Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Clustering approaches to dysarthria using spectral measures from the temporal envelope

Eugenia San Segundo[1], *Jonathan Delgado*[2], *Lei He*[3]

[1] Phonetics Laboratory, Spanish National Research Council (CSIC), Madrid, Spain
[2]Dept. of Developmental & Educational Psychology, La Laguna University, Spain
[3]Dept. of Computational Linguistics, University of Zurich, Switzerland
eugenia.sansegundo@cchs.csic.es, extjdelgado@ull.edu.es, lei.he@uzh.ch

Abstract

Several clustering techniques were used for finding subgroups of speakers sharing common characteristics within a sample of 14 dysarthric speakers and 15 non-dysarthric speakers. Our classifying variables were five spectral measures computed from the temporal envelope of each of the four sentences read by the participants. The unsupervised k-means clustering algorithm showed that the optimal number of clusters in this dataset is two, with Cluster 1 matching almost exactly the dysarthric population and Cluster 2 the non-dysarthric population. As for the importance of each variable, a PCA analysis revealed that centroid, spread, rolloff and flatness contribute equally to the first component, and entropy contributes to the second component. Hierarchical agglomerative clustering further supported the separation into two main clusters (highlighting the relevance of these rhythmic measures to characterize dysarthria), but also allowed us to detect possible subgroups within each main speaker group.

Index Terms: dysarthria, rhythm, temporal envelope, Spanish, clustering techniques

1. Introduction

Dysarthria is a speech disorder stemming from neurological factors that causes difficulties in both speech motor programming and execution [1]. Assessing this condition is difficult given the diversity of its symptoms [2] and requires the combination of objective and subjective tests [3]. This paper[i] is an exploratory investigation into the rhythmic characteristics of dysarthria using clustering techniques.

Speech rhythm is viewed here from MacNeilage's frame/content perspective, where the mouth opening-closing cycles (and thence the temporal modulations) constitute the rhythmic frame in speech [4]. As in [5], the method followed in this article for speech rhythmic characterization consists in the extraction of five common spectral measures (centroid, spread, rolloff, flatness, and entropy), computed from the temporal envelope of sentences in read speech of dysarthric and non-dysarthric speakers. In [5] a binomial logistic regression model showed that dysarthric speakers presented a significantly lower centroid and lower spread. In this paper different clustering approaches are explored in order to find subgroups of speakers sharing common rhythmic characteristics.

1.1. Previous studies on dysarthria

Rhythmic disturbances are one of the most common features of dysarthria. All types of dysarthria affect to a greater or lesser degree the articulation of consonants, and in the most severe cases vowel distortions are observed [6]. However, there are no phonological constraints, it is at the level of articulatory implementation where these motor speech disorders have their impact on the emergent flow of syllabic flow and the perceived rhythm of speech [7]. Although rhythm disturbance is a very common feature of dysarthria, speech rhythm is the least studied prosodic element [8]. The study of rhythm in these motor speech disorders generally aims to distinguish between healthy and dysarthric speech, and to establish the severity of dysarthria. For this purpose, acoustic measures of the duration of vowel and consonant intervals in continuous speech (%V) and of the variability of these durations, both raw (ΔV, ΔC, VarcoV, VarcoC, VarcoVC, rPVI-C, rPVI-VC) and normalized (nPVI-V, nPVI-VC) [6-9] have traditionally been used.

Despite the usefulness of these rhythmic metrics in distinguishing healthy speech from dysarthria [6-9] and in discriminating levels of dysarthria severity [6,8], results are sometimes disparate. For instance, [7] concluded that acoustic measures of vocalic and consonantal segment durations allow to distinguish control speech from dysarthria and to discriminate dysarthria subtypes, while another study [10] found that none of the rhythm metrics based on segmental durations could differentiate disordered from healthy speakers, "despite clear perceptual differences, suggesting that factors beyond segment duration impacted on rhythm perception" [10, p.1]. For this reason, a different acoustic method is proposed here, which analyses rhythm from the speech temporal envelope.

Although cluster analyses are common in recent dysarthric studies [11-15], they mostly focus on ataxic dysarthria [11-13], and on English speakers [11-14]. To the best of our knowledge this is the first study on a variety of dysarthric types, on Spanish language and following a particular approach to speech rhythm.

1.2. Speech rhythm: The frame/content perspective

Approaches to speech rhythm are diverse, focusing on different aspects of the (semi-)regularities and variabilities in speech production (see [16] for a general overview). The frame/content theory [4] provided a unifying perspective of these approaches [16], and is also the theoretical footing of the method used in this study. MacNeilage [4] argued that speech rhythm is a primordially developed phenomenon in speech production. It evolved from pre-existing cyclical mandibular movements in ancestral primates in the form of lip-smacking [4]. Subsequent research confirmed that such cyclical lip movements were important visuo-facial gestures in extant non-human primate communications [17]. It is also believed that the coupling between mouth opening-closing cycles and vocalization emerged in the course of human evolution: the sonority of

10.21437/SpeechProsody.2024-17

speech typically increases and decreases with mouth opening and closing gestures [17, 18]. Such opening-closing alternations are organized into syllable-sized units corresponding to the temporal modulations, which constitute the rhythmic frames; the open and closed phases are filled with vocalic and consonantal contents. In other words, the mouth opening/closing movements are mirrored in the temporal envelope of the speech signal, and should carry information about disorders in the motor control. The temporal regularities of an envelope can be characterized by analyzing its spectrum.

2. Method

2.1. Participants

Originally, 30 subjects voluntarily participated in this study: 15 with dysarthria (mean age 42.93, SD 10.31) and 15 neurologically healthy (mean age 41.86, SD 13.62). The two experimental groups (dysarthria and control) were sex matched. They were all speakers of Canarian Spanish. Within the dysarthric group, 10 participants presented ataxic dysarthria, 2 spastic dysarthria and 3 mixed dysarthria. After a preliminary analysis, one dysarthric participant presenting mixed dysarthria was discarded (Spkr #24). Her audio samples presented signal saturation and stammering. These aspects were deemed unfit for the type of acoustic analyses. Table 1 shows the general characteristics of dysarthric speakers.

Table 1: *Characteristics of dysarthric speakers. CVA: cerebrovascular accident; ALS: amyotrophic lateral sclerosis; CP: cerebral palsy; SCA-7: spino cerebellar ataxia-7; CBD: corticobasal degeneration; CCT: cranio-cerebral trauma*

Speaker	Sex	Age	Diagnosis	Dysarthria type
16	F	34	CVA	Ataxic
17	F	48	ALS	Spastic-flaccid
18	F	33	Tumor	Ataxic
19	M	21	CP	Spastic
20	F	30	CP	Spastic
21	F	51	CVA	Ataxic
22	F	40	Tumor	Ataxic
23	M	55	ALS	Spastic-flaccid
25	M	49	SCA-7	Spastic-ataxic
26	F	59	CBD	Ataxic
27	F	45	Tumor	Ataxic
28	M	39	CCT	Spastic-ataxic
29	F	47	CVA	Ataxic
30	F	52	Tumor	Ataxic

2.2. Recording setup and speech samples

All recordings were conducted in a soundproof booth with an AKG C544L head-mounted condenser microphone. They were digitized at a sampling rate of 44.1 kHz and 16 bits of resolution using the audio interface Alesis io2 express. The signal-to-noise ratio (SNR) was measured post hoc to check the level of environmental noise of the voice recordings. All samples were consistent with the recommended threshold proposed by [19]. The speech material consisted in reading aloud four phonetically balanced sentences of the Spanish Matrix Sentences Test [20].

2.3. Acoustic analysis

First, the acoustic signal per sentence was bandpass filtered between 700 and 1300 Hz (100 Hz smoothing) to keep the vocalic energy while removing the glottal energy and obstruent noise. This filter has been used to detect the P-centers or “beats” in the speech signal [21]. Then, the filtered signal was full-wave rectified and downsampled to the Nyquist frequency of 20 Hz, yielding the temporal envelope. Five common spectral measures (CENTROID, SPREAD, ROLLOFF, FLATNESS, and ENTROPY) [21] were calculated from the temporal envelope of each sentence. Among these five spectral measures, the CENTROID calculates the “balancing point” in the coherence and serves as a point estimate of the coherence. The SPREAD calculates to what extent the coherence disperses around the centroid. The ROLLOFF indicates the degree of skewness in the coherence. The FLATNESS and ENTROPY quantify the amount of unpredictability or disorder in the spectrum. Taken together, they provide an overview of the shape of the temporal envelope.

2.4. Statistical analysis

All statistical procedures were performed using R [22]. We ran two types of clustering analyses: k-means clustering first, then hierarchical clustering.

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given dataset into a set of *k* groups (i.e. *k* clusters), where *k* represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups, such that objects within the same cluster are as similar as possible, whereas objects from different clusters are as dissimilar as possible [23]. In k-means clustering, each cluster is represented by its center (i.e., centroid) which corresponds to the mean of points assigned to the cluster [23]. There are several k-means algorithms available. We used the standard algorithm (i.e. the Hartigan-Wong algorithm [24]), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid. Each observation is assigned to a given cluster such that the sum of squares distance of the observation to their assigned cluster centers is minimized. The *total within-cluster sum of square* measures the compactness of the clustering [23]. In order to determine the optimal number of clusters, we used the average silhouette method, which measures the quality of a clustering. This method computes the average silhouette of observations for different values of *k*. The optimal number of clusters *k* is the one that maximizes the average silhouette over a range of possible values for *k* [25].

Hierarchical agglomerative clustering allows the analyst to obtain a set of nested clusters that are organized as a tree. Each node (cluster) in the tree is the union of its children (subclusters), and the root of the tree is the cluster containing all the objects. This type of clustering starts with each point as a singleton cluster and then repeatedly merges the two closest clusters until a single, all-encompassing cluster remains. A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram, which displays both the cluster-subcluster relationship and the order in which the clusters were merged [26].

3. Results

3.1. K-means clustering

3.1.1. *Optimal number of clusters*

Using the silhouette method to determine the optimal number of clusters in our dataset, the results (Figure 1) show that two clusters maximize the average silhouette values, so they are the optimal number of clusters.

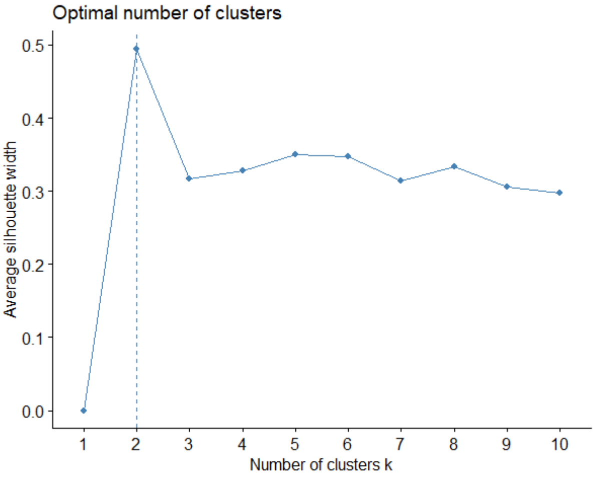

Figure 1: *Optimal number of clusters (silhouette method).*

3.1.2. *K-means with two clusters*

When “2” is specified as the number of centers in the k-means analysis, we obtain two clusters of sizes 56 and 60 (Figure 2), with the cluster means specified in Table 2.

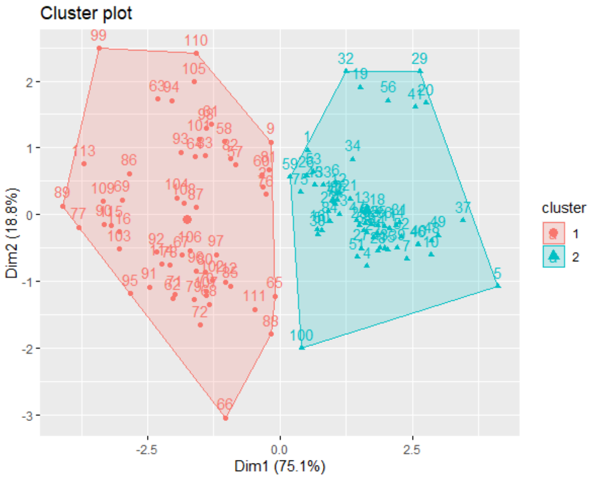

Figure 2: *Cluster plot using k-means with two centers. Each number represents a speaker.*

Table 2: *Cluster means*. Cl. = cluster

Cl.	Centroid	Spread	Flatness	Rollof	Entropy
1	-0.898	-0.838	-0.882	-0.888	-0.205
2	0.838	0.782	0.824	0.828	0.191

Figure 2 shows that cluster division with k-means is based on two dimensions. Dimension 1 accounts for 75.1% of the total variation. Dimension 2 accounts for 18.8% of the variation. Together they can explain 93.9% variation in the dataset.

3.2. Principal Component Analysis (PCA)

We ran a PCA to our dataset in order to find out the relative importance and contribution of each component in the cluster classification. The results (Table 3 and Figure 3) agree with the information shown in Figure 2 in that the two main dimensions or components explaining cluster division explain 93.9% of the variation in the dataset: dimension 1 explains 75.1% of the total variation; dimension 2 accounts for 18.8% of the variation. In addition, Table 3 shows the eigenvalues, variance percentage and cumulative variance percentage associated with the five main components in a PCA test. Figure 3 shows a biplot, which merges a PCA plot and a loadings plot: the former shows clusters of samples based on their similarity; the latter shows how strongly each characteristic influences a principal component. When vectors are close, forming a small angle, the variables that they represent are positively correlated (e.g. CENTROID, SPREAD, FLATNESS and ROLLOF). If they meet each other at 90°, they are not likely to be correlated (the four aforementioned variables and ENTROPY).

Table 3: *PCA. (Dim.= dimension; Cum.= cumulative)*

Dim.	Eigenvalue	Variance%	Cum. variance%
1	3.753	75.07	75.07
2	0.938	18.76	93.83
3	0.200	4.00	97.83
4	0.074	1.49	99.32
5	0.034	0.68	100

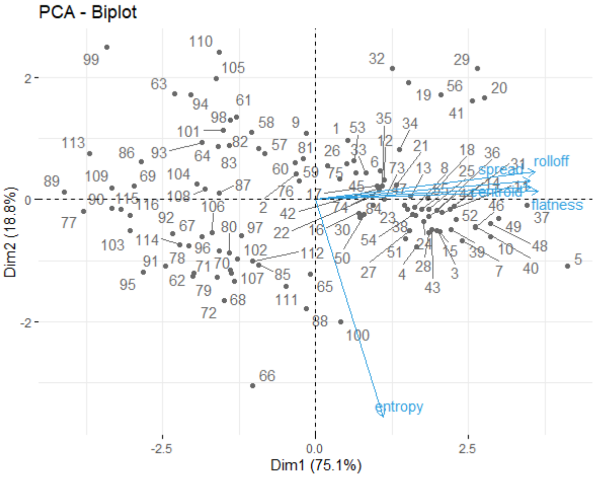

Figure 3: *PCA biplot. Bottom axis: PC1 score; left axis: PC2 score; top axis: loadings on PC1; right axis: loadings on PC2.*

3.3. Hierarchical clustering

The results of agglomerative hierarchical clustering are shown in Figure 4 and Figure 5. Dendrograms are the most important results of this type of cluster analysis. Dendrogram in Figure 4 and Figure 5 are the same, but Figure 4 shows whether the classified samples belong to the control group (C) or the dysarthric group (D), while Figure 5 lists all speakers. In both cases, the dendrograms indicate at what level of similarity any two clusters were joined. The position of the line on the scale indicates the distance at which clusters were joined.

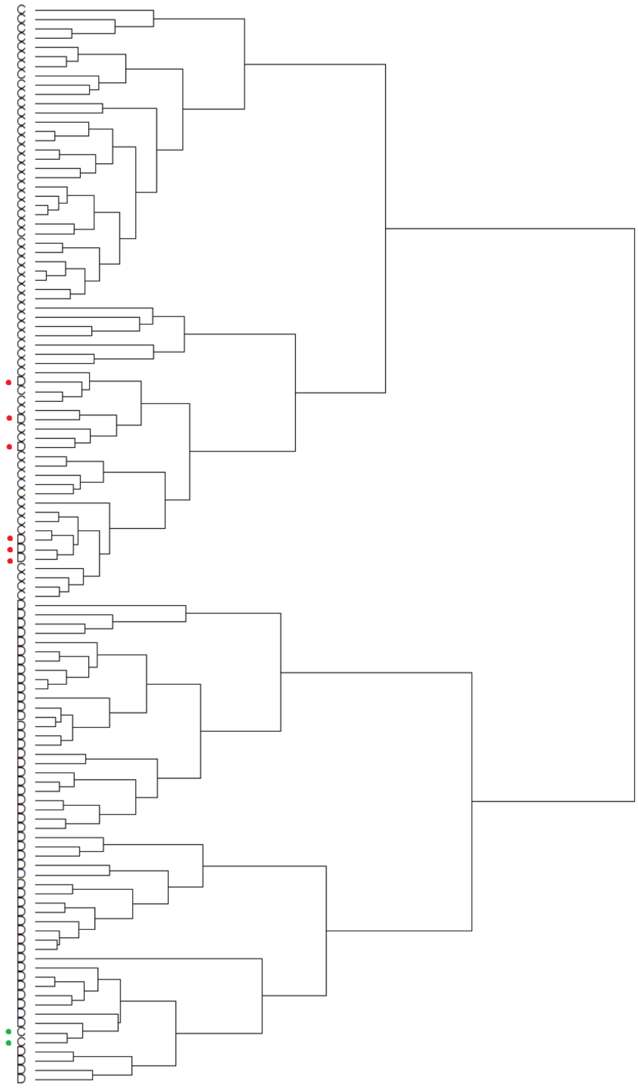

Figure 4: *Dendrogram (hierarchical agglomerative clustering). C: control; D: dysarthria. Red: D speakers classified in the C cluster; green: C speakers classified in the D cluster.*

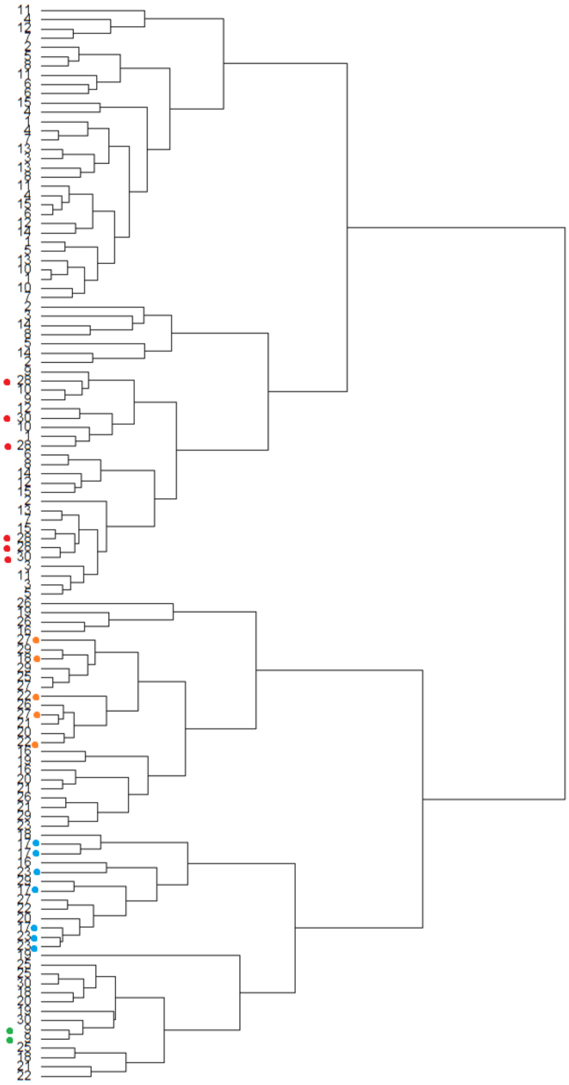

Figure 5: *Same plot as Fig.4 but showing Spkrs ID numbers. Orange marks speakers with ataxic dysarthria (tumor diagnosis); blue marks speakers with flaccid-spastic dysarthria (ALS diagnosis).*

4. Discussion and conclusions

K-means clustering (Fig. 2) shows that Cluster 1 matches well the 56 observations of dysarthric speakers (first 14 speakers * 4 sentences); i.e. datapoints from 61 to 116. Cluster 2 matches well the 60 observations of the control speakers (remaining 15 speakers * 4 sentences); i.e. datapoints 1-60. Cluster 1 is characterized by negative values in the five spectral measures, while Cluster 2 presents positive values. This basically means that dysarthric speakers present a stretched rhythmic *frame* and a centroid shifted towards lower frequencies, as well as a narrower spectral spread in the temporal envelope, in comparison with control speakers. These results agree with the fact that, in terms of openness, both the jaw and the mouth remain stationary throughout utterance production in speakers with this motor speech disorder.

Hierarchical clustering was then conducted to visualize possible outliers and to find possible subgroups within each main cluster. The dendrograms show that only two sentences of Control Spkr #9 (green) were misclassified as dysarthric. All the sentences of Dysarthric Spkr #28 and two of Dysarthric Speaker #30 (red) were classified in the control group. Besides, within the dysarthric group, at least two trends can be observed: Spkrs #18, #22 and #27 cluster together, and they all have ataxic dysarthria with a tumor diagnosis (blue points in Fig. 5). Spkrs #17 and #23 also cluster together, and they both present flaccid-spastic dysarthria, with an ALS diagnosis.

All in all, we can conclude that the five spectral measures computed from the temporal envelope of read sentences seem to separate well between dysarthric and non-dysarthric speakers, using two different types of clustering techniques. However, more studies are needed to explore why two control speakers were classified as dysarthric (maybe idiosyncratic slow rate or muffled voice quality). Likewise, it is necessary to explore which acoustic variables lie behind the clustering together of speakers with the same type of dysarthria; namely, ataxic dysarthria from a tumor diagnosis, on the one hand, and spastic dysarthria form an ASL diagnosis, on the other hand.

5. References

[1] N. Melle, *Guía de intervención logopédica en la disartria*. Madrid: Editorial Síntesis, 2007.

[2] Y. Kim, R.D. Kent, and G. Weismer, “An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria,” *J Speech Lang Hear Res*, vol. 54(2), pp. 417–429, 2011.

[3] S. Sapir, L. Ramig, J. Spielman, and C. Fox, “Formant centralization ratio (FCR) as an acoustic index of dysarthric vowel articulation: comparison with vowel space area in Parkinson disease and healthy aging,” *J Speech Lang Hear Res*, vol. 53, pp. 114–125, 2010.

[4] P. F. MacNeilage, “The frame/content theory of evolution of speech production,” *Behav. Brain Sci.*, vol. 21, pp. 499–511, 1998.

[5] E. San Segundo, J. Delgado, and L. He, “Characterizing rhythm in dysarthric speech using the temporal envelope,” in: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3907–3911). Guarant International, 2023.

[6] H. Dahmani, S.-A. Selouani, D. O’shaughnessy, M. Chetouani, and N. Doghmane, “Assessment of dysarthric speech through rhythm metrics,” *Journal of King Saud University - Computer and Information Sciences*, vol. 25, no. 1, pp. 43–49, 2013.

[7] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, “Quantifying Speech Rhythm Abnormalities in the Dysarthrias,” *J Speech Lang Hear Res*, vol. 52, no. 5, pp. 1334–1352, 2009.

[8] A. Hernandez, E.J. Yeo, S. Kim, and M. Chung, “Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics”. In INTERSPEECH, pp. 2897-2901, 2020.

[9] S. A. Selouani, H. Dahmani, R. Amami, and H. Hamam, “Using speech rhythm knowledge to improve dysarthric speech recognition,” *International Journal of Speech Technology*, vol. 15, no. 1, pp. 57–64, 2012.

[10] A. Lowit, “Quantification of rhythm problems in disordered speech: A re-evaluation,” *Philos. Trans. R. Soc. B.*, 369, 20130404, 2014.

[11] K. A. Spencer, J. Amaral, and K. Lansford, “Investigating perceptual subgroups in speakers with ataxic dysarthria: an auditory free classification approach,” *American Journal of Speech-Language Pathology*, 32(4S), pp. 1901–1911, 2023.

[12] A. Slis, R. Karlin, B. Parrell, “Unsupervised clustering reveals several subtypes in speakers with ataxia,” in: Radek Skarnitzl & Jan Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3952–3956). Guarant International, 2023.

[13] H. Bouchard, *Examination of Perceptual Subgroups of Ataxic Dysarthria Through Auditory Free Classification*, MSc.Thesis, University of Washington, 2023.

[14] D. Kim, S. Diehl, M. de Riesthal, K. Tjaden, S.M. Wilson, D.O. Claassen, and A.S. Mefferd, “Dysarthria Subgroups in Talkers with Huntington’s Disease: Comparison of Two Data-Driven Classification Approaches,” *Brain Sciences*, 12(4), p. 492, 2022.

[15] V. Illner, T. Tykalova, D. Skrabal, J. Klempir, and J. Rusz, “Automated Vowel Articulation Analysis in Connected Speech Among Progressive Neurological Diseases, Dysarthria Types, and Dysarthria Severities,” *Journal of Speech, Language, and Hearing Research*, 66(8), pp. 2600–2621, 2023.

[16] L. He, “Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements,” *J. Acoust. Soc. Am*, vol. 152, pp. 567-579, 2022.

[17] A. Ghazanfar, C. Chandrasekaran, and R.J. Morrill, “Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: Implications for the evolution of audiovisual speech,” *Eur. J. Neurosci*, vol. 31, pp. 1807–1817, 2010.

[18] R. J. Morrill, A. Paukner, P.F. Ferrari, and A.A. Ghazanfar, “Monkey lipsmacking develops like the human speech rhythm,” *Dev. Sci.* vol. 15, pp. 557–568, 2012.

[19] D.D. Deliyski, H.S. Shaw, and M.K. Evans, “Adverse effects of environmental noise on acoustic voice quality measurements,” *J Voice*, vol. 19, pp. 15–28, 2005.

[20] S. Hochmuth, T. Brand, M.A. Zokoll, F. Zenker, M. Wardenga, and B. Kollmeier, “A Spanish matrix sentence test for assessing speech reception thresholds in noise”, *Int J Audiol,* vol. 51, pp. 536–544, 2012.

[21] T. Giannakopoulos, and A. Pikrakis, *Introduction to Audio Analysis*, Academic Press, pp. 78–86, 2014.

[22] R Core Team, *R: A language and environment for statistical computing (R4.1.0)* [computer program], 2023.

[23] B. C. Boehmke, “K-means Cluster Analysis”, *UC Business Analytics R Programming Guide*, University of Cincinnati, 2023. URL: https://uc-r.github.io/kmeans_clustering#fn:kauf

[24] J.A. Hartigan, and M.A. Wong, “Algorithm AS 136: A k-means clustering algorithm”, *Journal of the royal statistical society. series c (applied statistics)*, vol. 28(1), pp. 100-108, 1979.

[25] L. Kaufman, and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[26] P.N. Tan, M. Steinbach, and V. Kumar, *Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining*, pp. 487-568, 2005.

Grant PID2021-124995OA-l00 funded by MICIU/ AEI/10.13039/501100011033 and by ERDF/EU