THE NUMBER OF SPECTRAL CHANGES IN SPEECH SEGMENTS FOR EVALUATION OF DYSFLUENT SPEECH

T. Lustyk[1)], *P. Bergl*[1)], *R. Čmejla*[1)], *J. Vokřál*[2)]

[1)]Faculty of Electrical Engineering Czech Technical University in Prague
[2)]Department of Phoniatrics, 1st Faculty of Medicine at Charles University and General Faculty Hospital in Prague

Abstract

This paper is focused on a measurement that could be useful for an automatic and objective evaluation of dysfluent speech. The parameter combines analysis of both the time and frequency domains. The experiment is based on the analysis of *read* audio recordings of stutterers. The highest correlation coefficient with control data is -0.821. This parameter could become part of an automatic and objective assessment system.

1 Introducing

Stuttering is one of the main fluency disorders in speech and appears in many forms and causes. Signs of stuttering are repetition (repetition of sounds, syllables, words), prolongation (unnatural extension of sounds), frequent pauses and broken words [1], [2]. Correct assessment of the disorder severity and follow-up treatment are very difficult tasks.

Determination of disorder severity is based on subjective speech assessment, one currently performed by clinicians. Hence a method, that would be able objectively and automatically to determine the degree of speech fluency disorder, would be useful. Application of such a method would be: 1) Determination of disorder severity. 2) Assessment of treatment results. 3) Comparison of treatment approaches.

The symptoms may be found in audio recordings to determine the degree of speech disorder. The article [3] concentrates on finding repetitions and prolongations in the speech signal. A simple VAD (Voice Activity Detector) and time threshold was used to detect repetitions. The detection of the formant frequencies was applied for finding of prolongations. More complex method involving HMM (Hidden Markov Model) was used for recognizing blockades with repetition and prolongations on fricative phonemes in [4].

The parameters do not have to search for the symptoms of speech disorder but they could instead process the signal as a whole. The group of parameters in the time and frequency domain has been described in thesis [5]. In the time domain, for example, these are: the average length of silence, the ratio of the total length of silence and speech and the parameters exploring speech signal energy. In the frequency domain, for example: the number of BACD (Bayesian change-point detector) maxima and the standard deviation of distance BACD maxima.

In this paper, a brief view is provided of one parameter which uses a combination of processing by means of spectral-change detector and VAD, and its results are discussed.

2 Database

The speech signal database was created in the past few years at the Department of Phoniatrics of the 1st Faculty of Medicine at Charles University and the General Faculty Hospital in Prague. The database contains recordings of approximately 160 Czech native speakers of different age and with different degree of speech fluency disorder. Utterances are *read* and *spontaneous*, recorded with and without DAF (Delayed Auditory Feedback). Thirty utterances are control recordings of healthy speakers. The range of DAF varies from 10 to 110 ms. The sampling

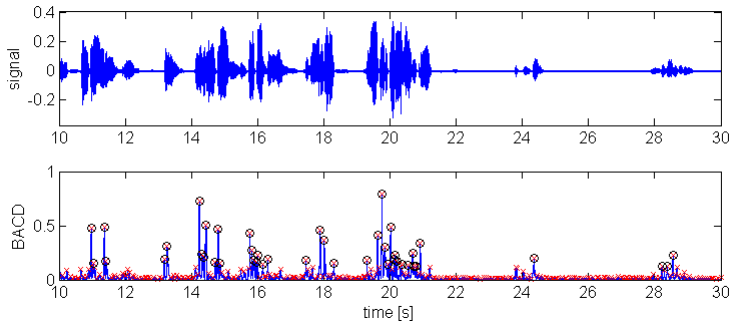

Figure 1: An example of speech signal (*above*) and BACD output (*below*).

frequency was 44 kHz when recording. The signals were down-sampled to 16 kHz for following analysis.

The *read* text consists of about 70 words. Each utterance takes approximately 60 s. *Spontaneous* utterances are formed as an image description. The *read* part of the database (121 audio recordings) was used in all experiments described in this paper.

Subjective assessment of stuttering and dysfluency was conducted in 2008 by two clinicians at the Phoniatrics Clinic in Prague. The degree of speech fluency disorder has been described on a five-level scale, from 0 - no occurrence of dysfluent speech to 4 - very severe degree of speech fluency disorder. This data serves as control data for all of our analysis.

3 Methods

The method for dysfluent speech evaluation could be imagine as a black box. At the beginning there is a speech signal, which is processed by an algorithm (black box) and at the end there is a number representing the level of speech fluency disorder.

The introduced algorithm uses two instruments for processing. The first is a voice activity detector and the second is a detector of spectral changes. VAD based on Mel-spectrum is used for analysis in the time domain. A brief view on its procedure: 1) Estimation of power spectra (computed by the Welch's method). 2) Application of the triangular Mel-frequency filter bank. 3) Decision about speech activity in each frequency band by means of adaptive threshold. 4) Final decision about speech activity (speech/silence).

Detection using Bayesian approach is used for analysis in the frequency domain. The detector (BACD) searches for spectral changes in the audio signal. Spectral changes should correspond to the phoneme boundaries. See [6] for more details. An example of speech signal and BACD output is shown in Figure 1. All positions of spectral changes are indicated by the red x-mark as local maxima. Many local maxima do not correspond to significant changes at phoneme boundaries. These maxima are excluded from further analysis by applying a threshold.

The analysis carried out on the detector outputs from different participants showed that the threshold should not be the same for all signals in the database. In thesis [5] an original method was presented for threshold extraction. The threshold is determined as a fragment of one selected maximum. The adaptive approach has been studied in thesis [7], but experiments did not have results as good as the fixed threshold. The significant spectral changes (maxima) are marked by black circles in Figure 1.

Other methods such as GLR (General Likelihood Ratio) and cepstral distance, or analysis by HTK (the Hidden Markov Model Toolkit), could be used for finding spectral changes. The following analysis will be identical.

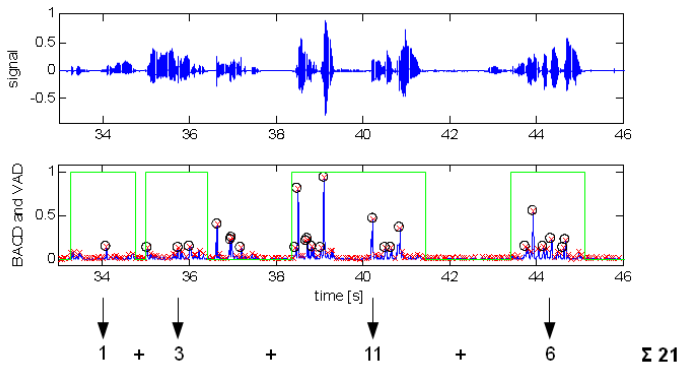

Figure 2: Procedure of calculation of *the number of spectral changes in speech segments*. *Above:* Speech signal. *Below:* BACD (*blue*) and VAD (*green*) output.

4 The number of spectral changes in speech segments

Automatic methods for evaluating speech fluency for second language learners' have been introduced in [8] and [9]. The aims of our work are very similar in spite of the apparent difference between fields of interest, i.e. to develop parameters that are able to measure the speech fluency (in our case the speech dysfluency).

The parameters *the rate of speech* and *the articulation rate* were found to be the most significant. Authors have defined the rate of speech as the number of phonemes/total duration of speech including sentence-internal pauses and the articulation rate as the number of phonemes/total duration of speech without pauses. These two parameters served as an inspiration for the measurement.

The automatic speech recognizer was utilized to find the number of phonemes in articles [8] and [9], but a different approach is used in this experiment - the combination of two instruments (VAD and a detector of spectral changes). These instruments do not measure directly the number of phonemes but the number of spectral changes in speech segments. Significant spectral changes represent phoneme boundaries and the measurement should corresponds to the number of phonemes.

The measurement could be divided into three steps. The analysis by means of BACD (finding of significant spectral changes in speech signal) in the first step, speech/silence detection in the second and the combination of previous procedures follows in the third step. The combination is based on calculation of the number of spectral changes in each speech segment. Spectral changes out of speech segments are excluded. Then the total number of spectral changes is simply computed and divided by the length of the signal because of signal length independence. The procedure is given in Figure 2.

The procedure based on the removal of short speech segments is applied to emphasize the difference between fluent and dysfluent speech. The removal is successive, at first segments shorter than 125 ms are removed then segments shorter than 150 ms, the final threshold is 1700 ms. See Figure 3. It is considered that the healthy speaker should have more unit speech segments and the removal do not involved the final speech/silence distribution in opposite to pathological speech (the speech is broken by repetitions, pauses within words). These pathological parts of speech are expected to be removed.

5 Results

The subjective assessment of phoniatrics experts has been noted in section concentrating on the database. A score provided by doctors serves as the control data for all experiments. The outputs of the algorithm are compared in several aspects with the subjective assessment. The correlation

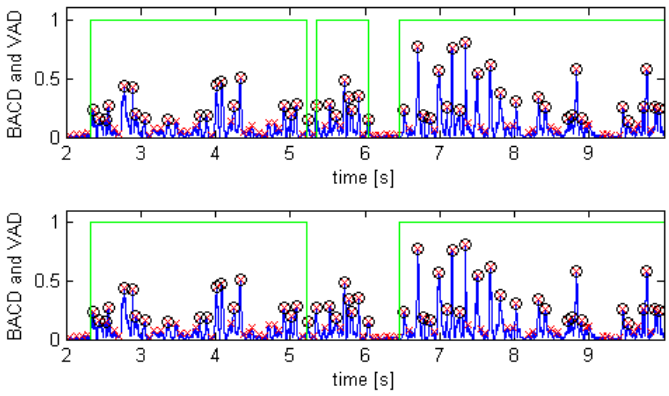

Figure 3: Procedure of removal of short speech segments. VAD and BACD output after removing speech segments shorter than 700 ms (*up*) and VAD and BACD output after removing speech segments shorter than 900 ms (*down*).

Table 1: CORRELATION COEFFICIENT BETWEEN *the number of spectral changes in speech segments* AND THE SUBJECTIVE ASSESSMENT OF FIRST DOCTOR. VALUE OF k DEFINES WHICH MAXIMUM IS USED FOR THRESHOLD DETERMINING, M_k IS THE VALUE OF MAXIMUM.

k	$0.10 \cdot M_k$	$0.15 \cdot M_k$	$0.20 \cdot M_k$	$0.25 \cdot M_k$	$0.30 \cdot M_k$
1	**−0.805**	−0.785	−0.735	−0.679	−0.609
1	**−0.818**	−0.808	−0.772	−0.724	−0.674
3	**−0.813**	**−0.807**	−0.774	−0.733	−0.687
4	**−0.816**	**−0.811**	−0.782	−0.744	−0.699
5	**−0.814**	**−0.811**	−0.791	−0.751	−0.719
6	**−0.817**	**−0.815**	−0.796	−0.761	−0.730
7	**−0.815**	**−0.815**	**−0.801**	−0.769	−0.737
8	**−0.819**	**−0.818**	**−0.808**	−0.783	−0.750
9	**−0.821**	**−0.819**	**−0.814**	−0.789	−0.761
10	**−0.820**	**−0.821**	**−0.816**	−0.794	−0.767

coefficient is the first and Wilcoxon rank sum test is the second. Correlation coefficient shows the level of agreement between outputs of the algorithm and assessments of clinicians, while the Wilcoxon test serves to determine whether the signals are partitionable into classes. From the *read* part of the database, 121 signals were used to verify the suitability of the parameter for evaluation of the speech disorder severity. The value of *the number of spectral changes in speech segments* is computed for each signal.

The results of the algorithm are very large because of the high number of possible settings, one part of settings deals with the BACD threshold and the second with the VAD. The threshold is determined as a fraction of one of the highest maximum in BACD output (see Section 3 for details). We used ten values for k (to specify the order of maxima) from the first to tenth highest maximum (k= 1, 2, ... 10). Values for the multiple were from 0.1 to 0.3 (the step 0.05). The settings dealing with VAD are derived from the procedure that was used to emphasize the difference between fluent and dysfluent speech. The threshold values are from 125 to 1700 ms.

Correlation coefficients with the first clinician assessment and various setting of the threshold are given in Table 1, the threshold for the removal of short speech segments was 1100 ms. Rows specify from which maximum the threshold is derived, columns define the multiple of maximum. Correlation coefficients higher than -0.8 are highlighted in bold.

The high values of coefficients can be seen for the multiples 0.10 and 0.15. The best achieved score is -0.821 for a threshold specified by k= 9 and the multiple 0.10, as well as k= 10

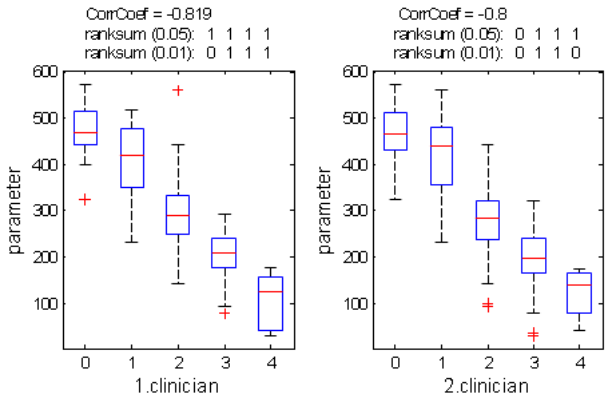

Figure 4: *The number of spectral changes in speech segments* for threshold determined k= 9, multiple 0.15 and the threshold for removal of speech segments 1100 ms. Boxplot, correlation coefficient and Wilcoxon test.

and multiple 0.15. Detailed results of one of the best setting are shown in Figure 4 (k= 9, multiple 0.15), the correlation coefficients exceeded value of -0.8 for both doctors.

The result is displayed using the Matlab function *boxplot*. Five boxes are plotted in figure, one for each level. If the parameter is indicative of the speech disorder severity, the boxes should be of different height and their size should be minimized. The correlation coefficients with assessment of doctors and results of Wilcoxon test for two significance levels ($\alpha = 0.05$ and $\alpha = 0.01$) are written in the header.

The Wilcoxon test performs a rank sum test of the hypothesis that two independent samples come from distributions with equal medians. Note that "0 0 1 0" means that the hypothesis about equal medians is rejected between groups 3 and 4. Then, the parameter could be useful for differentiation these two groups and ideally with an achievement of eight. The parameter presented in Figure 4 achieved seven of the eight ones; it turns out that the algorithm is able to differentiate between groups. The results of other algorithm settings suggest that decision between peripheral groups (especially 0 and 1) is a difficult task.

The comparison of parameter with the second clinician achieved similar results. The coefficient values are lower about 0.02 on average, approximately the same number of settings exceeds coefficient -0.79. The best results were obtained for the speech removal threshold 1100 ms for for both doctors. Another VAD based on SVM (Support Vector Machine) was also employed, but the value of -0.821 was not improved.

6 Conclusion

This paper deals with a method that could be used in automatic evaluation of dysfluent speech. The algorithm processes an audio signal in the time and frequency domain. The parameter was inspired by the papers [8] and [9] where the evaluation of speech fluency for second language learners' has been introduced.

VAD is used to find speech/silence segments in the time domain while Bayesian change-point detector is employed as a detector of spectral changes in the frequency domain of the audio signal. The following procedure is based on calculation of the number of spectral changes in each speech segment. The suitability of the parameter *the number of spectral changes in speech segments* to assess the severity of the speech fluency disorder was tested on 121 *read* utterances.

Parameter results were compared with the evaluation of two clinicians. The highest correlation coefficient was achieved for one of the doctors -0.821. Several different algorithm settings were also tested, some of which exceeded the coefficient of -0.8. This suggests that the algorithm

is very robust.

Results of comparing with clinicians suggest that the examined parameter could be useful in evaluating the severity of speech fluency disorders. Future work may take as its focus the use of other spectral changes detectors.

Acknowledgement

We would like to thank MUDr. M. Hrbková and MUDr. J. Černý for evaluation speech signals. This work was supported by the research plan "Transdisciplinary research in biomedical engineering" No. MSM6840770012 and grant "Analysis and modelling of biological signals" GACR No. 102/08/H008, and "Speech recognition under Real-World Conditions" GACR No.102/08/0707 and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS10/180/OHK3/2T/13.

References

[1] E. Škodová, I. Jedlička, and colective. *Klinická logopedie*. Portál, Prague, 2003. (in Czech).

[2] V. Lechta. *Poruchy plynulosti řeči*. Scriptorium, Prague, 1999. (in Czech).

[3] P. Howell, A. Hamilton, and A. Kyriacopoulos. Automatic detection of repetitions and prolongations in stutterred speech. *Speech Input/Output: Techniques and Applications*, pages 252–256, 1986.

[4] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W. Suszyński. Automatic detection of disorders in a continuous speech with the hidden markov models approach. *Computer Recognition Systems 2*, vol. 45 of Advaces in Soft Computing:445–453, 2007. Springer, Berlin, Germany.

[5] P. Bergl. *Objektivizace poruch plynulosti řeči*. PhD thesis, Faculty of Eletrical Engineering, CTU in Prague, 2010. (in Czech).

[6] R. Čmejla and P. Sovka. Audio signal segmentation using recursive bayesian change-point detectors. In *WSEAS International Conferences [CD-ROM]*, volume 1, pages 1087–1091. New York : WSEAS Press, 2004.

[7] T. Lustyk. Analýza neplynulé řeči. Master's thesis, Faculty of Eletrical Engineering, CTU in Prague, 2010. (in Czech).

[8] C. Cucchiarini, H. Strik, and Boves L. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *J. Acoustic Soc. Am*, 107:989–999, 2000.

[9] C. Cucchiarini, H. Strik, and Boves L. Quantitative assessment of second language learners' fluency: Comparision between read and spontaneous speech. *J. Acoustic Soc. Am*, 111:2862–2873, 2002.

Tomáš Lustyk
lustytom@fel.cvut.cz

Petr Bergl
berglpet@fel.cvut.cz

Roman Čmejla
cmejla@fel.cvut.cz