AUTOMATIC DETECTION OF VOICE ONSET TIME IN DYSARTHRIC SPEECH

Michal Novotný[1], Jakub Pospíšil[1], Roman Čmejla[1], Jan Rusz[1, 2]

[1]Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Technická 2, 160 00 Prague 6, Czech Republic

[2]Charles University in Prague, First Faculty of Medicine, Department of Neurology and Centre of Clinical Neuroscience, Kateřinská 30, 120 00 Prague 2, Czech Republic

{novotm26, pospij27, cmejla, ruszjan}@fel.cvut.cz

ABSTRACT

Although a number of speech disorders reflect varying involvement of brain areas, recently published automatic speech analyses have primarily been limited to hypokinetic dysarthria in Parkinson's disease (PD). Therefore, the aim of the present study was to provide an automatic algorithm suitable for the assessment of voice onset time (VOT) in various dysarthria types. Twenty-four PD participants with hypokinetic dysarthria and 40 Huntington's disease (HD) subjects with hyperkinetic dysarthria were included. These two types of dysarthria were selected in the design of a robust algorithm as they contain most of the dysarthric patterns found among all dysarthria subtypes. For a 10 ms threshold, the proposed algorithm reached approximately 90% accuracy in PD speakers and 80% accuracy in HD speakers. The accuracy of 80% obtained in HD was superior to the performance of 55% achieved by a previous algorithm designed particularly for hypokinetic dysarthria in PD.

Index Terms— Voice Onset Time, Dysarthria, Parkinson's disease, Huntington's disease, Speech disorder

1. INTRODUCTION

Neurodegenerative disorders are associated with the progressive damage of nerve cells and motor complications represent one of the most severe sequelae in the lives of patients. Parkinson's disease (PD) and Huntington's disease (HD) represent two well-known neurodegenerative disorders associated with various motor disruptions. Although both diseases primarily affect the basal ganglia, their differing mechanisms lead to contrasting effects on patient motor performance [1, 2]. Dopamine depletion in the course of PD leads to bradykinesia, muscular rigidity, gait instability and a 5 Hz resting tremor, whereas HD primarily leads to chorea characterized by extensive semi-directed, non-rhythmic movements [1, 2]. The motor speech disorder termed dysarthria is common to both PD and HD. In PD, hypokinetic dysarthria may be characterized by an increased rate of speech, monopitch, monoloudness, hypophonia and reduced stress [3]. The manifestation of hyperkinetic dysarthria in HD, in contrast to hypokinetic dysarthria, includes the presence of slow rate, excessive pitch, excessive loudness, excessive and equal stress, variable rate, inappropriate voice breaks and audible inspirations [3]. Together, these dysarthria subtypes encompass the majority of speech patterns appearing across all types of dysarthria, and therefore provide a suitable model for the development of robust automatic methods for the assessment of various types of dysarthria.

Although dysarthria is primarily an articulatory disorder, previous studies have mainly focused on automatic assessments of sustained phonation [4]. Recently, the automatic assessment of articulatory disorders has also been introduced [5]; however, it has been limited to hypokinetic dysarthria due to the relatively high incidence of PD. To the best of our knowledge, there is no tool available for the automatic assessment of speech disorders applicable to different dysarthria subtypes.

Voice onset time (VOT) represents a common method for the evaluation of articulatory deficits in the course of different types of dysarthria [6]. VOT is used as a marker of laryngeal and supralaryngeal coordination [7]. The diadochokinetic (DDK) task is particularly suitable for the assessment of VOT as it is based on fast, steady syllable repetition, which is a demanding task for dysarthric patients. Accordingly, participants perform utterances at their maximal speed and are not able to voluntarily compensate articulatory deficits. The commonly used DDK task, the repetition of /pa/-/ta/-/ka/ syllables, combines bilabial, alveolar and velar places of articulation.

Therefore, goal of the present study was to develop a robust algorithm for automatic VOT estimation applicable to different dysarthria subtypes and to demonstrate its applicability by evaluating PD and HD speakers.

This work was primarily supported in part by the Czech Grant Agency under grant No.102 /12/ 2230 and Charles University in Prague under grant No. PRVOUK-P26/LF1/4. The secondary support was provided by Nadace " Nadání Josefa, Marie a Zdeňky Hlávkových" and Nadace Český literární fond.

2. METHODS

2.1. Subjects

Hypokinetic dysarthria was represented by 40 utterances of 24 participants (20 men and 4 women) who fulfilled the diagnostic criteria for PD [8]. The mean age of the PD group was 60.1 ± standard deviation (SD) 12.6 years. All PD participants were assessed by a specialist using the Unified Parkinson's Disease Rating Scale (UPDRS) [9]; the mean UPDRS score was 17.4 ± SD 7.1. The PD utterances were originally collected as part of a previous study [10]. Hyperkinetic dysarthria was represented by 77 utterances from 40 speakers (20 men and 20 women) diagnosed with HD. The mean age of the HD group was 48.6 ± SD 13.4 years. All HD participants were assessed by a specialist using the Unified Huntington's Disease Rating Scale (UHDRS) [11]; the HD group obtained a UHDRS motor score of 26.9 ± SD 11.6. The HD utterances were originally collected as part of a previous study [12].

2.2. Recording

Recording took place in a quiet place with low ambivalent noise using a condenser microphone at a distance of approximately 5 cm from the subject's mouth. Data were recorded with a 48 kHz sampling frequency and 16 bit quantization. All participants were recorded during a single session by a speech-language pathologist who instructed them to perform rapid /pa/-/ta/-/ka/ syllable repetitions as consistently and rapidly as possible. No time limits were imposed and participants could repeat the task.

2.3. Labeling

In order to evaluate algorithm performance, manually-placed labels for the initial burst and vowel onset were used as reference positions. However, manually labeling dysarthric speech may be a challenging task and therefore two labeling rules were set in compliance with previously established guidelines [5]. First, in the case of multiple bursts, the first burst was set as the initial burst of consonants [13]. Second, vowel onset was defined by the presence of the fundamental and first two formant frequencies [14].

2.3. Algorithm

The VOT is defined as the difference between the initial burst and vowel onset [15] (see Fig. 1). Therefore, both positions must be detected in each syllable of the utterance. Nevertheless, the unknown number of syllables makes any segmentation difficult and therefore each utterance was first segmented into single syllables.

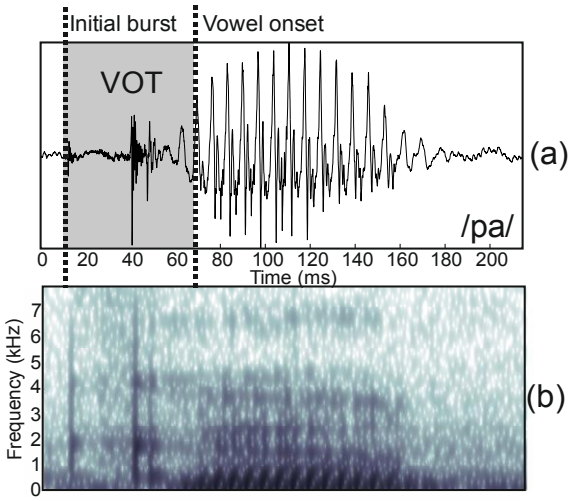

Fig. 1. Syllable /pa/ pronounced by HD speaker and the highlighted area of VOT and its borders including the initial burst and the vowel onset. Section (a) represents the time domain, whereas section (b) represents frequency domain.

Even though the recently presented algorithm showed sufficient robustness in hypokinetic utterances it was prone to highly variable rate caused by forced inspirations and expiration present in HD utterances [5]. Therefore to address this aspect of hyperkinetic speech an algorithm based on analysis of the linear prediction (LP) residual was used for the purposes of rough segmentation [16]. This approach uses the LP residual for the detection of voiced parts of the utterance. The LP residual was estimated from the signal which was down sampled to 8 kHz and filtered by a 500 Hz FIR filter with an order of 100. The 500 Hz FIR filter removed majority of high frequency signals including consonants and inspirations and preserved fundamental frequency included in voicing. Subsequently, the Hilbert envelope of the LP residual was estimated and smoothened by a moving average filter with an order of 500 [16].

Positions of peaks in the smoothened envelope were set as the positions of vowel nuclei. To detect these peaks, an envelope slope was computed using the first-order difference and every positive to negative zero crossing was marked as vowel nuclei. To eliminate false detections due to low signal-to-noise ratio or intensity fluctuations along single vowels, the minimal distance between two vowel nuclei was set to 10 ms. When two peaks were found to be within 10 ms distance, the higher peak was selected as vowel nucleus (see Fig. 2).

To detect syllable borders, the local minima of the smoothed Hilbert envelope were detected. Subsequently, each syllabic nucleus was associated with the nearest local minimum. The position of syllable beginning or end was determined according to the relative position of local minima to the nucleus. The second border was then chosen according to this decision.

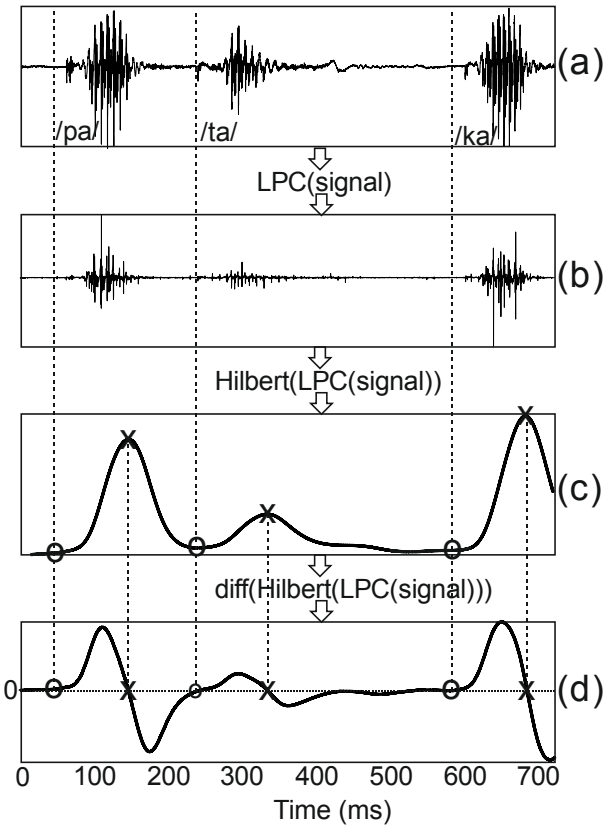

Fig. 2. Detection of syllable nuclei based on [13]. The utterance (a) is analyzed using LP residua (b), its smoothened Hilbert envelope (c) and slope of this envelope (d).The nucleus positions are marked as peaks in (c) and positive to negative crossing in (d).

Signal within these borders was evaluated to determine the position of the initial burst. Initial burst detection was based on spectral characteristics. First, a spectrogram with 6 ms window length was estimated from the signal sampled at 20 kHz. Then a filtration matrix **T** was computed according to

$$\mathbf{T}(i, 1 \dots n) = 0.8 \frac{1}{n} \sum_{j=1}^{n} \mathbf{P}(i, j), \qquad (1)$$

where *i* is the index of each frequency bin, *n* is the number of time bins and **P** is the power spectral density estimation matrix. The spectrogram matrix **P** was than filtered by comparison with **T** as is defined in

$$\mathbf{P}_{filtered} = \begin{cases} 1 & P(i,j) \geq T(i,j) \\ 0 & P(i,j) < T(i,j) \end{cases}. \qquad (2)$$

The envelope, given by summing all values in each time bin in $\mathbf{P}_{filtered}$, was used to emphasize widespread bursts over the formant frequency centered vowel (see Fig. 3).

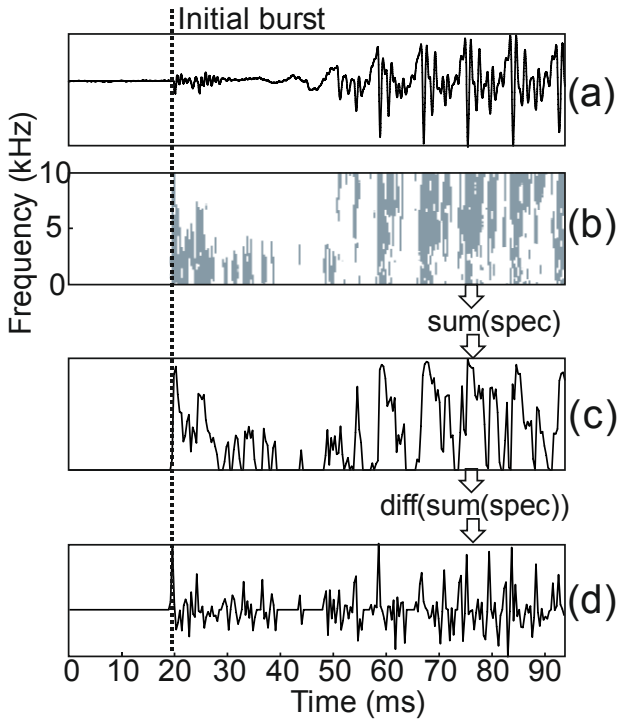

Fig. 3. Process of the initial burst detection. The syllable (a) is processed using filtered spectrogram (b). Then using summation along frequency axis the energy envelope (c) is computed and the initial burst is detected in the difference (d) of this envelope.

The quasi-periodic character of vowels with an abrupt onset of energy was detected using a Bayesian Step Changepoint Detector (BSCD) [17]. The BSCD assumes the signal to be composed of two constant values (e.g., 0.05 and 0.3 in our algorithm) and computes the *a posteriori* probability of changes in the signal using Bayesian marginalization. The character of the BSCD model based on two constant values highlights the boundary between two different signals (see Fig. 4).

2.4. Algorithm performance estimation

Based on the methods presented by Stouten and Van Hame [18], the cumulative distributions of absolute differences between reference and automatically detected positions were used for the purposes of algorithm performance estimation. Additionally, the 10 ms threshold was chosen as the representative threshold value with respect to previous studies [5, 18]. The cumulative distributions were computed using all syllables contained in the PD or HD groups and falsely detected or missed syllables were always set as erroneous. Cumulative distributions were estimated separately for PD and HD participants.

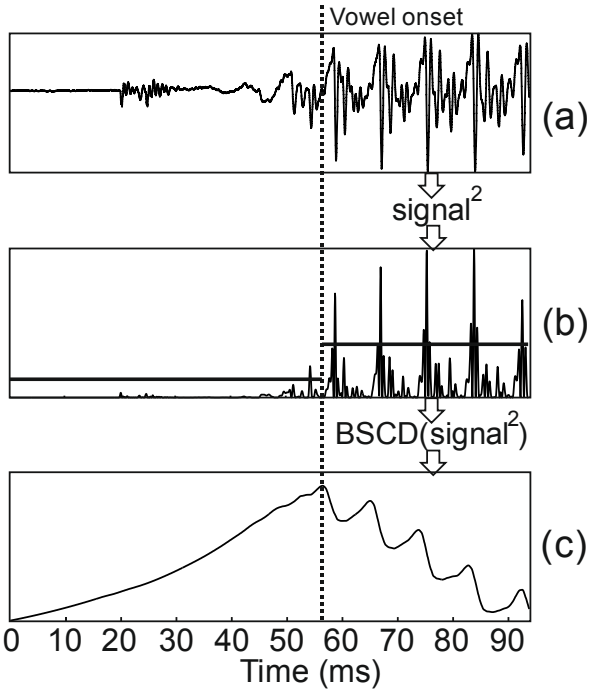

Fig. 4. The vowel onset detection is based on processing of signal (a), which is squared (b) and modeled using BSCD represented by two constant values (b). The vowel onset is detected in BSCD output (c).

For the purposes of comparison, data for both hypokinetic and hyperkinetic dysarthria subtypes were also analyzed by an algorithm designed previously for the automatic estimation of articulatory deficits in PD [5].

3. RESULTS

Figure 5 illustrates algorithm performance. VOT boundary detection is illustrated by solid lines, which shows algorithm performance using cumulative distributions of absolute differences between detected and reference positions. Furthermore, for the purposes of comparison, the dashed lines in figure 5 represent the performance of an algorithm designed previously for PD subjects [5]. Considering a 10 ms threshold and PD condition, the present algorithm achieved a slightly improved score of 81.5% for the initial burst and 89.5% for vowel onset, in comparison to a score of 78.2% for the initial burst and 88.6% for vowel onset achieved by the previous algorithm. Considering a 10 ms threshold and HD condition, the present algorithm achieved 77.8% for the initial burst and 80.1% for vowel onset, significantly outperforming scores of 45.8% for initial burst and 55.1% for vowel onset achieved by the algorithm designed previously for PD subjects.

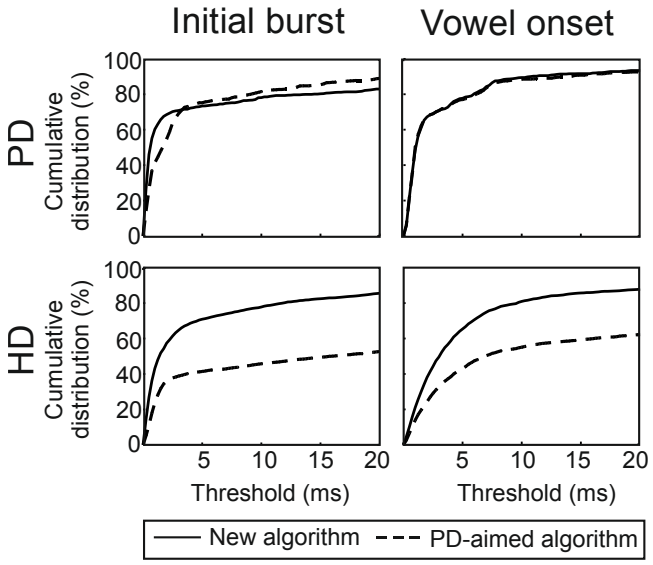

Fig. 5. Cumulative distributions of absolute difference between detected and reference values. The results obtained for the Initial burst (1st column) and Vowel onset (2nd column) and for PD (1st row) and HD (2nd row) speakers. For the purposes of comparison, the dashed line represents results of PD-aimed algorithm presented in [5].

4. CONCLUSION

We present a new, automatic algorithm for the estimation of VOT in dysarthria. We achieved a high performance score of up to 90% in PD speakers and up to 80% in HD speakers for a 10 ms threshold. In the case of hyperkinetic dysarthria in HD, the current approach was superior to a previous algorithm designed particularly for hypokinetic dysarthria in PD [5], with increased performance by over 25%. Indeed, the previous approaches [5, 18] were not sufficient in the evaluation of HD speech as hyperkinetic dysarthria may be particularly associated with audible inspirations and inappropriate voice breaks, which could affect the detection of syllable nuclei and their borders, and therefore leads to an increased occurrence of false detections. The robustness of the present algorithm to uncontrollable confounding effects in HD speech seems very promising for the automatic detection of VOT in different types of dysarthria due to various neurological conditions.

12. REFERENCES

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *J. Neurol. Neurodegener. Dis.*, vol. 5, pp. 368-376, 2012.

[2] A. Bernardelli, J. Noth, P.D. Thompson, E.L. Bollen, A. Curra et al., "Pathophysiology of Chorea and Bradykinesia in Huntington's disease," *Mov. Disord.*, vol. 14, pp. 398-403, 1999.

[3] J.R. Duffy, *Motor speech disorders. Substrates, differential diagnosis and management*, 2nd Ed., Elsevier Mosby, St. Louis, MO, 2005.

[4] M.A. Little, P.M. McSharry, E.J. Hunter, J. Speielman, and L.O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 4, pp. 1015-1022, 2009.

[5] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic Evaluation of Articulatory Disorders in Parkinson's Disease," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 366-378, 2014.

[6] E. Fischer, and A.M Goberman, "Voice onset time in Parkinson's Disease," *J. Commun. Disord.,* vol. 43, pp. 21-34, 2010.

[7] R.D. Kent, G. Weismer, J.F. Kent, J.K. Vorperian, and J.R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *J. Commun. Disord.*, vol. 32, pp. 141-186, 1999.

[8] A.J. Hughes, S.E. Daniel, L. Kilford, and A.J. Lees "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinicopathological study of 100 cases," *J. Nurole. Neurosur. Ps.*, vol. 55, pp. 181-184, 1992.

[9] G. Stebbing, and C. Goetz, "Factor structure of the Unified Parkinson's Disease Rating Scale: Motor Examination section," *Mov. Disord.*, vol. 13, pp. 633-636, 1998.

[10] J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, V. Majerová, J. Picmausová, J. Roth, and E. Růžička, "Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test," *Mov. Disord.*, vol. 26, no. 10, pp. 1951-1952, 2011.

[11] Huntington Study Group, "Unified Huntingto's Disease Rating Scale: reliability and consistency," *Mov. Disord.*, vol. 11, pp. 136-142, 1996.

[12] J. Rusz, J. Klempíř, T. Tykalová, E. Barborová, R. Cmejla, E. Růžička, and J. Roth, "Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication," *J. Neral. Transm*., vol. 121, pp. 655-664, 2014.

[13] Y. Wang, R. Kent, J. Duffy, J. Thomas, and G. Weismer, "Alternating motion rate as an index of speech motor disorder in traumatic brain injury," *Clin. Linguist. Phonet.*, vol. 18, no. 1, pp. 57-84, 2004.

[14] L. Volaitis, and J. Miller, "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," *J. Acoust. Soc. Am.*, vol. 92, pp. 723-735, 1992.

[15] J.H. Hansen, S.S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/, /t/, /k/) with application to accent classification," *Speech. Commun.*, vol. 52, pp. 777-789, 2010.

[16] S.R. Mahadeva Prasanna, B.V. Sanddep Reddy, and P. Krishnamoorty, "Vowel Onset point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 556-565, 2009.

[17] R. Čmejla, J. Rusz, P. Bergl, and J. Vokřál, "Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders," *Speech. Commun.*, vol. 55, pp. 178-189, 2013.

[18] V. Stouten, and H. Van Hame, "Automatic voice onset time for reassignment spectra," *Speech Commun.,* vol. 51, pp. 1194-1205, 2009.