INTERSPEECH 2015

A syllable-based analysis of speech temporal organization: a comparison between speaking styles in dysarthric and healthy populations

Brigitte Bigi[1], *Katarzyna Klessa*[2], *Laurianne Georgeton*[1], *Christine Meunier*[1]

[1]Laboratoire Parole et Langage, Aix-Marseille Université, CNRS
5, avenue Pasteur 13100 Aix-en-Provence, France
[2]Institute of Linguistics, Adam Mickiewicz University, Poznań
{brigitte.bigi,laurianne.georgeton,christine.meunier}@lpl-aix.fr, klessa@amu.edu.pl

Abstract

A comparison of how healthy and dysarthric pathological speakers adapt their production is a way to better understand the processes and constraints that interact during speech production in general. The present study focuses on spontaneous speech obtained with varying recording scenarios from five different groups of speakers. Patients suffering from motor speech disorder (dysarthria) affecting speech production are compared to healthy speakers. Three types of dysarthria have been explored: Parkinson's Disease, Amyotrophic Lateral Sclerosis and Cerebellar ataxia. This paper first presents general figures based on syllable-level annotation mining, including detailed information about healthy/pathological speakers variability. Then, we report on the results of automatic timing parsing of interval sequences in speech syllable annotations performed using TGA (Time Group Analysis) methodology. We observed that mean syllable-based speaking rates in time groups for the healthy speakers were higher than those measured in the recordings of dysarthric speakers. The variability in timing patterns (duration regression slopes, intercepts, and nPVI) depended also on the speaking styles in particular populations.

Index Terms: syllables, healthy speech, pathological speech, spontaneous speech, speaking styles, TGA

1. Introduction

Temporal organization of speech production is a major parameter conditioned by numerous factors including language specificity, speaker's characteristics, speaking styles, etc. Accurate speech timing is crucial for an optimal intelligibility and thus for communicative interaction. Moreover, speech timing provides subtle information about intentions, emotions, strategies, etc. However, when speakers suffer from language pathological damage, this major parameter can be affected and may lead to decrease of intelligibility. In order to better understand speech timing distortion we conducted analyses related to speech timing organization for healthy and dysarthric speakers.

Patients suffering from motor speech disorder (dysarthria) affecting speech production are compared to healthy speakers. Since dysarthria refers to different types of pathology, we examined three of them in order to inspect a possible distinction relative to syllable timing. Moreover, two types of speech styles for healthy speakers have also been compared in order to investigate the possible differences between speaking styles variation and variations due to pathology. Basic tenet is that observation of disordered speech can provide clues about the way normal speech is produced and vice-versa. Several studies have reported on specific speech rate organization within dysarthric populations [1, 2, 3]. [4] analyzed timing variability in a corpus of controlled dysarthric speech (isolated phrases, read speech) [5] by applying rhythm metrics as defined by [6, 7] to durational characteristics of vocalic and intervocalic intervals and Pairwise Variability Index (PVI), and found rhythm metrics to be sensitive to differences between groups of dysarthric speakers. The objective of the present work is to compare syllable-level temporal organization in spontaneous speech produced by healthy and dysarthric speakers using semi-automatized and automatized methods of data processing and annotation mining.

Syllable is one of the most fundamental units of speech temporal organization and an important structural unit in language production and perception. Phonetics gives no exact or straightforward specification of syllables. The feeling of syllable boundaries, although usually very strong, is subjective and often not unique [8]. While there are no phonetic definitions for the syllable which are universally agreed upon, a syllable may be defined linguistically as a sequence of speech sounds having a maximum or peak of inherent sonority between the two minima of sonority. The syllable is then credited as a linguistic unit conditioning both segmental (e.g., consonant or vowel lengthening) and prosodic phonology (e.g., tune-text association, rhythmical alternations). As such, the syllable was used as a basic unit in speech rhythm investigation and in many models of timing in speech (an overview in [9]), cf. also the distinctions between syllable- vs. stress-timed languages [10, 6], prosodic prominence investigation [11], durational variability and timing patterns in interpausal syllable groups [12]. The syllable has been also reported to provide a viable basis for segmental duration modeling. Such model was proposed by [13] for speech synthesis purposes due to the role of syllables in the structural and rhythmic organization of the utterances (segmental durations being calculated at a secondary stage and fitted to the higher level framework). Depending on the task, purpose and language in question, other base units are also applied, as well as multilevel approaches [9, 7, 14, 12]. Syllable timing information, and especially syllable durations and intersyllable pausing schemes are considered as important elements in the inventory of measurements of speech monitor control in speech dysarthria [2, 3].

In the present study, three types of dysarthria have been explored and compared to two groups of healthy speakers, all of them for spontaneous speech productions (Section 2). The corpus was time-aligned and syllables were generated automatically thanks to SPPAS software [15]. The functionality of the SPPAS tool has been extended for the present purpose with regard to the way of dealing with filled pauses. In Section 3, general figures for the syllable items were analyzed, including detailed information about healthy/pathological speakers variability. Section 4 of the paper provides a report on TGA (Time Groups Analysis) [16] results performed with the use of Annotation Pro + TGA software [17], used as a solution for automatic timing parsing of interval sequences in speech syllable-based annotations.

2. Corpus description

2.1. Corpus overview

In order to better understand variations due to speech disorder, three types of dysarthria have been explored: Parkinson's Disease (PAR), Amyotrophic Lateral Sclerosis (ALS) and Cerebellar ataxia (CER). Parkinsons Disease is a consequence of basal ganglia damage. It causes stiffness or slowing of movement. Parkinsonians production is often perceived as scanning speech. Amyotrophic Lateral Sclerosis is caused by upper and lower motor neuron damage and speech is characterized by slowing speaking rate. Cerebellar ataxia results from cerebellar damage which disrupts coordination of muscular activity leading to a slow speech rate.

The recordings of dysarthric speech have been acquired from 8 speakers with Cerebelar Ataxia (CER), 5 speakers with Parkinsons Disease (PAR) and 11 speakers with Amyotrophic Lateral Sclerosis disease (ALS). The speakers were recorded in a recording room using an external microphone. They were asked to tell about their everyday routines or about a typical day in the hospital.

For the sake of comparisons between the dysarthric (DYS) and healthy (HEA) populations, as well as with a view to inspect the role of speaking styles, 12 healthy speakers (6 HNC and 6 HNI) were recorded with head-mounted microphones in an anechoic chamber (HNC) and in a silent room (HNI). The two groups of speakers were recorded according to two different scenarios: HNI were requested to talk about their professional career or personal events while HNC were recorded in a narrative process within a relaxed conversation. The main difference between healthy and dysarthric corpora is the duration of the recordings. Narration is quite long for healthy speakers, while the dysarthric ones speak less (Table 1). Due to the physical and social distress in pathological speech, DYS tend to avoid speaking situations and usually do not speak for a long time.

Table 1: *Corpus recordings and patients severity degree in a range from 0 (normality) to 3 (high severity).*

Pop.	Nb spks	Degree of severity (mean <min-max>)	Rec. time (in sec.)	Speech (in sec.)
HNC	6		1846	1420
HNI	6		4455	3341
CER	8	1.30 <0.8-2.3>	663	475
PAR	5	0.99 <0.4-1.6>	339	211
ALS	11	2.02 <1.2-2.7>	1095	862

2.2. Corpus processing

First, each audio signal was automatically segmented into IPUs (Inter-Pausal Units). IPUs are understood as blocks of speech bounded by silent pauses over 250 ms, and time-aligned on the speech signal. This IPU-segmentation was then manually verified, and all noises (laughing, breathes, etc.) were manually segmented. For each of the speakers, an orthographic transliteration has been provided at the IPUs-level. The transliterations include a wide variety of phenomena that can occur in spontaneous speech. Conversational speech refers to an informal activity without specific preparation or planning and, as a consequence, numerous phenomena appear such as hesitations, repetitions, back-channel noises, etc. Phonetic phenomena such as non-standard elisions, reduction phenomena, truncated words, and more generally, non-standard pronunciations are also very frequent in the transcriptions of the present material. Transcribers were instructed to provide a transcription, which includes manually annotating non-standard events phonetized in SAMPA, consequently, the resulting transcription is pseudo-orthographic and pseudo-phonetic:

- elisions between parenthesis: (d)o (k), i(l), t(y), d(@)sy
- other specific realizations between brackets: [ils, iz], [heure, 2R2],
- proper names and acronyms are also transcribed in SAMPA.

This convention was designed to improve the quality of the Grapheme-To-Phoneme converter (all unknown words and irregular entries have been manually phonetized). Here is an example of a sentence extracted from the corpus:
"et a(l)OR i(l) [dit, de] m- euh qu'est-ce qu' i(l) m(@) veut"
(and then he say m- hum what he wants me).

This corpus was automatically time-aligned with signal at phone- and token-levels. Phonetization (or grapheme-to-phoneme conversion), which is based on the manual transcription was dictionary-based and performed by the phonetic segmentation tool. Short pauses included in speech segments were not indicated in the transcription and added automatically by the aligner. Finally, the automatic alignments were manually verified by two of the authors of this paper.

2.3. Automatic syllabification

Syllable boundaries were generated thanks to the automatic syllabification system described in [15] and included in SPPAS [18]. The task this system deals with is the syllabification of time-aligned phoneme sequences. The phoneme-to-syllable segmentation system is based on two main principles: (a) one syllable contains one vowel, and only one; and (b) a pause is a syllable boundary. These two principles focus the problem on the task of finding a syllabic boundary between two vowels. Phonemes are grouped into six classes (Vowels, Occlusives, Fricatives, Liquids, Nasals and Glides) and a set of rules were established to deal with these classes. The rules this system is using follow usual phonological statements for most of the corpora; and this system is reported to achieve good results on spontaneous speech.

For the needs of the present study, we extended the SPPAS system by adding an /fp/ entry to represent filled pauses and considered it as a syllable-break: each /fp/ is isolated into a syllable. Three new annotation layers were then created automatically and time-aligned for all the sub-corpora: syllables, syllable classes and syllable structures, as shown in Figure 1.

3. Description of syllable structures

3.1. Frequency of syllable structures

First, we examined the frequency of syllable structures. Results showed that syllable structures are strongly similar between populations. As observed in Figure 2, the most common

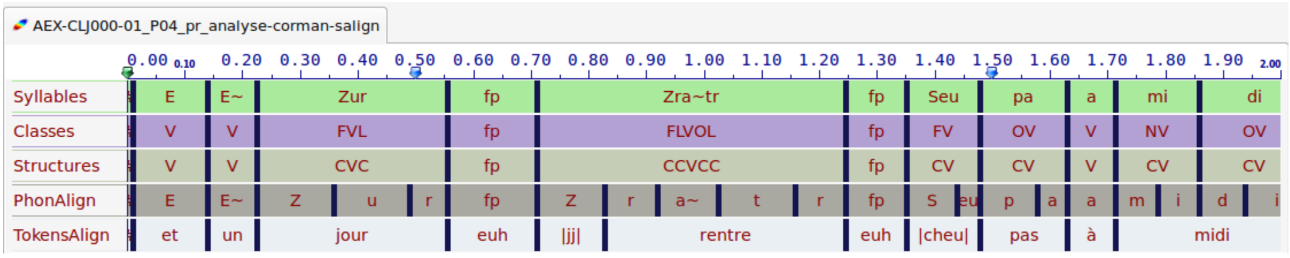

Figure 1: *Example multi-layer annotation including sample results of SPPAS automatic syllabification.*

syllable structures are CV (59% of occurrences), CVC (14% of occurrences), V (11% of occurrences) and CCV (11% of occurrences). Taken together, these syllable structures represent 95% of occurrences, the remaining 5% are divided into 10 categories: CCVC (2%), VC (1%), CCCV (0.6%), CVCC (0.4%), CCVCC (0.09%), CCCVC (0.08%), VCC (0,06%), CCCCV (0.01%), CCCCVC (0.01%), CVCCC (0.01%). These results are in accordance with [19, 20] and suggest that syllable distribution is stable across populations.

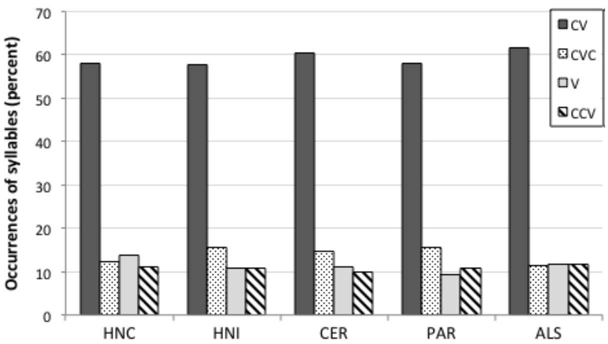

Figure 2: *Occurrences of syllable structures (CV, V, CVC, CCV) according to populations considered in this study (HNC, HNI, CER, PAR, ALS), (in percent).*

3.2. Durations of syllable structures

In order to examine the duration of syllable structures, we focused on the four most frequent syllable structures i.e., CVC, CCV, CV, and V. As illustrated in Figure 3, for all populations, we observed similar organization of duration, with the longest duration for CVC, and the smallest duration for V (CCV, CV are intermediate). Results showed that durations of syllables are dependent on populations. Indeed, duration of syllables are particularly high for ALS as compared to other populations (HNC, HNI, CER and PAR). CER also showed longer syllables than PAR or healthy groups (HNC or HNI). Durations of syllables are longer for HNI as compared to HNC and PAR for which the durations are similar.

4. Time Group Analysis

Our final set of experiments focused on the analyses of timing patterns using the TGA (Time Group Analysis) approach proposed by [16] and tools developed by [17]. TGA on-line tool enables automatic parsing and grouping of syllable sequences in speech annotations into Time Groups (TG), i.e. interpausal syllable groups, or into units based on deceleration models (con-

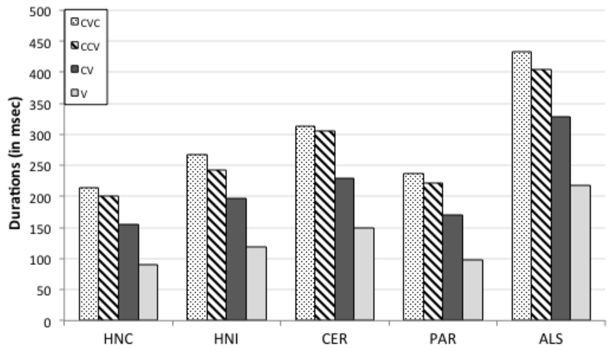

Figure 3: *Durations of syllables structure (CVC, CCV, CV, V) according to populations considered in this study (HNC, HNI, CER, PAR, ALS), in msec.*

sistent slowing down) or acceleration models (consistent speeding up). One of the novelties of the TGA was the use of duration difference slope (representing acceleration or deceleration) and intercept linear regression values for investigation of timing variability. Selected TGA options were implemented into Annotation Pro, a desktop software enabling annotation of linguistic and paralinguistic features of both single files and large collections of separate files [21]. As it was observed in several recent studies [12, 22], the variability of durational patterns, e.g., syllable duration difference slope patterning over interpausal time groups might contribute to differentiation between speaking styles.

In order to perform the TGA for the present set of healthy/dysarthric speech data, the syllable-based annotations were first imported to Annotation Pro and automatically divided into interpausal syllable groups. Then, the values of duration difference regression slope and intercept, as well as syllable-based nPVI [23], and speaking rates (in syll. per sec.) were automatically calculated using Annotation Pro + TGA plugin [17]. Altogether, a total of 2258 interpausal time groups were analyzed. We ignored segments including pause or noise labels as well as non-transcribed/not understandable stretches of speech. All other types of segments were included in the analysis, i.e. the segments including syllable labels as well as filled pauses labels (fp).

The mean values of slope obtained for the five groups of speakers (Figure 4) are close to the measurements reported by [12] for Aix-MARSEC corpus of French speech for genre categories A, B, C and D (news broadcast and lectures), with the only exception of PAR speakers who tend to produce slightly more deceleration patterns in their utterances (higher slopes on average). In case of mean intercepts, the values were higher for the CER and especially for the ALS speakers than for all other groups (and the above mentioned study of Aix-MARSEC data). The ALS group was also peculiar as regards the mean values of syllable-based nPVI (41) which were lower for this group than for all the others (HNC 49, HNI 50, CER 46, PAR 51) thus showing slightly weaker syllable-based pairwise durational variability (cf. also [6]).

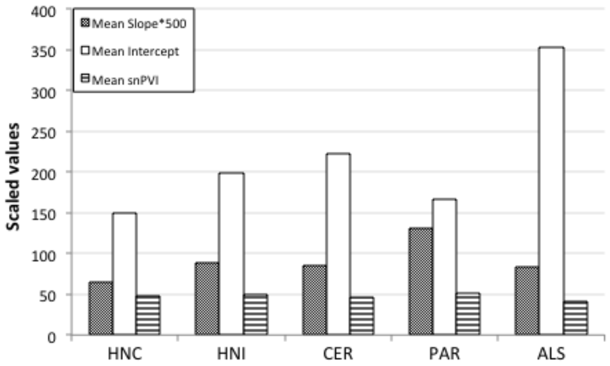

Figure 4: *TGA results. A comparison of quantitative measures for healthy and dysarthric speakers (mean duration difference slopes, mean intercepts, mean syllable nPVI).*

Mean syllable-based speaking rates (in syll. per sec.) in time groups for the healthy speakers were higher than those observed in the recordings of dysarthric speakers (5.37 on average for all healthy speakers, with the average for HNC at 6.03 syll. per sec. and for HNI at 4.72). This difference is consistent with the speaking style since speaking situation is an interview for HNI and a conversation for HNC. The lowest mean rates were obtained for the ALS speakers (3.41 syll. per sec.), and only slightly higher rates were observed for CER (3.92 syll. per sec.) while the result for PAR (5.39 syll. per sec.) was close to the average result of healthy speakers (similarly as in the case of syllable-based nPVI). Apart from the differences in mean rates, the PAR and ALS groups of speakers were characterized by significantly more inter-speaker variability which is illustrated by the confidence intervals displayed in Figure 5.

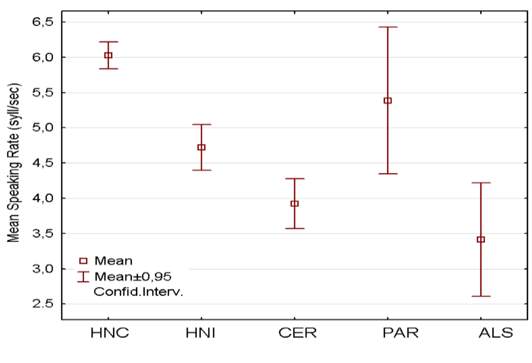

Figure 5: *Mean speaking rate (number of syll. per sec.) for each population.*

5. Discussion and conclusions

Overall, the results show that the populations in question can be distinguished according to syllable timing organization.

A first interesting point is that the distribution of syllable structures is strongly similar across populations. Motor constraints considerably affect dysarthric speakers and an avoidance strategy could be expected from these speakers, i.e. they could try to avoid the more complex syllabic structures including e.g., difficult consonant clusters. This is actually not the case. The strong predominance of CV structure is confirmed for all of the investigated populations and we do not note any decrease in the use of CCV structure for dysarthric speakers. Similarly, no differences are observed according to the relative duration of the four most frequent syllable structures. The only difference is the longer syllable durations for ALS which will be discussed below.

Time Group Analyses provides more subtle results. Mean speaking rates for each population is consistent with several observations on dysarthric populations [24, 25]: faster speech for PAR and a slow speaking rate for ALS, CER being intermediate. We also note a high dispersion around PAR mean value which suggests a strong variation for syllable durations. The utterances produced by PAR speakers are characterized by higher mean slopes than the other populations which suggests a pattern with more deceleration. This point is interesting: if we consider both the greatest deceleration value (slope) and the high speaking rate (correlating also with a low mean regression intercept), PAR seem to produce a relatively strong syllable time contrast (i.e. short syllables at the beginning of IPU, followed by relatively strong deceleration at the end of IPU). This pattern is clearly opposite to the ALS one. The significantly higher intercept for ALS speakers is expected to correlate strongly with their slower speaking rates and longer durations. Thus, ALS produce long syllables (speaking rate and mean intercept) but their mean slope value is similar to HNI one, which suggests a lower contrast in syllable time within the IPU.

With regard to speaking styles (comparison between HNI and HNC) both healthy populations show different speaking rates: 6 syll. per sec. for HNC and less that 5 syll. per sec. for HNI. This difference obviously results from the differences in speaking styles (conversation vs. interview). A difference in duration slopes (slightly higher mean values for HNI) is also observed and could also be due to the interview versus conversation context. Indeed, a high speaking rate in conversation may provide less timing contrast within each IPU.

To conclude, the analyses performed by TGA provide a relevant and interesting distinction between populations with regard to syllable timing organization. Distinct profiles have been highlighted by the analyses: parkinsonian speakers have a high speaking rate with an important syllable time contrast leading to strong deceleration within the IPU. At the opposite, ALS show low speaking rate with normal deceleration. CER appears intermediate between PAR and ALS and does not differ clearly from the healthy group except as concerns speaking rate. These results are highly consistent with recent ones [26] obtained on the basis of phone-level investigations. Finally, although each of the five populations can be distinguished from the remaining ones, the dysarthric group, treated as a whole, is not clearly distinct from the healthy group. These results suggest the complex boundary between healthy and pathological profiles as well as draw attention to the role of speaking styles.

6. Acknowledgements

This work was granted by the French National Agency TYPALOC Project: *Normal and abnormal speech variations: TYPology, Adaptation, LOCalisation* (ANR-12-BSH2-0003).

7. References

[1] L. O. Ramig, "The role of phonation in speech intelligibility: A review and preliminary data from patients with parkinsons disease," *Intelligibility in speech disorders: Theory, measurement and management*, pp. 119–155, 1992.

[2] R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *Journal of communication disorders*, vol. 32, no. 3, pp. 141–186, 1999.

[3] K. Bunton, R. D. Kent, J. F. Kent, and J. R. Duffy, "The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria," *Clinical Linguistics & Phonetics*, vol. 15, no. 3, pp. 181–193, 2001.

[4] H. Dahmani, S.-A. Selouani, D. Oshaughnessy, 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.

[5] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE, 1996, pp. 1962–1965.

[6] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[7] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proceedings of Speech Prosody Conference*, Aix-en-Provence, France, 2002, pp. 115–120.

[8] A. Content, R.-K. Kearns, and U.-H. Frauenfelder, "Boundaries versus onsets in syllabic segmentation," *Journal of Memory and Language*, vol. 45, pp. 177–199, 2001.

[9] W. Jassem, D. R. Hill, and I. H. Witten, "Isochrony in english speech: its statistical validity and linguistic relevance," *Intonation, Accent and Rhythm*, vol. 8, pp. 203–225, 1984.

[10] P. Roach, "On the distinction between stress-timedand syllable-timedlanguages," *Linguistic controversies*, pp. 73–79, 1982.

[11] Z. Malisz and P. Wagner, "Acoustic-phonetic realisation of polish syllable prominence: a corpus study." *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem.*, vol. 14, 2012.

[12] D. Gibbon, K. Klessa, and J. Bachan, "Duration and speed of speech events: a selection of methods," *Lingua Posnaniensis*, vol. 56, 2014.

[13] W. N. Campbell, "Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[14] D. Arnold, P. Wagner, and B. Möbius, "Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners," in *International Congress of the Phonetic Sciences*, 2011.

[15] B. Bigi, C. Meunier, I. Nesterenko, and R. Bertrand, "Automatic detection of syllable boundaries in spontaneous speech," in *Language Resource and Evaluation Conference*, La Valetta (Malta), 2010, pp. 3285–3292.

[16] D. Gibbon, "Tga: a web tool for time group analysis," in *Proc. of the Tools and Resources for the Analysis of Speech Prosody*, D. Hirst and B. B. (Eds.), Eds., Aix-en-Provence, France, 2013.

[17] K. Klessa and D. Gibbon, "Annotation pro+ tga: automation of speech timing analysis," in *Proceedings of Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*, 2014, pp. 26–31.

[18] B. Bigi, "SPPAS: a tool for the phonetic segmentations of Speech," in *The eighth international conference on Language Resources and Evaluation, ISBN 978-2-9517408-7-7*, Istanbul, Turkey, 2012, pp. 1748–1755.

[19] J.-P. Goldman and U. H. Frauenfelder, "Comparaison des structures syllabiques en français et en anglais," in *Actes des XXIèmes Journées d'Étude de la Parole*, 1996, pp. 119–122.

[20] I. Rousset, "Structures syllabiques et lexicales des langues du monde données, typologies, tendances universelles et contraintes substantielles," Ph.D. dissertation, Université Stendhal-Grenoble III, 2004.

[21] K. Klessa, "Annotation pro [software tool] version 2.2.1.2," Retrieved from: http://annotationpro.org/ on 2015-03-20.

[22] J. Yu, D. Gibbon, and K. Klessa, "Computational annotation-mining of syllable durations in speech varieties," in *Proceedings of 7th Speech Prosody Conference*, 2014, pp. 20–23.

[23] L. E. Ling, E. Grabe, and F. Nolan, "Quantitative characterizations of speech rhythm: Syllable-timing in singapore english," *Language and speech*, vol. 43, no. 4, pp. 377–401, 2000.

[24] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor Speech Disorders*, S. Philadelphia, Ed., 1975.

[25] K. Forrest, G. Weismer, and G. S. Turner, "Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2608–2622, 1989.

[26] L. Georgeton and C. Meunier, "Spontaneous speech production by dysarthric and healthy speakers: Temporal organisation and speaking rate," in *18th International Congress of Phonetic Sciences*, Glasgow, UK, submitted 2015.