Filled Pauses and Lengthenings Detection Based on the Acoustic Features for the Spontaneous Russian Speech

Vasilisa Verkhodanova[1] and Vladimir Shapranov[2]

[1] SPIIRAS, 39, 14th line, 199178, St. Petersburg, Russia
[2] Betria Systems, Inc, 50, Building 11, Ligovskii Prospekt, St. Petersburg, Russia
verkhodanova@iias.spb.su, equidamoid@gmail.com

Abstract. The spontaneous speech processing has a number of problems. Among them there are speech disfluencies. Although most of them are easily treated by speakers and usually do not cause any difficulties for understanding, for Automatic Speech Recognition (ASR) systems their appearance lead to many recognition mistakes. Our paper deals with the most frequent of them (filled pauses and sound lengthenings) basing on the analysis of their acoustical parameters. The method based on the autocorrelation function was used to detect voiced hesitation phenomena and a method of band-filtering was used to detect unvoiced hesitation phenomena. For the experiments on filled pauses and lengthenings detection an especially collected corpus of spontaneous Russian map-task and appointment-task dialogs was used. The accuracy of voiced filled pauses and lengthening detection was 80%. And accuracy of detection of unvoiced fricative lengthening was 66%.

Keywords: speech disfluencies, filled pauses, lengthenings, hesitation, speech corpus, spontaneous speech processing, speech recognition.

1 Introduction

There are a number of factors such as speech variation and different kinds of speech disfluencies that has a bad influence on automatic speech processing [1, 2]. Mainly such phenomena are characteristic to a spontaneous speech. Speech disfluencies are any of various breaks or irregularities that occur within the flow of otherwise fluent speech. These are filled pauses, sound lengthenings, self-repairs, etc. The occurrence of these phenomena may be caused by exterior influence as well as by failures during speech act planning [3]. Hesitations are breaks in phonation that are often filled with certain sounds. Filled pauses are those hesitations that are filled with certain sounds. The nature of sound lengthenings is also hesitational. Such phenomena are semantic lacunas and their appearance means that speaker needs an additional time to formulate the next piece of utterance [4]. In oral communication filled pauses and lengthenings may play a valuable role such as helping a speaker to hold a conversational turn or to express the speaker's thinking process of formulating the upcoming utterance fragment.

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 227–234, 2014.

These phenomena are an obstacle for processing of spontaneous speech as well as its transcriptions, because speech recognition systems are usually trained on the structured data without speech disfluencies, this decreases speech recognition accuracy and leads to inaccurate transcriptions [5, 6]. Nowadays there are two main types of methods of dealing with speech disfluencies: methods that process them by means of only acoustic parameters analysis, such as fundamental frequency transition and spectral envelope deformation [7] and methods that process them by means of combined language and acoustic modeling [8].

There are lots of works devoted to speech disfluencies modeling within the systems of automatic speech recognition [9, 10]. Also there are approaches that deal with speech disfluencies at the stage of signal preprocessing [11], as well as speech disfluencies removal using speech transcriptions [10, 12].

In [13] authors describe a method for automatic detection of filled pauses. This method detects filled pauses and word lengthening on the basis of two acoustical features: small F0 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 a Japanese spontaneous speech corpus consisting of 100 utterances by five men and five women (10 utterances per subject). Each utterance contained at least one filled pause. Experimental results for a Japanese spoken dialogue corpus showed that the real-time filled-pause-detection system yielded a recall rate of 84.9% and a precision rate of 91.5%.

In [14] authors focus on the identification of disfluent sequences and their distinct structural regions, based on acoustic and prosodic features. For the experiments a speech corpus of university lectures in European Portuguese “Lectra” with a relatively high percentage of disfluencies (7.6%) was used. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consists of spontaneous speech mostly, and its current version contains about 32h of manual orthographic transcripts. Several machine learning methods have been applied, and the best results were achieved using Classification and Regression Trees (CART). The set of features which were most informative for cross-region identification encompasses word duration ratios, word confidence score, silent ratios, pitch, and energy slopes. 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. Presented results confirm that knowledge about filled pauses and fragments has a strong impact on the performance. Without it, the performance decayed to 66% precision and 20% recall.

There are number of publications aimed to raise speech disfluencies recognition quality by means of additional knowledge sources such as different language models. In [8] three types of speech disfluencies are considered: repetition, revisions (content replacement), restarts (or false starts). A part of Switchboard-I as well as its transcription (human transcriptions and ASR output) was taken for research. Normalized word and pause duration, pitch, jitter (undesirable phase and/or random frequency deviation of the transmitted signal), spectral tilt, and the ratio of the time, in which the vocal folds are open to the total length of the glottal cycle, were taken as the prosodic features. Also three types of language models were used: (1) hidden-event word-based language model that describes joint appearance of the key words and speech disfluencies in spontaneous speech; (2) hidden-event POS-based language model that uses statistics on part-of-speech (POS) to capture syntactically generalized patterns, such as the tendency to repeat prepositions; (3) repetition pattern language model for detection of repetitions.

For the application of disfluencies detecting methods based on language modeling a large corpus of transcriptions is needed while for rule-based approaches there is no need for such corpus. Also rule-based approaches have an advantage of not relying on lexical information from a speech recognizer. For this research we decided further to test the effectiveness of rule-based approach for detecting filled pauses and lengthenings in Russian spontaneous speech.

This paper is organized as follows: in Section 2 the methodology for corpus recording and the collected corpus description are given. Section 3 is devoted to description of the method of filled pauses and lengthenings detection. In Section 4 the experimental results of hesitations and sound lengthenings are presented.

2 Corpus of Russian Spontaneous Speech

Usually corpora with Rich Transcription [12] are used for studying the speech disfluencies of different kind. Czech Broadcast Conversation MDE Transcripts [15] is an example of such corpus. It consists of transcripts with metadata of the files in Czech Broadcast Conversation Speech Corpus [16], and its annotation contains such phenomena as background noises, filled pauses, laugh, smacks, etc [17].

For our purposes a corpus of spontaneous Russian speech was collected based on the task methodology: map-tasks and appointment-task [18]. Thus, the recorded speech is informal and unrehearsed, and it is also the result of direct dialogue communication, what makes it spontaneous [19].

For example, in Edinburgh and Glasgow the HCRC corpus was collected, which consists only of map-task dialogs [20], and half of the another corpus, corpus of German speech Kiel, consists of appointment tasks [21].

Map task dialogs in the collected corpus represent a description of a route from start to finish, basing on the maps. Pair of participants had a map which had various landmarks drawn on it with one of them having a map with a route. And the task was to describe the route to the other participant, who had to draw this route onto his/her own map. After fulfilling this task participants switched their roles and dialogue continued. Several pairs of maps of varied difficulty were created, the criterion of difficulty being the number of unmatched landmarks. For dialogs based on appointment task, a pair of participants tried to find a common free time for: a) telephone talk (at least 15 minutes), b) meeting (1 hour) based on their individual schedules. Participants could not see maps or schedules of each other. Due to maps and schedules structure they had to ask questions, interrupt and discuss the route or possible free time. This resulted in speech disfluencies and artifacts appearance.

The recorded corpus consists of 18 dialogs from 1.5 to 5 minutes. Recording was performed in the sound isolated room by means of two tablets PCs Samsung Galaxy

Tab 2 with Smart Voice Recorder. Sample rate was 16 kHz, bit rate - 256 Kbit/s. All the recordings were made in St. Petersburg in the end of 2012 - beginning of 2013. 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 in the Wave Assistant [22] on two levels: those disfluencies and artifacts that were characteristic for one speaker were marked on the first level, those that were characteristic for the other speaker - on the second level. During annotation 1042 phenomena such as filled pauses (for example pauses filled with [ə] and [ɐ] sounds), artifacts (as laugh, breath), self-repairs and false-starts as well as word-fillers were marked.

Sighs and loud breath, filled pauses [ə] and [m], self-repairs and lengthening of sound /i/ appeared equally often in the speech of all 12 speakers. In the speech of 11 speakers also lengthening of /a/ and filled pause [ɐ] were common. And almost everyone used such fillers as /vot/ (“there”) and /nu/ (“well”).

Due to the fact that certain disfluencies are communicatively significant and hardly can be distinguished from normal speech, on this stage of research we have confined ourselves to the most frequent elements of in speech disfluencies: filled pauses and sound lengthenings.

3 Method of Filled Pauses and Lengthenings Detection

The basic idea of our method is to find acoustical features of filled pauses and sound lengthenings in speech signals by using spectrum analysis. Our method assumes that filled pauses and lengthenings contain a continuous voiced sound of an unvaried phoneme; due to this the neighboring instantaneous spectra are similar. For these phenomena such characteristics as unvaried value of pitch and duration of about 150-200ms are peculiar. This duration value is a reliable threshold for perception of speech pauses, because it is close to the value of mean syllable duration [23].

This paper logically follows from [18]. The method for voiced hesitational phenomena detection is based on the algorithm described in [24]. In the following, we describe the main procedure of our method (Figure 1).

The main idea of voiced hesitational phenomena detection algorithm is to find continuous segments in the signal where formants position remains the same. The “formant spectrum” for every moment T_n, where $T_n = d_t\, n$ was calculated basing on the algorithm described in [24]. The value of d_t is constrained by the duration of sound transitions and also by the needed temporal resolution.

Window length for autocorrelation procedure was 128 samples or 8 ms. The consequential spectra were compared by means of the K, that is the formants stability coefficient:

$$K = \frac{\sum_k A_k B_k}{(\sum_k A_k^2)(\sum_k B_k^2)} \quad , \qquad (1)$$

where A_k and B_k are neighboring formant spectra.

The coefficient K reflects the constancy of formants position. The value of K is close to 1 when the formants position does not change. The function of K is presented in the Figure 2. There are intervals in the function K where its value exceeds the experimentally set threshold value (0.79). The boundaries of the intervals are then checked whether there are occlusive consonants. And if these consonants are word's beginning/end the interval length is increased by the consonant duration. Intervals shorter than 0.16 sec are ignored. All the rest are then compared to the markup.

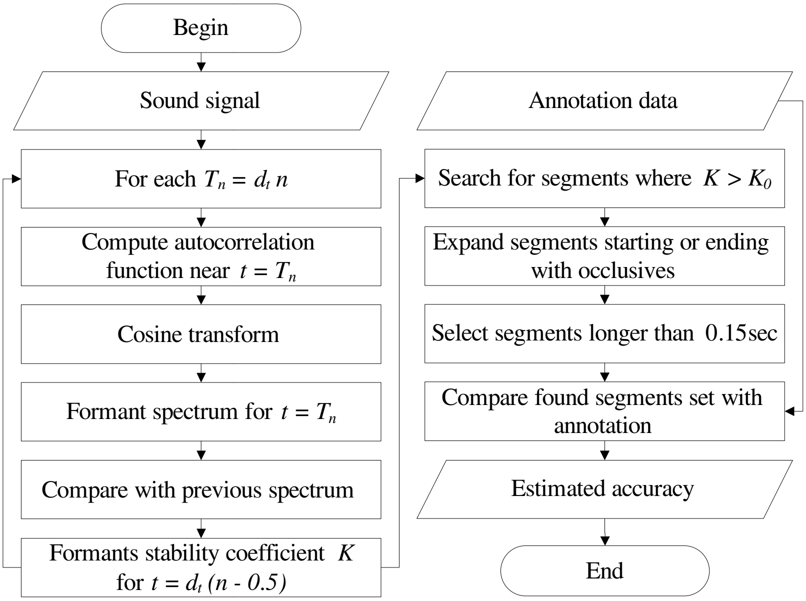

Fig. 1. Scheme of the method of voiced hesitational phenomena detection

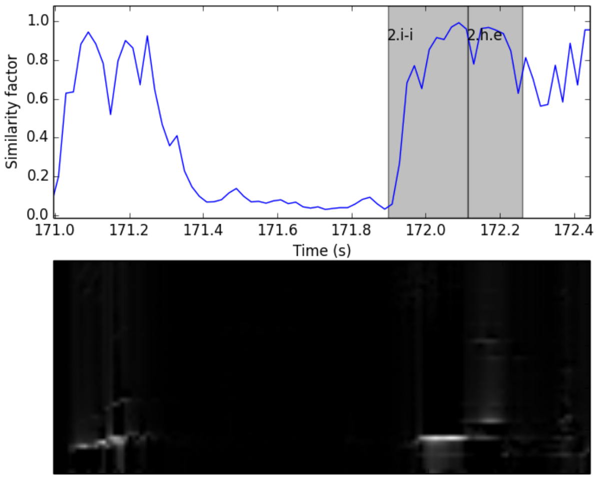

Fig. 2. The diagram of similarity function K (above), with the gray background indicating mark in the annotation, and the resampled spectrogram (below) of the same signal part for filled pause /e/

The method described above does not perform well on the lengthenings of unvoiced fricatives. Due to the small amount of these elements (about only 1% of all annotated phenomena), almost all of them being sibilants lengthenings, we relied on the fact that they are characterized by wide bands of certain frequencies ("fricative factors"). The situation of such bands for each unvoiced fricative sound is independent from the speaker. At this stage to detect unvoiced fricative lengthenings the following temporal series were computed: the ratios of the mean value of instantaneous spectrum samples in the band to the mean value of samples of the spectrum. Those intervals, where the series value exceeds a certain constant (more than 3), presumably contain the sound in question [23]. For fricatives separation the following actions were performed. For the found intervals values of "fricative factors" were examined by turns to detect among them those intervals that are corresponding to consonant lengthenings. The rest of the found elements were considered as breath.

The detected filled pause and lengthening events were compared to the markup. For each event we looked for a mark that overlapped it, with the common part of these intervals being sufficiently large (the value of 0.4 was defined experimentally):

$$L_{Ev \cup Mark} > 0.4 min(\mathrm{L}_{\mathrm{Ev}}, \mathrm{L}_{\mathrm{Mark}}), \quad (2)$$

where $L_{Ev \cup Mark}$ is the length of the common part, L_{Ev} is the length of the event, and L_{Mark} is the length of the mark. If the type of the mark matches the type of the event then the event was considered as match, otherwise it was considered as a false positive. All marks that were not matched during the events processing were treated as a false negative result.

4 Experimental Results

The filled pauses and sound lengthening algorithm based on the method described above was implemented and tested on a collected spontaneous Russian speech corpus. The development set consisted of 3 dialogs (4 speakers of different specialization): two map-task and one appointment-task dialogue. The testing set was the other part of the corpus (15 dialogs). The accuracy of voiced filled pauses and lengthenings detection was 80%. And accuracy of detection of unvoiced fricative lengthenings was 66%.

The main reasons for "misses" were the disorder of harmonic components in hoarse voice and the laryngealized filled pauses and lengthenings: the algorithm has found parts of laryngealized filled pauses and lengthening the duration of which was not enough to overcome the threshold for correctly found elements. Another reason for misses was filled pauses consisting of two different sounds, such as /ae/. In such a case algorithm detected two lengthenings /a/ and /e/ ignoring the transition part, and both these lengthenings appeared to be too short to overcome the threshold. On the other hand, false alarms were mainly caused by lengthenings that were missing in the annotation and by noises and overlappings. For example the paper riffle sometimes is very similar to lengthening of a /s/ consonant and can be detected incorrectly.

5 Conclusions

This paper presents the method of filled pauses and sound lengthening detection by using the analysis based on the autocorrelation function. The experiments were based on the corpus of spontaneous Russian speech that was especially collected and manually annotated taking into account speech disfluencies and artifacts. The criterion of matching with the annotation marks was used as estimation of algorithm work. The accuracy achieved for the voiced filled pauses and lengthenings detection was 80%. And the accuracy of the unvoiced fricative lengthening detection was 66%.

Further experiments are planned to be focused on more precise physical boundaries detection, on dealing with laryngealized sounds as well as on performing similar experiments with other Russian speech corpora within the other domain. Another stage of investigation will be devoted to context of filled pauses and lengthenings. This would help to detect more precisely their physical boundaries, which are of different nature, such possible sounds as glottal stops in the beginning of filled pauses, transition parts between two sounds, etc. We also are going to apply our method to Russian speech recognition at the stage of signal preprocessing. Future work will also include an integration of the method with speech dialogue systems to make full use of filled pauses communicative functions in the various domains [25].

Acknowledgements. This research is supported by the Committee on Science and Higher Education of the Administration of St. Petersburg.

References

1. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Communication 56, 213–228 (2014)
2. Kipyatkova, I.S.: Software complex for recognition and processing of Russian conversational speech. Information-control systems 53, 53–59 (2011) (in Rus)
3. Podlesskaya, V.I., Kibrik, A.A.: Speech disfluencies and their reflection in discourse transcription. In: Proc. of VII International Conference on Cognitive Modelling in Linguistics, pp. 194–204 (2004)
4. Clark, H.H., Fox Tree, J.E.: Using uh and um in spontaneous speaking. Cognition 84, 73–111 (2002)
5. Verkhodanova, V.O., Karpov, A.A.: Speech disfluencies modeling in the automatic speech recognition systems The Bulletin of University of Tomsk 363, 10–15 (2012) (in Rus.)
6. Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M.: Analysis of Long-distance Word Dependencies and Pronunciation Variability at Conversational Russian Speech Recognition. In: Proc. of Federated Conference on Computer Science and Information Systems, pp. 719–725 (2012)
7. Veiga, A., Candeias, S., Lopes, C., Perdigao, F.: Characterization of hesitations using acoustic models. In: Proc. of 17th International Congress of Phonetic Sciences, pp. 2054–2057 (2011)

8. Liu, Y., Shriberg, E., Stolcke, A., et al.: Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. IEEE Transactions on Audio, Speech and Language Processing 1(5), 1526–1540 (2006)
9. Verkhodanova, V.O.: Algorithms and Software for Automatic Detection of Speech Disfluencies in an Audio Signal. SPIIRAS Proceedings 31, 43–60 (2013)
10. Lease, M., Johnson, M., Charniak, E.: Recognizing disfluencies in conversational speech. IEEE Transactions on Audio, Speech and Language Processing 14(5), 1566–1573 (2006)
11. Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A.: Automatic Detection and Removal of Disfluencies from Spontaneous Speech. In: Proc. of 13th Australasian International Conference on Speech Science and Technology, pp. 98–101 (2010)
12. Liu, Y.: Structural Event Detection for Rich Transcription of Speech. PhD thesis, Purdue University and ICSI, Berkeley (2004)
13. Masataka, G., Katunobu, I., Satoru, H.: A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In: Proc. of 6th European Conference on Speech Communication and Technology, pp. 227–230 (1999)
14. Medeiros, R.B., Moniz, G.S., Batista, M.M., Trancoso, I., Nunes, L.: Disfluency Detection Based on Prosodic Features for University Lectures. In: Proc. of 14th Annual Conference of the International Speech Communication Association, pp. 2629–2633 (2013)
15. Corpus "Czech Broadcast Conversation MDE Transcripts", LDC., http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T20 (accessed January 5, 2014)
16. Corpus "Czech Broadcast Conversation Speech", LDC., http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009S02 (accessed January 5, 2014)
17. Kolar, J., Svec, J., Strassel, S., et al.: Czech Spontaneous Speech Corpus with Structural Metadata. In: Proc. of 9th European Conference on Speech Communication and Technology, pp. 1165–1168 (2005)
18. Verkhodanova, V., Shapranov, V.: Automatic detection of speech disfluencies in the spontaneous russian speech. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS (LNAI), vol. 8113, pp. 70–77. Springer, Heidelberg (2013)
19. Zemskaya, E.A.: Russian spoken speech: linguistic analysis and the problems of learning, Moscow (1979) (in Rus.)
20. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G.M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R.: The HCRC Map Task Corpus. Language and Speech 34, 351–366 (1991)
21. Kohler, K.J.: Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech. In: Kohler, K.J. (ed.) Proc. of 4th International Conference on Spoken Language, vol. 3, pp. 1938–1941 (1996)
22. Wave Assistant, the speech analyzer program by Speech Technology Center, http://www.phonetics.pu.ru/wa/WA_S.EXE (accessed October 6, 2013)
23. Krivnova, O.F., Chadrin, I.S.: Pausing in the Natural and Synthesized Speech. In: Proc. of Conference on Theory and Practice of Speech Investigations (1999) (in Rus)
24. Nelson, D.: Correlation based speech formant recovery. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997, vol. 3, pp. 1643–1646 (1997)
25. Meshcheryakov, R.M., Kostyuchenko, E., Yu, B.L.N., Choinzonov, E.L.: Structure and database of software for speech quality and intelligibility assessment in the process of rehabilitation after surgery in the treatment of cancers of the oral cavity and oropharynx, maxillofacial area. SPIIRAS Proceedings 32, 116–124 (2014)