Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Speech rate estimation in disordered speech based on spectral landmark detection

Hernandez-Diaz Huici[a,*], Hector A. Kairuz[b], Heidi Martens[a], Gwen Van Nuffelen[a,c], Marc De Bodt[a,c]

[a] Faculty of Medicine and Health Sciences, Dept. of ENT, Head & Neck and Communication Disorders, Antwerp University, Wilrijkstraat 10, 2650 Edegem, Belgium
[b] Faculty of Electrical Engineering, Central University Las Villas, C. Camajuani km 5.5, Santa Clara 54830, Cuba
[c] Faculty of Social Medicine and Health Sciences, University of Ghent, De Pintelaan 185, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 6 May 2015
Received in revised form 9 January 2016
Accepted 10 January 2016

Keywords:
Speaking rate estimation
Landmarks
Disordered speech
Dysarthria

ABSTRACT

Speech rate (SR) plays an important role in the assessment of disordered speech. Clinicians rely primarily on manual or semi-automatic methods to determine SR. The reported algorithms are designed for normal speech and show many restrictions with respect to disordered speech that are predominantly characterized by slow SR. This research presents an algorithm that in addition to energy and pitch, relies on information regarding the spectral characteristics of the borders of the syllables (landmarks). Speech samples (three sentences per speaker) for 66 healthy and dysarthric speakers were analyzed with four algorithms (*Mrate*, robust SR estimation method, *Praat*'s script and the proposed algorithm based on landmark detection). The landmark approach is demonstrated to be more accurate for speakers with slow SR. The Pearson correlation coefficient between the calculated SR and the reference remains over 0.84 for the 198 sentences analyzed, while the other algorithms' correlations are below the values reported in literature for fluent speech. In samples where SR is high, the algorithm shows similar limitations *versus* other algorithms due to the merging of syllables. The landmark-based algorithm is an adequate method for determining SR in disordered speech.

1. Introduction

Speech rate (SR) is defined as the number of phonetic units, such as syllables or words, uttered per time unit. SR can be expressed as the number of syllables per second that includes pauses and interjections, while articulation rate is expressed as the number of syllables per second excluding pauses [1].

A syllable is defined by Roach [2] as "consisting of a center which has little or no obstruction to airflow and which sounds comparatively loud; before and after that center (. . .) there will be a greater obstruction to airflow and/or less sound". This definition allows for a plausible way for detecting syllables in speech [3].

Speech rate is an important prosodic feature of speech that plays a major role in intelligibility and comprehensibility [4]. It is acknowledged widely that dysprosody is a hallmark of dysarthria [5]. Dysarthrias account for 54% of all acquired neurologic communication disorders at the Mayo Clinic Speech Pathology service [6]. Nearly all types of dysarthria are characterized by speech rate disorders. As such, speech rate plays a major role in clinical diagnosis and in therapy [7–9]. Together with the treatment of intonation and stress, the treatment of SR became an important aspect of therapy in dysarthria because rate control can improve intelligibility [7,10,11].

Auditory perceptual judgments of prosody are considered as the gold standard. Spectrographic analysis remains a principal tool to assist in analyzing prosodic aspects of normal and abnormal speech, but these methods are time consuming and their judgments are also susceptible to a variety of sources of error and bias [12]. Therefore, algorithms based on acoustic analysis can be powerful tools for improving the precision of diagnosis and for providing objective measures to document treatment progress [5,13,14].

Acoustic analysis of dysarthric speech struggles with a number of inherent limitations, such as phonatory disruptions, hypernasality, imprecise articulation, and low intensity. This is the reason why the algorithms based on normal speech fail for the analysis of dysarthric speech [15]. Moreover, objective measures in dysarthric speech face difficulties related to the assumptions inherent to signal processing. The variability of the signal amplified by the speech disorder and these difficulties increase when more complex

[*] Corresponding author. Tel.: +32 3 821 3485.
E-mail address: mariae.hernandez-diaz@uantwerpen.com (H.-D. Huici).

http://dx.doi.org/10.1016/j.bspc.2016.01.005
1746-8094/© 2016 Elsevier Ltd. All rights reserved.

language units, from phonemes to running speech, are analyzed [16]. To explore prosodic and paralinguistic aspects in dysarthric speech, Kent et al. [15] suggest the use of measures related to the fundamental frequency (especially intonation), duration of syllables and other units, and intensity. Acoustically based methods may provide additional information that is not easily obtained from auditory evaluation.

Direct speaking rate estimators, based on energy and periodicity measurements, were used in earlier studies because they are robust to speaker, language and genre, while estimators based on automatic speech recognition systems perform well only if the training and the test data are from the same speech group regarding language, genre or other particular characteristics [17]. Mermelstein [18] located syllable boundaries on energy between 0.5 to 4.0 kHz using a convex hull algorithm. With this precedent, a vowel landmark detector was introduced by Howitt [19], highlighting the importance of the frequency band between 300 to 900 Hz and the energy at the first formant for mapping vowel landmarks. A modified version of the same algorithm is applied by Xie and Niyogi [20] on both normalized energy and periodicity. The Mrate algorithm for syllable detection [21] introduced the use of sub-band energy correlation for this purpose. Wang and Narayanan [3] extended this concept by including temporal correlation and the use of a pitch detector. Both detection algorithms were tested on short and fluent speech from the Switchboard database and showed correlations with the transcribed syllable rate of 0.673 and 0.745, respectively. Authors of both studies reported limitations with slow or high rate segments. Combining energy and pitch contours, a syllable nuclei detector is reported in de Jong and Wempe [22]. It was implemented in Praat and tested on segments from the database reported in [23], resulting in a correlation of 0.71 between the calculated SR and the reference. Liu [24] proposed an algorithm to detect abrupt landmarks: stop closures and releases, nasal closures and releases, and vocal fold vibration start and stop. In each step, an energy waveform is constructed from each of the six bands, the derivative of the energy is computed (rate of rise: ROR), and peaks in ROR are detected. The localized peaks represent times of abrupt spectral change in the six bands, and from them, three types of landmarks are localized. The names and definitions by Liu [24] are:

- *Glottis (g):* marks a time when there is a transition of freely vibrating vocals to a condition where the vocal folds are not freely vibrating or vice versa;
- *Sonorant (s):* marks sonorant consonantal closures and releases;
- *Burst (b):* marks stop or affricate bursts and points where aspiration or frication ends due to a stop closure.

The feasibility of the landmark system to characterize syllabic clusters on repetitions of the syllable/ka/as produced by Parkinson's patients was reported in [25]. Kairuz et al. [26] applied Liu's algorithm to two sentences produced by 18 dysarthric patients to detect glottal activity. The error was 13.4%, demonstrating the need of adjustments of the algorithm to work with running speech from dysarthric patients.

A broad phonetic class recognizer for syllable detection and speech rate estimation is presented by Yuan and Liberman [17]. Its performance is comparable with the state-of-the-art algorithm reported by Wang and Narayanan [3] and shows an advantage in handling pauses and non-speech segments. They highlight the benefits of using distinct spectral characteristics, but the system needs training data. Another non-direct method for SR estimation is reported by Mujumdar and Kubichek [27], demonstrating a correlation of 0.88 between the estimated SR and the reference SR for the Switchboard data. The same algorithm was applied on a set of sentences extracted from the Grandfather Passage and the diadochokinetic rate task recorded from dysarthric patients [28]. The error in SR estimation on the sentences was 27% due to the presence of liquids and nasals in syllable borders, syllables with low intensity of short duration, continuous voicing, consecutive vowels and audible inspiration. This work illustrated the difficulties that disordered speech adds to the problem of syllable detection.

Disordered speech is often characterized by slow SR (*e.g.*, spastic, ataxic and hyperkinetic dysarthria) or variable SR (*e.g.*, hypokinetic dysarthria) [29]. Normal speech rate depends on cultural, demographic, linguistic and physiological variables [17]. As such, the language of as well as the listener as the speaker may influence the subjective evaluation of rate. Pellegrino et al. [30] reported different speech rates (syllables/s) on spontaneous speech for different languages, moving from German (5.97) to Japanese (7.84). The English mean speech rate was 6.19 syllables/s. The mean SR for Dutch in Belgium (4 syllables/s) was reported by Verhoeven et al. [31]. The SR classification used in this article is based on the transcribed speech rate presented by Wang and Narayanan [3]: fast (>5 syllables/s), normal (between 3 and 5 syllables/s) and slow (<3 syllables/s). An objective measure of SR can facilitate a comparison among clinical studies and avoids subjective bias.

The goal of this research is to construct an SR algorithm based on periodicity, intensity and landmark detection, which is sufficiently robust to manage variable and slow speech rates that characterize disordered speech.

2. Materials and methods

2.1. Speech samples

All samples used in this study were selected from the CATRIS database [32] that was developed for research on prosodic features. All subjects are native speakers of Dutch (Belgium). A total of 66 speakers: 33 healthy speakers (average SR is 4.88 syllables/s; SD is 0.84) and 33 dysarthric speakers (average speech rate is 4.09 syllables/s; SD is 1.47) were selected for further analysis. For all of the samples together (198), the reference SR varies between 1.01 to 7.9 syllables/s. The group of healthy speakers is younger than the pathological group, representing higher speech rates, and implies a challenge for the algorithms. Histograms for both groups are shown in Fig. 1. Three isolated sentences from each speaker were selected to evaluate the algorithms: sentence 1: "Het is bijna tijd," (It is almost time) containing five syllables; sentence 2: "Ik denk dat alles in orde is," (I think everything is O.K.) containing nine syllables; and sentence 3: "Zij wil geen telefoon meer krijgen" (She does not want any phone calls) also containing nine syllables.

The selection of these samples is based on the difficulties summarized in the introduction and to explore whether the proposed algorithm can manage these limitations. Sentence no. 1 is characterized by continuous voiced segments where two syllables are separated by a nasal ("bijna"). Sentence no. 2 includes liquids ("l" and "r") and consecutive vowels ("orde-is"). The last sentence contains a short syllable ("le") and a low intensity syllable at the end and liquids and nasals. Thanks to the short length of the sentences, deletions and insertions cannot compensate for each other. The characteristics of the database are represented in Table 1.

2.2. Manual determination of SR

All 198 sentences were annotated by an experienced speech language pathologist (SLP) to determine the reference SR. With the Praat software, the number of syllables and the duration of each sentence were checked by visual inspection on the spectrographic representation and the play back facility. SR was defined as the quotient of the number of syllables and the sentence duration.