Effects of Dysarthria on Speech Prosody – A Case Study

László Tóth[1], Gábor Gosztolya[1,2], and Andrea Deme[3,4,5]

[1] University of Szeged, Institute of Informatics, Szeged, Hungary
[2] HUN-REN–SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[3] ELTE Eötvös Loránd University, Budapest, Hungary
[4] MTA–ELTE NYTK Lendület "Momentum" Neurophonetics Research Group
[5] ELTE Research Centre for Linguistics
tothl @ inf.u-szeged.hu

Abstract. Dysarthria is a medical umbrella term that covers several motor speech disorders which are mostly caused by neurological illnesses or injuries and may affect multiple aspects of speech. Although medicine discriminates subtypes, this categorization is mainly based on the type of the neurological damage, not on the actual speech properties. Due to this, the observable features are quite various from a speech technologists point-of-view. Here we study the effect of dysarthria on one speaker's voice, from whom we have samples both from before the neurological injury and after that. Moreover, as instrumental analysis is rare in clinical practice, we are going to focus on the suprasegmental factors of speech, and on the methods speech technology offers for their objective evaluation. We propose several measurements and demonstrate them on both the healthy and dysarthric recordings of our actual speaker.

Keywords: Dysarthria · Prosody · Case study.

1 Introduction

Speech technology – e.g. speech synthesis – traditionally separately discusses the production of speech phones from the step of superimposing prosody on the signal [14]. This discrimination quite probably roots in phonetics, which also separates a 'segmental' and a 'suprasegmental' (or 'prosodic') level of speech [10]. Traditionally, syllables are not considered being a relevant unit of speech in speech technology from a modeling point of view, while theories in linguistics and cognitive neuroscience claim that speech comprehension parallelly operates at both a larger (syllabic) scale and a finer (phonetic) scale – this is the so-called AST (asymmetric sampling in time) hypothesis [21]. With respect to speech technology solutions, prosody, and the contribution of syllable-sized units to the synthesized speech or the effectiveness of automatic speech recognition are relatively underresearched. Therefore, in the present paper, we focus our attention to this domain of speech primarily.

Medically, speech production is a very complex process that requires the very finely tuned collaboration of several organs and muscles. Any slight flaw in this sophisticated execution results in so-called dysarthric speech. So dysarthria is a medical umbrella term for all kind of neurological damage that weakens, paralyzes, or impairs coordination of the muscles used for speech [9]. Although neurologists discriminate several subtypes [5], this categorization is usually based on the type of neurological trauma, not on the actual speech properties. Because of this, the observable symptoms may be quite various, making them nearly impossible to handle similarly in speech processing. Moreover, dysarthria frequently comes with dysphonia [2] – an impairment in laryngeal function, i.e., voice production –, which just adds one more factor to the already complex situation. (Not to mention the condition called aphasia, i.e. a language disorder affecting higher level linguistic functions such as word retrieval or the formulation of sentences). A speech-language therapist should be able to separate these factors, and our observation is that they typically rely on their auditory perception, without performing any instrumental analysis. Quoting Cloud et al., "speech naturalness and prosodic abnormality must be assessed perceptually due to a lack of validated objective assessment measures for clinical use" [3]. Conclusively, in this paper we are suggesting some possible measurements that could be performed on a dysarthric speech signal. In the present paper we focus on the suprasegmental level, while issues regarding the segmental level were covered elsewhere [7, 8].

We present our suggestions in the form of a 'case study', that is, not through a statistical analysis of a lot of dysarthric speakers, just one speaker. Our reason for this is twofold. First, as we already said, the medical term 'dysarthria' is so wide that handling different cases together may lead to inconclusive results. Second, for the actual speaker of this study we had a parallel corpus of his 'healthy' and 'dysarthric' speech, which gave us the very rare and unique opportunity of comparing the two.

2 Data

Our data consists of the recordings of a male speaker, who suffered a cerebellar stroke at the age of 50. He read 200 sentences from the PPBA corpus (sentences 479-678), which was originally invented for speech synthesis purposes [20]. The original recording were made approximately 3 years before the stroke, and re-recorded again about another 3 years after the neurological event (under similar acoustic conditions, although with a different hardware)[§]. A subset of these parallel recordings (each one containing one sentence) were partially phonetically annotated, so for the experiments that required a labeling we used only 95 files.

As an example, Fig. 1 displays the spectrogram of the shortest sentence in the corpus, which actually consists of just two words (we applied the Praat software [1] to create the figures). One might immediately observe the most common property of dysarthric speech: slowness, as the duration of the sentence has increased more than twice. One might also notice some timing and length issues [7]. However, otherwise the phones are roughly correct, so it is not easy to

[§] Although he has been receiving speech therapy ever since, the re-recordings were made in a very early phase.

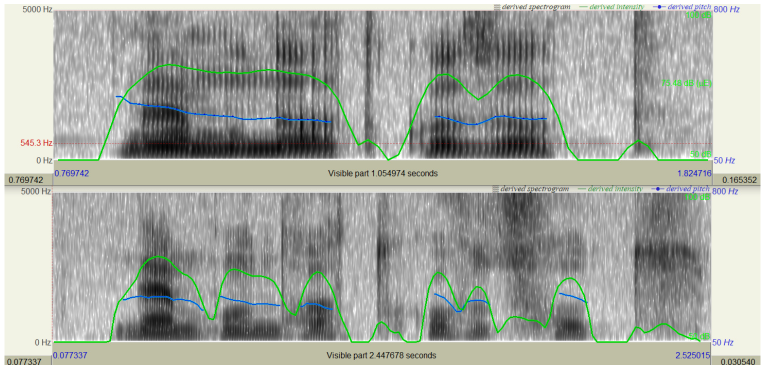

Fig. 1. Spectrograms of the same short sentence with a healthy (top) and dysarthric (bottom) pronunciation. The intensity and pitch curves are also shown superimposed, as calculated by the Praat software. Notice the different resolution on the time axis.

tell exactly what is the main reason behind the decreased intelligibility – which is yet another general property of dysarthric speech.

As already mentioned, in the present study we concentrate on the suprasegmental features, and Praat immediately displays two such factors: intensity and pitch. In general, we can say that their curves are much steadier for the healthy pronunciation, while much more variable and ragged in the case of the dysarthric speech. These are two examples of the prosodic factors that we seek to analyze automatically in the following.

3 Elements of Prosody

According to a classic speech technology textbook, the main factors of speech prosody are [14]:

- **Loudness (intensity) and stress.** Firstly, speech intensity patterns have a natural pulsation, which corresponds to the syllabic pace of speech. The reason behind this behavior is that in most languages the syllabic nuclei can only be vowels, which are usually louder (or more sonorous) than the surrounding consonants that form the onset and the coda [12]. Secondly, as our corpus mostly consists of declarative sentences, we expect mainly stable or just slightly falling intensity curves at the sentence level (see again Fig. 1 for an example). The speaker also may emphasize (stress) certain words or word groups (phrases) in order to convey additional information.
- **Pitch.** Every speaker has her/his natural and habitual fundamental frequency, that is, a frequency their vocal folds are vibrating at during speech production (which we perceive as ‘the pitch’), more precisely, during the pronunciation of voiced speech parts. One may also use pitch changes to emphasize certain phrases or to indicate sentence types, but otherwise the

pitch is relatively stable for declarative sentences (see again the upper panel of Fig. 1).

– **Timing and rhythm (rate/durations and pauses).** The most obvious manifestation of timing and rhythm is the duration of phones, which is very easy to measure supposed we have a phonetic segmentation of the input. Similarly, in order to quantify the fragmented nature of our signals, we measured the number and average duration of intra-sentence silences (in healthy speech, these typically occur at phrase boundaries, for example). Looking for any longer pauses, like inter-sentence pauses did not make sense in this case, as our recordings contained one sentence each. With regards to the pace of speech (pace of syllables), it is frequently examined via the modulation of intensity mentioned above [12].

The intensity pattern was investigated via calculating the modulation spectrum of the speech envelope – which gives information about the syllable rate. However, we could not come up with a simple, automatic procedure to analyze phrase stresses (although the word stress is fixed on the first syllable in Hungarian, unfortunately not all words are necessarily stressed). As regards the duration measurements, phonetic segmentation was obtained partially automatically as follows: because we had read speech, we could easily estimate the phonetic content. For the normal speech samples, we aligned the phonetic transcripts with the speech signals via forced alignment, with partial manual checking. Regarding the speech samples with dysarthria, forced alignment yielded poor results full of segmentation inconsistencies, so we had to perform (a partial) manual segmentation. For simplicity, we will report duration measurements only for some vowels, words, and the whole sentences. Lastly, calculating pitch-related statistics obviously makes sense only for voiced segments. Vowels, for example, as they are typically voiced. However,in the dysarthric samples we found many vowels to be partially or completely devoiced. Hence, all pitch-related statistics were extracted only from those frames that were found to be voiced by our software tools (apart from Praat [1], we used Librosa [19] with Python scripts).

In the following sections we discuss each element of the above points, and present the pertaining measurement(s), along with their results.

4 The Modulation Spectrum

The (amplitude) modulation spectrum corresponds to a frequency analysis of the time-domain evolution of the speech intensity (either broadband or narrowed to frequency bands) [4]. Its possible use in automatic speech recognition was mostly popularized by Hynek Hermansky and his collaborators in the late 90's [15], who found that the linguistically most important events happen in the low (0-16 Hz) modulation frequency range, and the examination should prioritize the 3-5 Hz region, which corresponds to the average syllable rate of human speech [12, 11]. Much more recently it was successfully applied in discriminating dysarthria subtypes [17].

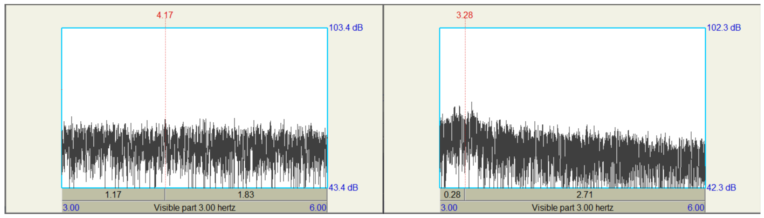

Fig. 2. Amplitude modulation spectra gained from the healthy (left) and the dysarthric (right) recordings. Only the 3-6 Hz frequency range is showed, as this is thought to be the most important for conveying the meaning.

For extracting the modulation spectrum we employed the script of He and Dellwo [13] in Praat. As the input, we concatenated all the sentences to get a long-term average. Fig. 2 compares the modulation spectrum of the healthy recordings with that of the dysarthric files. At first look, there seems to be no obvious peak around 4-5 Hz (the supposed syllabic speed) for the healthy samples. At closer inspection, there is a peak at 4.17 Hz, although not convincingly outstanding. Repeating the same examination for the dysarthric recordings, we see a clearer peak around 3.1–3.3 Hz. Firstly, this phenomenon shows that the speech became obviously slower. Secondly, the fact that there is no clear peak (just a wide protrusion) might indicate that the speech rate became much less stable. The different spectral tilt of the two images is yet another difference, but we have no plausible explanation for it.

5 Duration Measurements

As already mentioned, duration statistics are very easy to collect, assuming we have phonetic segmentation information. Because we had only a mostly automatic segmentation for the healthy recordings and a partial manual one for the dysarthric files, we report duration statistics only for certain units, namely the sentences, the words, and a couple of vowels. By comparing these statistics for the healthy and dysarthric recordings and for short and longer units (i.e. phones-words-sentences) we can study the effect of dysarthria and whether it is dependent on the length of phones.

Fig. 3 shows pretty similar changes for vowels and words, corresponding to about a twofold increase. The increase for sentences is slightly larger, quite probably due to the increased number of pauses. We will return to this issue later.

As regards the vowels, here we report duration statistics for only four of them (‘a’, ‘e’, ‘i’ and ‘u’). We chose these vowels because they are relatively language-independent and span the vowel chart of the language, but Hungarian, of course, has many more vowels. Quite uniquely, ’length’ is a discriminative feature in Hungarian, which means that most phones have their short and long

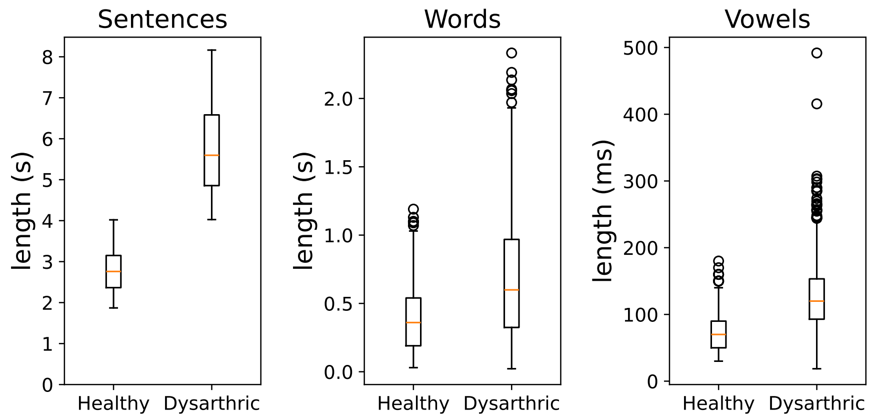

Fig. 3. Duration changes across units of different length.

version, where close and close-mid vowels differ primarily in their duration, while open-mid and open vowels differ also in their quality [6]. Thus, preserving the contrast between them is vital for any dysarthric speaker. Here, we worked only with the mentioned four short vowels, while a more detailed analysis of the vowels of our particular speaker can be found in our parallel study [7].

Fig. 4 shows the duration statistics of the mentioned four vowels. In general, we can see an approximately two-fold duration increase for all the vowels, and that it is relatively uniform across vowels.

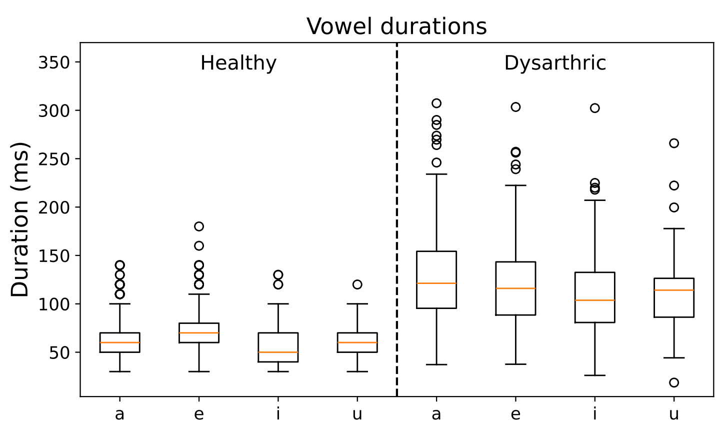

Fig. 4. Duration statistics for four vowels.

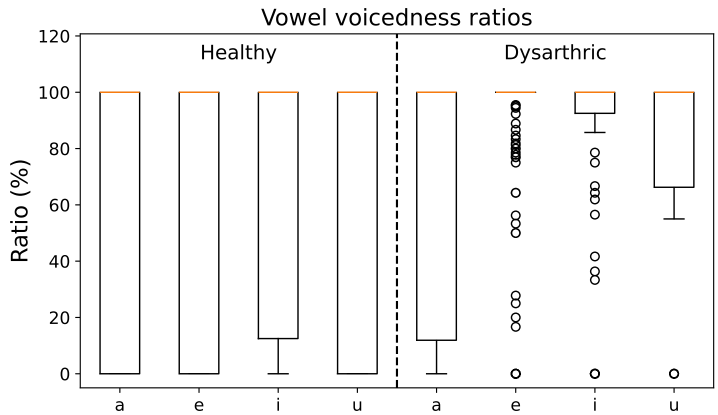

Fig. 5. The distribution of the ratio of voiced frames for 4 vowels of the healthy and the dysarthric samples, respectively.

6 Voicedness

Naturally, pitch measurements are only possible for voiced speech fragments. Therefore, before examining the pitch range, we examined the ratio of voiced frames, which, of course, also depends on the actual content. But, as the content was the same in the two sub-corpora, we expected to obtain the same ratio for both set of recordings. Compared to this, we got 64.74% for the 'healthy' corpus, which reduced to 58.94% for the dysarthric samples. This means that dysarthria had an obvious negative effect on vocal fold functions, and our speaker has difficulties with producing voiced phones.

However, the lack of voicing might also sign the failure of the pitch detection algorithm we used to detect voiced frames and estimate pitch. To examine this possibility, we repeated the whole process using only a few vowels. As vowels are typically voiced, we expected to get results close to 100% in the case of the 'healthy' recordings. Our finding are shown in Fig. 5 for the selected vowels ('a', 'e', 'i' and 'u') in the form of whisker plots. The left hand side shows our results for the healthy case, and they clearly contradict our previous expectation. Upon a deeper perceptual examination we found that our speaker was inclined to produce an irregular phonation [18] even before the stroke, and our pitch detector detected these segments as voiceless. This is the explanation for the large voiceless rates for the healthy vowels. As regards the dysarthric samples, we practically found that dysarthria can make even vowels devoiced.

7 Pitch Range

Based on Fig. 1, one may suspect that the pitch production of the subject became much less stable. To examine this, we performed two experiments. First, we examined the mean pitch (more precisely, F_0 value) over the previously studied vowels. Fig. 6 summarizes our findings.

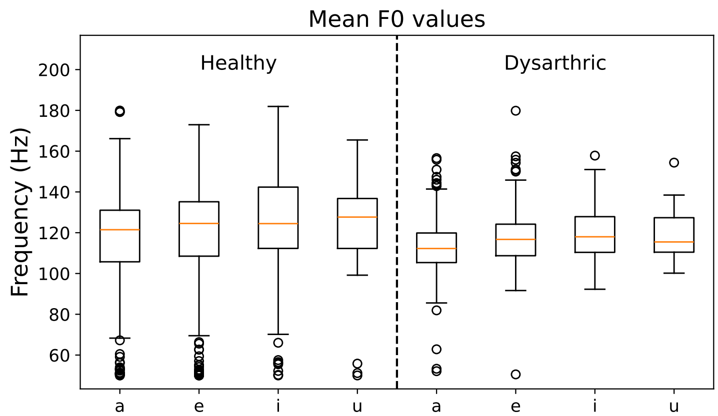

Fig. 6. The distribution of the mean F_0 value over 4 vowels of the healthy and the dysarthric samples, respectively.

One can see that the mean F_0 value indeed became somewhat lower. However, the mean value does not reflect the fact that the pitch became much less stable. To express this, we measured the average variance of the pitch over the voiced segments (these were defined by the continuity of voicing – so they does not necessarily consist of vowels only). Although the number and duration of voiced segments obviously depends very much on the actual content, but in our case *it remained the same*, so ideally we should get the same numbers before and after the neural event.

Table 1 summarizes what we found instead. Remember that earlier we said that the increase in the sentence durations was about twofold. However, the total increase in the voiced duration is much smaller, about a factor of 1.5. Also, the number of voiced segments significantly increased, while their average duration remained approximately the same. The pitch variance has also increased tremendously. This gives the impression that the earlier relatively steady pitch contour became fragmented and squiggly (cf. again Fig. 1). In the last subsection we try to construct a method to measure it.

8 Fragmentedness

In healthy speech, pauses within a sentence are quite rare and short; they typically appear at phrase boundaries of long sentences. As opposed to this, dysarthric persons tend to talk in short burst, which is mostly related to insufficient breath control [22]. In order to quantify this behaviour, we measured the

Table 1. Statistics of voiced segments.

	Healthy	Dysarthric
total voiced duration:	318.47 sec.	498.88 sec.
number of voiced segments:	1997	2887
average duration of a segment:	159.5 msec.	172.8 msec.
average pitch variance within a segment:	142.28 (Hz^2)	204.90 (Hz^2)

number and average duration of silences; Fig. 7 summarizes the results. We see that the silent segments show more than a threefold increase, both with respect to their number and their duration. Altogether, in total duration this means more than a tenfold increase.

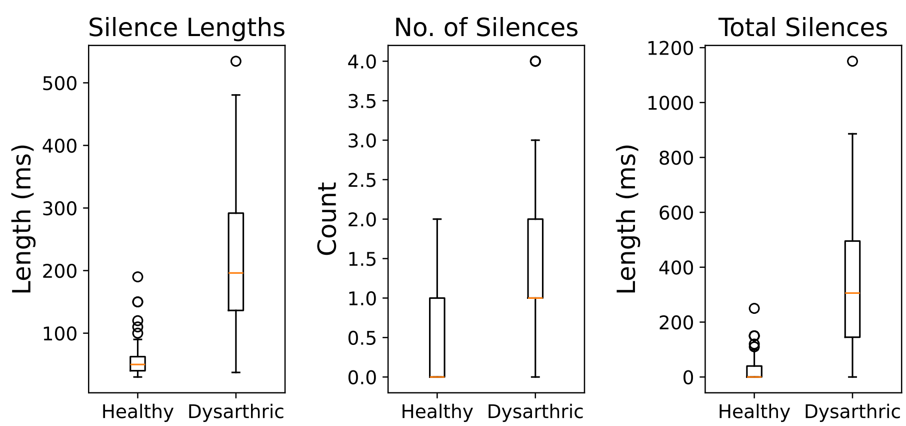

Fig. 7. The statistics of silent segments.

9 Steadiness

The intensity and pitch curves of Fig. 1 clearly demonstrate that our speaker faces difficulties with keeping a stable intensity and pitch. Jitter and shimmer was invented exactly to quantify this instability [16]. However, they are interpreted only for voiced signals - and usually evaluated over sustained vowels. Moreover, they quantify cycle-to-cycle variations of pitch and amplitude, while we needed something that would work on a longer term. Thus, we used the built-in ability of Praat to stylize the original intensity and pitch curves [1]. Figs. 8 and 9 show two examples of this – for the same samples that we saw earlier in Fig. 1. As we see in these figures, the stylization attempts to approximate the two curves

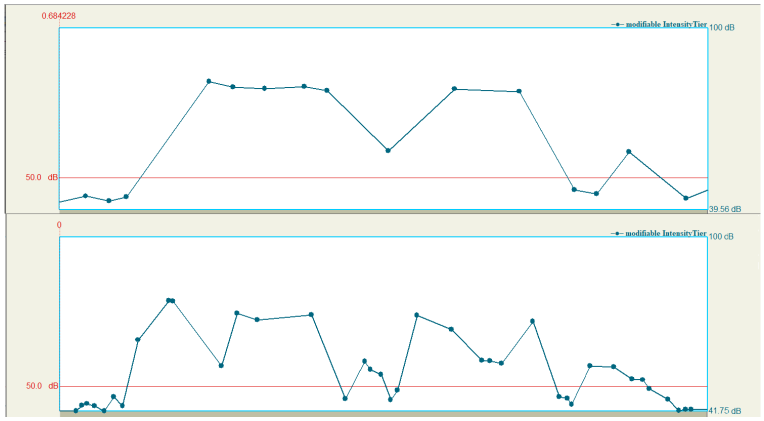

Fig. 8. The stylized intensity curves of Fig. 1, as calculated by the Praat software.

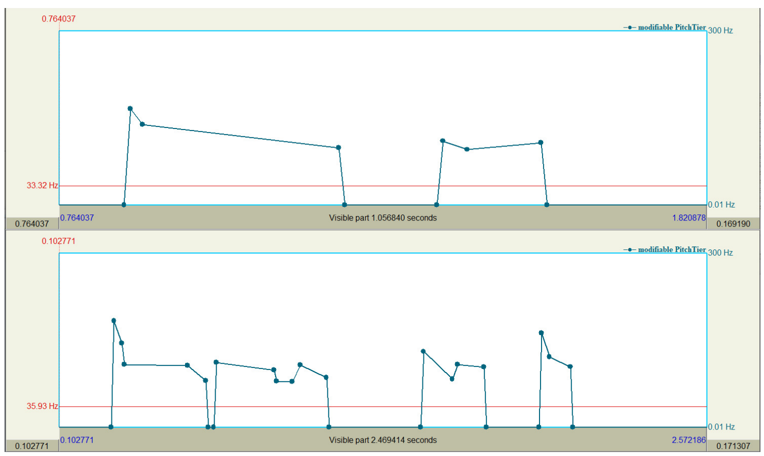

Fig. 9. The stylized pitch curves of Fig. 1, as calculated by the Praat software.

by linear segments, and inserts a breakpoint when the deviation from the line is above a small threshold.

Using the above stylization, we can simply estimate the steadiness of the two curves by the number of line segments (i.e., tier points) required for the stylization. We did exactly this (to exclude the leading and ending silences, we applied a 50dB threshold for the intensity, while these parts were naturally devoiced in the case of pitch), and Table 2 summarizes the number of tiers required for the stylization of the intensity and pitch curves of the healthy and the dysarthric recordings, respectively. The large increases clearly reflect the instability of the two curves (even if we keep in mind the appoximately twofold durational change).

Table 2. The number of tier points required for the stylization of all the sentences.

	Healthy	Dysarthric
intensity (above 50dB):	6758	12120
pitch (with a resolution of 2 semitones):	10500	14326

10 Conclusion

In this paper we presented several measurement methods that could be performed on dysarthric speech to evaluate its quality, especially with respect to its prosodic properties. A limitation of the present study is, of course, that it is based on data from only a single speaker. At the same time, the study is unique in that it analyzes both the healthy and dysarthric speech of the same speaker. In the future we would like to evaluate the usefulness of the presented methods with the voice samples of other dysarthric people, probably with different etiologies.

Acknowledgments. Andrea Deme was supported by the Bolyai János Research Scholarship of the Hungarian Academy of Sciences. This study was also supported in part by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the a 2024-1.2.3-HU-RIZONT funding scheme (project no. 2024-1.2.3-HU-RIZONT-2024-00017). The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (2026), https://www.fon.hum.uva.nl/praat/, accessed: [2026.05.05.]
2. Camillo, L., Ortiz, K.: Vocal analysis (auditory-perceptual and acoustic) in dysarthrias. Pro-Fono Revista de Atualizacao Cientifica **19**(4), 381–386 (2007)
3. Cloud, C., Georgen-Schwartz, K., Allison, H.: The contributions of pitch, loudness, and rate control to speech naturalness in cerebellar ataxia. American Journal of Speech-Language Pathology **33**(5), 2536–2555 (2024)
4. Crouzet, O.: A recent history of the 'modulation spectrum'for speech science: on the contribution of communication between research fields. In: Proc. HSCR 2025. pp. 167–175 (2025)
5. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. Journal of speech and hearing research **12**(2), 246–269 (1969)
6. Deme, A.: A magánhangzók változatossága a magyarban [Variability of vowels in Hungarian]. No. 173 in Nyelvtudományi értekezések, Akadémiai Kiadó, Budapest, Hungary (2025), scientific monograph (in Hungarian)
7. Deme, A., Tóth, L.: The discrimination of long and short vowels based on their duration in dysarthria after a celebellar stroke – a case study (in Hungarian). In: Proceedings of the MFFLT Congress. Budapest, Hungary (2026), talk presented on May 30, 2026

8. Deme, A., et al.: Hungarian long and short vowels' duration in speakers with Down syndrome (in Hungarian). In: Proceedings of the MFFLT Congress. Budapest, Hungary (2026), talk presented on May 30, 2026
9. Duffy, J.R.: Motor speech disorders: third edition. Elsevier (2013)
10. Fletcher, J.: The Prosody of Speech: Timing and Rhythm, chap. 15, pp. 521–602. John Wiley & Sons, Ltd (2010)
11. Goswami, U., Leong, V.: Speech rhythm and temporal structure: Converging perspectives? Laboratory Phonology **4**, 67–92 (2013)
12. Greenberg, S., Carvey, H., Hitchcock, L., Chang, S.: Temporal properties of spontaneous speech — a syllable-centric perspective. Journal of Phonetics **31**(3-4), 465–485 (2003)
13. He, L., Dellwo, V.: A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the hilbert transform. In: Proc. of Interspeech. pp. 530–534 (2016)
14. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice Hall (2001)
15. Kanedera, N., Hermansky, H., Arai, T.: Desired characteristics of modulation spectrum for robust automatic speech recognition. In: Proc. of ICASSP'98. pp. 613–616 (1998)
16. Koffi, E.: A comprehensive review of jitter, shimmer, and HNR: Linguistic and paralinguistic applications. Linguistic Portfolios **14**, 2 (2025)
17. Liss, J., LeGendre, S., Lotto, A.: Discriminating dysarthria type from envelope modulation spectra. Journal of Speech, Language, and Hearing Research **53**(May), 1246–1255 (2010)
18. Markó, A., Deme, A., Bartók, M., Gráczi, T.E., Csapó, T.G.: Word-initial irregular phonation as a function of speech rate and vowel quality in hungarian. In: Qiang, F., Dang, J., Perrier, P., Wei, J., Wang, L., Yan, N. (eds.) Studies on Speech Production: 11th International Seminar, ISSP 2017, Revised Selected Papers. pp. 134–145. Springer, Cham, Switzerland (2018)
19. McFee, B., Raffel, C., Liang, D., Ellis, Matt, P.D., McVicar, Battenberg, E., Oriol, N.: Librosa: Audio and music signal analysis in python. In: Proc. 14th python in science conference. pp. 18–25 (2015)
20. Olaszy, G.: The development and aims of a high-precision, parallel Hungarian speech database [in Hungarian]. In: Beszédkutatás (Speech Research). pp. 261–270 (2013)
21. Poeppel, D.: The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication **41**(1), 245–255 (2003)
22. Yorkston, K., Spencer, K., Duffy, J.: Behavioral management of respiratory/phonatory dysfunction from dysarthria: A systematic review of the evidence. Journal of Medical Speech-Language Pathology **11**, 13–38 (06 2003)