Evaluation of disfluent speech by means of automatic acoustic measurements

Tomas Lustyk,[a] Petr Bergl, and Roman Cmejla
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Technicka 2, 166 27, Prague, Czech Republic

(Received 30 May 2013; revised 20 November 2013; accepted 15 January 2014)

An experiment was carried out to determine whether the level of the speech fluency disorder can be estimated by means of automatic acoustic measurements. These measures analyze, for example, the amount of silence in a recording or the number of abrupt spectral changes in a speech signal. All the measures were designed to take into account symptoms of stuttering. In the experiment, 118 audio recordings of read speech by Czech native speakers were employed. The results indicate that the human-made rating of the speech fluency disorder in read speech can be predicted on the basis of automatic measurements. The number of abrupt spectral changes in the speech segments turns out to be the most appropriate measure to describe the overall speech performance. The results also imply that there are measures with good results describing partial symptoms (especially fixed postures without audible airflow). © *2014 Acoustical Society of America.*
[http://dx.doi.org/10.1121/1.4863646]

PACS number(s): 43.70.Dn, 43.70.Kv [SSN] Pages: 1457–1468

I. INTRODUCTION

Stuttering is a chronic speech fluency disorder characterized by "abnormally high frequency and/or duration of stoppages in the forward flow of speech" (Guitar, 2006). The symptoms mainly occur in speech: Repetitions (of sounds, syllables, words, or phrases), prolonged sounds, interjections, revisions, incomplete phrases, and broken words (Bloodstein and Bernstein Ratner, 2008). These symptoms impair the natural fluency of speech production (Conture, 2001). There is also an element of disorder that influences the psychological and social state of a person who stutters (Kalinowski, 2003; Ezrati-Vinacour and Levin, 2004).

Developmental stuttering typically starts between 2 and 7 yr of age with a prevalence of about 5% in preschool children (Yairi and Ambrose, 1999; Mansson, 2000). The symptoms persist into adulthood in approximately 1% of the population. The ratio between females and males is estimated to be 1:3 for 2–10 yr olds, and the ratio does not remain stable with age (1:4 for 11–20 yr olds, 1:2 between 21–49 yr, and 1:1.4 for the population over age 50) (Craig and Tran, 2005; Bloodstein and Bernstein Ratner, 2008).

The diagnosis and evaluation of the severity of a speech disorder are traditionally performed by clinical experts. Several stuttering scales have been introduced, such as the Lidcombe Behavioral Language of Stuttering (Teesson *et al.*, 2003) and the Stuttering Severity Instrument (Riley, 1972), but there has been a need for automatic and objective methods. Such a method would be helpful in diagnosis, the choice of treatment approach, and the evaluation of treatment progress and results (Metz and Samar, 1983; Van Borsel *et al.*, 2003).

The application of acoustical analysis could provide an objective and quantitative intrument to mark the presence of stuttering symptoms and/or describe the severity, characteristics, and progress of the disorder and its treatment (Kent *et al.*, 1999). Studies (Di Simony, 1974; Metz and Samar, 1983; Adams, 1987) have focused on the temporal characteristics of stuttered speech, investigating, for example, vowel duration and voiced stop consonant intervocalic intervals. The rate of speech (manually measured) has been also recognized as a helpful tool for the evaluation of stuttering (Johnson, 1961; Ryan, 1992; de Andrade *et al.*, 2003).

Methods based on digital signal processing may offer insight into stuttered speech. A great effort has been devoted to studying the behavior of formant frequencies, the fundamental frequency, and the voice onset time (VOT). The transition of the second formant frequency has been studied in Yaruss and Conture (1993), formant frequency fluctuation in Robb *et al.* (1998), fundamental frequency and fluent VOT in Healey and Gutkin (1984), fluent VOT and phrase duration in Healey and Ramig (1986), and fundamental frequency, jitter, and shimmer in Hall and Yairi (1992). Computer programs can be efficiently applied to the objective analysis of pathological speech. The computer system multi-dimensional voice program developed by Kay Elemetrics Corp. (Kay Elemetrics Corp., 2003), and the freely available PRAAT (Boersma, 2002), are among these programs and provide several measures for speech evaluation. However, the disadvantage of these programs is mostly the need for user control of the analysis. This can be avoided by using methods that process the entire signal without user control. An approach simply using temporal characteristics to find repetition and prolongation can be seen in Howell *et al.* (1986). Advanced digital signal processing methods have been employed for identifying manually selected stuttered parts of speech: Mel frequency cepstral coefficients in Ravikumar *et al.* (2009) and linear predictive cepstral

[a]Author to whom correspondence should be addressed. Electronic mail: lustytom@fel.cvut.cz

coefficients in Hariharan *et al*. (2012). Hidden Markov models (HMM) have been utilized in Noth *et al*. (2000), Wisniewski *et al*. (2007a), and Wisniewski *et al*. (2007b) to reveal repeated or prolonged parts of disfluent speech. A method does not have to look for symptoms of speech disorder: It could process the signal in another way. Such a method could investigate the energy of the speech signal (speech envelopes) (Kuniszyk-Jozkowiak, 1995, 1996) or could utilize Kohonen networks for the detection of speech nonfluency (Szczurowska *et al*., 2009).

Research on other speech disorders and in different areas of acoustics could supply interesting results and ideas. Maier *et al*. (2011) used automatic methods for evaluation of reading disorder in children's speech where the total reading time is one of the most useful features. Articulation disorder in children with a cleft lip or palate were investigated in Maier *et al*. (2009b), and patients who have had their larynx removed due to cancer and children with a cleft lip or palate in Maier *et al*. (2009a). Study (Godino-Llorente and Gomez-Vilda, 2004) has used short-term cepstral parameters to identify vocal fold impairment due to cancer. Acoustical methods have been applied to non-invasive biomarkers of patients with Parkinson's disease (Sapir *et al*., 2010; Rusz *et al*., 2011). Cucchiarini *et al*. (2000) applied a continuous speech recognizer to the quantitative assessment of second language learners' fluency. Nine automatic measurements based on temporal features of speech, such as the rate of speech, articulation rate, or the total duration of the pauses, have been employed.

The aim of the present study is to investigate whether the level of speech fluency disorder in audio recordings of read speech can be estimated by means of automatic acoustic measurements. Investigating recordings of read text could be a step toward spontaneous speech, which is more common in clinical practice, and clinical experts would appreciate a method that could help with evaluation. The database of recordings from Czech native speakers with different levels of the speech fluency disorder and two different evaluation scales used as expert rating are introduced in Sec. II. Four automatic acoustic measurements are described in the same section. The reliability of the evaluation and the results of a comparison between four automatic measurements and the expert ratings are given in Sec. III, along with an additional feature, the total reading time for comparison with other studies. Section IV provides a discussion of our main findings. Finally, Sec. V concludes the article with a short summary.

II. METHOD

The methodology of this study is divided into four stages: (1) participants and speech data, (2) the rating of the speech recordings, (3) algorithms of the automatic measurements, and (4) statistics.

A. Participants and speech data

The speech signal database was created in the past few years at the Department of Phoniatrics of the First Faculty of Medicine at the Charles University and the General Faculty Hospital in Prague. The database contains recordings of 118 Czech native speakers (28 women and 90 men) with different ages and levels of speech fluency disorder. The age structure of the whole database is as follows: Mean age 18.1 yr [± standard deviation (SD), 9.9 yr], the youngest participant was 8 yr old, the oldest was 50 yr old. Fifteen recordings (5 women and 10 men) are utterances of speakers without speech fluency disorder [mean age 27.37 yr [±SD, 7.4 yr)] and speakers with speech fluency disorder were in age, mean 16.73 yr (±SD, 9.4 yr). All participants read the standard text used by Czech speech therapists, the text is about 70 word-long, it is phonetically non-balanced, and it does not include tongue twisters. The average length of a recording is 66.1 s (±SD, 33.3 s).

The utterances were recorded with a sampling frequency of 44 kHz. The signals were down-sampled to 16 kHz for the subsequent analysis.

B. The rating of the speech recordings

To verify the suitability of a measure, reliable expert rating is necessary (Cordes and Ingham, 1994). Two different evaluation scales were used in the experiment. The first is the modified Kondas's scale, which is a system used by Czech speech therapists for rating stuttering (Lechta, 2004). The scale consists of five stages (from 0 to 4): 0 is normal healthy speech (without frequent signs of disfluency), 1 is mild disfluency (up to 5% stuttered words), 2 is moderate disfluency (6%–20% disfluent words), 3 is severe disfluency (20%–60% disfluent words), and 4 is very severe disfluency (more than 60% disfluent words). The evaluation was performed by two professional speech pathologists using the Kondas's scale. They evaluated recordings independently, and each participant got score according to her/his performance and the best knowledge of the evaluator. Then the judgment of both therapists was merged for further evaluation. In case their ratings differed (for example, if the first assigned the level 2 and the second 3), the higher level was adopted. The structure of speech fluency disorder according to the modified Kondas's scale (merged judgment of two therapists) is as follows: The groups of 0–4 include 15, 24, 41, 31, and seven recordings.

To have more insight and information about the extent of speech disfluencies, the second set of expert ratings was produced by means of the Lidcombe Behavioral Data Language of Stuttering (LBDL) (Teesson *et al*., 2003), which is a behaviorally based taxonomy of stuttering. The LBDL was developed to be both valid and reliable. It can be used to describe stuttering across all ages. The LBDL considers seven descriptors of stuttering symptoms: Syllable repetition (SR), incomplete syllable repetition (ISR), multisyllable unit repetition (MSUR), fixed posture with audible airflow (FPWAA), fixed posture without audible airflow (FPWOAA), superfluous verbal behaviors (SVB), and superfluous nonverbal behavior (SNB). All the categories are detectable in a speech signal except the descriptor SNB, which should be looked for in video recordings. The descriptor SNB is not used in these experiments. The descriptors *overall* (all descriptors except SNB), *repeated*

(SR + ISR + MSUR), and *fixed* (FPWAA + FPWOAA) are also considered in this paper.

One evaluator listened to the recordings and for each one wrote down the number of all symptom occurrences when evaluating by means of the LBDL. The final score was computed as the number of occurrences divided by the number of all words in the recording.

The reasons why this taxonomy was adopted in the experiment are: First, the results are valid and reliable when the system is used by experienced judges (Teesson *et al.*, 2003). Second, the taxonomy is easy to use. Also, when using the LBDL with all its categories, the measures that fit the most for particular descriptors can be found. An example of using the LBDL in a quite similar problem can be seen in the research on stuttering symptoms in Parkinson's disease (Goberman *et al.*, 2010).

C. Automatic measurements

This stage is dedicated to four automatic measurements designed to measure the level of the speech fluency disorder. Table I lists all the measures. For each measure, the table describes how it works, which symptoms it covers or takes into account, and which assumptions were made.

1. The average length of silence (ALS)

Subjects with stuttering have more silences and pauses than do healthy subjects. Thus it can be assumed that the higher level of fluency disorder results in higher amounts of silence in the speech signal.

A voice activity detector (VAD) is employed to split the signal into speech and silent parts. The VAD based on the Mel frequency filter bank was used. The first step in the procedure is the estimation of power spectra (computed by Welch's method) followed by application of the triangular Mel-frequency filter bank. Then the decision about the speech activity in each frequency band by means of an adaptive threshold is made. The last step of computing speech activity is the final decision about the speech activity (speech/silence) in the whole frequency band.

The average length of the silent parts can be simply calculated when knowing placement of speech/silence parts,

$$\mathrm{ALS} = \frac{1}{N_{SIL}} \sum_{i=1}^{N_{SIL}} T_{\mathrm{SILENCE}}(i), \tag{1}$$

where $T_{\mathrm{SILENCE}}(i)$ is the duration in seconds of the ith segment of silence and N_{SIL} is the number of segments of silence, see Fig. 1, where in part (a), the speech signal and, in (b), the detected voice activity are depicted. The final value of the ALS is modified by summing up with 1 (to avoid the situation when the ALS equals 0) and then using the logarithm.

To make the difference between fluent and disfluent speech more visible, an innovative procedure was developed. Short speech parts, such as repetition, superfluous verbal behavior, and parts of incorrectly pronounced words, could be removed by this method and the amount of silence increased in a speech signal with such disfluencies. The procedure uses successive removals of short segments of speech and silence. First, the speech segments shorter than 125 ms are removed (silent segments shorter than 125/2 ms are removed at the same time), next, there follows 150 ms, and this process continues up to the value 5000 ms. The procedure is depicted in Figs. 1(b) and 1(c) for demonstration. In this example, the ALS (without taking the logarithm and adding 1) is approximately 0.2 s [in Fig. 1(b)], and after removing intervals shorter than 150 ms, the ALS rises to 1 s [Fig. 1(c)]. The results presented in the results section are for the time limit values from 125 to 1500 ms.

2. The extent of speech fluency (ESF)

The speech rate has been found to be an important indicator of speech fluency (Johnson, 1961; Ryan, 1992; de Andrade *et al.*, 2003), and there have been experiments to measure the rate of speech automatically (Cucchiarini *et al.*, 2000). The measure ESF is very close to the speech rate. Abrupt spectral changes (AC) correspond to phoneme boundaries or transitions from speech to silence (and vice versa) in a speech signal. They can be captured by a detector of abrupt changes in the spectrum.

The Bayesian autoregressive changepoint detector (BACD) is employed to identify spectral changes in this study, and all the remaining measures are based on its use. The detector is based on the analytical solution of the changepoint problem between two autoregressive models (Ruanaidh and Fitzgerald, 1996) and assumes that a speech signal can be represented by an autoregressive (AR) model of a certain order. Then the identification of abrupt spectral changes is accomplished by detecting changes in that AR model. The algorithm of the sliding window is applied to the signal (Cmejla and Sovka, 2004) or more in depth (Cmejla *et al.*, 2013). This window is shifted through the signal, and the unknown changepoint is considered to be in the middle of the window (between the right and left parts of the window). The probability that the change in the AR model lies

TABLE I. List of the measures used as indicators of speech fluency.

Measure	Symptoms	Description
The average length of silence (ALS)	Overall performance, pauses	Average duration of silent parts identified by voice activity detector
The extent of speech fluency (ESF)	Overall performance	Number of abrupt spectral changes
The average number of spectral changes in short intervals (SCSI)	Pauses	Average number of abrupt spectral changes in short time windows
The number of spectral changes in speech segments (NSI)	Pauses, overall performance	Number of abrupt spectral changes included in speech segments

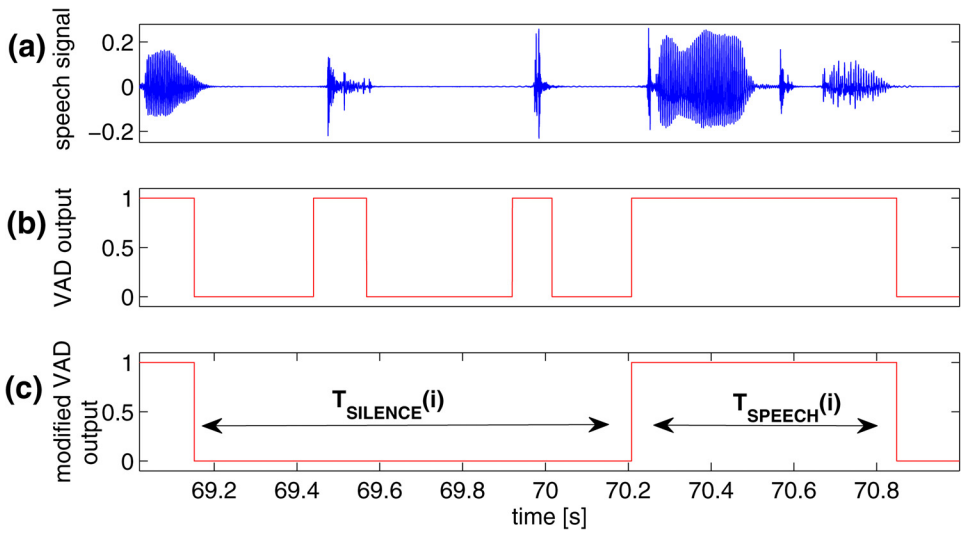

FIG. 1. (Color online) Steps of the calculation of the ALS. (a) Speech signal, repetition of the Czech word "k k kvitky" (flowers). (b) VAD detection. (c) Modified VAD output after successive removing speech segments shorter than 150 ms and silences shorter than 150/2 ms.

in the middle of the window is computed in each step. This procedure continues until the end of the signal is reached with one sample step. The result is a series of probabilities. High values of these probabilities should refer to abrupt spectral changepoints, see Fig. 2: The speech signal is in Fig. 2(a) and its BACD output curve in Fig. 2(b). The higher the probability is, the larger is the abrupt change. But in our case, the size of the change does not play a key role [as it does in articulation problems of patients with Parkinson's disease (Rusz *et al.*, 2011)]. The placement and distance of the maxima are more important for the analysis of disfluent speech.

Because the BACD output curve includes both significant and less significant abrupt changes, the following procedure is needed. The output of the BACD is filtered by a low-pass filter with the cutoff frequency at 20 Hz (to smooth the BACD output curve). The local minima are calculated in the smoothed output curve; thereafter the local maxima are found in the appropriate segments (between two local minima). Many of those local maxima do not correspond to significant spectral changes (phoneme boundaries), and they should be excluded. A threshold is utilized to separate these maxima [Fig. 2(b)]. Then the significant abrupt changes are obtained [Fig. 2(c)], their number is determined, and the ESF is calculated by the formula

$$\mathrm{ESF} = \frac{\sum_{i=1}^{N_{AC}} AC(i)}{T_{\mathrm{SIGNAL}}}, \tag{2}$$

where $AC(i)$ is an abrupt spectral change, N_{AC} is the number of abrupt spectral changes, and T_{SIGNAL} is the length of the speech signal in seconds. For example, in the figure there are 27 abrupt changes and the duration of this part of signal is 3 s, so the ESF of this fluently pronounced speech is 9.

The analysis, carried out on the detector outputs from different participants, showed that we are not able to use one threshold for the entire database. Hence a method of adaptive threshold extraction for each signal was used. The threshold is determined as a fraction of the kth highest maxima. Several algorithm settings were tested: From 1 to 9 for k and from 0.1 to 0.3 for the multiplication constant, and their results are shown in Sec. III in comparison to speech specialist evaluation. The best setting ($k = 4$ and 0.15 for the multiple) was established experimentally by comparing with the expert rating. The BACD of a sixth order AR model with

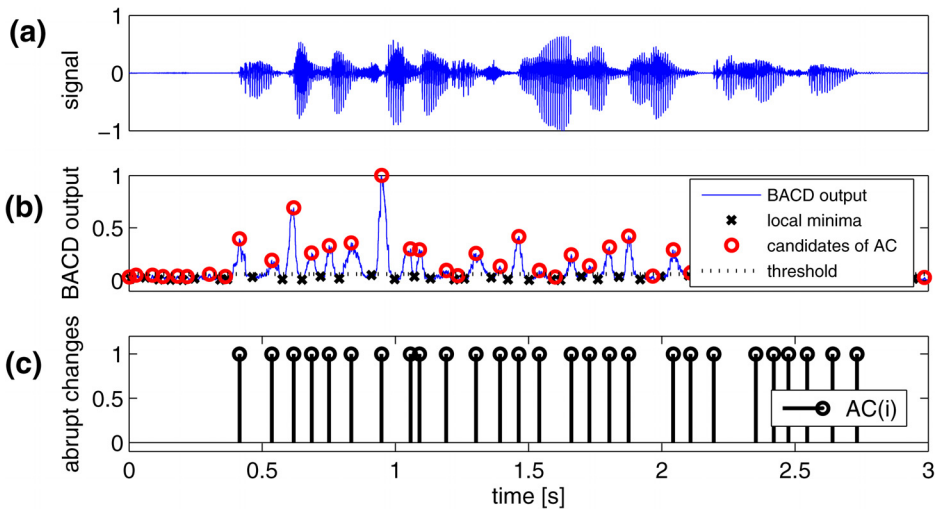

FIG. 2. (Color online) Identifying abrupt spectral changes. (a) Speech signal, Czech sentence "ozdoba sadu uschovana byla v komore" (garden adornment was kept in the pantry). (b) BACD output, local minima, and candidates of abrupt spectral changes. (c) Abrupt spectral changes.

a window length 60 ms was used in the whole experiment (Cmejla *et al.*, 2013; Bergl, 2010).

3. The average number of spectral changes in short intervals (SCSI)

In Sec. II C 1, it was mentioned that disfluent speech consists of many prolongations, frequent pauses, and broken words. The SCSI tries to capture these phenomena by processing the BACD output in short windows. If the output of the BACD is processed in short segments, the difference in the number of abrupt changes could be significant for segments with speech activity as opposed to segments with silence, taking into account the comparison of disfluent speech to healthy speech. For participants with disfluencies, it is expected that more silence appears in stuttered than in fluent speech (the average number of changes in the window is smaller). The number varies and in many cases is zero. Conversely, the number of changes for healthy speakers is more stable and the average should be higher.

The procedure of analysis using the average number of BACD changes in short interval as a parameter begins with identification of significant abrupt spectral changes. It is followed by the processing of the detector output in a short window. The number of spectral changes is found in each window, and the average number of abrupt spectral changes in the windows is quantified. The logarithm is used for the final value. An example of this calculation can be seen in Fig. 3, where the value of the logarithm of the SCSI is 0.84 (it is a part of a disfluent speech signal with severe disfluency) and the window length is 2 s.

The tested window lengths were 1, 2, and 4 s with half-overlap and all used window lengths reached very similar results. The studied BACD settings were as for the ESF. The window length 2 s and all settings of BACD are presented in Sec. III.

4. The number of spectral changes in speech intervals (NSI)

This measure makes the same assumptions as the measure ESF. It uses the VAD in combination with the BACD in addition to the ESF.

The beginning of the procedure is the same as in the previous BACD algorithms up to the point of identifying the relevant spectral changes at which point one then implements the following step: Applying the VAD and spectral changes in speech segments are identified.

The number of spectral changes in the ith speech segment $N_{\mathrm{AC}_{\mathrm{speech}}}(i)$ is determined, and finally the number of spectral changes in all speech segments is summed up and divided by the length of the speech signal T_{SIGNAL} in seconds. See Fig. 4 for a short demonstration of the method. When N_{speech} is the number of speech segments, the measure NSI can be written as follows:

$$\mathrm{NSI} = \log_{10} \frac{\sum_{i}^{N_{\mathrm{speech}}} N_{\mathrm{AC}_{\mathrm{speech}}}(i)}{T_{\mathrm{SIGNAL}}}. \tag{3}$$

There is also an additional step to this procedure, as at the measurement of the ALS, it is the successive removal of short speech segments that increases the difference between fluent and disfluent speech. The example in the figure shows a short part of the speech signal "Podzim na starem belidle" (autumn at the old bleachery) where the value of the NSI would be 0.43. All tested settings of the BACD are shown in Sec. III; the time limit for removing short speech and silence segments is 1000 ms based on the ALS algorithm results.

D. Statistics

The ability to recognize levels of speech disfluency was examined using the Pearson product-moment correlation, the classification with the linear discriminant analysis (LDA), and the statistical method analysis of variance (ANOVA) with *post hoc* Bonferroni adjustment. First, all settings of each algorithm are examined by means of correlations and deviations with respect to the expert ratings. Second, the ANOVA analysis is performed for one selected setting of each of the four acoustic measures to find significant differences between fluency levels. Then the relationship between the acoustic measures and all categories of the LBDL is evaluated by the Pearson product-moment correlation. The Kolmogorov–Smirnov test was used to examine the normality of the distribution of the data.

To demonstrate how the algorithms are able to separate all subjects into disfluency levels, the LDA is used. The

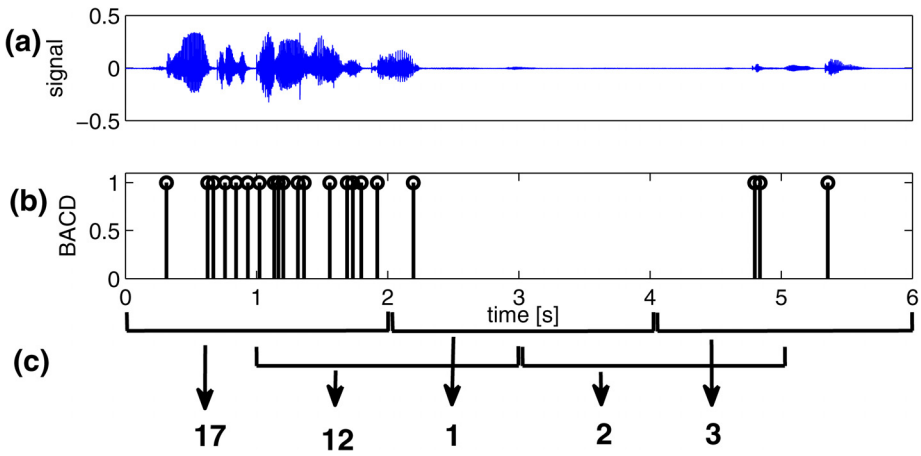

FIG. 3. (Color online) Procedure for calculating the average number of BACD changes in a short interval. (a) Speech signal, a part of the Czech sentence "chomace stareho listi buh vi kam" (bunch of old leaves God knows where). (b) Output of Bayesian detector. (c) Processing by means of a window with marked number of abrupt spectral changes.

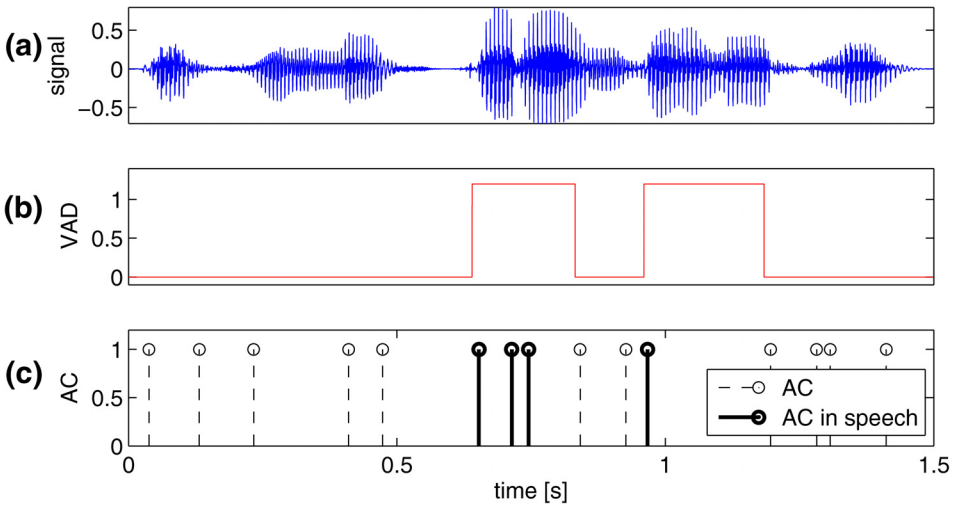

FIG. 4. (Color online) Calculation of the number of spectral changes in speech segments (NSI). (a) Speech signal. (b) Speech activity output with successive removed speech and silence parts (1, speech activity; 0, silence/pause). (c) Abrupt changes (dashed line) and ACs included in speech segments (thick line).

LDA (Harrington and Cassidy, 1999), a statistical technique, takes the knowledge that an element from training data set belongs to a certain group/level. On the basis of the elements' mean and standard deviation, the discriminant function is determined for each group from training data set. These discriminant functions could be then used for classification of a new element. Because the number of participants in the experiment is rather lower, especially in peripheral levels 0 and 4, we decided to perform the leave-one-out cross-validation instead of dividing the database into test and validation group. When using this method, all elements of the data set except one serve as the training set, and the one element is used as the validation data. This is repeated for each element of the data set, thus each element is used as the validation data.

The deviation Δ is defined to assess the success of classification,

$$\Delta = \sum_{i=1}^{N} (|o_i - \hat{o}_i|), \tag{4}$$

where o_i is merged evaluation of speech specialists for the ith subject in the database, $\hat{o}_i$ represents the estimated level for the same subject, the difference o_i - $\hat{o}_i$ represents the classification error for one subject, and N is the number of subject in the database. When inspecting the results, we can follow the theorem, the smaller deviation Δ, the better result of classification achieved.

III. RESULTS

A. Reliability of the stuttering rating

All the read speech recordings were evaluated by two evaluators using the modified Kondas's scale and by one evaluator using the LBDL. As appears from the Pearson correlation and Cronbach's alpha, the expert rating (the Kondas's scale) shows a very high relationship between both therapists, a correlation of 0.91 ($p < 0.001$) and Cronbach's alpha 0.95. The evaluation made by the speech therapists was also compared to the subjective evaluation made by means of the LBDL (overall), and here the Pearson correlation coefficient was 0.93 ($p < 0.001$) for the first expert, 0.92 ($p < 0.001$) for the second, and for the merged evaluation: 0.93 ($p < 0.001$). The logarithm of the LBDL evaluation was used because the Kondas's scale is rather logarithmic (Cmejla *et al.*, 2013).

The LBDL reports a high level of agreement in describing stuttering events and shows consistent results for intra- and inter-judge agreement (Teesson *et al.*, 2003). Thirty recordings (20% of the 118 recordings) were assessed twice to obtain the intra-judge agreement; the same 30 recordings were assessed by the second evaluator to obtain the inter-judge agreement, similar to Goberman *et al.* (2010). The lowest correlation coefficients across all descriptors for intra-judge agreement were 0.87 ($p < 0.001$) and 0.89 ($p < 0.001$), the others exceeded 0.94 ($p < 0.001$). The results for inter-judge agreement also seem to show good results (>0.76) except for descriptor superfluous verbal behaviors (SVB). The agreement for descriptor SVB was 0.32 ($p = 0.08$). These events are important but not as much so as the other events (repetition, prolongation, pauses), therefore we decided to use this evaluation but the results related to the descriptor SVB are viewed carefully.

The results of the intra- and inter-judge reliability in the experiment achieved a high level of agreement, and we can conclude that the evaluation is reliable and applicable for the purpose of this experiment.

B. Automatic measurements for estimation of the speech fluency disorder level

To establish whether automatic measurements can be successfully used as an indicator of disfluency/fluency in read speech, the following methods are used: The correlation between the automatic measurements and the stuttering rating, the classification using LDA with the leave-one-out cross-validation, and the ANOVA analysis. First, we present correlations and the deviation for all settings of four algorithms to the speech specialist rating (the Kondas's scale). The scope of investigation is to find the most appropriate setting of the algorithm for description of the level of the speech fluency disorder. Second, the typical range of values of each measure (according to the levels of speech fluency)

are shown, followed by the results of the statistical analysis ANOVA, and finally, the correlations between the measures and the LBDL stuttering scale. For a little comparison with other studies, the results of the feature the total reading time RT (the duration of the recording in seconds) are also displayed.

The correlation and results of classification represented by the deviation Δ [Eq. (4)] can be viewed in Tables II–V for algorithms ALS, ESF, SCSI, and NSI, respectively.

The measures reveal very good agreement with speech specialist evaluation as can be seen in tables. The ALS measure based on speech and silence segments recognition yielded the best correlation 0.64, the correlation rises with increasing time limit (the scope of setting) achieving its top at 1000 ms followed by a slow decline. The trend of classification deviation is opposite; the deviation decreases with rising time limit (the classification is more successful), one of the local minima is reached at the time limit 1000 ms $\Delta = 89$, but the smallest deviation is 82 at the time limit 1500 ms (for details, see the Table II). Many of the BACD based algorithms' settings reached correlation −0.75, with the highest correlation, −0.78, for the NSI algorithm, and −0.77 for the ESF and the SCSI; that could be a good sign of the algorithms' robustness. Deviation of the classified data from the expert rating are reduced in comparison to the ALS, the smallest is 72, 75, and 64 for the ESF, SCSI, and NSI, respectively. Thus the classification was more efficient. The comparative measure RT (total reading time) correlates with expert fluency rating with coefficient of 0.77 and the deviation $\Delta = 64$.

According to the results of correlations and classification, the following settings were chosen for further detailed analysis. Those are highlighted in bold in tables: The time limit 1000 ms for the ALS; $k = 4$ (fourth highest maximum) and multiplication constant 0.15 for the ESF; $k = 6$, multiplication constant 0.15, and the window length 2 s for the

TABLE II. The Pearson correlation and results of classification using the LDA (the deviation Δ from specialist evaluation) for all algorithm settings of the ALS in comparison to the merged evaluation of both speech specialists. The time limit for successive removing of short speech and silence segments is the subject of setting.

Settings (ms)	Correlation (deviation Δ)
125	0.35 (172)
150	0.34 (153)
200	0.36 (146)
300	0.40 (123)
400	0.46 (124)
500	0.50 (121)
700	0.56 (109)
800	0.59 (105)
900	0.62 (94)
1000	0.64 (89)
1100	0.64 (91)
1200	0.62 (98)
1300	0.62 (102)
1400	0.62 (112)
1500	0.62 (82)

TABLE III. The Pearson correlation and results of classification using the LDA (the deviation Δ from specialist evaluation) for all algorithm settings of the ESF in comparison to the merged evaluation of both speech specialists. To set the algorithm, the *k*th highest maximum and multiplication constant are used.

	Multiplication constant, correlation (deviation Δ)				
k	0.10	0.15	0.20	0.25	0.30
1	−0.75 (85)	−0.73 (95)	−0.67 (98)	−0.61 (105)	−0.53 (126)
2	−0.76 (80)	−0.76 (82)	−0.72 (96)	−0.65 (102)	−0.60 (105)
3	−0.75 (74)	−0.76 (79)	−0.72 (102)	−0.66 (120)	−0.61 (135)
4	−0.75 (79)	**−0.77 (74)**	−0.73 (95)	−0.68 (112)	−0.62 (127)
5	−0.74 (74)	−0.76 (72)	−0.74 (92)	−0.69 (111)	−0.65 (121)
6	−0.74 (87)	−0.77 (74)	−0.75 (87)	−0.70 (112)	−0.66 (125)
7	−0.72 (97)	−0.77 (73)	−0.75 (86)	−0.71 (106)	−0.67 (121)
8	−0.72 (93)	−0.76 (75)	−0.76 (82)	−0.73 (106)	−0.68 (121)
9	−0.72 (83)	−0.76 (75)	−0.76 (82)	−0.74 (101)	−0.70 (108)

TABLE IV. The Pearson correlation and results of classification using the LDA (the deviation Δ from specialist evaluation) for all algorithm settings of the SCSI in comparison to the merged evaluation of both speech specialists. To set the algorithm, the *k*th highest maximum and multiplication constant are used, the window for processing is 2 s.

	Multiplication constant, correlation (deviation Δ)				
k	0.10	0.15	0.20	0.25	0.30
1	−0.75 (82)	−0.73 (88)	−0.68 (112)	−0.63 (120)	−0.58 (120)
2	−0.76 (79)	−0.76 (80)	−0.72 (93)	−0.67 (99)	−0.62 (99)
3	−0.75 (80)	−0.75 (83)	−0.72 (105)	−0.67 (120)	−0.62 (134)
4	−0.75 (76)	−0.77 (77)	−0.73 (101)	−0.69 (119)	−0.64 (123)
5	−0.74 (75)	−0.77 (76)	−0.74 (92)	−0.70 (111)	−0.66 (122)
6	−0.74 (87)	**−0.77 (75)**	−0.75 (92)	−0.71 (110)	−0.67 (123)
7	−0.72 (100)	−0.76 (77)	−0.75 (89)	−0.71 (114)	−0.67 (118)
8	−0.72 (97)	−0.77 (75)	−0.76 (89)	−0.73 (106)	−0.68 (118)
9	−0.71 (104)	−0.77 (78)	−0.76 (82)	−0.73 (105)	−0.69 (111)

TABLE V. The Pearson correlation and results of classification using the LDA (the deviation Δ from specialist evaluation) for all algorithm settings of the NSI in comparison to the merged evaluation of both speech specialists. To set the algorithm, the *k*th highest maximum and multiplication constant are used, the time limit for successive removing of short speech and silence segments is set to 1000 ms.

	Multiplication constant, correlation (deviation Δ)				
k	0.10	0.15	0.20	0.25	0.30
1	−0.77 (64)	−0.77 (72)	−0.74 (82)	−0.70 (89)	−0.66 (95)
2	−0.77 (65)	−0.78 (68)	−0.76 (69)	−0.73 (82)	−0.69 (93)
3	−0.76 (70)	−0.77 (68)	−0.76 (76)	−0.73 (87)	−0.69 (94)
4	−0.77 (72)	−0.78 (70)	−0.77 (70)	−0.74 (80)	−0.71 (93)
5	−0.76 (68)	−0.77 (70)	−0.77 (73)	−0.74 (80)	−0.72 (91)
6	−0.76 (69)	**−0.78 (64)**	−0.77 (75)	−0.75 (78)	−0.73 (93)
7	−0.76 (71)	−0.77 (66)	−0.77 (75)	−0.75 (82)	−0.73 (95)
8	−0.76 (70)	−0.77 (66)	−0.78 (74)	−0.76 (76)	−0.74 (87)
9	−0.76 (68)	−0.77 (64)	−0.78 (73)	−0.77 (77)	−0.74 (84)

TABLE VI. The mean $\bar{x}$ and standard deviation SD of fluency measures and statistical significance by means of the ANOVA analysis with comparison between levels by the *post hoc* Bonferroni adjustment.

	ALS		ESF		SCSI		NSI		RT	
	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Normal healthy speech (0)	0.19	0.06	7.46	1.14	1.23	0.08	0.87	0.06	34.8	8.0
Mild disfluency (1)	0.21	0.05	6.60	1.34	1.17	0.08	0.81	0.09	43.9	11.2
Moderate disfluency (2)	0.26	0.10	5.09	1.22	1.06	0.09	0.67	0.12	58.2	14.2
Severe disfluency (3)	0.38	0.11	3.79	0.80	0.95	0.09	0.51	0.13	92.0	25.7
Very severe disfluency (4)	0.60	0.26	3.32	0.41	0.90	0.04	0.28	0.22	140.0	40.3
					Comparison between the levels					
ANOVA F(4, 117)		27.97*		42.84*		43.33*		48.89*		60.08*
0 vs 1		NS		NS		NS		NS		NS
1 vs 2		NS		$p<0.001$		$p<0.001$		$p<0.001$		$p<0.05$
2 vs 3		$p<0.001$		$p<0.001$		$p<0.001$		$p<0.001$		$p<0.001$
3 vs 4		$p<0.001$		NS		NS		$p<0.001$		$p<0.001$

NS = not significant; *$p<0.001$

SCSI; $k=6$, multiplication constant 0.15, and time limit for removal of short speech and silence segments 1000 ms.

Table VI shows the typical range of the algorithms' values (for selected settings), the mean value $\bar{x}$ and the standard deviation SD of the measures according to the level of the speech fluency scale. The results of the ANOVA analysis are attached at the bottom part of the table. It can be seen that the measures ALS and RT increase with the level of the speech fluency disorder. On the other hand, the measures ESF, SCSI, and NSI decrease with the growing level of the disorder.

The best results in the ANOVA analysis with the *post hoc* Bonferroni adjustment were achieved by the NSI algorithm. It is able to find significant differences ($p<0.001$) between levels 1 vs 2 vs 3 vs 4 (all levels except 0 vs 1). The measures ESF and SCSI recognized differences ($p<0.001$) between the mild and moderate levels of disfluency (1 vs 2), and between moderate and severe disfluency (2 vs 3). Also, the ALS measures found significant differences between levels 2 vs 3 vs 4 ($p<0.001$). The comparative measure RT identified significant differences among the moderate, severe, and very severe levels of disfluency (2 vs 3 vs 4) with $p<0.001$, and between mild and moderate disfluency (1 vs 2) with $p<0.05$. No measure found a statistically significant difference between the normal healthy level of speech and mild disfluency (0 vs 1).

Figures 5 and 6 depict the output values of two characteristics in comparison to the LBDL evaluation to show the range of values and dependency of the characteristics on the level of the speech fluency disorder. Figure 5 includes the values of the ALS characteristic and the *overall* LBDL descriptor (the measures increases with the level of the disorder, the correlation coefficient is 0.68). Figure 6 is for the NSI characteristic in comparison to the FPWOAA (pauses) characteristic (the measure decreases with the level of the speech fluency disorder, the correlation coefficient is −0.82).

Table VII gives detailed results for the Pearson product-moment correlation for selected settings of described measures in comparison to all categories of the LBDL. The algorithms achieved the best results for *overall* grade (summary descriptor), *fixed* postures (summary descriptor for prolongations and pauses), and the FPWOAA descriptor (fixed postures with- out audible airflow, i.e., pauses). The magnitude of correlation coefficients exceeded 0.70 in some cases: The highest were for the measure NSI versus FPWOAA (0.84), *fixed* (0.85), and *overall* (0.82) descriptors. The comparative measure RT achieved correlation equivalent to those introduced, 0.68, 0.75, 0.74, and 0.86 for categories FPWOAA, repeated, fixed, and overall, respectively.

Higher correlation coefficients were also obtained for summary descriptor *repeated* values for some automatic measurements: About 0.65, for the measures ESF, SCSI, and NSI. The consensus between automatic measurements and individual repeated characteristic (SR, ISR, and MSUR) is rather moderate. The smallest agreement between automatic measures and individual characteristic of the LBDL is when considering the category SVB.

To summarize these partial results: The automatic measurements are able to indicate with a very good agreement the characteristics overall, fixed, and FPWOAA (the measure

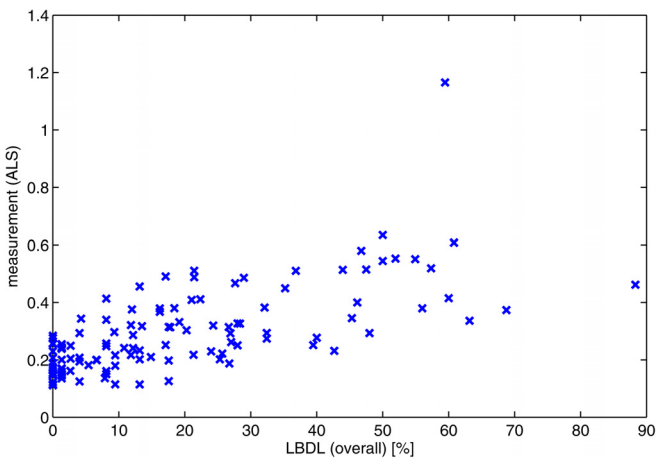

FIG. 5. (Color online) The comparison of the ALS to the subjective rating (*overall* score). The measure increases with the level of the speech fluency disorder.

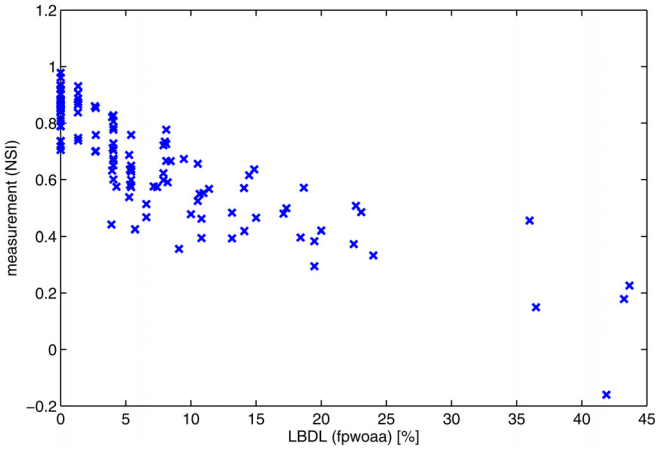

FIG. 6. (Color online) The comparison of the NSI to the subjective rating (the FPWOAA characteristic). The measure decreases with the level of the speech fluency disorder.

NSI demonstrates the best results, the other measures based on the BACD also show high correlations, and the results of the ALS measure could be considered as good).

An interesting issue would be the cross-correlation between all measures, which are given in Table VIII. It is obvious that some of the automatic measures are highly correlated with each other, but there are exceptions. The characteristics ESF, SCSI, and NSI (all based on the BACD) are cross-correlated with coefficients >0.88. Some of the cross-correlation coefficients exceed 0.9. There is also a stronger relationship between the measure ALS (based on the VAD) and the BACD-based measure NSI (0.85). The measures ESF and SCSI report correlation with the ALS about 0.60. We can consider that there might be a possibility of combining several measures into one with better results in the case of smaller correlations between them.

A small experiment was carried out with a combination of all measures (ALS, ESF, SCSI, and NSI). The procedure was very simple. First, normalization of the measured values between 0 and 1 was done, then the normalized values were summed up, and these values were compared to the fluency rating. Simply combining the measures this way achieved a Pearson correlation coefficient of 0.82 with the overall characteristic (LBDL) and 0.80 with the speech therapists using the Kondas's scale.

TABLE VII. The Pearson correlation coefficients and the levels of significance (in parentheses when $p > 0.001$) for one selected setting of each measure in comparison to the LBDL descriptors and the merged evaluation of speech pathologists.

	Measure				
Descriptor	ALS	ESF	SCSI	NSI	RT
SR	0.38	−0.49	−0.48	−0.48	0.54
ISR	0.44	−0.51	−0.54	−0.52	0.65
MSUR	0.28	−0.54	−0.57	−0.50	0.60
FPWAA	0.25	−0.46	−0.48	−0.38	0.49
FPWOAA	0.73	−0.67	−0.72	−0.84	0.68
SVB	0.28	−0.31	−0.32	−0.29	0.60
Repeated	0.49	−0.63	−0.65	−0.63	0.75
Fixed	0.72	−0.73	−0.78	−0.85	0.74
Overall	0.68	−0.76	−0.80	−0.82	0.86
Specialists (merged)	0.64	−0.77	−0.77	−0.78	0.77

TABLE VIII. Correlations among all automatic speech measures.

	Measure			
Measure	ESF	SCSI	NSI	RT
ALS	−0.58	−0.61	−0.85	0.69
ESF		0.99	0.88	−0.77
SCSI			0.90	−0.80
NSI				−0.81

IV. DISCUSSION

The study presents four automatic and objective measures applied to the analysis of audio recordings of stutterers. The measures are based on the voice activity and detection of abrupt spectral changes. The main goal is to find out whether these automatic measurements are able to estimate the level of the speech fluency disorder in read speech.

The expert ratings are very important when comparing automatic measurements to subjective assessments. To have more information about the extent of the speech fluency disorder, two different evaluation scales were applied: The first is the modified Kondas's scale (Lechta, 2004) and the second is the LBDL taxonomy (Teesson *et al.*, 2003). All 118 audio recordings of read speech were evaluated by two experienced phoniatric experts using the Kondas's scale. The Pearson correlation coefficient and Cronbach's alpha showed a very high relationship between both speech therapists. The second subjective evaluation was made by one evaluator who assessed all recordings by means of the LBDL taxonomy. The evaluation of 30 recordings for the second time and by another judge was used for intra- and inter-judge reliability. The same procedure was used in Goberman *et al.* (2010). The Pearson correlation coefficient showed a strong agreement between the original and the repeated evaluation using the LBDL, which is consistent with Teesson *et al.* (2003) and Goberman *et al.* (2010), where very high intra-judge agreement was achieved. When we consult the inter-judge agreement, the lowest correlation (0.32) was found for superfluous verbal behaviors; the other categories of the LBDL report significant positive correlations. Because of the low correlation of the characteristic superfluous verbal behaviors, the results dealing with this characteristic are viewed carefully. When comparing the individual or merged evaluations by experts (Kondas's scale) and the descriptor overall of the LBDL, the conclusion can be adopted that these two evaluations report very strong relationships (the Pearson correlations for the individual experts and the merged evaluation with the LBDL surpasses 0.9), these results of assessment suggests that the expert ratings are reliable and useful for the purposes of this experiment.

Our main findings dealing with automatic measurements of audio recordings for the evaluation of speech disfluency can be expressed as follows. First, the measures are able to

indicate the overall level of the speech fluency disorder (at least in read speech). This finding is supported by the results where three of four measures have magnitudes of the correlation coefficient with two experienced speech pathologists higher than 0.77 and with the LBDL evaluation *overall* score exceeding 0.76 (the highest 0.82). The comparative measure total reading time achieved very similar correlation (0.77 for speech experts); it surpasses introduced algorithms when looking at the overall LBDL score (correlation of 0.86). The correlation are supported by results of classification using the linear discriminant analysis with the leave-one-out cross-validation when the selected setting of the NSI algorithm classified 61 subjects (52%) into the correct level of the Kondas's scale, 50 subjects (42%) with the classification error 1 (the estimated level by algorithm differs by one level from the subjective evaluation), and seven participants (6%) with classification error 2; the total deviation from the speech therapists evaluation is 64. For comparison, the total reading time classified 59 subjects correctly (50%), 54 subjects with the classification error 1 (46%), and five subjects (4%) with the classification error 2 (the total deviation from subjective evaluation is 64). Both measures show very similar results. The algorithms ALS, SCSI, NSI, and also the comparative measure total reading time tend to assign rather lower levels of the speech disorder than the speech therapists, the ESF algorithm does the opposite. Assessment of group differences confirms that the measure NSI is able to find statistically significant differences ($p < 0.001$) between the groups mild and moderate, moderate and severe, and severe and very severe. The measures ALS, ESF, and SCSI can separate one group less. In comparison, the total reading time can differentiate levels moderate, severe, very severe ($p < 0.001$), and mild and moderate ($p < 0.05$). A major problem is distinguishing between normal fluent speech and mild disfluencies: No measure is able to recognize a statistically significant difference here (the similar phenomenon can be observed in classification). This is probably caused by the definition of the levels of the modified Kondas's scale, where the level 0 (normal healthy speech—without frequent signs of disfluency) and the level 1 (mild disfluency, up to 5% disfluent words) are very close. These two groups often overlap because normal fluent speakers usually exhibit some signs of disfluencies (Johnson, 1961; Yairi and Clifton, 1972; Goberman *et al.*, 2010), and it is difficult to recognize the difference (Onslow *et al.*, 1992).

Second, some measures are able to describe individual or summary characteristics of the LBDL. The best results can be found for the fixed postures without audible airflow: Three measures achieved a Pearson product-moment correlation higher than 0.7 in magnitude (the highest was 0.84 for the measure NSI). This finding suggests that a large part of the fluency evaluation in read speech may lie in the pauses, which is in line with Cucchiarini *et al.* (2000). Also Noth *et al.* (2000) found pauses very important for automatic evaluation of stuttered speech. This finding led us to examine the cross-correlations between all characteristics of the LBDL and a strong relationship between overall and fixed postures without audible airflow was found (Pearson correlation of 0.81), which means that pauses constitute a large part of the subjective evaluation of read speech at least in this case. Thus the measures that obtained a good agreement with the fixed postures without audible airflow have a strong relation with the overall subjective evaluation based on the LBDL. On the contrary, the total reading time has balanced results for all individual categories and manages to achieve a very good results for the overall score. The results for the other individual categories of LBDL do not reach those for pauses.

The total reading time was found distinctive for evaluation of disfluencies in read speech (Maier *et al.*, 2011). This measure was added to the experiment to have a comparison to other possibility of how to measure stuttering severity. It turned out to be a very good instrument for the evaluation even though it is very simple. The results are comparable and in some cases better than those of introduced algorithms, and it could be possible to replace the algorithms with the total reading time. But we would like to use these algorithms for evaluation of spontaneous speech where the utterances are mostly limited by time and the total time of a recording will not be as influential as in recordings of read speech.

Because of the basic method used for the larger part of the measures (the Bayesian abrupt spectral changes detector), it is appropriate to investigate the relationships between these measures, and a strong relationship can be expected as in Cucchiarini *et al.* (2000). Examining these results, we can see that all the measures based on the BACD are strongly correlated (some of the coefficients exceed 0.9). In case of lesser correlation, there exists a high probability that a combined measure created from less correlated measures will be more successful. A small experiment was carried out to see whether this is so by a simple combination (summing up the normalized values of measures), and a correlation coefficient of 0.8 with speech pathologists and 0.82 with the overall characteristic was achieved; this is higher than that for any single measure. A suitable combination and selection of measures could be a future focus of research.

A possible limitation of the algorithms is that they are able to describe fixed postures without audible airflow with good agreement and the other individual characteristics of the subjective evaluation, such as syllable and incomplete syllable repetitions or prolongations, to a limited extent. The results of this study for these symptoms do not reach the results of Noth *et al.* (2000), Wisniewski *et al.* (2007a) or Wisniewski *et al.* (2007b), but on the other hand, we are not aware of other studies concentrating on automatically measured temporal speech characteristics in stuttered speech that do not use hidden Markov models. The database could be considered a weak point of the present study, and especially its gender imbalance and its distribution of participants across the levels of the disorder. There were only a few participants at the very severe level, and most participants were located at the mild, moderate, or severe levels. However, the database reflects the situation in common practice (Yairi and Ambrose, 1999; Bloodstein and Bernstein Ratner, 2008).

An advantage of our methods could be the possibility to exchange one instrument for another. In other words, it provides the opportunity to apply other reliable abrupt spectral changes detectors or voice activity detectors. The BACD (Cmejla *et al.*, 2013) applied in this study was tested using

synthetic and real speech signals (Bergl and Cmejla, 2007) or for stuttered speech (Bergl, 2010) in comparison to other divergence metrics with very good results. Algorithms, from simpler ones such as spectral or cepstral distance to more complex ones, such as general likelihood ratio (Appel and Brandt, 1983) and Kullback–Leiber divergence, could be employed. A great advantage of BACD- and VAD-based measures could be that they are language independent, and there is no need for a training database as in the case of systems based on hidden Markov models. They could be considered for use in experiments with second language learning as in Cucchiarini *et al.* (2000, 2002) and Maier *et al.* (2009c). Another VAD was also tested, one based on parameters (Atal and Rabiner, 1976) in cooperation with the support vector machine making the decision about speech vs silence. When this VAD was applied, very similar results were obtained.

V. CONCLUSION

An experiment was carried out to determine whether the level of the speech fluency disorder can be objectively estimated by means of automatic acoustic measurements of read speech. On the basis of the results, the following conclusions can be drawn. First, automatic measurements based on the detection of abrupt spectral changes using the Bayesian detector, and also voice activity detection, are able to indicate the overall level of the speech fluency disorder in read speech. Second, some measures can describe individual symptoms of stuttering—the best results were obtained for fixed postures without audible airflow (pauses in speech). An advantage of all the measures presented is that there is no external intervention, the measures are fully automatic and the methods can be replaced with other reliable algorithms. Future research could focus on the analysis of spontaneous speech by means of the measures introduced.

ACKNOWLEDGMENTS

We would like to thank Jan Vokral for providing the signal database and clinical data; we would also like to thank Tereza Tykalova, Jan Cerny, and Miroslava Hrbkova for the evaluation of the speech signals. This research was supported by Project No. GACR P102/12/2230 and by the Grant Agency of the Czech Technical University in Prague, Grant No. SGS12/185/OHK4/3T/13.

Adams, M. R. (**1987**). "Voice onsets and segment durations of normal speakers and beginning stutterers," J. Fluency Disord. **12**, 133–139.

Appel, U., and Brandt, V. A. (**1983**). "Adaptive segmentation of piecewise stationary time series," Inform. Sci. **29**, 27–56.

Atal, B., and Rabiner, L. (**1976**). "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. Acoust. Speech Signal Process. **24**, 201–212.

Bergl, P. (**2010**). "Objektivizace poruch plynulosti reci (Objectification of speech disfluen- cies)," Ph.D. thesis, Czech Technical University in Prague, 135 pp (in Czech).

Bergl, P., and Cmejla, R. (**2007**). "Improved detection of boundaries of phonemes in speech databases," in *Proceedings of the Fifth IASTED International Conference: Biomedical Engineering (BIEN'07)* (ACTA Press, Anaheim, CA), pp. 171–174.

Bloodstein, O., and Bernstein Ratner, N. (**2008**). *A Handbook on Stuttering*, 6th ed. (Delmar, Cengage Learning, New York), Chap. 1.

Boersma, P. (**2002**). "PRAAT, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

Cmejla, R., Rusz, J., Bergl, P., and Vokral, J. (**2013**). "Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders," Speech Commun. **55**, 178–189.

Cmejla, R., and Sovka, P. (**2004**). "Recursive Bayesian autoregressive changepoint detector for sequential signal segmentation," in *EUSIPCO-2004-Proceedings [CD-ROM]* (Technische Universitat, Wien, Austria), pp. 245–248.

Conture, E. (**2001**). *Stuttering: Its Nature, Diagnosis, and Treatment*, 1st ed. (Allyn and Bacon, Boston), Chap. 1.

Cordes, A. K., and Ingham, R. J. (**1994**). "The reliability of observational data. II. Issues in the identification and measurement of stuttering events," J. Speech Lang. Hear. Res. **37**, 279–294.

Craig, A., and Tran, Y. (**2005**). "The epidemiology of stuttering: The need for reliable estimates of prevalence and anxiety levels over the lifespan," Int. J. Speech-Lang. Pathol. **7**, 41–46.

Cucchiarini, C., Strik, H., and Boves, L. (**2000**). "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," J. Acoust. Soc. Am. **107**, 989–999.

Cucchiarini, C., Strik, H., and Boves, L. (**2002**). "Quantitative assessment of second language learners' fluency: Comparison between read and spontaneous speech," J. Acoust. Soc. Am. **111**, 2862–2873.

de Andrade, C. R. F., Cervone, L. M., and Sassi, F. C. (**2003**). "Relationship between the stuttering severity index and speech rate," Sao Paulo Med. J. **121**, 81–84.

Di Simony, F. G. (**1974**). "Some preliminary observations on temporal compensation in the speech of children," J. Acoust. Soc. Am. **56**, 697–699.

Ezrati-Vinacour, R., and Levin, I. (**2004**). "The relationship between anxiety and stuttering: A multidimensional approach," J. Fluency Dis. **29**, 135–148.

Goberman, A. M., Blomgren, M., and Metzger, E. (**2010**). "Characteristics of speech disfluency in Parkinson's disease," J. Neurol. **23**, 470–478.

Godino-Llorente, J., and Gomez-Vilda, P. (**2004**). "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," IEEE Trans. Biomed. Eng. **51**, 380–384.

Guitar, B. (**2006**). *Stuttering, an Integrated Approach to its Nature and Treatment*, 3rd ed. (Lipptincott Williams and Wilkins, Baltimore), Chap. 1, p. 13.

Hall, K. D., and Yairi, E. (**1992**). "Fundamental frequency, jitter, and shimmer in preschoolers who stutter," J. Speech Hear. Res. **35**, 1002–1008.

Hariharan, M., Chee, L. S., Ai, O. C., and Yaacob, S. (**2012**). "Classification of speech dysfluencies using LPC based parameterization techniques," J. Med. Syst. **36**, 1821–1830.

Harrington, J., and Cassidy, S. (**1999**). *Techniques in Speech Acoustics* (Kluwer Academic, Dordrecht, Netherlands), Chap. 9, pp. 239–277.

Healey, E. C., and Gutkin, B. (**1984**). "Analysis of stutterers' voice onset times and fundamental frequency contours during fluency," J. Speech Hear. Res. **27**, 219–225.

Healey, E. C., and Ramig, P. R. (**1986**). "Acoustic measures of stutterers' and nonstutterers' fluency in two speech contexts," J. Speech Hear. Res. **29**, 325–331.

Howell, P., Hamilton, A., and Kyriacopoulos, A. (**1986**). "Automatic detection of repetitions and prolongations in stuttered speech," in *Speech Input/Output: Techniques and Applications* (IEE Publications, Bochum, Germany), pp. 252–256.

Johnson, W. (**1961**). "Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers," J. Speech Hear. Disord. **7**, 1–20.

Kalinowski, J. (**2003**). "Self-reported efficacy of an all in-the-ear-canal prosthetic device to inhibit stuttering during one hundred hours of university teaching: An autobiographical clinical commentary," Disabil. Rehabil. **25**, 107–111.

Kay Elemetrics Corp. (**2003**). *Multi-Dimensional Voice Program (MDVP): Software Instruction Manual* (Kay Elemetrics, Lincoln Park, IL).

Kent, R., Weismer, G., Kent, J., Vorperian, H., and Duffy, J. (**1999**). "Acoustic studies of disartric speech: Methods, progress, and potential," J. Commun. Dis. **32**, 141–186.

Kuniszyk-Jozkowiak, W. (**1995**). "The statistical analysis of speech envelopes in stutterers and non-stutterers," J. Fluency Disord. **20**, 11–23.

Kuniszyk-Jozkowiak, W. (**1996**). "A comparison of speech envelopes of stutterers and non- stutterers," J. Acoust. Soc. Am. **100**, 1105–1110.

Lechta, V. (**2004**). *Diagnoza Narusene Komunikacni Schopnosti (Diagnostics of Impaired Communication Ability)* (Portal, Prague), pp. 317–332 (in Czech).

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., and Noth, E. (**2009a**). "PEAKS—A system for the automatic evaluation of voice and speech disorders," Speech Commun. **51**, 425–437.
Maier, A., Honig, F., Bocklet, T., Noth, E., Stelzle, F., Nkenke, E., and Schuster, M. (**2009b**). "Automatic detection of articulation disorders in children wit cleft lip and palate," J. Acoust. Soc. Am. **126**, 2589–2602.
Maier, A., Honig, F., Steidl, S., Noth, E., Horndasch, S., Sauerhofer, E., Kratz, O., and Moll, G. (**2011**). "An automatic version of a reading disorder test," ACM Trans. Speech Lang. Process. **7**(4), 17.
Maier, A., Honig, F., Zeissler, V., Batliner, A., Korner, E., Yamanaka, N., Ackermann, P., Peter, D., and Noth, E. (**2009c**). "A language-independent feature set for the automatic evaluation of prosody," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, England, pp. 600–603.
Mansson, H. (**2000**). "Childhood stuttering: Incidence and development," J. Fluency Disord. **25**, 47–57.
Metz, D. E., and Samar, V. J. (**1983**). "Acoustic analysis of stutterers' fluent speech before and after therapy," J. Speech Hear. Res. **26**, 531–536.
Noth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., and Wittenberg, T. (**2000**). "Automatic stuttering recognition using hidden Markov models," in *Sixth International Conference on Spoken Language Processing*, Beijing, China, Vol. 4, pp. 65–68.
Onslow, M., Gardner, K., Bryant, K., C. M. Stuckings, and Knight, T. (**1992**). "Stuttered and normal speech events in early childhood: The validity of a behavioral data language," J. Speech Hear. Res. **35**, 79–87.
Ravikumar, K. M., Rajagopal, R., and Nagaraj, H. C. (**2009**). "An approach for objective assessment of stuttered speech using mfcc features," ICGST Int. J. Digital Signal Process. **9**, 19–24.
Riley, G. D. (**1972**). "A stuttering severity instrument for children and adults," J. Speech Hear. Disord. **37**, 314–322.
Robb, M., Blomgren, M., and Chen, Y. (**1998**). "Formant frequency fluctuation in stuttering and nonstuttering adults," J. Fluency Disord. **23**, 73–84.
Ruanaidh, J., and Fitzgerald, W. (**1996**). *Numerical Bayesian Methods Applied to Signal Processing* (Springer-Verlag, New York), Chap. 5, pp. 96–101.
Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E. (**2011**). "Quantitative acoutic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," J. Acoust. Soc. Am. **129**, 350–367.
Ryan, B. P. (**1992**). "Articulation, language, rate and fluency characteristics of stuttering and nonstuttering preschool children," J. Speech Hear. Res. **35**, 333–342.
Sapir, S., Ramig, L. O., Spielman, J. L., and Fox, C. (**2010**). "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," J. Speech Lang. Hear. Res. **53**, 114–125.
Szczurowska, I., Kuniszyk-Jozkowiak, W., and Smolka, E. (**2009**). "Speech nonfluency detection using Kohonen networks," Neural. Comput. Appl. **18**, 677–687.
Teesson, K., Packman, A., and Onslow, M. (**2003**). "The Lidcombe bahavioral data language of stuttering," J. Speech Lang. Hear. Res. **46**, 1009–1015.
Van Borsel, J., Reunes, G., and Van den Bergh, N. (**2003**). "Delayed auditory feedback in the treatment of stuttering: Clients as consumers," Int. J. Lang. Commun. Disord. **38**, 119–129.
Wisniewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., and Suszynski, W. (**2007a**). "Automatic detection of disorders in a continuous speech with the Hidden Markov Models approach," in *Comp. Recognition System 2, 45 of Advances in Soft Computing* (Springer, Berlin), pp. 445–453.
Wisniewski, M., Niewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., and Suszynski, W. (**2007b**). "Automatic detection of prolonged fricative phonemes with the Hidden Markov models approach," J. Med. Inform. Technol. **11**, 293–297.
Yairi, E., and Ambrose, N. (**1999**). "Early childhood stuttering. I: Persistency and recovery rates," J. Speech Lang. Hear. Res. **42**, 1098–1112.
Yairi, E. and Clifton, N. F., Jr. (**1972**). "Disfluent speech behavior of preschool children, high school seniors, and geriatric persons," J. Speech Hear. Res. **15**, 714–719.
Yaruss, J. S., and Conture, E. G. (**1993**). "F2 transitions during sound/syllable repetitions of children who stutter and predictions of stuttering chronicity," J. Speech Hear. Res. **36**, 883–896.