# CHARACTERIZING RHYTHM IN DYSARTHRIC SPEECH USING THE TEMPORAL ENVELOPE

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

[1]Dept. of Spanish Language & General Linguistics, UNED (Spain), [2]Dept. of Developmental & Educational Psychology, La Laguna University (Spain), [3]Dept. of Computational Linguistics, University of Zurich (Switzerland)

[1]esansegundo@flog.uned.es, [2]extjdelgado@ull.edu.es, [3]lei.he@uzh.ch

## ABSTRACT

This study investigated the rhythmic characteristics of 15 dysarthric speakers and 15 non-dysarthric speakers. Speech rhythm was viewed from MacNeilage's frame/content theory. Specifically, the temporal envelope was analysed. The subjects, native speakers of Canarian Spanish, read four phonetically-balanced sentences. Five common spectral measures (centroid, spread, rolloff, flatness, and entropy) were computed from the temporal envelope of each sentence. A binomial logistic regression model was built to evaluate how well the five spectral measures can characterize speech rhythm in dysarthric speakers versus the control group. The results show that the dysarthric group present a significantly lower centroid and lower spread. Possible explanations for these results are discussed in relation to previous phonetic studies into the rhythmic speech patterns of dysarthric speakers.

**Keywords**: dysarthria, rhythm, temporal envelope, Spanish, binomial logistic regression.

## 1. INTRODUCTION

This paper is an exploratory investigation into the rhythmic characteristics of dysarthria. 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 [1]. In particular, the method followed in this article for speech rhythmic characterization consists in the extraction of five spectral measures common in audio engineering (centroid, spread, rolloff, flatness, and entropy), computed from the temporal envelope of sentences in read speech.

Dysarthria is a speech disorder with neurological origin that causes difficulties in speech motor programming and execution [2]. Its evaluation is complex due to the heterogeneity of its symptoms [3] and requires the combination of a battery of objective and subjective tests [4].

In the upcoming section, literature is reviewed in terms of: (1) previous phonetic studies on dysarthric speech and (2) the frame/content approach to the study of rhythm.

## 2. PREVIOUS STUDIES

### 2.1. Phonetic studies on dysarthria

Different acoustic analyses of dysarthric speech have been conducted in order to obtain objective data that distinguish individuals with this type of speech disorder, or to determine the severity of dysarthria by looking for measures that correlate with the level of speech intelligibility. Acoustic analyses are also useful to distinguish among different dysarthria subtypes. Commonly analysed acoustic measurements include: the area of the vowel space [3, 4, 5], the range and slope of the second formant [3, 6, 7], VOT [5, 8], temporal measures, such as speech rate [3, 6, 9], or a combination of temporal, spectral, and cepstral parameters [3, 10].

Summarizing the findings of these investigations is not a simple task, as some of them focus on one particular type of dysarthria or analyse patients of a particular language. Hence results cannot be extrapolated to other dysarthria subtypes or to different languages. While some processes affecting motor speech planning and execution in dysarthria could be largely language-independent, others could be prone to cross-language differences [11].

In terms of rhythmic analyses, since Holmes's historical paper [12], dysarthria has been described as presenting reduced articulation rate, irregular duration contrasts between stressed and unstressed syllables and a scanning rhythm (*staccato*) [10]. However, this description refers to one particular type of dysarthria: ataxic dysarthria. Recent investigations tend to study a wide range of dysarthria subtypes, although results are sometimes disparate. As a case in point, while Liss et al. [9] concluded that rhythm metrics (acoustic measures of vocalic and consonantal segment durations) allow to distinguish control speech from dysarthria and to discriminate dysarthria subtypes, another study [13] concluded 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" [13, p.1].

Investigations into the perceptual characterization of dysarthria are rarer, although perceptual assessment in clinical studies is still considered the gold standard against which acoustic measures are compared [14]. In [15] a simplified version of the Vocal Profile Analysis (VPA) protocol [16] is used to analyse the voice quality of dysarthric speakers perceptually. The authors found that the perceptual settings 'vocal tract tension' and 'laryngeal tension' were the most useful to characterize this disorder. In particular, trained raters were more likely to score a voice with high vocal tract tension if it belonged to the dysarthric group than if it belonged to the control group of neurologically healthy speakers. The authors suggest that acoustic-perceptual assessment through this protocol could be an important complement to other types of evaluations, especially because several settings of the VPA refer to supraglottic structures commonly affected in dysarthria, a speech disorder in which the muscles used to produce speech are damaged, paralyzed, or weakened. However, more research is necessary into the acoustic correlates of perceptual settings such as 'vocal tract tension'. San Segundo et al. [17] suggested that low inter-rater agreement for this perceptual dimension could be due to the different salience of prosodic aspects in each rater. They showed that the perceptual ratings of one evaluator correlated with the mean intensity variability across syllables. In contrast, the ratings of the other evaluator correlated positive and significantly with two rhythmic measures related to mean consonant duration.

Rhythm is a key element of prosody, together with intonation and tempo. Since dysprosody, or prosody degradation, is considered a hallmark of dysarthria [11], the current investigation has focused on the rhythmic characterization of this speech disorder. This will complement previous investigations, such as [15], where 13 acoustic features (mainly spectral and cepstral measures) were analysed, together with voice quality perceptual settings, with the aim of describing dysarthria phonetically.

Due to the fact that traditional acoustic approaches to rhythm (i.e. durational variability in different phonetic intervals) in dysarthric speakers have given rise to disparate results [9, 13], a different acoustic method is proposed here, which analyses rhythm from the speech temporal envelope.

### 2.2. Speech rhythm: The frame/content perspective

The key idea which lies behind the concept of rhythm is cyclicity or regularity. The first phonetic studies of speech rhythm distinguish two major types of regularity: isochronous syllables and isochronous feet [18, 19]. However, it has not been possible to corroborate strict syllable- or foot-isochrony in traditionally considered syllable-timed and stress-timed languages [20-21]. Different rhythmic metrics have since been developed by calculating durational variabilities in different phonetic intervals, usually consonantal and vocalic intervals [22]. These metrics have been used to characterize the rhythm of languages or speaking styles, but also to evaluate speech disorders [9, 13]. Other approaches to speech rhythm can be roughly grouped into 'prominence', 'modulation' and 'coupling strengths' [23], which basically depend on which aspect gives rise to different rhythmicities: syllable intensity, peak frequencies in the amplitude envelope spectrum, or coupling oscillations.

A recent investigation [23] suggests that MacNeilage's frame/content theory of speech evolution unifies the different approaches to the study of rhythm: "MacNeilage held that speech rhythm evolved from pre-existing cyclical mandibular movements in ancestral primates in the form of lip-smacking [1]. It is also believed that the coupling between mouth opening-closing cycles and vocalization emerged en route to human evolution: the sonority of speech typically waxes and wanes with mouth opening and closing gestures [24, 25]. 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. More recently, Fitch [26] endorsed the frame/content mechanism and argued that it affords the evolution of the language system at large" [23, p. 568].

As explained in [23], the rhythm metrics quantifying the durational variability of vocalic and consonantal intervals typically focus on the *content* perspective, whereas the modulation-based approaches target the *frame* perspective with different emphases: recurring frequencies in the temporal envelope [27, 28] and coordination between envelopes at slower and faster rates [29]. This is how the different approaches to speech rhythm are related to the frame/content theory. Furthermore, Erickson and colleagues [30] demonstrated that the jaw displacement well explained the metrical structure and subjective prominence ratings in a number of languages in which a more prominent syllable was typically associated with a lower jaw position. To integrate both temporal modulation and the opening and closing cycles of the mouth for the characterization of the rhythmic frame, spectral coherence has been used [23, 31]. The coherence is calculated from the spectra of two signals to show the

connectivity between them in terms of their correlation in the frequency domain. For instance, He [23] uses spectral coherence between the temporal envelope and the mouth opening and closing kinematics to characterize speech rhythm in L2 English speakers of Mandarin. He found that the native group was significantly higher than the non-native group in terms of spectral centroid and spread; two of the five variables extracted from the spectral coherence, which can also be simply extracted from the temporal envelope (see section 3.3).

## 3. METHOD

### 3.1. Participants

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; with different medical origins. After a preliminary analysis, one dysarthric participant presenting mixed dysarthria was discarded. Her audio samples presented signal saturation and stammering. These aspects were deemed unfit for this type of acoustic analyses.

### 3.2. 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 [32]. The speech material consisted in reading aloud four phonetically balanced sentences of the Spanish Matrix Sentences Test [33].

### 3.3. 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 [35]. Then, the filtered signal was full-wave rectified and downsampled to the Nyquist frequency of 20 Hz, yielding the temporal envelope. Five spectral measures (CENTROID, SPREAD, ROLLOFF, FLATNESS, and ENTROPY) [34] were calculated from the temporal envelope of each sentence. 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. For details please look at the supplementary material ( https://osf.io/3z9uq ).

### 3.4. Statistical analysis

All statistical procedures were performed using R [35] and associated packages (*lme4*, *car*, *effects*, and *ggplot2*). The generalized linear mixed model (binomial logit) was used for data analysis. The five spectral variables extracted from the temporal envelope were modeled as the fixed-effect numeric predictors; GROUP (control and dysarthric speakers) was modeled as the dichotomous response variable. Both groups produced the same set of sentences, and thus SENTENCE was modeled as a random factor.

## 4. RESULTS

Table 1 summarizes the model-fitting results of all predictors. The reference of the response variable group is C (control speakers). The model explains between 66.54% and 88.76% of the dependent variable variation (Cox & Snell's $R^2$ = 0.6654; Nagelkerke's $R^2$ = 0.8876).

| Predictor | Estimate | Std. Error | z |
|---|---|---|---|
| CENTROID | –4.51 | 1.72 | –2.63 ** |
| SPREAD | –5.95 | 2.84 | –2.09 * |
| FLATNESS | 10.87 | 11.34 | 0.96 |
| ROLLOFF | 0.22 | 0.73 | 0.30 |
| ENTROPY | 3.11 | 1.88 | 1.66 . |

**Table 1:** Results of the fixed-effect predictors in the generalized linear mixed model (binomial logit). **p<0.01; *p<0.05; . p<1.

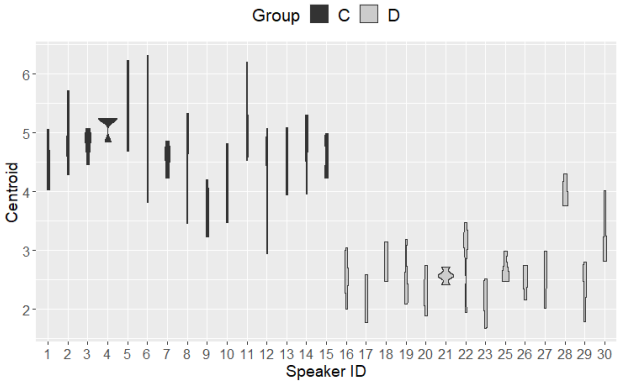


**Figure 1**: Boxplots showing the distribution of CENTROID values per speaker. C = Control; D = Dysarthria.

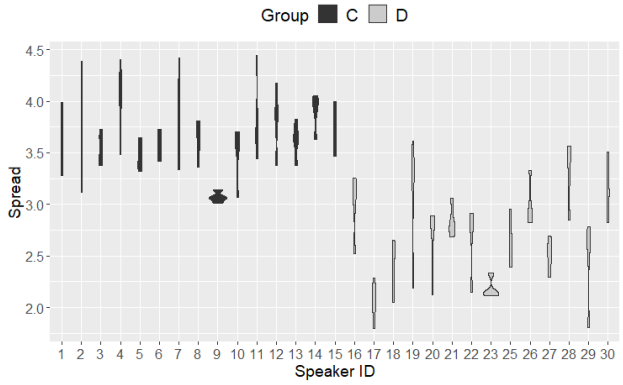


**Figure 2**: Boxplots showing the distribution of SPREAD values per speaker. C = Control; D = Dysarthria.

Fig. 1 and Fig. 2 show the distribution of CENTROID and SPREAD values, respectively, per speaker and group (black for the control group and grey for the dysarthric speakers).

## 5. DISCUSSION

This paper has investigated the rhythmic characteristics of read sentences in a group of Spanish speakers, distinguishing dysarthric speakers and non-dysarthric speakers. For this purpose, five variables capturing the temporal envelope of the sentences were computed and analysed as numeric predictors in a binomial logistic regression model.

Dysarthric speakers in general showed significantly lower CENTROID frequencies in the temporal envelope (estimate = –4.51, with $p<0.01$). Although this study was mainly exploratory, this finding agrees with our expectations. Dysarthric speech typically exhibits a slow speech rate due to slow articulatory movements. As a result, a stretched rhythmic *frame* was expected, together with a centroid shifted towards lower frequencies, in comparison with those of the control speakers.

A more restricted spread could also be anticipated in the dysarthric group. Our results corroborate that indeed dysarthric speakers in general showed a significantly narrower spectral SPREAD in the temporal envelope (estimate: –5.95, with $p<0.05$) in comparison with the non-dysarthric group. This suggests that the oscillations in the rhythmic frame were more regular around the centroid. A possible explanation for this could be that jaw displacements in the non-dysarthric population are variable while jaw oscillations in dysarthria do not exhibit large variation. Instead, in terms of openness, both the jaw and the mouth remain stationary in these speakers throughout utterance production. These results can be explained by the fact that dysarthria is a motor speech disorder in which the muscles used to produce speech are damaged, paralyzed, or weakened.

Our acoustic findings agree with impressionistic perceptual descriptions of dysarthria. For instance, Ziegler [10] highlights reduced articulation rate and 'drawling' speech as clinical symptoms of ataxic dysarthria. In future studies we will consider analysing ataxic, spastic and mixed dysarthric patients separately. Upon observation of the boxplots (Figs 1-2), there seems to be some differences among the speakers belonging to different subtypes of dysarthria. For instance, speakers 16, 18, 21, 22, 26, 27, 29 and 30 are ataxic speakers while the rest of dysarthric speakers present either mixed or spastic dysarthria. Possible explanations for these differences could be related to what Ziegler [10, p.14] states about ataxic dysarthria, namely that voice impairment in this disorder can be more irregular than in other dysarthrias: "vocal pitch or loudness may suddenly change and thereby interrupt the natural intonation pattern, and voice quality may change from strained-strangled to breathy or rough (…). Likewise, articulation may change between lenis and fortis consonant production, with staccato transitions between syllables or with syllable lengthening".

## 6. CONCLUSIONS

This study have shown that the dysarthric speakers analysed present a significantly lower centroid and lower spread in the temporal envelope of their read sentences, in comparison with a control group of non-dysarthric speakers. These variables can discriminate dysarthric and non-dysarthric speakers, although more studies, preferably with a larger number of subjects and in different languages, are necessary to ensure replicability and validation of these findings.

The literature has shown that *content*-based approaches to rhythm (durational variability in different phonetic intervals) have resulted in disparate findings in dysarthric studies. The method proposed here analyses rhythm from the speech temporal envelope and presents the advantage of not requiring audio transcriptions or phonetic interval segmentation. Previous L2 studies have shown that rhythmic characteristics may not be sufficiently explained from the traditional *content*-based perspective alone [23], highlighting the importance of the *frame* aspect in rhythm acquisition. In a similar way, investigations on speech disorders can benefit from this combined perspective. Since the sonorant/obstruent alternations typically follow the opening-closing mouth movements, according to the frame-content theory [1], future dysarthric studies could analyse the spectral coherence between the temporal envelope and mouth opening-closing articulatory data, as well as correlations with jaw openness ratings in the VPA perceptual protocol.

## 7. REFERENCES

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

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

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

[4] Sapir, S., Ramig, L., Spielman, J., Fox, C. 2010. 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.* 53, 114–125.

[5] Delgado Hernández, J. 2016. Medida de la severidad de la disartria atáxica a través del análisis acústico, *Estudios de fonética experimental*, 25, 149–166.

[6] Yunusova, Y., Green, J. R., Greenwood, L., Wang, J., Pattee, G. L., Zinman L. 2012. Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica*, 64(2), 94–102.

[7] Delgado Hernández, J., Izquierdo Arteaga, L. 2016. Relación entre las pendientes del segundo formante y las alteraciones motoras del habla en la disartria. *Revista de logopedia, foniatría y audiología* 36(2), 71–76.

[8] Auzou, P., Ozsancak, C., Morris, R. J., Jan, M., Eustache, F., & Hannequin, D. 2000. Voice onset time in aphasia, apraxia of speech and dysarthria: a review, *Clinical linguistics & phonetics*, 14(2), 131–150.

[9] Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., Caviness, J. 2009. Quantifying speech rhythm abnormalities in the dysarthrias, *J. Speech Lang. Hear. Res*. 52(5), 1334–1352.

[10] Ziegler, W. 2016. The phonetic cerebellum: cerebellar involvement in speech sound production. In: Marien, P., Manto, M. (eds), *The linguistic cerebellum*. Academic Press, 1–32.

[11] Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., Sadat, J. 2017. A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. *J Phonetics* 64, 156–167.

[12] Holmes, G. 1917. The symptoms of acute cerebellar injuries due to gunshot injuries. *Brain*, 40(4), 461–535.

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

[14] Ma, E.P.M., Yu, E.M.L. 2005. Multiparametric evaluation of dysphonic severity. *J. Voice* 20, 380–390.

[15] San Segundo, E., Delgado, J. 2021. A preliminary approach to the acoustic-perceptual characterization of dysarthria. *Proc. ISAPh 2021*, Universitat Rovira i Virgili, Tarragona (pp. 6-8).

[16] Laver, J. 1980. *The phonetic description of voice quality*. London: Cambridge Studies in Linguistics.

[17] San Segundo, E., Schwab, S., Dellwo, V., He, L., González, J. A. M. 2017. Perception of vocal tract tension: Exploring possible prosodic correlates. In *Tendencias actuales en fonética experimental: Cruce de disciplinas en el centenario del Manual de Pronunciación Española* (pp. 79-82). UNED.

[18] Jones, D. 1922. *An Outline of English Phonetics*. G. E. Stechert & Co.

[19]Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh University Press.

[20] Roach, P. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In: Crystal, D. (ed), *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F. R. Palmer*, Edward Arnold, 73–79.

[21] Dauer, M. 1983. Stress-timing and syllable-timing reanalyzed, *J. Phon*. 11, 51–62.

[22] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal, *Cognition* 73, 265–292.

[23] He, L. 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am*. 152(1), 567–579.

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

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

[26] Fitch, W. T. 2019. Sequence and hierarchy in vocal rhythms and phonology, *Ann. N.Y. Acad. Sci.* 1453, 29–46.

[27] Tilsen, S., Johnson, K. 2008. Low-frequency Fourier analysis of speech rhythm. *J. Acoust. Soc. Am.* 124 (2), EL34–EL39.

[28] Tilsen, S., Arvaniti, A. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages, *J. Acoust. Soc. Am*. 134, 628–639.

[29] Leong, V., Stone, M.A., Turner, R.E., Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception, *J. Acoust. Soc. Am.* 136, 366–381.

[30] Erickson, D., Kawahara, S. 2016. Articulatory correlates of metrical structure: Studying jaw displacement patterns, *Linguist. Vanguard* 2, 20150025.

[31] Chandrasekaran, C., Trubanova, C., Stillittano, S., Caplier, A., Ghazanfar, A.A. 2009. The natural statistics of audiovisual speech, *PLoS Comput. Biol.* 5, e1000436.

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

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

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

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