Applied Computer Science, vol.17, no. 4, pp. 45–54
doi: 10.23743/acs-2021-28

Submitted: 2021-07-11 | Revised: 2021-08-14 | Accepted: 2021-09-20

Keywords: stuttering, fillers disfluency, automatic recognition, fillers detection

Waldemar SUSZYŃSKI [0000-0003-2990-2078]*,
Małgorzata CHARYTANOWICZ [0000-0002-1956-3941]*,
Wojciech ROSA [0000-0002-7051-6008]**, *Leopold KOCZAN* [0000-0002-7775-1836]**,
Rafał STĘGIERSKI [0000-0001-7225-3275I]*

DETECTION OF FILLERS IN THE SPEECH BY PEOPLE WHO STUTTER

Abstract

Stuttering is a speech impediment that is a very complex disorder. It is difficult to diagnose and treat, and is of unknown initiation, despite the large number of studies in this field. Stuttering can take many forms and varies from person to person, and it can change under the influence of external factors. Diagnosing and treating speech disorders such as stuttering requires from a speech therapist, not only good professional preparation, but also experience gained through research and practice in the field. The use of acoustic methods in combination with elements of artificial intelligence makes it possible to objectively assess the disorder, as well as to control the effects of treatment. The main aim of the study was to present an algorithm for automatic recognition of fillers disfluency in the statements of people who stutter. This is done on the basis of their parameterized features in the amplitude-frequency space. The work provides as well, exemplary results demonstrating their possibility and effectiveness. In order to verify and optimize the procedures, the statements of seven stutterers with duration of 2 to 4 minutes were selected. Over 70% efficiency and predictability of automatic detection of these disfluencies was achieved. The use of an automatic method in conjunction with therapy for a stuttering person can give us the opportunity to objectively assess the disorder, as well as to evaluate the progress of therapy.

1. INTRODUCTION

The recognition of speech pathology on the basis of the acoustic analysis of the utterance enables simple, non-invasive diagnostics. Stuttering is a speech impediment that is very complex, difficult to diagnose and treat, and also not fully understood, despite the large number of studies in this field. It can take many forms and can vary from person to person. Moreover, it can ease off or intensify under the influence of external factors (Bloodstein 1995; Stromsta1993, Wingate 2012).

* Lublin University of Technology, Faculty of Electrical Engineering and Computer Science, Department of Computer Science, Poland, w.suszynski@pollub.pl, m.charytanowicz@pollub.pl, rafal.stegierski@gmail.com
** Lublin University of Technology, Faculty of Technology Fundamentals, Poland, w.rosa@pollub.pl, l.koczan@pollub.pl

Measurements of disfluent episodes in the speech of people who stutter are very important in diagnosing, monitoring the course and assessing the final effects of therapy. Currently, they are auditioned, which is associated with a considerable effort on the part of the speech therapist. Due to the subjectivity of such measurements and sometimes low agreement in the assessments of the audience, it is difficult to compare different therapeutic techniques. It would be desirable, therefore, to develop an objective method of detecting and measuring the duration of individual types of disfluencies, based on the acoustic characteristics of the speech signal (Kuniszyk-Jóźkowiak et al., 2003, 2004).

In this article, we focus on one type of stuttering – fillers. They are among the mildest symptoms of stuttering. In Polish, the vowel "y" is most often used as a filler. It is often an approach taken by people who stutter as a way to begin the utterance of a difficult word. Fillers are common in fluent speech as well, and if rare enough, they do not affect the overall judgment of the statement. In the case of people who stutter, this relatively mild form of disfluency should be controlled and gradually eliminated, as it often hides therapeutically improper treatment that does not eliminate the psychological basis of this extremely complicated disorder.

Automatic determination of the fluency disturbance level is very significant for diagnosing, forecasting and therapy, and the detection and duration measurements of stuttering episodes are of great importance in a logopaedist's work. However, there are not many studies aiming at automation of speech assessment of people who stutter. Still, such studies were carried out by Howell and colleagues (Howell A Sackin,1995; Howell et al., 1997; Czyżewski, Kaczmarek & Kostek, 2003; Kuniszyk-Jóźkowiak, Smołka & Suszyński, 2001). In other studies ware used fuzzy logic (Suszynski et al., 2003a, 2003b), correlation function (Suszynski et al., 2005), Hidden Markov Models (Wiśniewski et al., 2010; Wiśniewski & Kuniszyk-Jóźkowiak, 2015), Kohonenn Neural Network (Smołka et al., 2003) or Hierarchical AAN system (Świetlicka, Kuniszyk-Jóźkowiak & Smołka, 2013). It was also used label sequences to detect stuttering events in reading speech (Alharbia et al., 2020).

The aim of our study was to develop an algorithm for automatic recognition of fillers in the statements of people who stutter – doing so on the basis of their parameterized features in the amplitude-frequency space.

2. PREPARATION OF MATERIAL FOR RESEARCH

The acoustic classification and identification of stuttering was carried out by observing the spectral waveforms and parameters obtained from the developed computer procedures. Based on the results of the acoustic classification of disfluency, separate procedures have been developed for the recognition and classification of individual groups of these episodes. The acoustic features of particular types of disfluency and the limits of their variability were determined on the basis of a set of disfluent statements of stutterers and a comparison of these with their fluent counterparts.

The simplest and most frequently used graphic image of speech signals is an oscillograph record. The program for acquiring and processing recordings on the oscillograph enables the initial visualization of speech, marking or deleting specific fragments, adjusting the time scale and amplitude, etc. In the diagnosis of many speech disorders, as well as in work with people with hearing problems, amplitude recording alone is not sufficient. Full information

about speech signals is given by three-dimensional frequency characteristics that take in the variables of time, amplitude and frequency. This analysis shows the changes taking place in speech with distorted articulation. However, it requires some experience on the part of the analysing person to be accomplished in this field.

The research used the digital speech signals of people who stutter. The data were analyzed by FFT with a Hamming window at 20 ms time intervals using $N = 21$ one-third octave filters in the frequency range of 100-10000 Hz. Additionally, an A filter was used and a logarithmic amplitude scale was applied. This type of analysis is a certain approximation of the characteristics of sound processing by the human auditory system. This approach to analysis allowed the development of automatic methods similar to human analysis (Moore & Glasberg, 1983; Moore, Peters & Glasberg,1990). In addition, for the automation of the process, the average sound levels and the band in which the maximum sound level was located were determined.

A general block diagram of a set of computer procedures is shown in Fig. 1. We can distinguish here outputs for preliminary analyzes and an automatic detection block. In the test block of the program, the time courses of the average level and the location of the maximum spectrum were determined and visualized. In the automatic detection block, various types of disfluency were detected and classified. Herein, four main types of non-fluent episodes were distinguished: prolongation, stops, repetitions and fillers (Fig. 1).

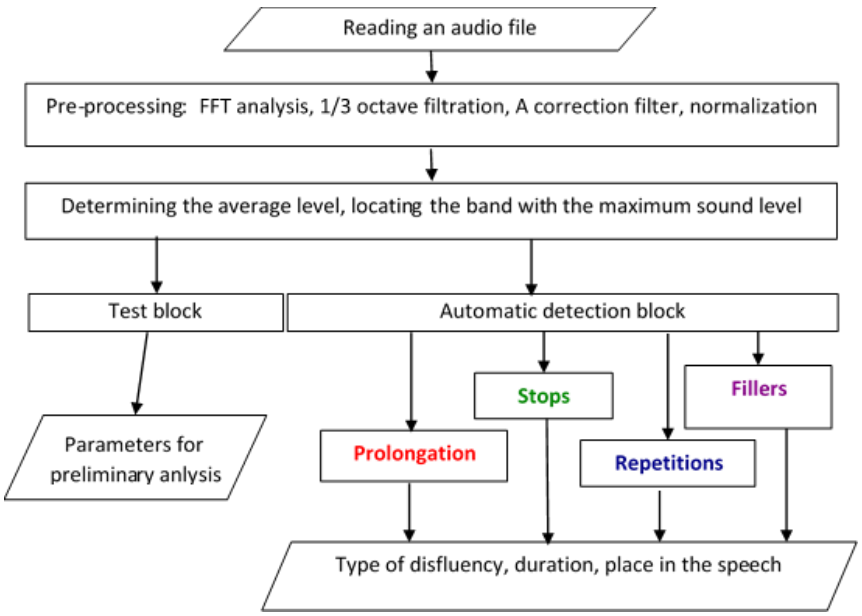

Fig. 1. General block diagram of the program for the analysis and automatic recognition of speech disfluency

3. ALGORITHM AND OPERATION OF THE FILLERS DETECTION PROGRAM

The fillers detection program includes the following procedures:

1. Calculation and visualization of the 1/3 octave spectrum.
2. Calculation and visualization of the average sound level.
3. Arbitrary stretching and narrowing of these waveforms and reproduction of the sound, the spectrum and average level of which are illustrated.
4. Readout of the cursor position on the average level waveform, which is set by the user of the program at the beginning of the filler considered being characteristic of the given person. Setting and reading the final position of the cursor, which practically comes down to choosing the width of the time window T.
5. Calculation of the correlation function according to formula (1), (2).

The principle of fillers detection was the correlation with the pattern marked by the examiner. Based on the conducted research, it can be concluded that in the majority of people who stutter, if there are any interferences, the outcome of the applied research refers to the same that is characteristic for a given person's sound (the "y" is most often inserted in the Polish language).

Calculation of the correlation function according to the following formulae:

$$R(t,T) = \frac{\sum_{i=1}^{N}\sum_{l=0}^{T-1}[x_i(t+l)-\mu(t)][x_i(t_w+l)-\mu(t_w)]}{\sqrt{\left(\sum_{i=1}^{N}\sum_{l=0}^{T-1}[x_i(t+l)-\mu(t)]^2\right)\left(\sum_{i=1}^{N}\sum_{l=0}^{T-1}[x_i(t_w+l)-\mu(t_w)]^2\right)}} \quad (1)$$

$$\mu(t) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t}^{t+T-1} x_i(t) \quad (2)$$

where: $x_i(t), x_i(t_w)$ – corrected sound levels in the i-th band for the time samples t and pattern t_w,
$\mu(t)$, $\mu(t_w)$ – average of all values in N bands in the window T,
t – current number of the time sample,
T – window width (number of samples in the time window),
l – number of the sample in the window.

Figure 2 shows an example of the obtained data set by pre-processing sound samples. The table contains the corrected and normalized sound levels in the bands 1/3-octave at consecutive moments of time. The columns represent successive moments of time (marked with t1, t2, ..., t29), lines – successive 1/3-octave bands (1–21). Expert selected patterns with a window width of two (t11–t12) are marked in red, while the current time window, shifting from the beginning to the end of the file is in green. Left arrow indicates the computation of the correlation for t = t1, right arrow for t = t17. The correlation coefficient described by formula 1 for a given pattern (width T) is only a function of time.

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24	t25	t26	t27	t28	t29
1	46	20	48	47	44	49	27	25	19	25	17	26	7	47	9	40	19	32	45	30	0	37	40	32	25	48	29	42	29
2	3	11	25	7	3	6	18	28	29	22	12	30	23	37	22	20	2	48	9	2	29	17	2	47	19	47	21	48	7
3	29	31	48	49	47	45	5	6	37	3	37	21	4	42	10	31	16	24	3	33	43	38	27	46	19	25	3	7	13
4	48	31	13	38	38	10	39	36	25	2	26	41	23	40	41	43	48	25	9	27	44	23	15	5	15	18	7	1	4
5	46	3	14	23	19	24	9	5	47	7	32	1	43	16	20	37	13	29	7	26	34	49	47	39	49	26	37	47	7
6	49	28	11	26	16	0	19	13	38	36	41	28	40	29	6	34	17	8	46	26	34	37	8	46	37	24	33	47	34
7	17	33	21	0	29	1	0	37	2	46	22	49	46	26	29	8	44	24	29	46	28	18	27	9	22	40	2	6	28
8	25	47	36	48	23	23	48	1	41	46	20	3	14	13	10	0	12	27	47	36	0	0	37	4	45	6	38	35	33
9	48	0	11	27	20	12	39	20	40	14	36	13	4	13	12	34	38	10	21	9	20	20	10	21	1	18	43	42	25
10	40	16	44	12	24	25	29	20	37	22	24	18	34	42	46	1	37	35	30	32	2	15	7	20	43	42	24	3	0
11	29	38	34	25	17	20	16	11	40	1	25	17	16	6	28	2	5	10	38	29	43	38	38	30	37	26	5	12	10
12	21	5	42	18	3	8	18	40	11	22	15	3	47	6	48	44	31	39	46	7	26	38	17	44	15	37	21	33	11
13	13	41	35	20	31	42	13	22	19	33	41	39	44	45	27	40	4	0	14	28	39	49	22	45	20	14	31	16	32
14	21	41	19	46	3	42	28	30	33	35	20	20	38	17	34	35	28	8	7	1	36	45	19	0	7	27	41	8	24
15	49	7	19	7	11	29	28	9	49	39	1	45	2	20	44	2	6	34	10	34	1	25	4	42	43	2	39	15	0
16	32	24	45	48	30	49	36	38	37	47	48	4	38	0	22	49	25	27	1	35	29	25	25	5	6	44	16	13	20
17	39	28	23	19	48	2	20	24	7	18	9	23	3	20	23	2	18	25	18	21	5	8	48	6	3	39	0	21	40
18	33	14	7	41	6	3	3	45	15	21	44	25	8	22	15	5	19	27	9	44	9	36	40	43	12	49	13	31	30
19	0	14	39	0	31	23	0	5	29	41	33	37	21	19	47	5	5	25	45	4	15	33	18	40	48	36	22	27	29
20	16	13	10	28	0	45	44	32	10	37	1	4	32	25	27	44	43	28	43	41	33	18	7	28	15	7	22	3	12
21	7	39	29	16	31	30	7	22	23	28	32	34	28	5	11	36	42	13	47	15	39	49	26	22	17	44	15	44	38

Fig 2. Schematic diagram of the correlation procedure
(vertical – consecutive spectra, horizontal – consecutive time moments)

In Figure 3, we can see a screenshot used to automatically search for fillers in a stutterer's statements.

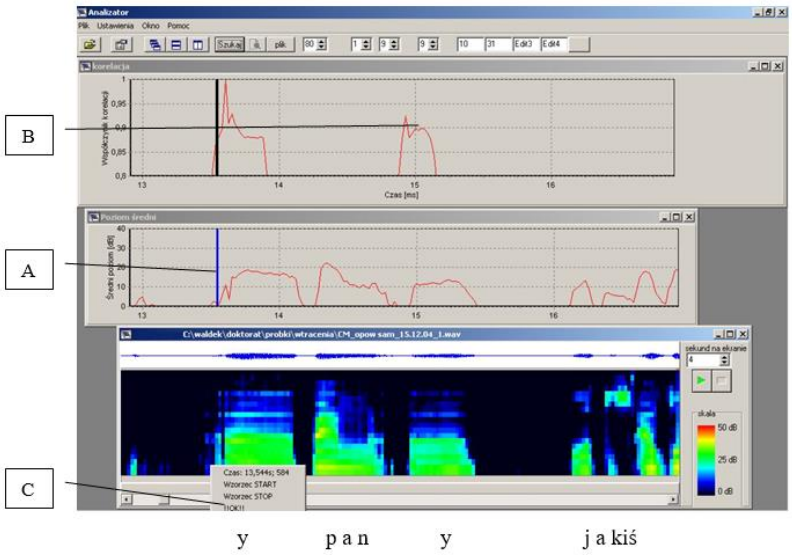

Fig. 3. Screen from the program for automatic fillers detection
(statement in Polish "pan jakiś" with fillers "y")

The top window – the value of the correlation coefficient; middle window – medium sound level; lower window – oscillogram and spectrogram. A – setting the beginning of the pattern, B – correctly identified repeated fills, C – time (from the beginning of the sound file) indicated by the cursor position.

Due to the variety of fillers present, initial expert input was necessary. The individual had to indicate in the audio file one characteristic occurrence of a given disorder. The program then automatically indicated other instances of this disfluency. The examiner had to mark the beginning of the filler and determine its length (Figure 3). A spectrogram (lower window) was presented in the program window facilitating speech analysis. There was also the possibility to play a fragment of a speech sample.

In Part B of this figure, the expert observed the operation of the algorithm – the chart presents the value of the correlation coefficient (at the beginning of the chart it can be observed – the filler is detected and the correlation coefficient reaches the maximum value of one). The second peak (approx. 15 s) indicates the next filler detected (the correlation coefficient is approx. 0.8). Measurements were made for different sizes of windows. The obtained data were saved and on their basis 3D charts were prepared (Figures 4–6).

The program was created in the Delphi environment and written in the Pascal language. All procedures are implemented directly in the code.

4. VERIFICATION

As part of the research, the sensitivity and predictability of the method were determined. The dependence on the border coefficient of correlation and the width of the time window on the above parameters was also assessed. Setting the pattern start also plays an important role in detecting disfluency. Through many experiments, it was found that the optimal positioning of the cursor at the beginning of the disfluency chosen as the reference is (formula 3 and formula 4).

$$sensitivity = \frac{\sum correctly\ detected\ fillers}{\sum correctly\ detected\ fillers + \sum\ undetected\ fillers} \tag{3}$$

$$predictability = \frac{\sum correctly\ detected\ fillers}{\sum correctly\ detected\ fillers + \sum\ false\ fillers} \tag{4}$$

In order to verify and optimize the proposed procedures, the statements of seven stutterers with total of 170 fillers were selected. Each file contained one or more fillers surrounded by fluently spoken words. The examiner was responsible for selecting the model disfluency. The aim of these studies was to select parameters so that both the sensitivity and predictability were as high as possible and the correlation coefficient was at a level clearly indicating the similarity of the fillers found. Contour charts were built on the basis of the obtained data in order to accurately trace and select the optimal parameters. The optimal parameters of the tested parameters were those for which sensitivity and predictability exceed at least 70%. In the drawings, they were located in the ranges covered by red lines and marked in yellow.

Figures 4–7 show exemplary contour graphs of sensitivity and predictability depending on the boundary value of the correlation coefficient and the width of the time window for different people and different fillers. Each of the figures at the top presents also an example of the settings of the start and end of the filler pattern on the average sound level, along with an example of the length of the time window for which the assumed predictability and sensitivity results were obtained.

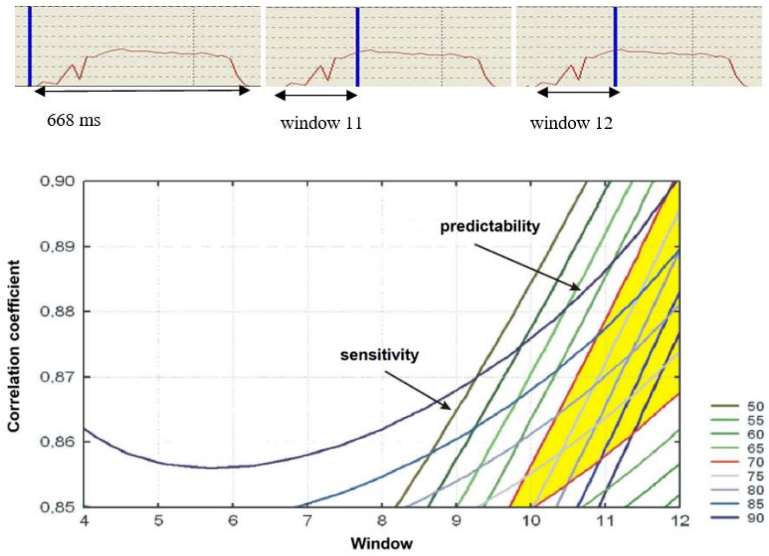

Fig. 4. Contour plots of the dependence of sensitivity and predictability on the limit value of the correlation coefficient and the width of the time window – case 1: correct identification for long windows

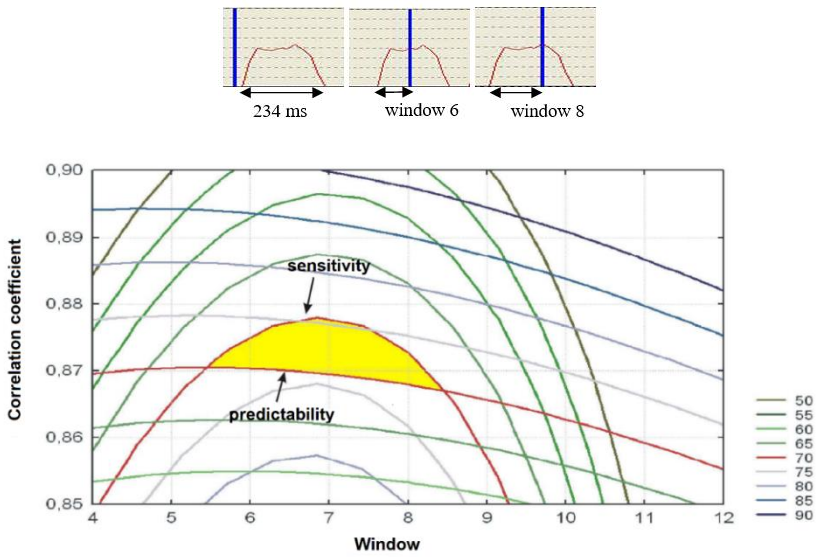

Fig. 5. Contour plots of the dependence of sensitivity and predictability on the limit value of the correlation coefficient and the width of the time window – case 2: correct identification for short windows

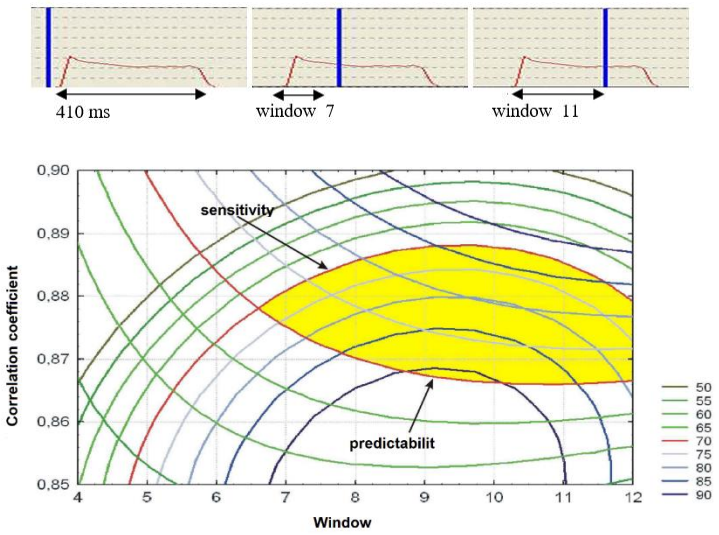

Fig. 6. Contour plots of the dependence of sensitivity and predictability on the limit value of the correlation coefficient and the width of the time window – case 3: correct identification for average windows

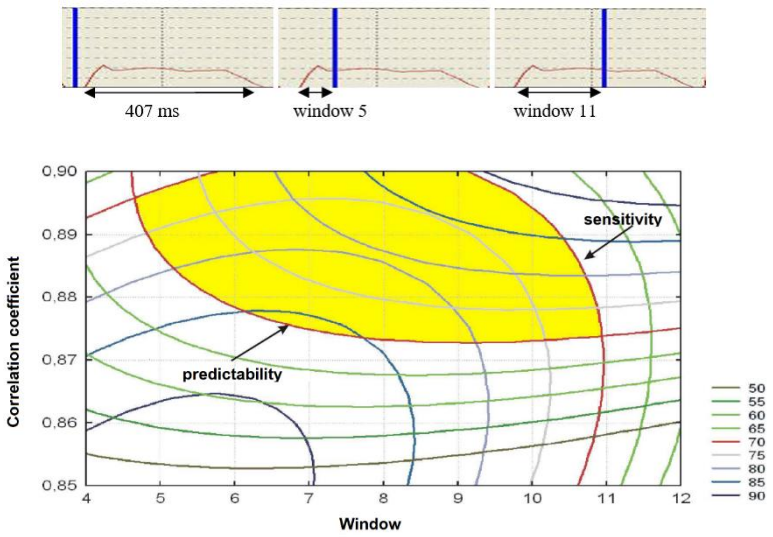

Fig. 7. Contour plots of the dependence of sensitivity and predictability on the limit value of the correlation coefficient and the width of the time window – case 4: correct identification for average windows with a simultaneous large value of the correlation function

The speech disorder described at work is difficult to detect automatically and requires a preliminary precise determination of the disfluency pattern and the method of searching for subsequent episodes. The presented drawings show a fragment of the research leading to the results, as well as to illustrate various cases. Figure 4 presents the results for a long filler which is correctly detected only for long windows. Figure 5 – a short filler – positive results

were obtained for short windows. In Figure 6, an example of average length of an filler – a wide range of windows for positive identification. Figure 7 – medium-length with a very large correlation function.

The width of the time window is specified as times of 23 ms. The optimal choice of the time window width was assumed so that both parameters exceed 70%. The areas corresponding to the optimal selection of parameters are highlighted.

Data analysis allowed to establish that the limit that should be set for the correlation coefficient should be 0.87–0.88. As can be seen, the width of the time window (the final position of the cursor on the pattern image) does not need to be precisely defined. However, it should be inserted in the middle of the pattern (between 1/3 and 2/3 of its duration).

5. SUMMARY

The developed and described procedures for automatic recognition of this type of disfluency can be used in continuous speech and do not require initial segmentation. According to verification, they also do not introduce erroneous classifications of the disfluency type and do not require perfect noise-free audio recordings.

In order to verify and optimize the procedures, the statements of seven stutterers (four boys and three girls aged 10 to 18) with duration of 2 to 4 minutes were selected. There were a total of 170 fillers in these statements (from 14 to 37 in the statements of individual people). Over 70% efficiency and predictability of automatic detection of these disfluencies was achieved.

The procedures presented in the paper using the correlation coefficient can also be applied to find other types of disfluency, e.g. repetitions or stops. After building a sufficiently large database, the fillers can be adjusted to fully automatically detect the set type of disorder. The use of an automatic method in conjunction with therapy for a stuttering person can give us the opportunity to objectively assess the disorder, as well as to evaluate the progress of therapy.

REFERENCES

Alharbia, S., Hasana, M., Simonsa, A. J. H., Brumfitt, S., & Green, P. (2020). Sequence labeling to detect stuttering events in read speech. *Computer Speech & Language, 62,* 101052. http://doi.org/10.1016/j.csl.2019.101052

Bloodstein, O. (1995). *A handbook on stuttering.* Singular Publishing Group, Inc.

Czyżewski, A., Kaczmarek, A., & Kostek, B. (2003). Intelligent processing of stuttered speech. *Journal of Intelligent Inform. Systems,* 143–171.

Howell, P., & Sackin, S. J. (1995). Automatic recognition of repetitions and prolongations in stuttered speech, Stuttering. *Proceedings of the First World Congress on Fluency Disorders* (pp. 372–374). Munich.

Howell, P., Sackin, S. J., Glenn, K., & Au-Yeung, J. (1997). *Automatic stuttering frequency counts, Speech Motor Production and Fluency Disorders*. Elsevier.

Kuniszyk-Jóźkowiak, W., Dzieńkowski, M., Smołka E., & Suszyński, W. (2003). Computer Diagnosis and Therapy of Stuttering. *Structures – Waves – Human Health, VIII*(2), 133–144.

Kuniszyk-Jóźkowiak, W., Smołka, E., & Suszyński, W. (2001). Acoustical characteristics alteration in persons who stutter resulting from therapy. *Structures-Waves-Biomedical Engineering, X(2),* 57–68.

Kuniszyk-Jóźkowiak, W., Smołka, E., Dzieńkowski, M., & Suszyński W. (2004). Computer therapy of speech non-fluency with automatic adaptation of individual person's difficulties. *Structures-Waves-Human Health, VIII*(2), 63–70.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter banwidths and excitation patterns. *The Journal of the Acoustical Society of America*, *74*, 750–753.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1990). Auditory filters shapes at low center frequencies. *The Journal of the Acoustical Society of America*, *88*, 132–149.

Smołka, E., Kuniszyk-Jóźkowiak, W., Suszyński, W., & Dzieńkowski, M. (2003). Speech syllabic structure extraction with application of Kohonen network. *Annales Informatica Universitatis Mariae Curie-Skłodowska, AI 1*,125–131.

Stromsta, C. (1993). The nature and management of stuttering. *Proceedings Abstracta, Congressus XVIII* (pp. 16–18). Societatis Phoniatricae Europaeae, Praga.

Suszyński, W., Kuniszyk-Józkowiak, W., Smolka, E., & Dzienkowski, M. (2003). Automatic Recognition of Nasals Prolongations in the Speech of Persons who Stutter. *Structures-Waves-Human Health, XII*(2), 175–184.

Suszyński, W., Kuniszyk-Jóźkowiak, W., Smołka, E., & Dzieńkowski, M. (2003). Prolongation detection with application of fuzzy logic. *Annales Informatica Universitatis Mariae Curie-Skłodowska, AI 1*, 133–140.

Suszyński, W., Kuniszyk-Jóźkowiak, W., Smołka, E., & Dzieńkowski, M. (2005). Speech disfluency detection with correlative method. *Annales Informatica Universitatis Mariae Curie-Skłodowska, AI 3*, 131–138.

Świetlicka, I., Kuniszyk-Jóźkowiak, W., & Smołka, E. (2013). Hierarchical ANN system for stuttering identification. *Computer Speech & Language*, *27*(1), 228–242. https://doi.org/10.1016/j.csl.2012.05.003

Wingate, M. E. (2002). *Foundation of stuttering*. Academic Press.

Wiśniewski, M, Kuniszyk-Jóźkowiak, W., Smołka, E., & Suszyński, W. (2010). Improved Approach to Automatic Detection of Speech Disorders Based the Hidden Markov Models Approach. *Journal of Medical Informatics & Technologies*, *15*, 145–152. http://doi.org/10.1007/978-3-540-75175-5_56

Wiśniewski, M., & Kuniszyk-Jóźkowiak, W. (2015). Automatic detection of stuttering in a speech. *Journal of Medical Informatics & Technologies*, *24*, 31–37.