Filled pauses and lengthenings detection using machine learning techniques

Vasilisa Verkhodanova, Vladimir Shapranov, Alexey Karpov
SPIIRAS, Saint Petersburg, Russia
https://doi.org/10.36505/ExLing-2016/07/0042/000301

Abstract

This paper addresses the issue of filled pauses and lengthenings detection and classification in Russian using machine learning techniques, such as ELM. We use such parameters as formants and energy variation and MFCC coefficients. The experiments on FPs detection and classification, that are carried out on the joint material of SPIIRAS task-based dialogs corpus, Russian casual conversations from Binghamton Open Source MultiLanguage Audio Database, reports from the appendix No5 to the phonetic journal "Bulletin of the Phonetic Fund" belonging to the Department of Phonetics of Saint Petersburg University and small part of SWITCHBOARD corpus. For evaluation of the experiments results we calculate the F1 score. The best achieved F1 score was 0.42.

Key words: speech disfluencies, filled pauses, spontaneous speech processing, Russian, ELM

Introduction

The need of detecting speech disfluencies 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. Since the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) (ComParE, 2013) appeared a lot of works on detection of fillers using the different machine learning approaches, since ComParE raised interest in automatic detection of fillers providing a standardised corpus and a reference system.

In (Medeiros et al., 2013) 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: for detecting words inside of disfluent sequences performance 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. In

(Prylipko et al., 2014) 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 ComParE (ComParE, 2013), extracted with the openSMILE toolkit (Eyben et al, 2010). 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.

Though evidence on filled pauses and lengthenings (further jointly referred as FPs) differs across languages, genres, and speakers, on average there are several disfluencies per 100 syllables, filled pauses being the most frequent disfluency type (O'Connell et al., 2004). In Russian speech filled pauses and lengthenings (jointly referred as FPs later on) 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 (Kibrik et al., 2014). In this paper we present the results of machine learning experiments on detection of FPs on the mixed and quality diverse corpus of Russian spontaneous speech with a addition of 20 minutes from SWITCHBOARD (Godfrey et al, 1992).

Corpus

The corpus we use for the experiments comprises various material. There are dialogs collected in St. Petersburg in the end of 2012 - beginning of 2013 (Verkhodanova et al., 2014). This part consists of 18 dialogs from 1.5 to 5 minutes, where people in pairs fulfilled map and appointment tasks. 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). There are also recordings from Multi-Language Audio Database (Zahorian et al., 2011), that consists of approximately 30 hours of sometimes low quality, varied and noisy speech in each of three languages, English, Mandarin Chinese, and Russian taken from open source public web sites, such as http://youtube.com. From the 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 (284 FPs:188 filled pauses and 96 sound lengthenings). There are also12 recorded scientific reports (linguistics, logic, psychology, etc) from the appendix No5 to the phonetic journal "Bulletin of the Phonetic

Fund" belonging to the Department of Phonetics of Saint Petersburg University (Dep. of Phonetics). 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. The number of manually annotated FPs is 285 (225 filled pauses and 60 lengthenings). Another part we added for making our corpus more quality diverse is the records from the SWITCHBOARD corpus (Godfrey et al., 1992): 3 telephone dialogues, approximately 6 minutes each. The number of manually annotated FPs is 113 (67 filled pauses and 46 lengthenings). In total, the data set we used is about 2.5 hours and comprises 1174 FPs, the duration of a single FP lies between 9ms and 2.3s, the average duration is 360ms.

Experiments on FPs detection using ELM

In this study we describe experiments on FPs 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 (Akusok et al., 2015), number of sigmoid neurons was 600.

The feature set used in the experiments consisted of 21 standard deviations (for F0 and first three formants, energy, voicing probability and its derivative, 14 MFCC coefficients), and of 3 mean values (for energy, voicing probability and its derivative). The formants value was taken from Praat (Boersma et al., 2016) and all other parameters – from openSMILE (Eyben et al., 2010). Parameters were calculated in a window of 100ms with a 10ms step, and within each 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". 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" (Prylipko et al., 2014). 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 this downsampled training set.

ELM method yields a real number for every sample that was classified as a FP event if this number exceeded a certain threshold. This threshold was

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

Conclusion

In this paper we presented experiments on detection of filled pauses and lengthenings using acoustic-only features for machine learning classification (Extreme Learning Machines). For the experiments we used diverse material differing in quality, recording sites and situations. The feature set consisted of 21 standard deviations (for F0 and first three formants, energy, voicing probability and its derivative, 14 MFCC coefficients), and of 3 mean values (for energy, voicing probability and its derivative). As the result we achieved F1 score of 0.42.

Acknowledgements

This research is supported by the grant of Russian Foundation for Basic Research (project No 15-06-04465).

References

Akusok, A., Bjork, K. M., Miche, Y., Lendasse, A. 2015. High-performance extreme learning machines. Access, IEEE, 3, 1011-1025.

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

Department of Phonetics of Saint Petersburg University. http://phonetics.spbu.ru/

Prylipko, D., Egorow, O., Siegert, I., Wendemuth, A. 2014. Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech. In Proc. of INTERSPEECH 2014, 1816-1820, Singapore.

Eyben, F., Wollmer, M., Schuller, B. 2010. OpenSMILE: the Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proc. 18th ACM International conference on Multimedia, 1459-1462.

O'Connell, D., Kowal, S. 2004. The History of Research on the Filled Pause as Evidenceof the Written Language Bias in Linguistics. Journal of Psycholinguistic Research, vol. 33(6), 459-474.

Kibrik, A., Podlesskaya, V. (eds.). 2014. Rasskazy o Snovideniyah: Korpusnoye Issledovaniye Ustnogo Russkogo Diskursa [Night dream stories: Corpus study of Russian discourse], Litres.

Godfrey, J.J., Holliman, E.C., McDaniel, J. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92). vol. 1, 517-520.

Verkhodanova, V., Shapranov, V. 2014. Automatic Detection of Filled Pauses and Lengthenings in the Spontaneous Russian Speech. In: Proc. 7th International Conference Speech Prosody, 1110-14, Dublin, Ireland.

Zahorian, S.A., Wu, J., Karnjanadecha, M., Vootkur, C.S., Wong, B., Hwang, A., Tokhtamyshev, E. 2011. Open-Source Multi-Language Audio Database for Spoken Language Processing Applications. In Proc. INTERSPEECH 2011, 1493-96, Florence, Italy.

Boersma P., Weenink D. 2016. Praat: doing phonetics by computer [Computer program]. Version 6.0.11, retrieved 20 January 2016 from http://www.praat.org/