Multi-factor Method for Detection of Filled Pauses and Lengthenings in Russian Spontaneous Speech

Vasilisa Verkhodanova[1](✉) and Vladimir Shapranov[2]

[1] SPIIRAS, 39, 14th Line, St. Petersburg, Russia
verkhodanova@iias.spb.su
[2] Betria Systems Inc, 50, Building 11, Ligovsky Prospekt, St. Petersburg, Russia
equidamoid@gmail.com

Abstract. Spontaneous speech contains high rates of speech disfluencies, most common being filled paused and lengthenings (FPs). Human language technologies are often developed for other than spontaneous types of speech, and disfluencies occurrence is the reason for many mistakes in automatic speech recognition systems. In this paper we present a method of automatic detection of FPs using linear combination of statistical characteristics of acoustic parameters variance, basing on a preliminary study of FPs parameters across the mixed and quality-diverse corpus of Russian spontaneous speech. Experiments were carried out on a corpus, consisting of the task-based dialogue corpus of Russian spontaneous speech collected in SPIIRAS and on Russian casual conversations from Open Source Multi-Language Audio Database collected in Binghamton University.

Keywords: Speech disfluencies · Filled pauses · Sound lengthenings · Automatic speech processing · Russian spontaneous speech

1 Introduction

Almost all speech we produce and comprehend every day is spontaneous. This type of oral communication is likely to be one of the most difficult forms of speech communication among people: during very dense time interval speaker has to solve several laborious cognitive tasks. One has to form the utterance and to choose the exact linguistic form for it by selecting words, expressions, grammatical forms, etc. This process leads to different flaws in spontaneous speech production, so called speech disfluencies, like self-repairs, repetitions, filled pauses and lengthenings, slips of the tongue and other mispronunciations. These phenomena indicate the mental processes of underlying speech generation and have been viewed as a sign of word-searching problem [5] or difficulties in conceptualization at major discourse boundaries [3]. Speech disfluencies are breaks or irregularities that occur within the flow of otherwise fluent speech. These are self-repairs, repetitions, filled pauses, lengthenings etc. There are evidence that

A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 285–292, 2015.
DOI: 10.1007/978-3-319-23132-7_35

they can affect up to one third of utterances [20]: for example, about 6 per 100 words are disfluent in conversational speech in American English [20,25].

In Russian speech filled pauses occur at a rate of about 4 times per 100 words, and at approximately the same rate inside clauses and at the discourse boundaries [11]. Though evidence on filled pauses differs across languages, genres, and speakers, it can be summarized that on average there are several filled pauses per 100 syllable [17]. They also are most frequent speech disfluencies: filled pauses occur more often than any other speech disfluencies (repetitions, word truncations, etc.) [17], signaling not only of breaks in speech production process, but also of explication of this process [11].

The need in coping automatically with speech disfluencies appeared along with the need of spontaneous speech processing, what brought up a lot of interesting challanges to speech science and engineering. Once seen as errors, along with other disfluencies filled pauses and lengthenings (jointly referred to as FPs in the rest of the paper) were acknowledged as integral part of natural conversation [11,21]. They may play a valuable role such as helping a speaker to hold a conversational turn or expressing the speaker's thinking process of formulating the upcoming utterance fragment [4,18,21]. According to [24] in the conversational Switchboard database [8], about 39.7 % of the all disfluencies contain a filled pause. Thus, the detection of vowel lengthening and filled pauses could be an important step towards locating the disfluent regions and evaluating the spoken fluency skills of a speaker. The problem of detecting filled pauses has been addressed from various perspectives. In computational linguistics speech disfluencies analysis is sometimes incorporated into syntactic parsing and language comprehension systems [6], as well as into automatic speech recognition systems [15]. However, FPs as well as other speech disfluencies, were always an obstacle for automatic processing of spontaneous speech as well as its transcriptions, because speech recognition systems are usually trained on the structured data without speech disfluencies, what decreases speech recognition accuracy and leads to inaccurate transcriptions [2,9,16,21]. The interest in automatic detection of fillers also has been raised by INTERSPEECH Computational paralinguistic challenge in 2013 [10]. Nowadays the most efficient methods coping with FPs can be roughly divided into those, that use only acoustic parameters and [2,9,16,26] and those, that combine language and acoustic modeling with the purpose of incorporating them into automatic speech recognition systems [13–15]. This division is caused by the unavailability of resources: the application of disfluencies detecting methods based on language modeling requires a large corpus with rich transcriptions, while for approaches based on acoustic properties there is no such need.

In this study we describe a method based on the gradient decent aimed to find the values of acoustic parameters of FPs that would maximize the harmonic mean of precesion and recall (F_1score) for FPs detection for Russian spontaneous speech. The results of the experiments based on a mixed different quality corpus of Russian spontaneous speech are also presented.

2 Material

The material we have used in this study consists of two parts, one being the corpus of task-based dialogs (2/3 of the corpus) collected at SPIIRAS in 2012–2013 [27], and the other one being 5 casual conversations recordings (1/3 of the corpus) taken from the Russian part of Open Source Multi-Language Audio Database collected in Binghamton University in 2010–2012 [28]. The whole dataset we used for the experiments is about 1.5 h with 17 speakers, 8 men and 9 women.

First corpus was collected specially for analysis of speech disfluencies in Russian. The task methodology was chosen to elicit speech as close as possible to normal conversational spontaneous speech. This approach towards spontaneous speech eliciting is well-known and used. For example the HCRC corpus collected Edinburgh and Glasgow consists only of map-task dialogs [1], and half Kiel corpus of German speech consists of appointment tasks [12]. We consider the recorded speech to be spontaneous since it is informal and unrehearsed and also it is the result of direct dialogue communication [29]. In recording the two types of tasks were used: map tasks and appointment tasks. Map task dialogs represent a description of a route from start to finish, basing on the maps with different sometimes unmatched landmarks. This task was fulfilled twice by pair of speakers switching roles. In appointment task dialogs, a pair of participants tried to find a common free time for a telephone talk and for a meeting based on their individual schedules. Participants could not see maps or schedules of each other. Due to maps and schedules structure participants had to ask questions, interrupt and discuss the route or possible free time, what resulted in higher rates of FPs and artefacts. The recorded corpus consists of 18 dialogs from 1.5 to 5 min. All the recordings were made in St. Petersburg in the end of 2012 - beginning of 2013 in the sound isolated room by means of two tablet PCs Samsung Galaxy Tab 2 with Smart Voice Recorder. Participants were students: 6 women speakers and 6 men speakers from 17 to 23 years old with technical and humanitarian specialization. Corpus was manually annotated into speech disfluencies, with 222 filled pauses and 270 sound lengthenings.

The second part of the corpus we used is part of Multi-Language Audio Database [28]. This database consists of approximately 30 h of sometimes low quality, varied and noisy speech in each of three languages, English, Mandarin Chinese, and Russian. For each language there are 900 recordings taken from open source public web sites, such as http://youtube.com. All recordings have been orthographically transcribed at the sentence/phrase level by human listeners. The Russian part of this database consists of 300 recordings of 158 speakers (approximately 35 h). The casual conversations part consists of 91 recordings (10.3 h) of 53 speakers [28]. From this Russian part we have taken the random 5 recordings of casual conversations (3 female speakers and 2 male speakers) that were manually annotated into FPs. The number of annotated phenomena is 266 (186 filled pauses and 80 sound lengthenings).

3 Method for Automatic FPs Detection

We have based our method on acoustical features of FPs that are peculiar to these events in Russian. We used gradient decent method to get optimal parameters to maximize the F_1 score for FPs detection.

Acoustic features of hesitation pauses in Russian speech was studied in [23], where author has found that filled pauses in Russian differ in terms of F1 and F2 values from vowels in stressed positions. In our study we have analyzed duration, F0, three first formants, energy and stableness of spectra across our corpus. Similar approaches have been applied for FPs detection in other languages and proved the relevancy of these acoustic properties [2,7,9].

In the literature the most commonly observed feature of FPs is the long duration [9,11,20,26]. In our corpus the average duration of FPs is 400ms (minimum and maximum durations of FPs are 129 ms and 2.3 s respectively). 87 % of the whole set of 758 FPs are longer than 200 ms (Fig. 1).

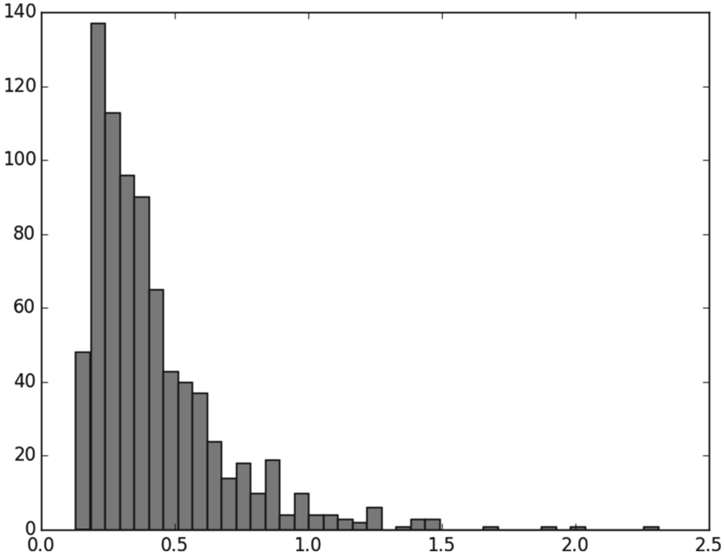

Fig. 1. The distribution of FPs duration

Another prominent feature of FPs is a gradual fall of fundamental frequency (F0). This tendency is a well-known fact, in [19] it has been shown that FPs tend to be low in F0 as well as displaying a gradual, roughly linear F0 fall [19]. To compare characteristics of FPs with surrounding signal, we have manually annotated the left and right context of each phenomena, marking at least one word (or two-three words) to gain minimum two syllables at either side of disfluency. We have checked the standard deviation for F0 and first three formants and the most obvious case of small variance of standard deviation in FPs were for F0 and energy (Figs. 2 and 3).

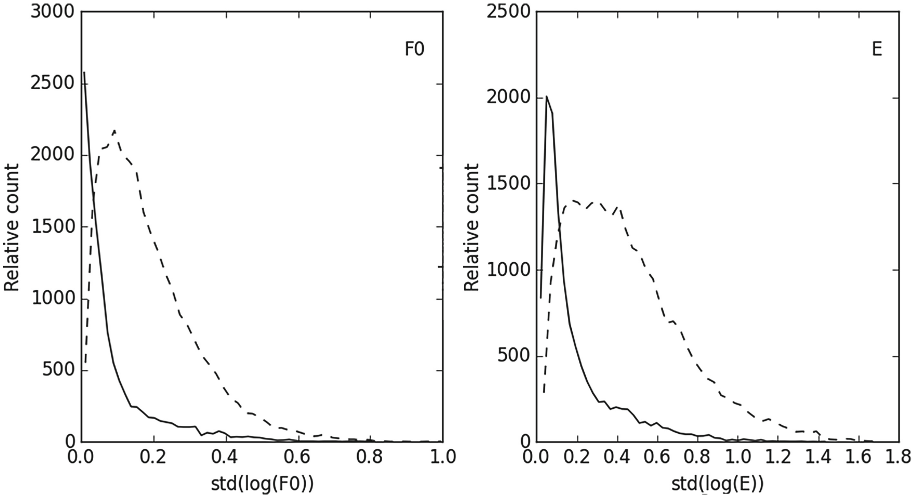

Fig. 2. The standard deviation of the logarithms of F0 (left) and energy (right) of FPs (thick line) and of neighboring words and phrases (dashed line).

The proposed algorithm has several parameters that should be optimized: w_n that are weights for stardard deviations of $log(E)$ and $log(F_N)$ and E_0 that is a minimal mean energy level. The following steps were performed:

- Calculate the standard deviations and means for the logarithms of F0, F1 and energy in 150-ms windows
- Obtain the optimal values of parameters w_n and E_0 for criterion $C = \sum_n w_n V_n < 1; E > E_0$ where w_n are weights for values V_n: stardard deviations of $log(E)$ and $log(F_N)$ and E_0 is a minimal mean energy level. The optimal values are those that maximize F_1score for the task of selection of 150ms windows that are part of the FPs.
- Find the consecutive intervals matching the criterion C.
- Compare the intervals with annotation.
- Calculate $F_1\text{score} = \frac{2\cdot\text{true positive}}{2\cdot\text{true positive}+\text{false negative}+\text{false positive}}$

The parameters were then optimized using gradient descent method [22].

At the stage of comparison with annotation, the intervals intersecting with the labeled ones are found. Here we calculate the intersection length $T_{int} = \mathbf{len}(I \cap L)$ and length of non-matching part of the interval $T_{ext} = \mathbf{len}(I \setminus S)$, where I is interval and L is label. If $T_{int} > 0.2\mathbf{len}(L)$ and $3T_{ext} < T_{int}$ the pair of label and interval is considered matching. After processing the whole signal the amount of non-matched intervals is considered false positive count and the amount of non-matched labels is considered false negative count.

The experiments were conducted on 85 % of the corpus with 15 % used as a test-set. The test-set was randomly taken from the whole corpus and the results were averaged. The performance was compared to the annotation of 758 FPs. The obtained F_1score is 0.46. For two parameters (F0 and Energy) and with grid search instead of gradient decent the F_1 score was 0.43.

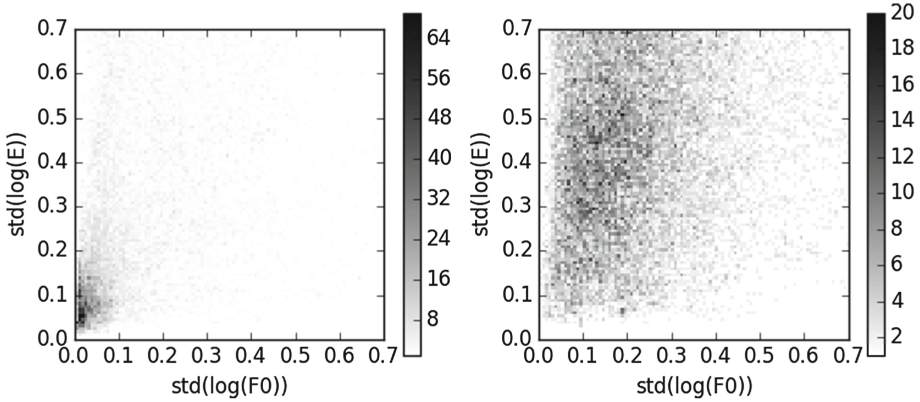

Fig. 3. The heatmap of standard deviation of the logarithms of energy and F0 of FPs (left) and of neighboring words and phrases (right).

The misses were mainly caused by the disorder of harmonic components in hoarse voice and the laryngealized filled pauses and lengthenings. In some cases the FP has an unstable expressive intonation contour, which was not flat or lowering, that it can be argued whether they are FPs or interjections. Few cases of misses were the result of small duration of annotated phenomena. Some false alarms were caused by lengthenings that were missing in the annotation. And noises (especially in the part from the open source multi-language database) and overlappings (in task dialogs part) caused number of false negatives.

4 Conclusions

This paper presents the method based on acoustic parameters of FPs in Russian spontaneous speech, that maximizes the F_1score for FPs detection by the gradient decent method. As the acoustic parameters we have used standard deviations of F0, F1, F2, F3 and energy. This gave us F_1score of 0.46, what, comparing to F_1score obtained with only two parameters and grid search is a small improvement. This could be due to the challenging quality data (in the part from the open source multi-language database) or because of small influence of F1, F2 and F3 on the FPs detection for our data, that should be tested, since it is not comply with studies on other languages. We also plan to use non-linear expression to separate FPs from other signal to test whether it would improve the obtained F_1score. The results of this study are also a step towards building a complex and reliable FPs search method, that could be used on data of different recording quality and conditions.

Acknowledgements. This research is supported by the grant of Russian Foundation for Basic Research (project No 15-06-04465) and by the Council of Grants of the President of Russia (project No MK-5209.2015.8).

References

1. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., et al.: The hcrc map task corpus. Lang. Speech **34**(4), 351–366 (1991)
2. Audhkhasi, K., Kandhway, K., Deshmukh, O., Verma, A.: Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 4857–4860. IEEE (2009)
3. Chafe, W.L. (ed.): The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Ablex, Norwood (1980)
4. Clark, H.: Using Language. Cambridge University Press, Cambridge (1996)
5. Eisler, F.G.: Psycholinguistics: Experiments in Spontaneous Speech. Academic Press, New York (1968)
6. Ferreira, F., Lau, E.F., Bailey, K.G.: Disfluencies, language comprehension, and tree adjoining grammars. Cogn. Sci. **28**(5), 721–749 (2004)
7. Garg, G., Ward, N.: Detecting filled pauses in tutorial dialogs (2006)
8. Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: telephone speech corpus for research and development. In: 992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92, vol. 1, pp. 517–520. IEEE (1992)
9. Goto, M., Itou, K., Hayamizu, S.: A real-time filled pause detection system for spontaneous speech recognition. In: Eurospeech, Citeseer (1999)
10. INTERSPEECH: computational paralinguistic challenge in 2013. http://emotion-research.net/sigs/speech-sig/is13-compare. Accessed 1 Apr 2015
11. Kibrik, A., Podlesskaya, V. (eds.): Rasskazy o snovideniyah: Korpusnoye issledovaniye ustnogo russkogo diskursa [Night dream stories: Corpus study of Russian discourse]. Litres (2014)
12. Kohler, K.: Labelled data bank of spoken standard german: the kiel corpus of read/spontaneous speech. In: Proceedings of Fourth International Conference on Spoken Language, ICSLP 96, vol. 3, pp. 1938–1941. IEEE (1996)
13. Lease, M., Johnson, M., Charniak, E.: Recognizing disfluencies in conversational speech. IEEE Trans. Audio, Speech Lang. Process. **14**(5), 1566–1573 (2006)
14. Liu, Y., Shriberg, E., Stolcke, A.: Automatic disfluency identification in conversational speech using multiple knowledge sources. In: 8th European Conference on Speech Communication and Technology Proceedings, INTERSPEECH, pp. 957-960 (2003)
15. Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M.: Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Trans. Audio, Speech Lang. Process. **14**(5), 1526–1540 (2006)
16. Medeiros, H., Moniz, H., Batista, F., Trancoso, I., Nunes, L., et al.: Disfluency detection based on prosodic features for university lectures. In: 14th Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2629–2633 (2013)
17. O'Connell, D., Kowal, S.: The history of research on the filled pause as evidence of the written language bias in linguistics (linell, 1982). J. Psycholinguist. Res. **33**, 459–474 (2004)
18. Ogden, R.: Turn-holding, turn-yielding and laryngeal activity in finnish talk-in-interaction. J. Int. Phonetics Assoc. **31**(1), 139–152 (2001)
19. O'Shaughnessy, D.: Recognition of hesitations in spontaneous speech. In: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92, vol. 1, pp. 521–524. IEEE (1992)

20. Shriberg, E.: Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkeley (1994)
21. Shriberg, E.: Spontaneous speech: how people really talk and why engineers should care. In: 9th European Conference on Speech Communication and Technology, INTERSPEECH, pp. 1781–1784 (2005)
22. Kober, J., Peters, J.: Introduction. In: Kober, J., Peters, J. (eds.) Learning Motor Skills. STAR, vol. 97, pp. 1–6. Springer, Heidelberg (2014)
23. Stepanova, S.: Some features of filled hesitation pauses in spontaneous Russian. Proc. ICPhS. **16**, 1325–1328 (2007)
24. Stolcke, A., Shriberg, E., Bates, R.A., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., Lu, Y.: Automatic detection of sentence boundaries and disfluencies based on recognized words. In: ICSLP (1998)
25. Tree, J.E.F.: The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. J. Mem Lang. **34**(6), 709–738 (1995)
26. Veiga, A., Candeias, S., Lopes, C., Perdigão, F.: Characterization of hesitations using acoustic models. In: International Congress of Phonetic Sciences-ICPhS XVII, pp. 2054–2057 (2011)
27. Verkhodanova, V., Shapranov, V.: Automatic detection of filled pauses and lengthenings in the spontaneous russian speech. In: 7th Speech Prosody conference, pp. 1110–1114 (2014)
28. Zahorian, S.: Open-source multi-language audio database for spoken language processing applications. Technical report, DTIC Document (2012)
29. Zemskaya, E.: Russian spoken speech: linguistic analysis and the problems of learning. Moscow (1979)