Detecting Filled Pauses and Lengthenings in Russian Spontaneous Speech Using SVM

Vasilisa Verkhodanova(✉) and Vladimir Shapranov

SPIIRAS, St. Petersburg, Russia
verkhodanova@iias.spb.su, equidamoid@gmail.com

Abstract. Spontaneous speech differs from any other type of speech in many ways. And the presence of speech disfluencies is its prominent characteristic. These phenomena are important feature in human-human communication and at the same time a challenging obstacle for the speech processing tasks. This paper reports the experiment results on automatic detection of filled pauses and sound lengthenings basing on the automatically extracted acoustic features. We have performed machine learning experiments using support vector machine (SVM) classifier on the mixed and quality diverse corpus of Russian spontaneous speech. We applied Gaussian filtering and morphological opening to post-process the probability estimates from an SVM classifier. As the result we achieved F1–score of 0.54, with precision and recall being 0.55 and 0.53 respectively.

Keywords: Speech disfluencies · Filled pauses · Lengthenings · Speech processing · Support vector machines

1 Introduction

Speech disfluencies are common in spontaneous speech. They consist of hesitations, self-repairs, repetitions, substitutions, etc. They do not add the semantic information to the speech signal, but may play a valuable role such as helping a speaker to hold a conversational turn or expressing the speakers thinking process of formulating the upcoming utterance fragment [4,17].

However human language technologies are often developed for other than spontaneous type of speech, and the occurrence of disfluencies is one of the factors that makes the spontaneous speech processing challenging [19]. The need of detecting them automatically emerged mainly from the problems of automatic speech recognition (ASR): disfluencies are known to have an impact on ASR results, they can occur at any point of spontaneous speech, thus they can lead to misrecognition or incorrect classification of adjacent words.

Speech disfluencies occur quite often: for example, in conversational speech in American English about 6 per 100 words are disfluent [21]. Though evidence on filled pauses and lengthenings (jointly reffered as FPs later on) differs across languages, genres, and speakers, on average there are several disfluencies per 100 syllables, FPs being the most frequent disfluency type [16]. According to [25] in

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 224–231, 2016.
DOI: 10.1007/978-3-319-43958-7_26

the conversational Switchboard database [7], about 39.7 % of the all disfluencies contain a filled pause. In the corpus of Portuguese lectures LECTRA filled pauses correspond to 1.8 % of all the words and to 22.9 % of all disfluency types being the most frequent type in the corpus [14]. In Russian speech FPs occur at a rate of about 4 times per 100 words, they also occur at approximately the same rate inside clauses and at the discourse boundaries [13].

FPs exhibit universal as well as linguistic and genre specific features. FPs are represented mainly by vocalizations with rare cases of prolonged consonants (which was shown to be a peculiarity of Armenian hesitational phenomena [12]). These vocalizations in FPs are usually phonetically different from the lexical items, since they are pronounced with minimal movements of the articulatory organs due to the articulatory economy [24]. However, it was also shown that phonological system of the language may influence the quality of FPs vocalizations [6]. Even universal characteristics of FPs, such as lengthenings being accompanied by creaky voice, may operate differently in different languages: e.g. in Finnish it was proposed that creaky voice may indicate turn-transitional locations [17], which is not the case for English [22].

Although the speech technologies, and particularly ASR systems, have to account for all types of disfluencies (filled pauses, prolongations, repetitions, deletions, substitutions, fragments, editing expressions, insertions, etc.), in the present study we focus on the detection of the most frequent disfluent categories: filled pauses and sound lengthenings. In this paper we present the results of experiments on detection of FPs on the mixed and quality diverse corpus of Russian spontaneous speech. We used an SVM classifier and two methods applied at the stage of post-processing: Gaussian filtering and morphological opening.

2 Related Work

During last years speech disfluencies and particularly FPs have received more attention in the field of speech processing due to speech recognition tasks: ASR systems are usually trained on the structured data without any types of speech disfluencies.

It has been shown that along with duration the prominent characteristic of FPs is a gradual fall of fundamental frequency (F0) [18]: FPs tend to be low in F0 as well as displaying a gradual, roughly linear F0 fall. In [23] it was shown that for fair detection of FPs these two characteristics and distance to a pause are enough. In [28] authors used duration and statistical characteristics of F0, first three formants and energy for the experiments based on gradient decent optimizing parameters for maximization of F1–score; achieved result was F1–score = 0.46.

In [8], filled pauses are detected on a basis of two features (small fundamental frequency transition and small spectral envelope deformation) using as material 100 utterances extracted from a Japanese spoken language corpus where 91.5 % precision and 84.9 % recall were achieved. In [26] authors developed a detection system in order to improve the speech recognizer performance, achieving precision of 85 % at a recall rate of 70 %.

In [15] authors focused on detection of filled pauses basing on acoustic and prosodic features as well as on some lexical features. Experiments were carried on a speech corpus of university lectures in European Portuguese Lectra. Several machine learning methods have been applied, and the best results were achieved using Classification and Regression Trees. The performance achieved for detecting words inside of disfluent sequences was about 91 % precision and 37 % recall, when filled pauses and fragments were used as a feature, without it, the performance decayed to 66 % precision and 20 % recall. Further experiments on filled pauses detection in European Portuguese were carried out using prosodic and obtained from ASR lexical features; the best results were achieved using J48, corresponding to about 61 % F-measure [14].

The INTERSPEECH 2013 Paralinguistic Challenge [11] raised interest in automatic detection of fillers providing a standardised corpus and a reference system. The winners of the Social Signals Sub-Challenge introduced a system, built upon a DNN classifier complemented with time series smoothing and masking [9].

In [20] authors presented a method for filled pauses detection using an SVM classifier, applying a Gaussian filter to infer temporal context information and performing a morphological opening to filter false alarms. For the feature set authors used the same as was proposed for [11], extracted with the openSMILE toolkit [5]. Experiments were carried out on the LAST MINUTE corpus of naturalistic multimodal recordings of 133 German speaking subjects in a so called Wizard-of-Oz (WoZ) experiment. The obtained results were recall of 70 %, precision of 55 %, and AUC of 0.94.

3 Material

The material we have used in this study consists of several parts. The first part is the corpus of task-based dialogs collected at SPIIRAS in St. Petersburg in the end of 2012 - beginning of 2013 [27]. It consists of 18 dialogs from 1.5 to 5 min, where people in pairs fulfilled map and appointment tasks. Recording was performed in the sound isolated room. Participants were students: 6 women and 6 men from 17 to 23 years old with technical and humanitarian specialization. Recordings were annotated manually into different types of disfluencies, the FPs being the majority - 492 phenomena (222 filled pauses and 270 lengthenings).

For the second part of our material we used part of Multi-Language Audio Database [29]. 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 hours). The casual conversations part consists of 91 recordings (10.3 h) of 53 speakers [29]. From this Russian part we have taken the random 6 recordings of casual conversations (3 female speakers and 3 male speakers) that

were manually annotated into FPs. The number of annotated phenomena is 284 (188 filled pauses and 96 sound lengthenings).

The third part is the corpus of scientific reports from seminar devoted to analysis of conversational speech held at SPIIRAS in 2011. Recordings of reports of 6 people (3 female and 3 male speakers) were manually annotated into speech disfluencies. Since speakers didn't base their reports on a written text, these recordings ontain considerable amount of speech disfluencies. 951 FPs were manually annotated: 741 filled pauses and 210 lengthenings.

Another part we added for making our corpus more quality and situation diverse is the the records from the appendix No5 to the phonetic journal "Bulletin of the Phonetic Fund" belonging to the Department of Phonetics of Saint-Petersburg University [1]. The 12 recorded reports concerned different scientific topics (linguistics, logic, psychology, etc.). They were all recorded in 70s-80s in Moscow except one that was recorded in Prague. All speakers (6 men and 6 women) were native Russian speakers, and were recorded while presenting on conferences and seminars. The number of manually annotated FPs is 285 (225 filled pauses and 60 lengthenings).

In total, the data set we used is about 3 h and comprises 2012 filled pauses. Distribution of FP duration over the corpus is shown on Fig. 1. The duration of a single FP lies between 6 ms and 2.3 s, the average duration is 388 ms.

The data has been separated into two classes: "FPs" and "Other". First one consists of FPs only, while the other comprises the rest of the frames. Each 10th file was selected for train set, then again each 10th - for development set, and the rest was used as the test set. This operation was performed 10 times producing 10 different triplets of train, development and test sets.

Since the classes were not balanced (there were about 12 times more Other instances than FPs ones) we downsampled the train set to avoid the bias towards the class Other [20]. Thus we created subset containing randomly chosen 8 % of the instances of the class Other and all the FPs data. To train the classifier we use these downsampled training set.

4 Filled Pauses and Lengthenings Detection

In our study we have followed [20], basing our experiments on support vector machine (SVM) classifier. The extreme learning machines (ELM) in our unreported study shows that SVM provides better detection accuracy with better harmonized mean of precision and recall (with ELM we got F1–score = 0.4). We used a scikit-learn Python library [2] implementation of SVM with polynomial kernel, that enables the probability estimates by means of C-Support Vector Classification, the implementation is build upon libsvm [3].

The feature set is based in the set that was used for the INTERSPEECH 2013 Social Signals Sub-Challenge [11]. Features were extracted with the openSMILE toolkit [5] on the frame-level basis (25 ms window, 10 ms shift). This set is derived from 54 low-level descriptors (LLDs): 14 mel-frequency cepstral coefficients (MFCCs), logarithmic energy as well as their first and second order delta

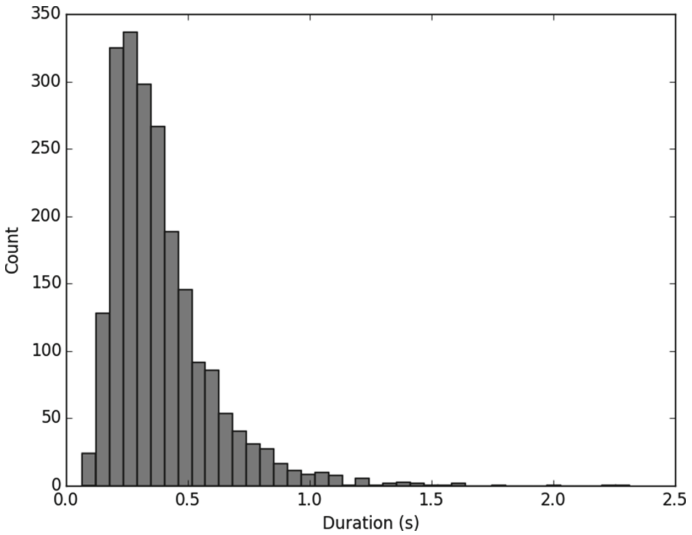

Fig. 1. The distribution of FPs duration

and acceleration coefficients; there are also voicing probability, F0 and zero-crossing rate, together with their deltas. For each frame-wise LLD the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after) are used as the actual features. As the result we have in 162 values per frame.

After training our SVM classifier, as the post-processing step we applied Gaussian filter and morphological opening [10,20] that proved to be reasonably efficient for improving both precision and recall rates due to the usage of contextual information. Both these techniques are applied in the signal and image processing tasks for noise removal. Gaussian filter was used to smooth the spikes and remove the outliers on the probability estimates. Morphological opening is proved to be useful for making the detection of FPs more balanced by filtering false alarms and improving F1–score [20]. The parameters for Gaussian and morphological opening, as well as the decision threshold were determined using grid search on the development set.

The Gaussian filter allows us ahieve 12 % improvement for F1–score (precision rate improving by 17 % and recall rate by 5 %). Morphological opening gave us only 2 % improvement for F1–score, precision and recall, reducing false alarm rate. The example of dependence of results from varying decision threshold on SVM output is shown on Fig. 2.

As the result we achieved F1–score = F1–score = 0.54 ± 0.027, with precision and recall being 0.55 ± 0.05 and 0.53 ± 0.04 respectively. Measures on the test set are reported in terms of mean and standard deviation over the ten evaluations using classifiers trained on ten training subsets.

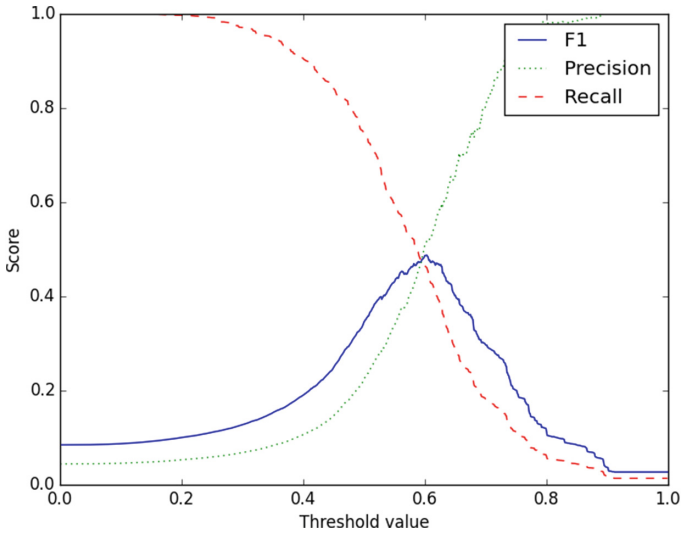

Fig. 2. The dependence of results from decision threshold

5 Conclusion

In this paper we present the experimental results on the detection of filled pauses and lengthenings in Russian spontaneous speech. For the purposes of this study we have united several corpora, that differ in quality, recording sites and situations, into one corpus that was used for the experiments. We used an SVM classifier and applied Gaussian filtering and morphological opening to post-process the probability estimates from an SVM, since these two techniques make use of contextual information. As the result we achieved F1–score $= 0.54 \pm 0.027$, with precision and recall being 0.55 ± 0.05 and 0.53 ± 0.04 respectively. The future work will be aimed at addressing the remaining problem of false positives and false negatives by tuning SVM and introducing more contextual information.

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

References

1. Department of Phonetics of Saint Petersburg University. http://phonetics.spbu.ru/
2. Scikit-Learn: Machine Learning in Python. http://scikit-learn.org
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**, 1–27 (2011). http://www.csie.ntu.edu.tw/cjlin/libsvm
4. Clark, H.: Using Language. Cambridge University Press, Cambridge (1996)

5. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International conference on Multimedia, pp. 1459–1462. ACM (2010)
6. Giannini, A.: Hesitation phenomena in spontaneous italian. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, pp. 2653–2656 (2003)
7. Godfrey, J.J., Holliman, E.C., McDaniel, J.: SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992), vol. 1, pp. 517–520. IEEE (1992)
8. Goto, M., Itou, K., Hayamizu, S.: A real-time filled pause detection system for spontaneous speech recognition. In: Proceedings of the Eurospeech, Budapest, Hungary, pp. 227–230. ISCA (1999)
9. Gupta, R., Audhkhasi, K., Lee, S., Narayanan, S.: Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In: Proceedings of the INTERSPEECH 2013, Lyon, France, pp. 173–177. ISCA (2013)
10. Heijmans, H.J.: Mathematical morphology: a modern approach in image processing based on algebra and geometry. SIAM Rev. **37**(1), 1–36 (1995)
11. INTERSPEECH: Computational Paralinguistic Challenge (2013). http://emotion-research.net/sigs/speech-sig/is13-compare
12. Khurshudian, V.: Hesitation in typologically different languages: An experimental study. In: Proceedings of the International Conference on Computational Linguistics Dialogue, pp. 497–501 (2005)
13. Kibrik, A., Podlesskaya, V. (eds.): Rasskazy o Snovideniyah: Korpusnoye Issledovaniye Ustnogo Russkogo Diskursa [Night dream stories: Corpus study of Russian discourse]. Litres (2014)
14. Medeiros, H., Batista, F., Moniz, H., Trancoso, I., Meinedo, H.: Experiments on automatic detection of filled pauses using prosodic features. In: Actas de Inforum 2013, pp. 335–345 (2013)
15. Medeiros, H., Moniz, H., Batista, F., Trancoso, I., Nunes, L., et al.: Disfluency detection based on prosodic features for university lectures. In: Proceedings of the INTERSPEECH 2013, Lyon, France, pp. 2629–2633 (2013)
16. O'Connell, D., Kowal, S.: The history of research on the filled pause as evidence of the written language bias in linguistics. J. Psycholinguist. Res. **33**(6), 459–474 (2004)
17. Ogden, R.: Turn-holding, turn-yielding and laryngeal activity in finnish talk-in-interaction. J. Int. Phonetics Assoc. **31**(1), 139–152 (2001)
18. O'Shaughnessy, D.: Recognition of hesitations in spontaneous speech. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 1992), vol. 1, pp. 521–524. IEEE (1992)
19. Ostendorf, M., Shriberg, E., Stolcke, A.: Human Language Technology: Opportunities and Challenges. Technical report, DTIC Document (2005)
20. Prylipko, D., Egorow, O., Siegert, I., Wendemuth, A.: Application of image processing methods to filled pauses detection from spontaneous speech. In: Proceedings of the INTERSPEECH 2014, Singapore, pp. 1816–1820. ISCA (2014)
21. Shriberg, E.: Spontaneous speech: how people really talk and why engineers should care. In: Proceedings of the INTERSPEECH 2005, Lisbon, Portugal, pp. 1781–1784. ISCA (2005)
22. Shriberg, E.: To 'Errrr' is human: Ecology and acoustics of speech disfluencies. J. Int. Phonetic Assoc. **31**(1), 153–169 (2001)

23. Shriberg, E., Bates, R.A., Stolcke, A.: A prosody only decision-tree model for disfluency detection. In: Proceedings of the 5th European Conference on Speech Communication and Technology Eurospeech 1997, Rhodes, Greece, pp. 2383–2386 (1997)
24. Stepanova, S.: Some features of filled hesitation pauses in spontaneous Russian. In: Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrucken, Germany, vol. 16, pp. 1325–1328 (2007)
25. 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)
26. Stouten, F., Martens, J.P.: A feature-based filled pause detection system for Dutch. In: Workshop on Automatic Speech Recognition and Understanding, ASRU 2003, pp. 309–314. IEEE (2003)
27. Verkhodanova, V., Shapranov, V.: Automatic detection of filled pauses and lengthenings in the spontaneous russian speech. In: Proceedings of the 7th International Conference Speech Prosody, Dublin, Ireland, pp. 1110–1114 (2014)
28. Verkhodanova, V., Shapranov, V.: Multi-factor method for detection of filled pauses and lengthenings in russian spontaneous speech. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 285–292. Springer, Heidelberg (2015)
29. Zahorian, S.A., Wu, J., Karnjanadecha, M., Vootkur, C.S., Wong, B., Hwang, A., Tokhtamyshev, E.: Open-source multi-language audio database for spoken language processing applications. In: Proceedings of the INTERSPEECH 2011, Florence, Italy, pp. 1493–1496 (2011)