSTERING REVEALS SEVERAL SUBTYPES IN SPEAKERS WITH ATAXIA

Anneke Slis[1], Robin Karlin[1], Benjamin Parrell[1,2]

[1]Waisman Center, University of Wisconsin–Madison, [2]Communication Sciences and Disorders, University of Wisconsin–Madison
slis@wisc.edu, rkarlin@wisc.edu, bparrell@wisc.edu

ABSTRACT

While ataxic dysarthria is traditionally considered a single diagnostic group, recent work suggests that distinct subtypes exist within this population. Here, we explore whether an unsupervised cluster analysis of diadochokinetic speech can detect the presence of potential subgroups, without a priori knowledge about individual speakers. Thirty-eight individuals with degenerative spinocerebellar ataxia produced [oj] repetitively. Fifteen acoustic markers were extracted from each production, consisting of temporally changing energy patterns that capture durational patterns and vocal tract shape changes. Principal component analysis yielded three dimensions (capturing 86% of the total variance), which were used as input for clustering analysis. Across clustering algorithms and evaluation criteria, 2-3 clusters were consistently found. Although it remains to be tested how these clusters align with clinical diagnoses, these results suggest that distinct subgroups of speech impairments do exist within individuals with ataxia, and that these subgroups may be automatically detectable from a simple speech task.

Keywords: ataxic dysarthria, clustering, subtypes

1. INTRODUCTION

Ataxic dysarthria is a disorder of the speech motor system, caused by a range of conditions involving the cerebellum [1-3]. Speech is characterized by distorted phonation (e.g., [3]), impaired articulatory control, and deficits in temporal coordination [2]. Articulatory difficulties surface as imprecise consonants as well as distorted vowel productions, while temporal impairments include elongated speech segments, reduced speaking rate, scanning speech (overly isochronous syllable durations), and irregular alternating motion rates [3, 4].

Although ataxic dysarthria is a single clinical diagnosis, a growing body of evidence suggests that distinct subgroups exist within this population [4-7]. Two distinct patterns within ataxic dysarthric speech have been observed – inflexible and instable speech productions [5-7, 16]. These classifications cover both the temporal and articulatory domain. The subgroup with instability presents with increased variability in both articulation (e.g., variably imprecise consonants) and temporal control (e.g., variable stress and rhythm). Conversely, the group with inflexibility displays extremely persistent speech patterns, including consistent articulatory errors, scanning speech, and isochronous durations. In a series of studies [5-7], the two categories have been used to perceptually identify subgroups within ataxic dysarthria. For example, out of 10 speakers in [5], five were classified with instability, one with inflexibility, and four with impairments combining features of both instability and inflexibility. Some evidence suggests that these subtypes align with specific causes of cerebellar damage [4]. Regularity of diadochokinetic speech has been shown, for example, to be impaired in patients, diagnosed with spinocerebellar ataxia type SCA3, whereas rate and prosodic modulation were affected in type SCA6 [4].

To date, identifying potential subgroups within ataxic dysarthria has relied largely on perceptual ratings by experienced, clinically trained speech language pathologists. Despite extensive experience of clinicians, however, identifying perceptual subgroups shows only moderate inter-rater reliability [5]. Conversely, a recent study using naïve listeners to classify dysarthria subtypes without a priori knowledge of the speech characteristics of ataxic dysarthria showed higher reliability [7] but depended on a large number of listeners (>20).

Previous work examining potential subgroups within ataxic dysarthria has relied principally on perceptual characterizations, rather than objective acoustic measures. To validate whether these perceptually identified subgroups exist, an automated tool that provides empirical and reliable identification of subgroups within ataxic dysarthria would be beneficial. Successful automated classification of ataxic dysarthria subtypes not only benefits clinical diagnosis but also lends theoretical support to the existence of these subtypes themselves.

As a step towards such an automated tool, the current study attempts to identify subgroups within speakers diagnosed with degenerative spinocerebellar ataxia using unsupervised clustering. Although a similar approach, using linear discriminant analysis,

has previously been shown to identify clusters within this population [4], this study relied only on temporal and amplitude variation in the speech signal, omitting the clinically critical aspect of articulatory precision. Here, we include this factor in an unsupervised clustering analysis by using acoustically derived measures, quantifying changes in spectral energy over time [8-9]. Importantly, these energy changes correlate well with articulatory trajectories [8] and have recently been shown to be effective in separating hypokinetic, spastic, and flaccid dysarthria [9, 10]. A particularly beneficial characteristic is that this method separates supra-laryngeal articulation from the phonatory source. Consequently, it is possible to examine laryngeal as well as supra-laryngeal contributions to ataxic dysarthria subtypes. Our results show that unsupervised clustering based on these measures, extracted from diadochokinetic speech, consistently identifies subgroups within speakers with cerebellar degeneration.

2. METHODS

2.1. Participants

38 participants (14 M, mean age: 60, range: 31-81; 24 F, mean age: 60, range 23-77) participated in the study. Individuals were recruited via CoRDS [14] and were previously diagnosed with degenerative spinocerebellar ataxia (three SCA1; nine SCA2; ten SCA3; twelve SCA6; four SCA8). All speakers were native speakers of American English and reported no additional history of hearing, speech, or language disorders. The participants were compensated monetarily for participating in the study. The study was approved by the Institutional Review Board at the University of Wisconsin – Madison.

2.2 Materials and equipment

Diadochokinetic speech was collected during an assessment of dysarthria, including oral motor skill evaluation. For each task, participants were instructed to repeat a particular syllable as quickly and accurately as possible. In the current study, we examined the repetitive production of the syllable [oj]. This sequence targets anterior-posterior tongue motions, potentially revealing possible articulatory deficiencies in lingual control. The speech material was collected online via Zoom. The participants used their home computer and equipment (see 4. DISCUSSION about possible limitations). They were sitting in front of their computer at an individually preferred distance from the microphone, in the same fashion as during Zoom meetings. The recordings were sampled at 32 kHz (bit rate: 126 kbps) in a m4a format. All data processing, except where noted, was conducted in MATLAB [17].

2.3 Acoustic parameter extraction

The procedure has been described in detail previously [8, 9]. A summary follows here. For each participant, the .m4a sound files were converted to .wav files (32 Khz, 16-bit PCM) in Audacity [18]. The mean peak intensity of the extracted sounds was 70.5 dB, with a low standard deviation across participants (1.92 dB) indicating a relatively consistent intensity level across the recording sessions. Inspecting the spectrum and the closest zero crossings, the start and end of the target sequences were labelled in PRAAT [11]. Only sequences that were produced without an audible breath or a large pause were selected. In the case of multiple good productions, the sequences with the most repetitions of the target syllable were selected. The resulting annotated speech signals were imported in MATLAB, resampled to 16 kHz, and pre-emphasized to account for the high-frequency roll-off of the glottal source. Speech spectra were generated using 25 ms windows with 2 ms steps between samples. Frequencies between 300 and 4000 Hz were considered for Mel-frequency cepstral coefficients (MFCC) analysis. 20 Mel-frequency filter banks processed these spectra to emphasize the characteristics of the human ear [12]. 15 MFCCs were then extracted by log-transforming the resulting spectral values and applying a discrete cosine transformation.

The first coefficient reflects the average spectral energy, containing source information, and was analysed to quantify speech amplitude changes. To capture articulatory changes over time, the first differential of each of the remaining MFCCs (2-15) was computed, and these values were summed at each time point. This resulted in a trajectory with “spectral energy difference” values with peaks indicating large changes in energy (e.g., changes from [o] to [j] and [j] to [o]) and valleys indicating small changes in energy (e.g., during the steady states of vocoids; see Figure 1). These changes in spectral energy are highly correlated with vocal tract shape changes, which indirectly represent articulatory movement patterns [8]. Consequently, the magnitude of spectral energy changes indicates the speed of articulatory motions (i.e., from [o] to [j] and vice versa) and variability in these changes relates to consistency in articulatory movements. Temporal distances between maxima and minima reflect articulatory timing patterns, from which timing variability can be extracted. In total, 15 measures were extracted from the trajectories (Table 1 and Figure 1 and 2).

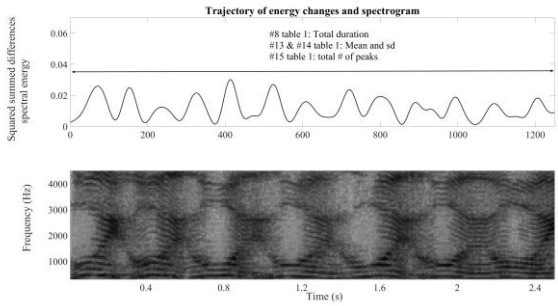

Figure 1: Diadochokinetic speech sample. Top: the sum of the first differential of MFCC 2-15 over time, reflecting a global metric of spectral change related to supra-laryngeal articulation. Bottom: time-aligned speech spectrogram.

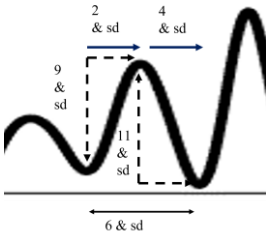

Figure 2: Temporal and articulatory parameters extracted from the spectral energy difference signal in Figure 1 (appr. 200 to 420 ms). See Table 1 for details.

2.4 Data processing and clustering algorithms

To determine the relevant principal components for the cluster analysis, the matrix containing the raw acoustic parameter values for each participant was normalized by centering and z-scoring each parameter. Resulting values for the 15 extracted parameters were transformed into three principal components (PCs), using the PCA algorithm in MATLAB (see 3. RESULTS). The output loadings were further rotated with ROTATEFACTORS on these three PCs, with a standard varimax rotation. The three PCs served as input for the unsupervised clustering process.

To assess the reliability of our results, two separate clustering algorithms were used: k-Means clustering, which assumes equal variances across dimensions and groups (KMEANS in MATLAB) and the Expectation-Maximization (EM) algorithm for Gaussian Mixture models, allowing variances to differ [15]. Potential clusters were generated using each algorithm for one to seven groups (inclusive). The optimal number of clusters was then determined using the Calinski-Harabasz criterion [13] with EVALCLUSTERS. Because unsupervised clustering often identifies local, rather than global minima, and thus has the potential to yield different results across iterations depending on initial conditions, this process was repeated 3000 times for each clustering algorithm.

	Amplitude descriptor
1	Amplitude coefficient of variation of the first MFCC (the standard deviation divided by the mean).
	Temporal descriptors
2	Duration from valley (i.e., steady state) to following peak (transition from [o] to [j] and vice versa): large values indicate slower vocal shape changes.
3	The standard deviation of these durations.
4	Duration from peaks to following valleys.
5	The standard deviation of these durations.
6	Duration from valleys to valleys.
7	The standard deviation of these durations.
8	The total sequence duration.
	Articulatory descriptors
9	The magnitude of the difference between valleys and following peaks (analogous to the speed of vocal tract shape changes from steady state (valley; no vocal tract shape change) to peak (fast vocal tract shape change from [o] to [j]).
10	Standard deviation of these differences.
11	The magnitude of the difference between peaks and following valleys (analogous to the rate of vocal tract shape (peak) compared to steady state (valley).
12	The standard deviation of these differences.
13	The mean of all the values of spectral energy changes across the full sequence.
14	The standard deviation of these magnitudes
15	Number of peak-to-peak cycles in sequence.

Table 1: Temporal and articulatory parameters extracted from the spectral energy difference signal.

3. RESULTS

Observing the eigenvalues of the PCA, only three factors were retained, as after the third dimension the curve tapered off. The three dimensions together explained the 86% of the variance: 40%, 32%, and 14% for the first, second and third PCs respectively. The first PC weighted heavily towards the articulatory descriptors, the second PC towards the temporal descriptors, and the third towards the total sequence duration, number of syllable repetitions, and variation in speech amplitude (see Table 2). Using these PCs, the EM clustering analysis identified two to three clusters (46% and 37% of the iterations respectively), while the k-Means analysis more consistently identified three clusters (76% of the iterations).

		PC 1	PC 2	PC 3
1	COV first MFCC	-0.11	0.09	**0.25**
	Temporal descriptors			
2	M valley-peak	0.08	**0.41**	-0.10
3	SD #2	-0.03	**0.37**	0.16
4	M peak-valley	0.04	**0.42**	-0.10
5	SD #4	-0.08	**0.35**	0.20
6	M valley-valley	0.05	**0.44**	-0.11
7	SD #6	-0.05	**0.40**	0.13
8	Total duration	0.09	-0.01	**0.62**
	Articulatory descriptors			
9	M valley to peak	**0.41**	0.01	-0.09
10	SD #9	**0.39**	0.05	0.08
11	M peak to valley	**0.41**	0.01	-0.09
12	SD #11	**0.39**	0.06	0.06
13	Mean contour	**0.38**	-0.11	-0.04
14	SD #13	**0.41**	-0.01	0.03
15	# peaks	0.06	-0.11	**0.64**

Table 2. Columns show loadings for each PC after Varimax rotation. Loadings with values equal or higher than 0.25 are shown in bold.

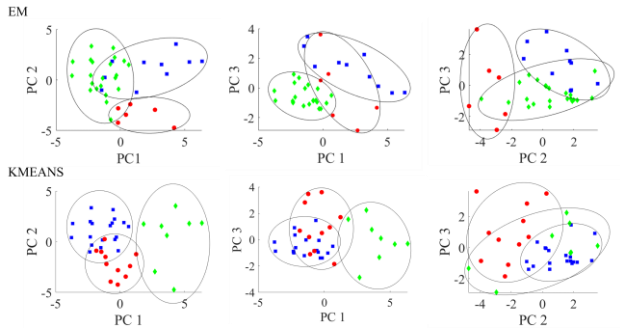

Figure 3. Clusters of speakers as grouped by EM (upper) and k-Means (lower) algorithms.

Several patterns can be observed (see Figure 3). For the k-Means algorithm, first and second PCs separated all three clusters efficiently; the first PC distinguishes between one group of two clusters (round (red) and squared (blue) dots) and a third group (diamond (green) dots), while the second and PC distinguishes between the red and blue groups. For the EM algorithm, the three clusters are less clearly distinguished by any single pair of PCs.

4. DISCUSSION

This study explored whether an unsupervised cluster analysis of diodochokinetic speech detected potential subgroups of dysarthria in speakers with spinocerebellar ataxia, without a priori knowledge about clinical diagnoses. The study showed that k-Means and, to a lesser extent EM, identified distinct clusters of speakers. Subgroups were distinguished along dimensions consistent with the first three PCs, representing articulatory features, temporal features, and utterance length/number of peaks/amplitude variation, respectively. These results are consistent with the notion that distinct subtypes of ataxic dysarthria exist [4-7] and can be distinguished based on unique clusters of symptoms. Observing the PC loading factors suggests that including supra-laryngeal parameters helps to distinguish these groups, in addition to the temporal and phonatory features employed in [4]. The emergence of three clusters is promising – an average of 3.9 clusters were recognized perceptually for diadochokinetic speech by naïve listeners in [7]; subsequent hierarchical clustering analyses extracted two main clusters from these data, which mapped onto the "inflexibility"-"instability" distinction. Future work will explore the extent to which the clusters detected with unsupervised learning in our study align with clinical diagnoses of inflexibility and instability.

Despite the promising results, the identified clusters do show overlap in some PCs; a part of the population likely shares characteristics of both subgroups and are thus not likely to form well-separated clusters [5]. With a larger sample, we expect that the groups will become more well-defined. A larger sample may enable us to determine the extent to which these clusters relate to different subtypes of spinocerebellar ataxia as well as other causes of ataxic dysarthria.

One potential methodological issue with the current study is that the material consisted of home recordings; the quality of the recordings is lower than in the lab and less consistent between participants, which influences the subsequent MFCC analysis [19, 20]. The data should thus be interpreted with caution, especially regarding the first parameter, intensity [19], which has been shown to be unstable with, especially, older versions of Zoom, and should be validated with material collected in the lab. That said, our results suggest well-defined groups even with possible distortions of the acoustic signal. Importantly, the ability to record speakers with rare disorders online provides a substantially larger sample size than would otherwise be available.

In sum, unsupervised clustering based on parameters extracted from energy changes over time is a promising approach in distinguishing dysarthria subgroups in ataxia. Though the extent to which the identified groups reflect the inflexibility-stability dichotomy remains to be tested, this method has potential to complement or replace perceptual assessments of different subtypes of ataxic dysarthria in clinical settings.

5. ACKNOWLEDGEMENTS

Work was supported by NIDCD/NIH awards R01 DC017091, F32DC0193535, and P50HD105353.

6. REFERENCES

[1] Duffy, R., *Motor Speech Disorders*, 4th ed., 2019. https://www.elsevier.com/books/motor-speech-disorders/duffy/978-0-323-53054-5 (accessed May 19, 2020).

[2] Darley, F. L., Aronson, A. E., and Brown, J. R., 'Differential diagnostic patterns of dysarthria', *J. Speech Hear. Res.*, vol. 12, no. 2, pp. 246–269, Jun. 1969, doi: 10.1044/jshr.1202.246.

[3] Kent, R. D., 'Research on speech motor control and its disorders: a review and prospective', *J. Commun. Disord.*, vol. 33, no. 5, pp. 391–427, Oct. 2000.

[4] 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, doi: 10.1007/s00415-014-7511-8.

[5] Spencer, K. A., and France, A. A., 'Perceptual ratings of subgroups of ataxic dysarthria', *Int. J. Lang. Commun. Disord.*, vol. 51, no. 4, pp. 430–441, Jul. 2016, doi: 10.1111/1460-6984.12219.

[6] Boutsen, F., Bakker, K., and Duffy, J., 'Subgroups in ataxic dysarthria', *J. Med. Speech-Lang. Pathol.*, vol. 5, pp. 27-36, Mar. 1997.

[7] K. A. Spencer, J. Amaral, and K. Lansford, 'Investigating perceptual subgroups in speakers with ataxic dysarthria: An auditory free classification approach', *Am J. Speech Lang Pathol.*, pp. 1-11, 2022, 10.1044/ 2022_AJSLP-22-00159.

[8] Goldstein, L., 'The role of temporal modulation in sensorimotor interaction', *Front. Psychol.*, vol. 10, p. 2608, Dec. 2019, doi: 10.3389/fpsyg.2019.02608.

[9] Slis, A., Lévêque, N., Fougeron, C., Pernon, M., Assal, F., and Lancia, L., 'Analysing spectral changes over time to identify articulatory impairments in dysarthria', *J. Acoust. Soc. Am.*, vol. 149, no. 2, p. 758, Feb. 2021, doi: 10.1121/10.0003332.

[10] Lévêque, N., Slis, A., Lancia, L., Bruneteau, G., and Fougeron, C., 'Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases', *J. Speech Lang. Hear. Res.*, vol. 65, no. 5, pp. 1767–1783, May 2022, doi: 10.1044/2022_JSLHR-21-00434.

[11] Boersma, P., and Weenink, D., Praat: doing phonetics by computer [Computer program], Version 6.1.03, retrieved Spetember 1, 2019 from http://www.praat.org/.

[12] O'Shaughnessy, D., *Speech communication: human and machine*. Addison-Wesley Pub. 6Co., 1987.

[13] Caliński, T. and Harabasz, J., 'A dendrite method for cluster analysis', *Commun. Stat..*, vol. 3, no. 1, pp. 1–27, Jan. 1974, doi: 10.1080/03610927408827101.

[14] CoRDS (https://research.sanfordhealth.org/rare - disease-registry/researchersj).

[15] Chen, M., 'EM Algorithm for Gaussian Mixture Model (EM GMM)', https://www.mathworks.com/matlabcentral/fileexchange/26184-em-algorithm-for-gaussian-mixture-model-em-gmm, Retrieved November 15, 2022.

[16] Schalling, E., Hammarberg, B., and Hartelius, L., 'A longitudinal study of dysarthria in spinocerebellar ataxia (SCA): aspects of articulation, prosody, and voice', *Journal of Medical Speech–Language Pathology*, 16, 103–117, 2008.

[17] The MathWorks, Inc., MATLAB version: 9.13.0 (R2022a). 2022.

[18] Audacity Team, Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.2.1 retrieved October 27, 2022, from https://audacityteam.org/.

[19] Zangh, C., Jepson, K., Loblink, G., and Arvanti, A., 'Comparing acoustic analyses of speech data collected remotely', *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp 3910-3916, 2021, doi: 10.1121/10.000513.

[20] Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., and Bowern, C., '(Don't) try this at home! The effects of recording devices and software on phonetic analysis', *Language*, vol. 97, no 4, 2021, pp e360-e382, doi:10.1353/lan.2021.0075.