Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 2013658, 8 pages
http://dx.doi.org/10.1155/2016/2013658

Research Article

Experiments on Detection of Voiced Hesitations in Russian Spontaneous Speech

Vasilisa Verkhodanova and Vladimir Shapranov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

Correspondence should be addressed to Vasilisa Verkhodanova; interiora@gmail.com

Received 31 July 2016; Accepted 31 October 2016

Academic Editor: Alexander Petrovsky

The development and popularity of voice-user interfaces made spontaneous speech processing an important research field. One of the main focus areas in this field is automatic speech recognition (ASR) that enables the recognition and translation of spoken language into text by computers. However, ASR systems often work less efficiently for spontaneous than for read speech, since the former differs from any other type of speech in many ways. And the presence of speech disfluencies is its prominent characteristic. These phenomena are an important feature in human-human communication and at the same time they are a challenging obstacle for the speech processing tasks. In this paper we address an issue of voiced hesitations (filled pauses and sound lengthenings) detection in Russian spontaneous speech by utilizing different machine learning techniques, from grid search and gradient descent in rule-based approaches to such data-driven ones as ELM and SVM based on the automatically extracted acoustic features. Experimental results on the mixed and quality diverse corpus of spontaneous Russian speech indicate the efficiency of the techniques for the task in question, with SVM outperforming other methods.

1. Introduction

Speech technologies are often developed for different types of speech and rarely for a spontaneous one. However, 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, and so on. This process leads to different flaws in spontaneous speech production, so-called speech disfluencies. These are self-repairs, repetitions, voiced hesitations (filled pauses and lengthening that are often referred together as FPs), slips of the tongue, and other breaks or irregularities that occur within the flow of otherwise fluent speech. These phenomena indicate the mental processes of underlying speech generation and have been viewed as a sign of word-searching problem [1] or difficulties in conceptualization at major discourse boundaries [2]. There is evidence that they can affect up to one-third of utterances [3]; for example, in conversational speech in American English, about 6 per 100 words are disfluent [3, 4].

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 [5]. Though evidence on filled pauses differs across languages, genres, and speakers, on average there are several filled pauses per 100 syllables [6]. They are also the most frequent speech disfluencies; filled pauses occur more often than any other speech disfluencies (repetitions, word truncations, etc.) [6], signalling not only of breaks in speech production process, but also of explication of this process [5]. According to [7] in the conversational Switchboard database [8], 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 [9].

The need in coping automatically with speech disfluencies appeared along with the need of spontaneous speech processing, which brought up a lot of interesting challenges to speech science and engineering. Once seen as errors, along with other disfluencies, hesitations were acknowledged as integral part of natural conversation [5, 10]. They 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 [10–12]. The comparison of prosodic patterns of stutterers and nonstutterers disfluencies was done in [13], where authors analysed spontaneous storytelling of 8 people: four stutterers and four nonstutterers. Results, as expected, showed that stutterers have significantly more disfluencies than nonstutterers and that disfluencies affected the adjacent tonal contexts and phrasing differently in these two groups of people, stutterers' disfluencies being accompanied by more prosodic irregularities. Details can be found in [13]. 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 hesitations has been addressed from various perspectives. In computational linguistics analysis of speech disfluencies is sometimes incorporated into syntactic parsing and language comprehension systems [14] and more often into automatic speech recognition systems [15]. Hesitations, as well as other speech disfluencies, were always an obstacle for automatic processing of spontaneous speech as well as its transcriptions; 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 [9, 10, 16, 17].

Hesitations exhibit universal as well as linguistic and genre specific features. Filled pauses and lengthenings are represented mainly by vocalizations with rare cases of prolonged consonants (which was shown to be a peculiarity of Armenian hesitational phenomena [18]). These vocalizations are usually phonetically different from the lexical items, since they are pronounced with minimal movements of the articulatory organs due to the articulatory economy [19]. However, it was also shown that phonological system of the language may influence the quality of FPs vocalizations [20]. Even universal characteristics of hesitations, such as lengthenings being accompanied by creaky voice, may operate differently in different languages; for example, in Finnish it was proposed that creaky voice may indicate turn-transitional locations [12], which is not the case for English [21].

Although the speech technologies and particularly ASR systems have to account for all types of disfluencies (filled pauses, lengthenings, repetitions, deletions, substitutions, fragments, editing expressions, insertions, etc.), in the present study, we focus on the detection of the most frequent disfluent category: voiced hesitations (filled pauses and sound lengthening) in Russian spontaneous speech.

2. Related Work

Various methods have been proposed for speech disfluencies detection. All of them can be roughly divided into the following: (1) those that use language modelling (LM) incorporating information on speech disfluencies into ASR systems and (2) those that take into account only acoustic parameters. The second group is more popular, since there is no need of additional large corpus of transcriptions for LM training, despite the possible way of dealing with the problem by including the filler as an ordinary word in the lexicon and ignoring it during LM-probability computation [22]. Although this inclusion may sound reasonable, it does not necessarily lead to a higher accuracy; too many filled pauses may be hypothesized due to the acoustic similarity between filled pauses and function words or single syllables of content words [23].

It has been shown that, along with duration, the prominent characteristic of voiced hesitations is a gradual fall of fundamental frequency ($F0$) [24]; they tend to be low in $F0$ and display a gradual, roughly linear $F0$ fall. In [25] it was shown that for fair detection of hesitations these two characteristics and distance to a pause are enough.

In [17], filled pauses are detected on a basis of two features (small fundamental frequency transition and small spectral envelope deformation) which are estimated by identifying the most predominant harmonic structure in the input. The method has been implemented and tested on 100 utterances extracted from a Japanese spoken language corpus. Each utterance contained at least one filled pause. The achieved results were 91.5% precision and 84.9% recall. However, the authors admit that these figures may be optimistic because in their corpus there were no low-voiced male speakers.

In [23] authors developed a detection system in order to improve the speech recognizer performance. As a classifier authors used the Multilayer Perceptron with one output. The features were segment duration, spectral stability, stable interval durations, silence before and after the hesitations, spectral centre of gravity, and simple filled pause model output (a 4-mixture GMM that was trained to model the frames belonging to a filled pause). On three Flemish parts of Spoken Dutch Corpus authors achieved precision of 85% at a recall rate of 70%.

In [9] authors focused on detection of filled pauses based 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 [26].

In 2013 the INTERSPEECH Paralinguistic Challenge [27] raised interest in automatic detection of fillers providing a standardized 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 [28]. In [29] 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 [27], extracted with the openSMILE toolkit [30]. 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

Usually for studying speech disfluencies researchers use corpora with Rich Transcription [31]. An example of such corpora is English CTS Treebank with Structural Metadata corpus of English telephone conversations with metadata annotation [32], which includes, for example, filled pauses and discourse markers. Another example is the corpus Czech Broadcast Conversation MDE Transcripts [33] that consist of transcripts with metadata of the files in Czech Broadcast Conversation Speech Corpus [34]. Its annotation contains such phenomena as background noises, filled pauses, laugh, smacks, and so on [35].

For our purposes we combined different material of diverse quality and recordings situation. Thus, the material we used in this study consists of several parts.

The first part is the corpus of task-based dialogues collected at SPIIRAS in St. Petersburg in the end of 2012-beginning of 2013 [36]. Thus, the recorded speech is informal and unrehearsed, and it is also the result of direct dialogue communication, what makes it spontaneous [37]. For example, in Edinburgh and Glasgow the HCRC corpus was collected, which consists only of map-task dialogues [38], and half of the other corpus, corpus of German speech Kiel, consists of appointment tasks [39]. This corpus consists of 18 dialogues from 1.5 to 5 minutes, where students (6 women and 6 men) from 17 to 23 years fulfilled map and appointment tasks in pairs. Recordings were annotated manually into different types of disfluencies, the voiced hesitations 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 [40]. This database consists of approximately 30 hours 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 hours) of 53 speakers [40]. 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 hesitations. 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 did not base their reports on a written text, these recordings contain considerable amount of speech disfluencies. 951 hesitations were manually annotated: 741 filled pauses and 210 lengthenings.

Another part we added for making our corpus more quality and situation diverse is 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 [41]. 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 hesitations is 285 (225 filled pauses and 60 lengthenings).

In total, the data set we used is about 3 hours and comprises 2012 filled pauses. Distribution of hesitations duration over the corpus is shown in Figure 1.

Distribution of ten most frequent hesitations across different parts of the joint corpus is shown in Figure 2.

The duration of a single hesitation lies between 6 ms and 2.3 s; the average duration is 388 ms. Among annotated hesitation the most frequent filled pause was [əː] with total 905 utterances, and the most frequent lengthening was that of vowel /a/ - 197 utterances.

4. Experiments on Hesitations Detection in Russian with Machine Learning Techniques

To develop a good hesitations detector a proper set of prosodic and acoustic cues that are likely to mark hesitations in speech signal is needed. As it was already mentioned above, in [25], it was shown that for fair detection of hesitations these two characteristics and distance to a pause are enough. Thus, at first we started testing rule-based approaches towards hesitation detection.

4.1. Rule-Based Approaches towards Hesitations Detection in Russian. The pilot step of experiments was to try a similar simple [25] approach on Russian speech, based our method on acoustical features of voiced hesitations that are peculiar to these events in Russian. To find the most prominent ones we have checked duration, $F0$, three first formants, energy, and stableness of spectra across the corpus. Similar approaches have been applied for filled pauses detection in other languages and proved the relevancy of these acoustic properties [16, 17, 42]. As a result, we used standard deviations of $F0$, $F1$, and energy as parameters, since they showed smaller variance in hesitations (the smallest were for $F0$ and energy (Figure 3)).

We obtained the optimal values of parameters a and b for criterion

$$C = aX + bY < 1, \tag{1}$$

where X is standard deviation of logarithm of $F0$ – std(log($F0$)) and Y is standard deviation of logarithm of energy E. The optimal values are those that maximize $F1$-score for the task of selection of 150 ms windows that are part of the

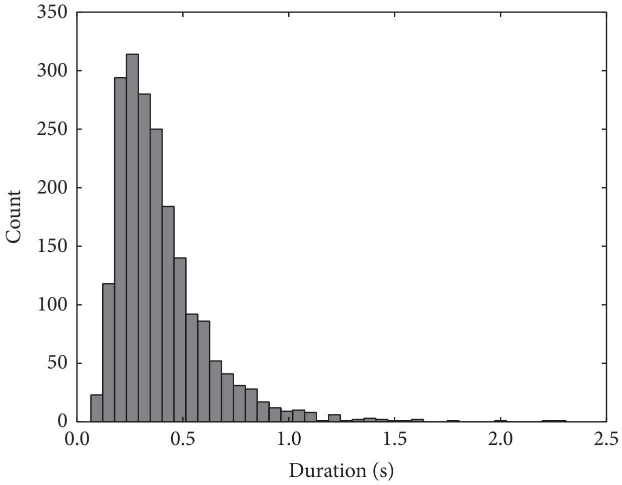


Figure 1: Distribution of hesitation duration over the joint corpus.

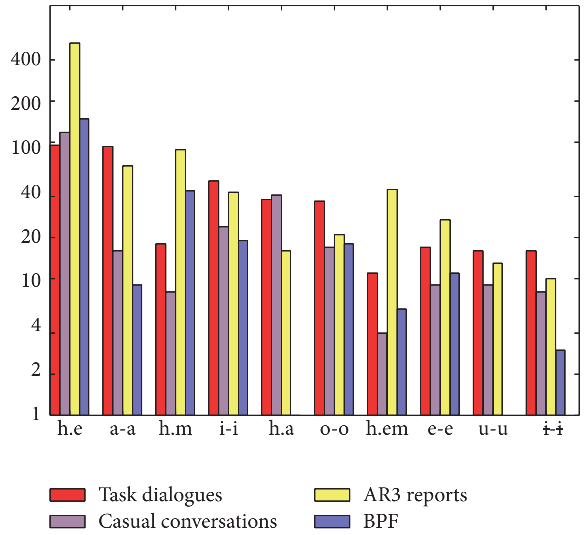


Figure 2: Distribution (in log scale) of the top ten frequent hesitations across the parts of joint corpus, where h.e, h.m, h.a and h.em are filled pauses (of [əː], [ɐː], [mː], and [əːmː] types, resp.), and others are lengthenings of certain vowels (/a/, /i/, /o/, /e/, /u/ and /ɨ/).

hesitations (Figure 4), the standard deviation of $F1$ logarithm – std(log($F1$)) being the additional threshold.

The experiments were conducted on 85% of the corpus with 15% used as a test set. The obtained $F1$-score was 0.41.

Then we have changed the criterion and maximization process. We obtained the optimal values of parameters w_n and $E0$ for criterion

$$C = \sum_n w_n V_n < 1, \qquad E > E0, \tag{2}$$

where w_n are weights for values V_n: standard deviations of $\log(E)$ and $\log(F_N)$; and $E0$ is a minimal mean energy level. The maximization of the $F1$-score for hesitations detection was made by the gradient descent method [43]. This gave us $F1$-score of 0.46 [44].

For both these approaches the stage of comparison with annotation was the same. At first we found the intervals intersecting with the labeled ones. Then we calculated the intersection length

$$T_{\text{int}} = \text{len}\,(I \cap L) \tag{3}$$

and length of nonmatching part of the interval

$$T_{\text{ext}} = \text{len}\,(L \cap I)\,, \tag{4}$$

where I is interval and L is label. If $T_{\text{int}} > 0.2\text{len}(L)$ and $3T_{\text{ext}} < T_{\text{int}}$, the pair of label and interval is considered matching. After processing the whole signal the amount of nonmatched intervals was considered false positive count and the amount of nonmatched labels was considered false negative count.

For these two approaches misses were mainly caused by the disorder of harmonic components in hoarse voice and the laryngealized filled pauses and lengthenings. In some cases hesitation had an unstable expressive intonation contour, which was not flat or lowering, that it can be argued whether they are hesitations or interjections. Few cases of misses were the result of small duration of annotated phenomena. And noises (especially in the part from the open source multi-language database) and overlaps (in task dialogues part) caused number of false negatives.

Thus, instead of accounting for all these phenomena in the rule-based methods, we decided to employ the data-driven approaches.

4.2. Data-Driven Approaches towards Hesitations Detection in Russian. In [45] we described experiments on hesitations detection using the Extreme Learning Machines (ELM), a particular kind of Artificial Neural Networks that solve classification and regression problems. We used the Python ELM implementation described in [46]. In our method the number of sigmoid neurons was 600. The feature set used in these experiments consisted of 21 standard deviations (for $F0$ and first three formants, energy, voicing probability and its derivative, and 14 MFCC coefficients) and of 3 mean values (for energy, voicing probability, and its derivative). The formants value was taken from Praat [47] and all other parameters from openSMILE [30]. Within each 100 ms window we calculated standard deviation for every parameter from the feature set and mean value for energy.

To create train and test sets out of the data we selected random 10% of the data for test set, and the rest was used as the train set. This operation was performed 10 times producing 10 different pairs of train and test sets. The data has been separated into two classes: FPs and Other, and since they were not balanced we downsampled the train set to avoid the bias towards the class Other [29]. This resulted in creating the subset containing randomly chosen 8% of the instances of the class Other and all the hesitation FPs data. We used this downsampled training set to train the classifier. ELM method yields a real number for every sample that was classified as a hesitation event if this number exceeded a certain threshold.

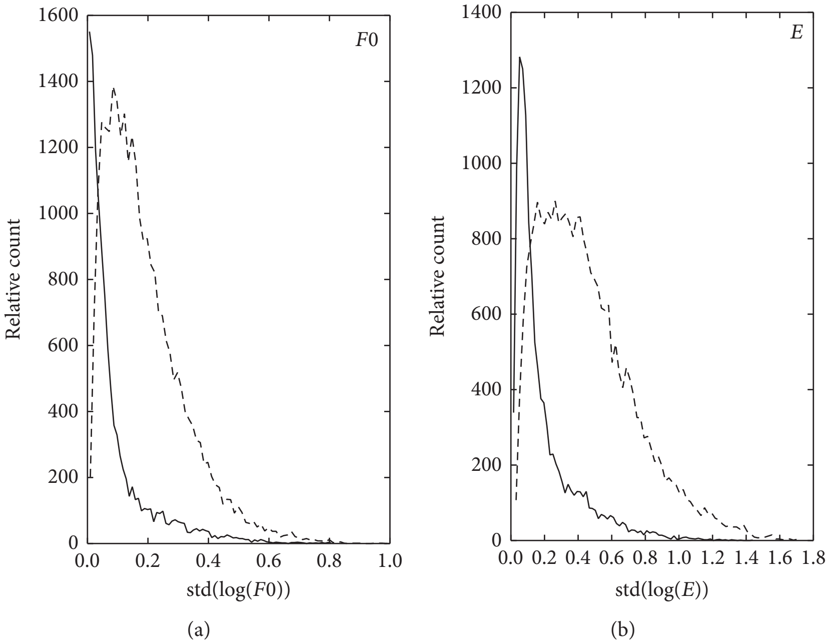


Figure 3: The standard deviation of the logarithms of $F0$ (a) and energy (b) of FPs (thick line) and of neighbouring words and phrases (dashed line).

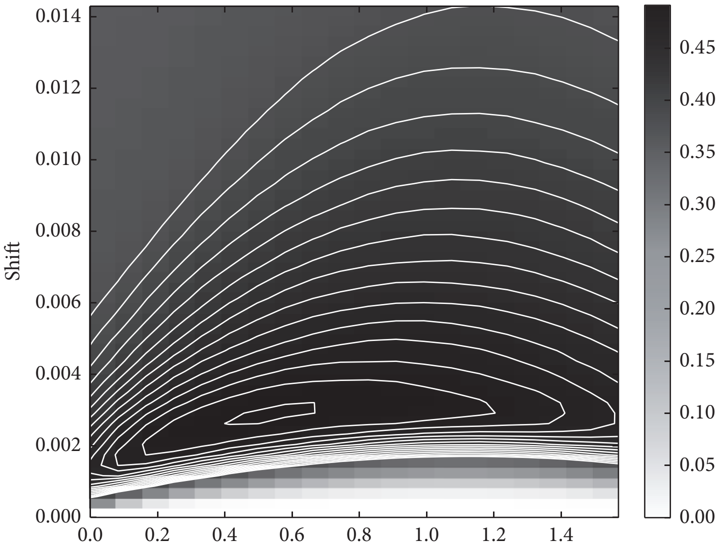


Figure 4: The $F1$-score dependence on C criterion parameters a and b, where *slope* is $\arctan(a/b)$ and *shift* is $\sqrt{a^2 + b^2}$.

This threshold was determined by a grid search in a way maximizing the $F1$-score on the training set. As the result we achieved $F1$-score of 0.42.

Our most recent experiments [48] are based on the Support Vector Machine (SVM) classifier, as we followed [29]. Compared with ELM, SVM provides better detection accuracy with better harmonized mean of precision and recall. For the experiments with SVM we used a Scikit-Learn Python library [49] implementation of SVM with polynomial kernel that enables the probability estimates by means of C-Support Vector Classification; the implementation is built upon LibSVM [50].

The feature set is based on the set that was used for the INTERSPEECH 2013 Social Signals Sub-Challenge [27]. Features were extracted with the openSMILE toolkit [30] 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

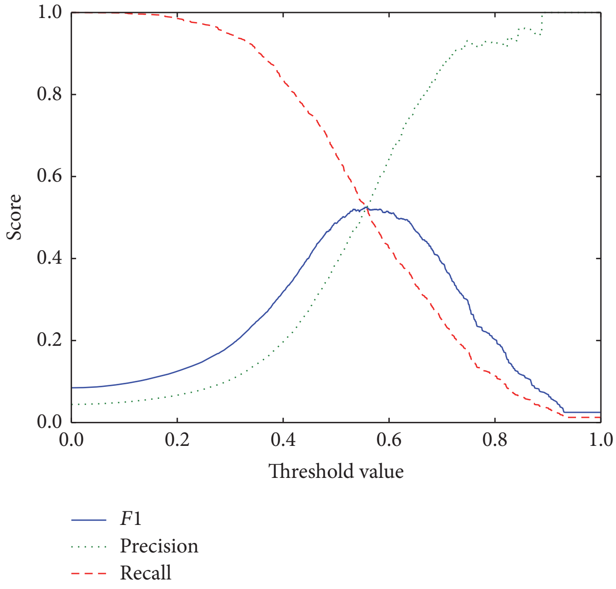


Figure 5: The dependence of results from the decision threshold.

as well as their first- and second-order delta, 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 a result, we have 162 values per frame.

As in [45] we also separated our data into two classes: "FPs" and "Other," but changed the process of separation. 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 to produce 10 different triplets of train, development, and test sets.

After training our SVM classifier, as the postprocessing step we applied Gaussian filter and morphological opening [29, 51] 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 is used to smooth the spikes and remove the outliers on the probability estimates, while morphological opening is useful for making the detection of hesitations more balanced by filtering false alarms and improving $F1$-score [29]. 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 to achieve 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 in Figure 5.

As a result we achieved $F1$-score = 0.54 ± 0.027, with precision and recall being 0.55 ± 0.05 and 0.53 ± 0.04, respectively [45]. 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.

The ongoing experiments are concerned with the broadening of the features set for SVM classifier by adding 4 formants with 2 derivatives for each, their standard deviations as well as means and standard deviations of their contexts, which gives us additional 36 features.

5. Conclusions and Future Work

Detection of speech disfluencies is important for many reasons from evaluating the spoken fluency skills to improving the performance of ASR systems. In this article we presented different approaches towards hesitation detection on the joint and quality diverse corpus of Russian spontaneous speech. We discussed the application of the rule-based and data-driven methods to the hesitation detection for Russian. We implemented different techniques from grid search and gradient descent in rule-based approaches to such data-driven ones as ELM and SVM based on the automatically extracted acoustic features. Experimental results on the mixed and quality diverse corpus of spontaneous Russian speech indicate the efficiency of the techniques for the task, with SVM outperforming other methods, at the moment giving us $F1$-score = 0.54 ± 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 problem of analysis of false positives and false negatives by tuning SVM, by expert analysis and by utilizing additional context levels.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This research is supported by the grant of Russian Foundation for Basic Research (Project no. 15-06-04465) and by the State Research no. 0073-2014-0005.

References

[1] F. G. Eisler, *Psycholinguistics: Experiments in Spontaneous Speech*, Academic Press, 1968.

[2] W. L. Chafe, Ed., *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, Ablex Publishing Corp, Norwood, Mass, USA, 1980.

[3] E. Shriberg, *Preliminaries to a theory of speech disfluencies [Ph.D. thesis]*, University of California at Berkeley, 1994.

[4] J. E. F. Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, no. 6, pp. 709–738, 1995.

[5] A. Kibrik and V. Podlesskaya, Eds., *Night Dream Stories: Corpus Study of Russian Discourse*, Litres, 2014.

[6] D. C. O'Connell and S. Kowal, "The history of research on the filled pause as evidence of the written language bias in linguistics (Linell, 1982)," *Journal of Psycholinguistic Research*, vol. 33, no. 6, pp. 459–474, 2004.

[7] A. Stolcke, E. Shriberg, R. A. Bates et al., "Automatic detection of sentence boundaries and disfluencies based on recognized

words," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '98)*.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switch board: telephone speech corpusfor research and development," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 1, pp. 517–520, San Francisco, Calif, USA, March 1992.

[9] H. Medeiros, H. Moniz, F. Batista, I. Trancoso, and L. Nunes, "Disfluency detection based on prosodic features for university lectures," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH '13)*, pp. 2629–2633, Lyon, France, August 2013.

[10] E. Shriberg, "Spontaneous speech: how people really talk and why engineers should care," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1781–1784, ISCA, Lisbon, Portugal, September 2005.

[11] H. Clark, *Using Language*, Cambridge University Press, Cambridge, UK, 1996.

[12] R. Ogden, "Turn-holding, turn-yielding and laryngeal activity in finnish talkin-interaction," *Journal of the International Phonetics Association*, vol. 31, no. 1, pp. 139–152, 2001.

[13] T. Arbisi-Kelm and S. A. Jun, "A comparison of disfluency patterns in normal andstuttered speech," in *Disfluency in Spontaneous Speech*, 2005.

[14] F. Ferreira, E. F. Lau, and K. G. D. Bailey, "Disfluencies, language comprehension, and tree adjoining grammars," *Cognitive Science*, vol. 28, no. 5, pp. 721–749, 2004.

[15] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1539, 2006.

[16] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 4857–4860, April 2009.

[17] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system forspontaneous speech recognition," in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, pp. 227–230, ISCA, Budapest, Hungary, 1999.

[18] V. Khurshudian, "Hesitation in typologically different languages: an experimentalstudy," in *Proceedings of the International Conference on Computational Linguistics Dialogue*, pp. 497–501, 2005.

[19] S. Stepanova, "Some features of filled hesitation pauses in spontaneous russian," in *Proceedings of the 16th International Congress of Phonetic Sciences*, vol. 16, pp. 1325–1328, Saarbrucken, Germany, 2007.

[20] A. Giannini, "Hesitation phenomena in spontaneous italian," in *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2653–2656, Barcelona, Spain, 2003.

[21] E. Shriberg, "To 'Errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[22] J. Peters, "LM studies on filled pauses in spontaneous medical dictation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 2, pp. 82–84, Association for Computational Linguistics, Edmonton, Canada, May 2003.

[23] F. Stouten and J. P. Martens, "A feature-based filled pause detection system for dutch," in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 309–314, IEEE, 2003.

[24] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 1, pp. 521–524, IEEE, 1992.

[25] E. Shriberg, R. A. Bates, and A. Stolcke, "A prosody only decision-tree model fordisfluency detection," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 2383–2386, Rhodes, Greece, 1997.

[26] H. Medeiros, F. Batista, H. Moniz, I. Trancoso, and H. Meinedo, "Experiments onautomatic detection of filled pauses using prosodic features," *Actas de Inforum*, pp. 335–345, 2013.

[27] INTERSPEECH: Computational Paralinguistic Challenge, 2013, http://emotion-research.net/sigs/speech-sig/is13-compare.

[28] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic timeseries smoothing and masking," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH '13)*, pp. 173–177, Lyon, France, August 2013.

[29] D. Prylipko, O. Egorow, I. Siegert, and A. Wendemuth, "Application of image processing methods to filled pauses detection from spontaneous speech," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association: Celebrating the Diversity of Spoken Languages (INTERSPEECH '14)*, pp. 1816–1820, Singapore, September 2014.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia ACM Multimedia (MM '10)*, pp. 1459–1462, ACM, Firenze, Italy, October 2010.

[31] Y. Liu, *Structural event detection for rich transcription of speech [Ph.D. thesis]*, Purdue University, 2004.

[32] LDC: English CTS treebank with structural metadata, http://catalog.ldc.upenn.edu/LDC2009T01.

[33] LDC: Czech broadcast conversation MDE transcripts, http://catalog.ldc.upenn.edu/LDC2009T20.

[34] LDC, "Czech broadcast conversation speech," http://catalog.ldc.upenn.edu/LDC2009S02.

[35] J. Kolár, J. Svec, S. Strassel, C. Walker, D. Kozlíková, and J. Psutka, "Czech spontaneous speech corpus with structural metadata," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, Lisbon, Portugal, September 2005.

[36] V. Verkhodanova and V. Shapranov, "Automatic detection of filled pauses and lengthenings in the spontaneous Russian speech," in *Proceedings of the 7th International Conference on Speech Prosody (SP '14)*, pp. 1110–1114, Dublin, Ireland, May 2014.

[37] E. Zemskaya, Russian Spoken Speech: Linguistic Analysis and the Problems of Learning. Moscow, 1979.

[38] A. Anderson, M. Bader, E. Bard et al., "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[39] K. J. Kohler, "Labelled data bank of spoken standard German—the Kiel Corpus of read/spontaneous speech," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)*, vol. 3, pp. 1938–1941, IEEE, October 1996.

[40] S. A. Zahorian, J. Wu, M. Karnjanadecha et al., "Open-source multi-language audio database for spoken language processing applications," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, pp. 1493–1496, Florence, Italy, August 2011.

[41] Department of Phonetics of Saint Petersburg University, http://phonetics.spbu.ru/.

[42] G. Garg and N. Ward, "Detecting filled pauses in tutorial dialogs," 2006.

[43] J. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, vol. 97, Springer Science & Business Media, 2005.

[44] V. Verkhodanova and V. Shapranov, "Multi-factor method for detection of filledpauses and lengthenings in Russian spontaneous speech," in *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20–24, 2015, Proceedings*, vol. 9319 of *Lecture Notes in Computer Science*, pp. 285–292, Springer, Berlin, Germany, 2015.

[45] V. Verkhodanova, V. Shapranov, and A. Karpov, "Filled pauses and lengthenings detection using machine learning techniques," in *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics ExLing*, pp. 175–178, Saint Petersburg, Russia, July 2016.

[46] A. Akusok, K.-M. Björk, Y. Miche, and A. Lendasse, "High-performance extreme learning machines: a complete toolbox for big data applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.

[47] P. Boersma and D. Weenink, Praat: doing phonetics by computer [computer program], version 6.0.11, http://www.praat.org/.

[48] V. Verkhodanova and V. Shapranov, "Detecting filled pauses and lengthenings in russian spontaneous speech using SVM," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23–27, 2016, Proceedings*, vol. 9811 of *Lecture Notes in Computer Science*, pp. 224–231, Springer, Berlin, Germany, 2016.

[49] Scikit-Learn: Machine learning in Python, http://scikit-learn.org.

[50] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," in *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 1–127, 2011.

[51] H. J. Heijmans, "Mathematical morphology: a modern approach in image processing based on algebra and geometry," *SIAM Review*, vol. 37, no. 1, pp. 1–36, 1995.

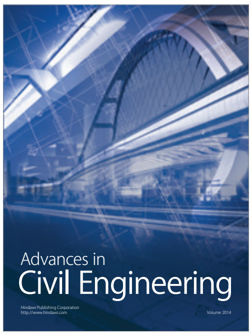
Journal of
Engineering

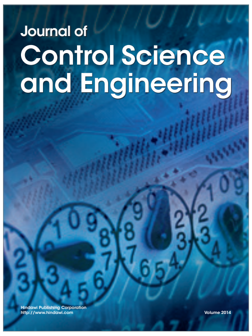
The Scientific World Journal

International Journal of
Rotating Machinery

Journal of
Sensors

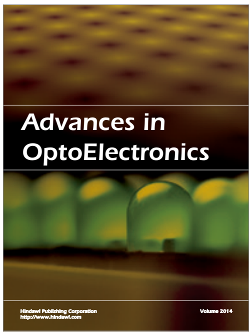
International Journal of
Distributed Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science and Engineering

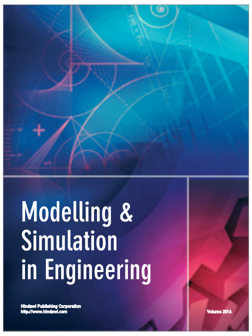
Journal of
Robotics

Journal of
Electrical and Computer Engineering

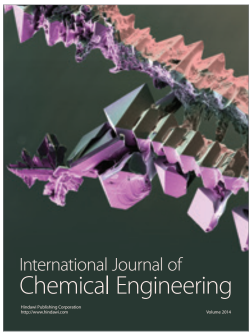

Advances in
OptoElectronics

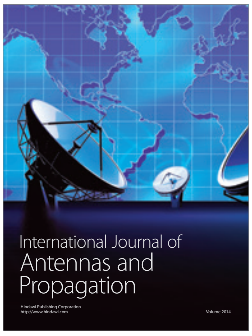
VLSI Design

International Journal of
Navigation and Observation

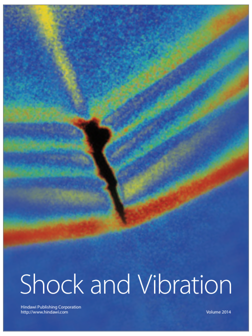

Modelling & Simulation in Engineering

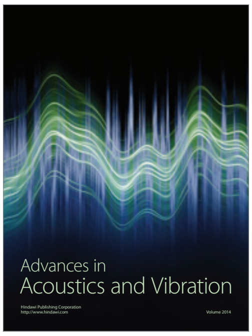
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration