Detection of disfluencies in speech signal

Katarzyna Barczewska[1], Magdalena Igras[2]

[1] Department of Automatic Control and Biomedical Engineering, [2] Department of Electronics
AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków
e-mail: kbarczew@agh.edu.pl, migras@agh.edu.pl

During public presentations or interviews, speakers commonly and unconsciously abuse interjections or filled pauses that interfere with speech fluency and negatively affect listeners impression and speech perception. Types of disfluencies and methods of detection are reviewed. Authors carried out a survey which results indicated the most adverse elements for audience. The article presents an approach to automatic detection of the most common type of disfluencies - filled pauses. A base of patterns of filled pauses (prolongated *I*, prolongated *e*, *mm*, *Im*, *xmm,* using *SAMPA* notation) was collected from 72 minutes of recordings of public presentations and interviews of six speakers (3 male, 3 female). Statistical analysis of length and frequency of occurrence of such interjections in recordings are presented. Then, each pattern from training set was described with mean values of first and second formants (F1 and F2). Detection was performed on test set of recordings by recognizing the phonemes using the two formants with efficiency of recognition about 68%. The results of research on disfluencies in speech detection may be applied in a system that analyzes speech and provides feedback of imperfections that occurred during speech in order to help in oratorical skills training. A conceptual prototype of such an application is proposed. Moreover, a base of patterns of most common disfluencies can be used in speech recognition systems to avoid interjections during speech-to-text transcription.

Keywords and phrases: speech processing, phoneme recognition, dynamics of speech, disfluencies of speech, elocution

Background

During public speeches, presentations or interviews, many factors (like stress, lack of confidence, insufficient preparation or lack in oratorical skills) cause recurrent disfluencies in speech. Speakers are often not aware of quantity and frequency of undesirable, unintentional parts of speech. Such imperfections interfere with content of speech and comfort of listening to the speaker and reception of his speech, as well as his image.

Disfluencies in speech

Different types and numbers of disfluencies in speech are distinguished in state-of-art literature. Roberts et al. propose 5 groups of them for non-stuttering adults [1] and these are: interjections, revisions, repetitions, prolongations and blocks. *Interjection* can be defined as negligible word, phrase or sound that doesn't affect meaning of the sentence. It depends on language, in English frequently occurs "*uh, well, like, you know*", in Polish "*yyy, hmm, to znaczy.*." (prolongated *I*, prolongated *e*, *mm*, *Im*, *xmm,* using *SAMPA* notation). Interjections are also called *filled pauses* or just *fillers* [2]. Second type of disfluencies are *revisions*. They appear when speaker corrects an error in utterance or begins it but does not complete. An example of revision is broken word, in Polish language an example can be "*Chciałbym zacz... rozpocznę od*", in English "*I'd like to sta.. I'll begin..*". Another type are *repetitions*. They take place if any part of sentence (e.g. word, syllable) is said more than once, with the exception of these repetitions which are intentionally used by the speaker to emphasize something. *Prolongation* is any sound in speaking which last longer than normally and, like repetition, is not used intentionally. Last type of speech disfluecies are *blocks* which appear when the sound is produced with too big force, or when speaker gets stuck trying to pronounce a syllable [1]. Prevalence of particular disfluency in speech is expressed by disfluency rate which is defined as "the number of disfluencies of that type divided by the total number of word tokens" [2].

The disfluencies which occur most frequently in spontanuous speaking are interjections. Their rate is about 15% while the rate of the second frequently occuring disfluency, revision, is about 5%. Summarizing total percentage of all disfluencies, interjections represent 60% of all [3].

Speakers use fillers in cases when they consider what to say next and they don't want to be interrupted by the others or when they hesitate [4]. Filled pauses can occur in any place of the utterance [2]. Occurrence of any disfluency disrupts the utterance, adversely affects the image of the speaker and has a negative influence on the reception of

the content of speech. Detecting it is very helpful not only in automatic speech recognizers (ASR) to skip fillers instead of transcribing them to text, but also can be an excellent tool in supporting speakers. The main focus of this research concerned detection of filled pauses for improving oratorical abilities.

Detection methods

Various approaches to the detection of filled pauses were applied, starting with simplest methods based on frequency analysis, ending with cepstral parameters analysis. O'Shaughnessy et al., to achieve this, use vowel identification and determination of their time duration. Their training set was used to determine filled pauses spectral characteristics which were patterns for detecting fillers in testing set. They classified signal frames as vowels or non-vowels basing on F0 stability. If time duration of indicated vowel exceeded threshold it was identified as speech disfluency [4]. Filled pauses in the speech signal were identified as long, steady vowels, which duration exceeded 120 ms. They had low F0, relative to the average F0 calculated for the speaker.

Likewise Audhkhasi and Kandhway [5] present frequency based approach. They detect filled pauses basing on the relative stability of the vocal tract shape during the production of filled pauses. They made an observation that vocal tract characteristics don't change and not only the pitch, but also formants remain stable during the course of a single filled pause, in contrast to their variation in normal speech. In the detection algorithm they used analysis of two first formants. The detection algorithm is preceded by excluding silent, low energy regions and weak fricatives from the speech signal. Then two first formants are computed (frame rate 10 ms), and for each of them stability is analyzed. The measure of formant stability is standard deviation computed for windows containing specific number of frames [5]. Filled pauses has completely different distribution of formants standard deviations than normal speech. To identify filled pauses, values of probability mass functions for both distributions are calculated, after comparing them (using Log Likelihood Ratio), it's possible to say with certain probability that a region of speech signal is a filled pause [5].

Another detection technique, proposed by Stouten and Martens includes analysis of MFCC parameters. After calculating cepstral parameters for signal segmentation, several features are calculated: segment duration, spectral stability, stable interval durations, presence of silence before and after segment, spectral center of gravity and simple filled pause model output. Classification (filled pause or non-filled pause) is performed with Multi-Layer perceptron [6].

Taking into account prototype of application which could help speakers improve their oratorical abilities, the process of disfluency recognition can't be very complicated because of the real-time performance. Authors proposed simple method based on only measure of two first formats stability in time. In section 3.2 method is described in details.

Material and Methods

In order to gather information about the speech elements that have most adverse effect on the reception of long utterances, a survey was conducted. After analysis of the survey results authors decided to recognize these disfluencies in speech signal which are the most undesirable from the listeners point of view.

Questionnaire

The survey was conducted on 35 people (Polish students, age 20-30). The questionnaire consisted of 2 points: first open question (*What is bothering you most in listening to long speeches, taking into account the speaker's way of speaking?*) at which respondents were asked to give a written response. In the second point respondents were asked to prepare the ranking of speech disfluencies, given in the list, in the order from that which is least desirable for listener. List of disfluencies from the second point of questionnaire is presented in Table 1.

Survey results clearly indicate which elements of the speech are undesirable for listeners. The answers given by respondents to the first question are divided into four groups. Mentioned elements were: speech disfluencies

Table 1. List of the disfluencies, which respondents used to make a ranking in order of the most disturbing in listening to long speeches.

A – syllable repetitions (e. g. *ro-rozpocznę od..*)
B – words repetitions (e. g. *rozpocznę od grupy grupy pierwszej*)
C – filled pauses (e. g. *yyyy, hmmm)*
D – revisions (e. g. *chciałem zacz... rozpocznę od*)
E – prolongations (e. g. *rozpocznę oooood...*)
F – additional words not affecting the meaning of the sentence (e. g. *to znaczy, w sensie...)*

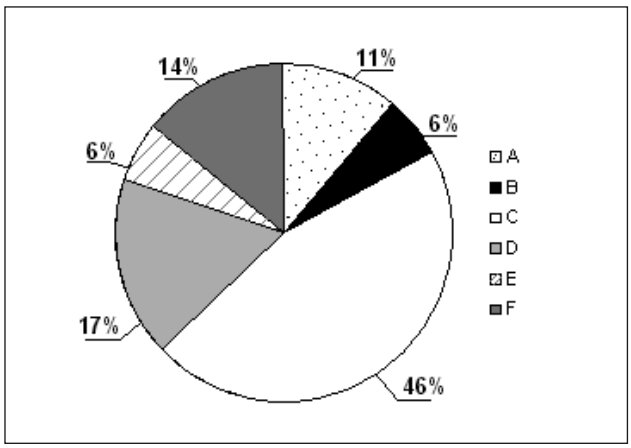

Figure 1. Summary of the percentage of respondents who indicated concrete type of the speech disfluency as the most negatively affecting the reception of speech. Designation of disfluencies like in the Table 1.

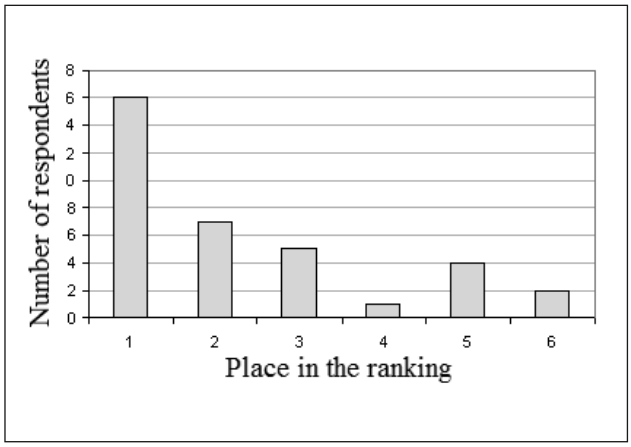

Figure 2. The number of respondents who placed filled pauses on each place (from 1 to 6) in the ranking of disfluencies from the most negatively affecting the reception of speech.

(43% of people), intonation (40%), lack of preparation (29%), and lack of interaction with audience (17%). In the group of speech disfluencies, the most frequently mentioned were filled pauses (17% of people) and stuttering (15%). From a group of elements related to the speaker's voice and intonation, respondents attributed the greatest importance to the monotonous tone of voice (29%) and too quiet speaking (17%). A large number of respondents (29%) associated undesirable speech elements with the lack of or poor preparation, mentioned reading a speech from a paper, mismatch the level of vocabulary to the audience. The results for the second survey task are shown in diagrams Figure 1 and 2.

After analyzing the survey results for the second task, it is found that filled pauses (described like: "*yyyy, hmmm*" in questionnaire): prolongated *I*, prolongated *e*, *m*, *Im*, *xmm*, using *SAMPA* notation, are most often placed by the listeners on the first place in the ranking of the disfluencies most adversely affecting listening to the speeches. 46% of respondents placed "*yyy*" (SAMPA *I*) at the beginning of the ranking. Eliminating this type of disfluency could be the first step in improving oratorical abilities. The algorithm described in this paper is then specified to recognize only these certain type of disfluencies.

Recordings

Research material consisted of recordings from the students presentation and radio broadcasts, which were prepared before processing by removing the voice of interviewer, leaving only the expressions of interviewees (using Audacity application). The length recordings after preparation was 12 minutes for each speaker (three woman and three men). Because of the fact, that authors didn't eliminate individual features from signals, first 4 minutes of each recording were used as training data and last 8 minutes from the same recording were added to testing set.

For investigating statistics of appearing different sorts of disfluencies, another set of recordings from radio broadcasts was prepared. The statistical set contained recordings of 6 speakers, 10 minutes for each speaker.

Finally, the database for training and testing algorithm of detection consisted of 72 minutes of good quality (16 bit, 16 000 Hz, SNR > 30 dB) recordings.

Analysis and results

Speech disfluencies in recordings

For each recording from statistical set, segments containing different types of disfluencies were manually annotated in mlf files in HTK standard using Anotator software [7]. All types of speech disfluencies mentioned in questionnaire where marked in speech signals. Summary information on selected segments are shown in Table 2.

Disfluencies that occur most frequently in all analyzed speech signals are definitely filled pauses, which account for 58,5% of total number of all disfluencies. In the set of filled pauses the most numerous is group of prolongated *I*. It accounts for 75% of filled pauses and 40% of total number of all disfluencies. Second frequent are repetitions (15,9%), and the third – prolongations (10.2%). In this study authors focused on the most frequent and the most undesirable (according to survey) filled pauses. For them, several parameters were calculated, describing time length of each filled pause segment annotated in training set. Parameters are shown in Tab.3. According to Wang et al [8], time length of such segments has gamma distribution, thus median and interquatile range were used in Tab.3 to describe data. Histogram of *I* segments' time lengths for whole training set is presented on Fig.3.

The average duration of phoneme *I* in Polish was proved to be 88 ms (with standard deviation 43 ms) [9]. In the case of prolongated *I* as a filled pause, the duration appears to be much longer (median: 370 ms), which determines the duration as one of disctinctive features for detection algorithm.

Table 2. Statistics of the types of disfluencies for the whole database (3 women, 3 men).

	M1	M2	M3	W1	W2	W3	**Mean [%]**
Filled pauses, total number:	16	96	95	88	178	84	**58.5**
- prolongated *I*	15	66	88	69	135	43	**43.9**
- prolongated e	0	25	1	4	0	14	**5.0**
- prolongated *m*	1	5	4	13	33	22	**7.9**
- *Im*	0	0	2	0	10	5	**1.5**
Prolongations	0	25	58	9	4	18	**10.2**
Revisions	8	7	17	12	8	19	**9.5**
Repetitions, total number:	9	1	45	42	10	31	**15.9**
- repeated word	4	1	21	24	5	23	**8.8**
- repeated syllable	5	0	24	18	5	8	**7.1**
Fillers (words)	3	12	26	0	1	2	**4.8**
Total number of disfluencies	39	141	241	151	199	154	**100**

Table 3. Calculated parameters of disfluencies I for training recordings (3 women, 3 men). IQR is interquartile range.

Parameter	M1	M2	M3	W1	W2	W3	**Mean**	**Standard deviation**
average number of segments *I* per minute	1,25	10,00	8,25	6,25	9,50	3,25	6,42	3,53
median time length [s]	0,38	0,43	0,45	0,35	0,27	0,36	0,37	0,06
IQR of segment time length [s]	0,09	0,24	0,21	0,27	0,12	0,31	0,21	0,09
shortest duration [s]	0,34	0,22	0,22	0,12	0,09	0,19	0,20	0,09
longest duration [s]	0,52	1,38	1,43	0,96	0,47	0,87	0,94	0,41
total length of all the segments *I* in the 4-minute speech [s]	2,03	20,61	15,92	9,55	9,90	5,96	10,66	6,71
overall ratio of *I* segments in recording [%]	0,85	8,59	6,63	3,98	4,12	2,49	4,44	2,79

Courses of variability of the fundamental frequency F0 were also determined for each segment with filled pause *I*. Courses for one person (W1) are presented in Figure 4. With few exceptions, where appeared episodes of rising or falling intonation, the course of F0 for segment is placed on a constant level, as opposed to fragments corresponding to the normal speech signal. The result is corresponding to Audhkhasi and Kandhway observation [5].

Figure 5 presents spectrogram which corresponds sequentially to filled pause of type *I* and the fragment of a signal representing a normal part of speech. One can observe the stability of the two first formants (F1, F2) in the spectrogram of the speech disfluency, which can not be noticed for the signal of normal, continuous speech, for which the formants are usually highly volatile over time.

Parameterization of phonemes

Taking into account the stability of the formants in these places in signal where filled pauses occurs throughout whole their duration, and the fact that the shortest disflu-

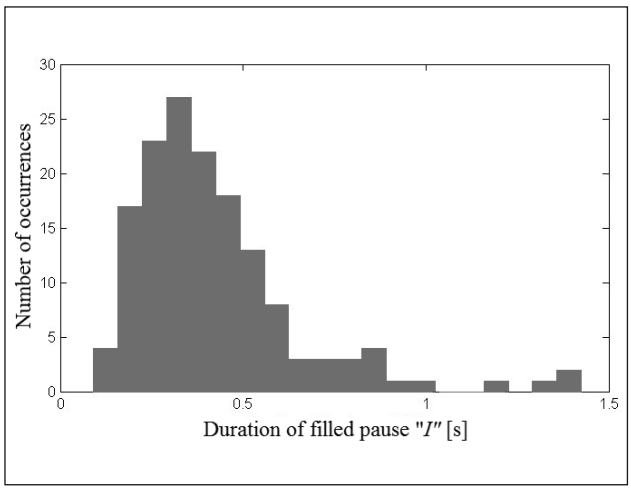

Figure 3. Histogram of I duration in the training set

ency of this type in training set lasted 200 ms, authors proposed a method for detecting filler *I* in speech, consisting of the following steps:

- vowels identification based on 2 first formants (F1, F2);
- measure of formants stability based on variability of standard deviation;
- checking duration time of identified vowels;

On the basis of the F1 and F2 formants values calculated for disfluencies which occurred in the statements of the training set using LPC method, a chart showing the position of points corresponding to disfluencies on the plane of formants was prepared. Comparing the location of points that characterize a set of fillers from training set to the map of vowels, it can be observed that disfluencies points appear in region which is close to the areas of vowel

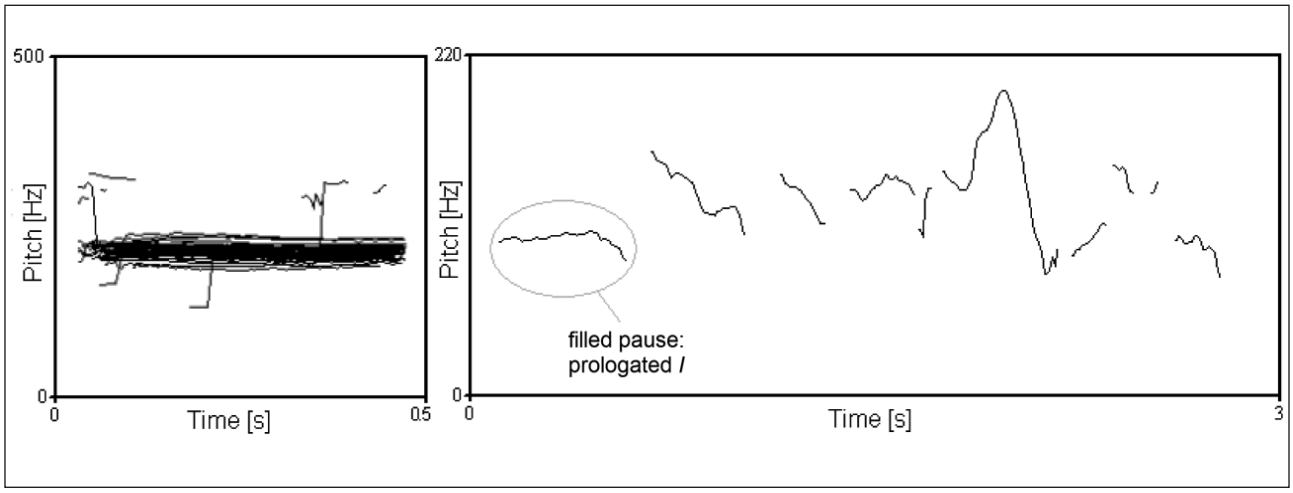

Figure 4. Pitch variability in segments with I filled pause. Left: pitch courses for all filled pauses annotated for W1. Right: comparison of pitch variability for filled pause and normal speech for M2. Graphs obtained using Praat application [10].

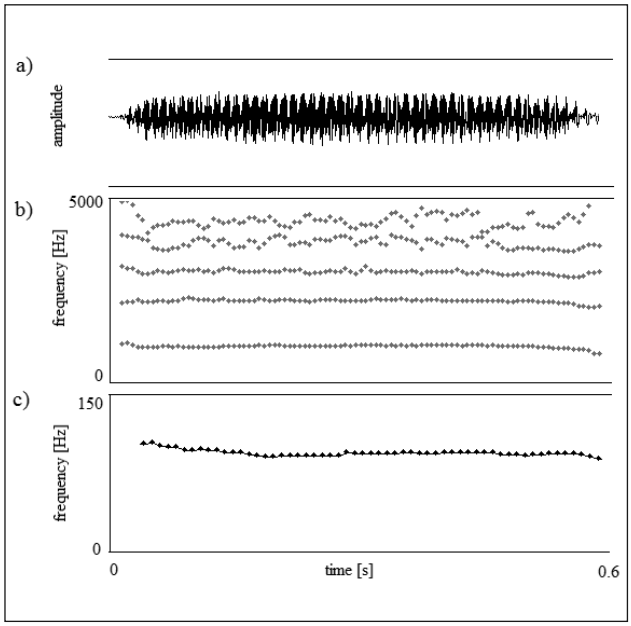

Figure 5. Characteristics of an example filled pause: a) waveform, b) formants, c) F0.

"y" and "e" (see Figure 6). On this basis, the boundaries of an area which corresponds to fillers can be defined. Boundaries of that region are used to recognize disfluencies in the testing set [11].

Patterns base

Using recordings from training set, patterns base of the first and second formants for the fillers *I* was prepared. The mean values with standard deviations for formants for each recorded person are given in Table 4. Their values were used in developing a fillers detection algorithm in the phase of formants stability evaluation in the analyzed time frames.

Values of the standard deviations with are the measure of formants stability in segment with filler *I* do not exceed 18% of the average value of F1 and 11% of the average values of F2 for women and 22% of the average 10% of the F1 and F2 mean value for men. Average values of F1 for a group of men are comparable, in contrary to its values in a group of women. Differences in its values can be caused by the differences in age of woman. Average F2 values are at

Table 4. Summary of formants F1 and F2 values [Hz] for filled pauses with prolongated I for each speaker (W: woman, M: man).

Parameter	M1	M2	M3	W1	W2	W3	**Mean**
average F1	520.3	473.2	473.6	469.8	529.0	478.6	**490,8**
standard deviation of F1	190.1	125.5	102.9	98.9	90.9	129.2	**122,9**
average F2	1768.2	1813.9	1845.9	1785.9	1589.7	1672.5	**1746,0**
standard deviation of F2	166.2	125.3	238.3	139.6	171.4	252.2	**182,2**

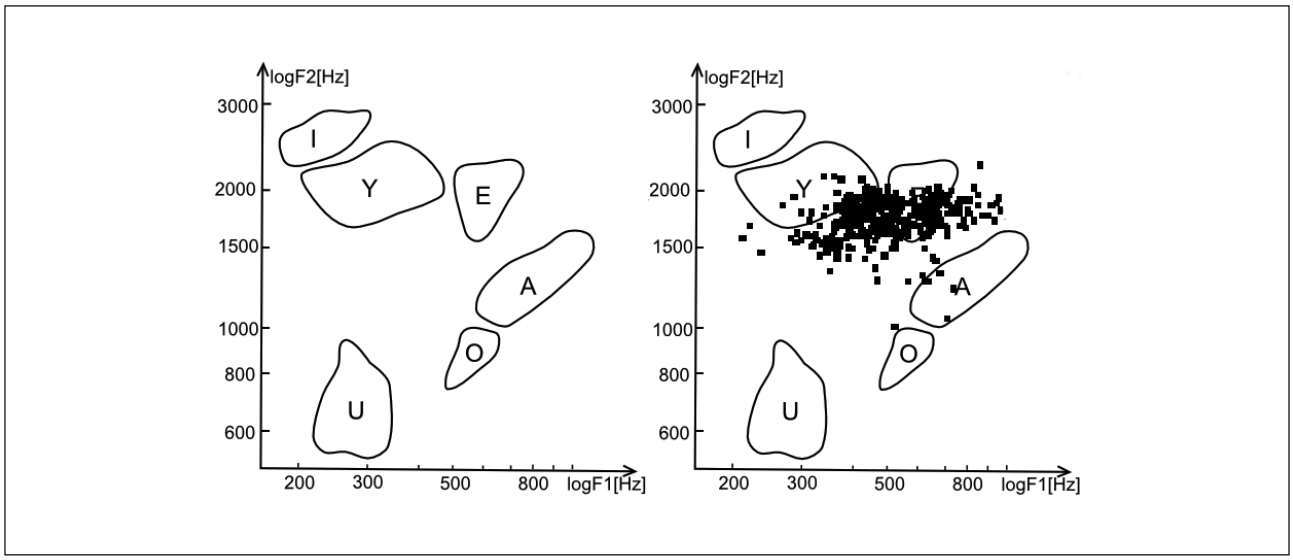

Figure 6. Polish vowels map on F1-F2 plane (by Tadeusiewicz), original figure from [11] and modified version with points corresponding to all disfluencies from training set.

the same level both for male and female voices.

Dependency graphs of the first two formants for all annotated segments of filler *I* were prepared for all speakers. A single point on the graph corresponds to a single disfluency. Cloud of the points corresponding to all disfluencies annotated in recordings of all speakers were presented on a Polish vowels map in the plane of formants F1 - F2 what is presented in Figure 6. [11]

Evaluation

To verify the method of parameterization, test were performed on a testing set (last 6 minutes of recording for each person). The training set and testing set were dependent (contained different recordings, but of the same speakers) on account of the plans of applying the results for system for speakers training which will be speaker-dependent and – as many speech racognition systems – will require training to adapt for the user.

Table 5. Results of testing algorithm.

	W1	W2	W3	M1	M2	M3	**Mean**
Number of correctly detected fillers per minute	7	6.25	4	0.5	1.25	3.75	**3.8**
Number of incorrectly detected fillers per minute	3	1.75	2.25	0.75	0.25	0.75	**1.46**
Number of non-detected fillers per minute	2.5	3.25	1.25	0.25	2	2.5	**1.96**
Ratio of correctly detected fillers to all fillers in signal [%]	82.3	65.8	76.1	66.7	38.5	60	**68**
Ratio of incorrectly detected fillers to all fillers in signal [%]	35.3	18.4	42.9	100	8	12	**36**

Algorithm consisted of following stages:

- tested speech signal was divided into frames (length 50 ms, with overlap 10 ms);
- values of F1 and F2 for each frame were obtained;
- the parameters were compared to pattern parameters;
- if parameters values are included in range of variability of pattern parameters, the frame is marked as potential part of filler;
- if length of adjacent frames marked as potential parts of fillers were longer than minimal length of filler from patterns base, a set of the frames was qualified as final filler;
- for each recognition of filler, beginning and end of the segment was obtained and compared to manually annotation of recording;
- quantity of correct recognition and recognition errors were counted separately for each speaker.

The results of fillers detection are shown in Table 5.

Mean ratio of correct recognition (68%) is sufficient but still has to be improved using additional information like detection of silence segment or using another classification method (e.g. k-nearest neighbours with treshold of confidence measure).

Application prototype

As a result of this work, prototype of an application that enables the detection of filled pauses in speech signal was prepared. The script was implemented in Matlab. According to observations, speakers are not aware of the fillers appearance in the speech. They realize the number of filled pauses after listening to the previously recorded speech. Ultimately, the application will run in real time so as to keep the speaker aware of the use of undesirable elements of speech and let him/her to correct them immediately. The interface will include the indicator that changes color (e.g. from gray to red) indicating the presence of undesirable filled pauses. The program detects occurring filled pauses, calculates their number in the duration time of training speech and saves the values corresponding to the successive training sessions, to let the speaker to analyze progress in the process of improving oratorical abilities. Following parameters are calculated:

- average number of the filled pauses per minute;
- average length of the filled pauses with a standard deviation [s];
- shortest and longest duration of the filled pauses [s];
- total length of the filled pauses in the entire utterance [s].

The application can be used not only as help in improving the speaking skills, but also its elements could be integrated with automatic speech recognition systems, supporting the elimination of those elements of the utterances which do not affect the meaning of the speech but may impede the recognition process.

Discussion

Polish language speakers often unconsciously put in their utterances elements called speech disfluencies. These are e.g. filled pauses, revisions, repetitions, prolongations, blocks. Authors carried out a survey which results indicated that for audience, the most adverse elements during listening to somebody's speech are disfluencies (next are: monotonous intonation, lack of preparation and interaction with listeners) and filled pauses are these disfluencies that have the most negative effect on reception of the speech. To make their utterances more attractive for listeners, speakers should improve their elocution skills. After review of different techniques used to detect speech disfluencies available in literature (based on temporal changes in the fundamental frequency, analysis of the first and second formant stability, feature vectors consisted of cepstral coefficients), authors implemented method based on formants F1 and F2 to detect filled pauses in speech signal.

Research material used in this work contained recordings of students lectures and interviews from radio programs, divided into training and testing set. Using training set, transcription and temporal annotation of speech disfluencies was prepared, as well as statistical analysis of occurring filled pauses in relation to the whole speech. Calculation of filled pause average time duration with standard deviation and average values of first two formants was made. Parameters obtained for the filled pauses from training set were used as a pattern base in detection algorithm. The effectiveness of the detection algorithm was examined on testing set (speaker-dependent, as a preliminary study), its measures were two ratios: ratio of the number of filled pauses detected correctly to all that signal contained (68%), and the ratio of incorrectly detected to all present in analyzed speech (36%). Authors proposed also a prototype of application which next to calculating parameters for speech signal, indicates occurrence of filled pause in real time of the speech, what let the speaker correct his utterance immediately.

Conclusions

The paper presents analysis of disfluencies in spontaneous speech, focusing on filled pauses. The significance of the issue for both speech technology systems (especially for automatic speech recognition) and improving public speaking skills, were presented. Also the statistics of occurrences of the main types of disfluencies was investigated. The results pointed on filled pauses, usually prolongated *I*, as most frequent disfluency (a half of all occurences). This type of disfluency was analyzed statistically and described with parameters covering its duration features and characterized by formants.

The algorithm of automatic detection of prolongated *I*, based on tracking stability of the first and second formant and segment duration, was implemented. Effectiveness of

the algorithm was about 68 % of true positives. The future works will be concentrated on improving the algorithm, in aspect of its precision and real-time performance. Authors consider tests of different classifiers (machine learning methods, k-nearest neighbor, affinity propagation), as well as enlargement of the database. We also plan to adapt the algorithm for speaker-independent performance.

The habit to abuse of such undesirable elements of speech can be relieved or eliminated through appropriate exercises for improving the diction. Methods of research leading to automatic detection of inarticulated speech elements proposed by the authors aim to provide the tools to gradually eliminate disfluencies and improve oratorical abilities through real-time indication of filled pause occurrence. Speech analysis for detection of disfluencies can be not only a tool that provides feedback to individuals who want to become better speakers, helpful in getting rid of unwanted habits of speech and improving communication during public speaking, but also can be a part of automatic speech recognition systems.

References

[1] Roberts P. M., Meltzer A., Wilding J., *Disfluencies in non-stuttering adults across sample lengths and topics*. Journal of Communication Disorders, 42 (2009) 414–427.

[2] Stouten F., Duchateau J., Martens J. P., Wambacq P., *Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation*. Speech Comunication, 48 (2006) 1590–1606.

[3] Myers F., Bakker K., St. Luis K. O., Raphael L. J., *Disfluencies in cluttered speech*. Journal of Fluency Disorders, 37 (2012) 9–19.

[4] O'Shaughnessy D., Gabrea M., *Automatic identification of filled pauses in spontaneous speech*. In Proc.: Canadian Conference on Electrical and Computer Engineering Conference, 2 (2000) 620-624.

[5] Audhkhasi K., Kandhway K., *Formant-based technique for automatic filled-pause detection in spontaneous spoken English*. In Proc.: International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.

[6] Stouten F., Martens J. P., *A feature-based filled pases detection system for Dutch*. In Materials of IEEE Workshop on Automatic Speech Recognition and Understanding, 2003.

[7] Ziolko B., Miga B., Jadczyk T.: *Semisupervised production of speech corpora using existing recordings*, International 24. Seminar on Speech Production (ISSP'11) , Montreal, 2011

[8] Wang X., Pols L.C.W, Ten Bosch L.F.M., *Analysis Of Context-Dependent Segmental Duration For Automatic Speech Recognition*, Proceedings ICSLP '94

[9] Ziółko B., Ziółko M., *Time durations of phonemes in Polish language for speech and speaker recognition*, Human language technology : challenges for computer science and linguistics : 4th Language and Technology Conference, LTC 2009 : Poznan, Poland, November 6–8, 2009.

[10] Boersma P., *Praat, A system for doing phonetics by computer*. Glot International 5:9/10, 341-345.

[11] Tadeusiewicz R., *Sygnał mowy*, WKiŁ, Warszawa, 1987.

[12] Ziółko M., Ziółko B., *Przetwarzanie mowy*, Wydawnictwa AGH, Kraków 2011.

[13] Ciota Z., *Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy*, Wyd. EXIT Warszawa 2010.