Automatic Detection of Filled Pauses and Lengthenings in the Spontaneous Russian Speech

Vasilisa Verkhodanova[1], *Vladimir Shapranov*[2]

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

Abstract

During automatic speech processing a number of problems appear, and among them there are such as speech variation and different kinds of speech disfluences. In this article an algorithm for automatic detection of the most frequent of them (filled pauses and sound lengthenings) based on the analysis of their acoustical parameters is presented. The method of formant analysis 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 a specially collected corpus of spontaneous Russian map-task and appointment-task dialogs was used. The accuracy of voiced filled pauses and lengthening detection was 82%. And accuracy of detection of unvoiced fricative lengthening was 66%.

Index Terms: speech disfluencies, filled pauses, lengthenings, speech corpus, automatic speech processing, automatic speech recognition.

1. Introduction

A number of factors such as speech variation and different kinds of speech disfluences has a bad influence on automatic speech processing. 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. Another problem close to speech disfluencies are speech artifacts such as cough, laugh or sighs. The occurrence of these phenomena may be caused by exterior influence as well as by failures during speech act planning [1]. Hesitations are breaks in phonation that are often filled with certain sounds. Filled pauses are those hesitations that are filled with certain sounds, and the nature of sound lenghtenings 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 [2]. In oral communication filled pauses and lengthenings may play a valuable role such as helping a speaker to hold a conversational turn or expressing the speaker's thinking process of formulating the upcoming utterance fragment. Self-repairs appear when speakers want to change partly or entirely some piece of their utterances, and may be online: when speaker changes a piece of utterance immediately, or retrospective: speaker changes it post factum.

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, what decreases speech recognition accuracy and leads to inaccurate transcriptions [3,4].

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 [5,6] and methods that process them by means of combined language and acoustic modeling [7,8].

There are lots of works devoted to speech disfluencies modeling within the systems of automatic speech recognition [5,7,9]. Also there are approaches that deal with speech disfluencies at the stage of signal preprocessing [10], as well as speech disfluencies removal using speech transcriptions [9,11].

Thus, in [10] an algorithm, which defines and eliminates filled pauses and repetitions from the speech signal, is proposed. For detection of boundaries of filled pauses the following characteristics were applied: duration, pitch, spectral and formant characteristics. For extraction and further elimination of repetitions the proposed algorithm used duration and frequency of the repeated segments as well as the Euclidian distance between the logarithms of the Linear Predictive Coding (LPC) spectra of each pair of the voiced sections around a long pause. Also the fact that repetitions are usually accompanied by a pause was taken into account.

In [12] authors describe a method for automatic detection of filled pauses. They propose a method that 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 [13] 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” was used. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consist 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

10.21437/SpeechProsody.2014-211

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 rise speech disfluencies recognition quality by means of additional knowledge sources such as different language models. In [7] 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 disfluences 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 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 the 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 lengthening are presented.

2. Corpus of Russian Spontaneous Speech

Nowadays, for studying speech disfluencies corpora with Rich Transcription [11] are used. As example such corpus as Czech Broadcast Conversation MDE Transcripts [14] may be cited. This corpus consists of transcripts with metadata of the files in Czech Broadcast Conversation Speech Corpus [15], and its annotation contains such phenomena as background noises, filled pauses, laugh, smacks, etc [16].

For our purposes a corpus of spontaneous Russian speech was collected based on the task methodology: map-tasks and appointment-task. Thus, we have recorded speech that is informal and unrehearsed, and it is also the result of direct dialogue communication, what makes it spontaneous [17]. For example, in Edinburgh and Glasgow the HCRC corpus was collected, which consists only of map-task dialogs [18], and half of the another corpus, corpus of German speech Kiel, consists of appointment tasks [19].

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. One participant also had a route marked on their map. And the task was to describe the route to the other participant, who had to draw this route onto their own map. After fulfilling this task participants switched their roles and dialogue continued. For our investigation several pairs of maps of varied difficulty were created. As the criterion of difficulty the number of unmatched landmarks was used. An example of difficult maps is shown on the Figure 1. 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.

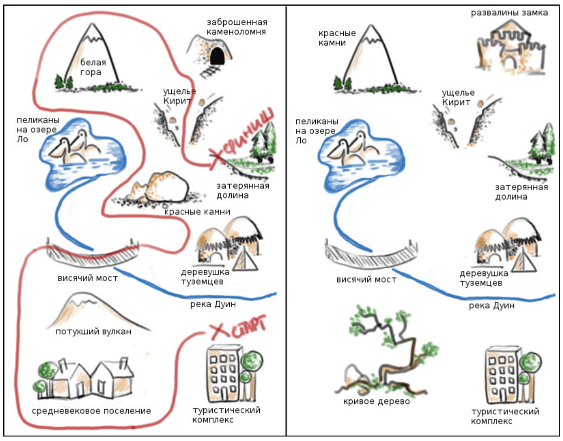

Figure 1: *An example of maps with the route (left) and without the route (right), used in map-task dialogs recording.*

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 16kHz, 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 [18] 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 [ ] and [ɐ] sounds), artifacts (as laugh, breath), self-repairs and false-starts as well as word-fillers were marked. The most frequent elements are shown on the Figure 2.

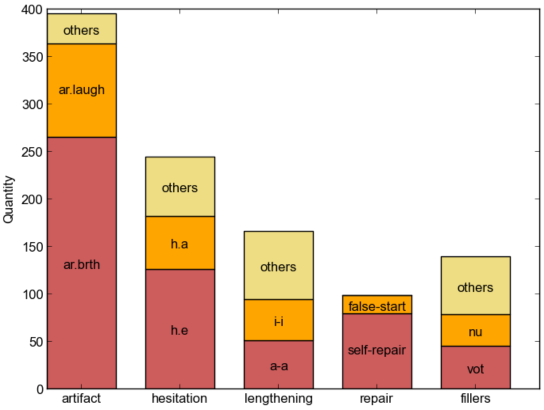

Figure 2: *A diagram of most frequent speech disfluencies and artifacts in the collected corpus, where ar.laugh - laugh, ar.brth – sighs and loud breath, h.a – hesitation [ɒ], h.e – hesitation [ə], i-i – lengthening of /i/, a-a – lengthening of /a/, “nu” and “vot”: are common fillers in Russian.*

Sighs and loud breath, filled pauses [ ] and [m], self-rapairs and lengthening of sound /i/ appeared equally often in the speech of all 12 speakers. For 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 disfluences – 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 [21].

Taking into account only pitch change and duration it is possible to confuse sonorant sounds with filled pauses. For example, in such Russian word as “налево” /nalevo/ (“to the left”) the pitch movement is almost horizontal as in filled pauses and lengthenings (Figure 3).

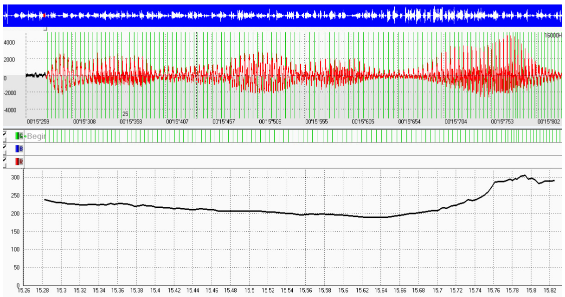

Figure 3: *A diagram of pitch movement for the word “налево” /nalevo/ (“to the left”) with averaging interval of 50ms.*

As the measure of their similarity we have used a criterion of formant similarity between neighboring spectra. We also implemented the preliminary detection of lengthening of unvoiced fricatives.

In the following, we describe the main procedure of our method (Figure 4). First step was to calculate the Fourier transform to acquire a spectrogram with window length of 512 frames and step of 256 frames. This window length provides both reasonable spectral and temporal resolution.

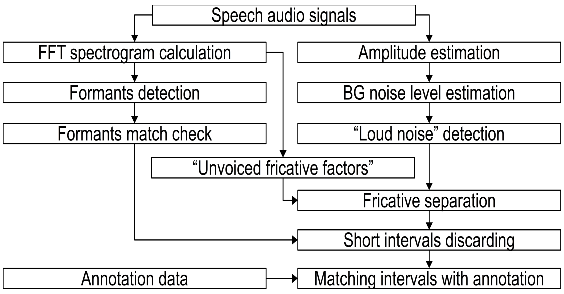
Speech audio signals
FFT spectrogram calculation
Amplitude estimation
Formants detection
BG noise level estimation
Formants match check
“Loud noise” detection
“Unvoiced fricative factors”
Fricative separation
Short intervals discarding
Annotation data
Matching intervals with annotation

Figure 4: *Scheme of hesitations and breath detection method.*

The next stage was formants detection. First, spectrum was resampled to obtain exponential scale on frequency axis. This was done to acquire reasonable resolution in the middle- and high-frequency parts of the spectrum. Then we have searched for formants: the value of maximum and values of two surrounding samples were interpolated with quadratic curve and the position and value of the curve maximum was used as formant frequency and amplitude.

For formants match check we compared two neighboring spectra and estimated the coefficient of matching *c* for these two spectra (1). For every such pair of spectra the sum of formants' amplitudes multiplied by weight was estimated. Weight was calculated as a function of amplitude change and relative formant shift for every formant. Then this sum was divided by the sum of all amplitudes.

$$c=\frac{\sum_{match} A_n * F_{match}\left({A_n}/{A'_n}, {F_n}/{F'_n}\right)}{\sum A_k} \qquad (1)$$

where A_n, F_n - are amplitude and frequency in one neighboring spectrum and A'_n, F'_n - are amplitude and frequency in the other neighboring spectrum.

C reflects the sound invariableness in the current moment of time, if spectra are equal $c=1$, and if they are completely different, $c=0$. The diagram of *c* is shown on Figure 5 (the upper part). Those intervals where this function was above certain threshold for a long period of time are considered as filled pauses and lengthenings.

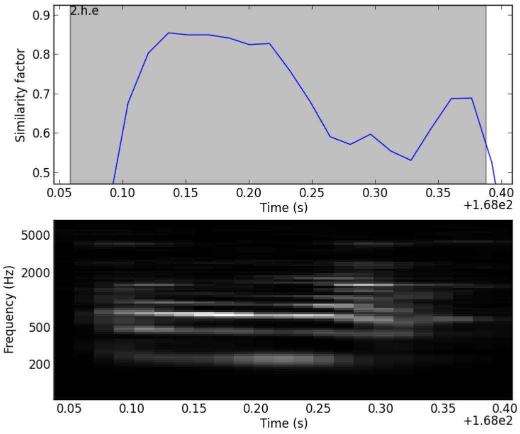

Figure 5: The diagram of s*imilarity function (above), with the gray background indicating mark in the annotation, and the resampled spectrogram (below) of the same signal part for filled pause /e/.*

To estimate an amplitude the signal was divided into overlapping frames, where the root mean square of samples in each frame is taken as the amplitude in correspondent moment of time.

To estimate a BG noise the signal was smoothed using the rectangular window with the length of 200ms, that is significantly less than characteristic length of silence intervals and greater than amplitude estimation window length. Minimum of this function was taken as background noise level.

The method based on the formant similarity described above doesn't 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 lenghtenings, 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 lenghtenings the following temporal series were computed: the ratio 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 [22].

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 lenghtenings. The rest of the found elements were considered as breath.

Then the detected filled pause and lenghtening events were compared to the markup. For each event we looked for a mark that overlaped it, with the common part of these intervals being sufficiently large (the value of 0.4 was defined experimentally) (2):

$$L_{Ev\cup Mark} > 0.4\min\left(L_{Ev}, L_{Mark}\right) \qquad (2)$$

where the $L_{Ev\cup Mark}$ – is 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 training 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 82%. And accuracy of detection of unvoiced fricative lengthenings was 66%.

The main reasons for “misses” were the disorder of harmonic components in of hoarse voice and by 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 formant analysis. The experiments were based on the corpus of spontaneous Russian speech that was specially collected and manually annotated taking into account speech disfluencies and artifacts. The criterion of matching with the annotation marks was used as algorithm work estimation. The accuracy achieved for the voiced filled pauses and lengthenings detection was 82%. And the accuracy of the unvoiced fricative lengthening detection was 66%.

Further experiments will focus on more precise physical boundaries detection as well as 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, that are of different nature, so there are such possible sounds as glottal stops in the beginning of filled pauses, transition parts between two sounds, etc. We also plan to apply our method to a Russian speech recognizer at a stage of signal preprocessing. Future work will also include an integration of the method with a speech dialogue system to make full use of the of filled pauses communicative functions.

6. References

[1] Podlesskaya, V.I., Kibrik, A.A., “Speech disfluencies and their reflection in discourse transcription”, VII International Conference on Cognitive Modelling in Linguistics Proc., 1: 194–204, 2004.

[2] Clark, H.H., Fox Tree, J.E., “Using uh and um in spontaneous speaking”, Cognition 84: 73–111, 2002.

[3] 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.)

[4] Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M., “Analysis of Long-distance Word Dependencies and Pronunciation Variability at Conversational Russian Speech Recognition”, Federated Conference on Computer Science and Information Systems Proc., 719–725, 2012.

[5] Masataka, G., Katunobu, I., Satoru, H., “A real-time filled pause detection system for spontaneous speech Recognition', 6th European Conference on Speech Communication and Technology Proc., 227–230, 1999.

[6] Veiga, A., Candeias, S., Lopes, C., Perdigao, F., “Characterization of hesitations using acoustic models”, 17th International Congress of Phonetic Sciences Proc., 2054–2057, 2011.

[7] Liu, Y., Shriberg, E., Stolcke, A., “Automatic Disfluency Identication in Conversational Speech Multiple Knowledge Sources”, 8th European Conference on Speech Communication and Technology Proc., 957–960, 2003.

[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] Lease, M., Johnson, M., Charniak, E., “Recognizing disfluencies in conversational speech”, IEEE Transactions on Audio, Speech and Language Processing, 14(5): 1566–1573, 2006.

[10] Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A., “Automatic Detection and Removal of Disfluencies from Spontaneous Speech”,13th Australasian International Conference on Speech Science and Technology Proc., 98–101, 2010.

[11] Liu, Y., “Structural Event Detection for Rich Transcription of Speech”, PhD thesis, Purdue University and ICSI, Berkeley, 253 p., 2004.

[12] Masataka, G., Katunobu, I., Satoru, H, “A Real-time Filled Pause Detection System for Spontaneous Speech Recognition”, 6th European Conference on Speech Communication and Technology Proc., 227–230, 1999.

[13] Medeiros, R.B., Moniz, G.S., Batista, M.M., Trancoso, I., Nunes, L., “Disfluency Detection Based on Prosodic Features for University Lectures”, 14th Annual Conference of the International Speech Communication Association, 2629 – 2633, 2013.

[14] Corpus “Czech Broadcast Conversation MDE Transcripts”, LDC. Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T20, accessed 5 Oct 2013.

[15] Corpus “Czech Broadcast Conversation Speech”, LDC. Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009S02, accessed 5 Oct 2013.

[16] Kolar, J., Svec, J., Strassel, S., et al.,”Czech Spontaneous Speech Corpus with Structural Metadata”, 9th European Conference on Speech Communication and Technology Proc.,1165–1168, 2005.

[17] Zemskaya, E.A., “Russian spoken speech: linguistic analysis and the problems of learning”, Moscow, 1979. (in Rus.)

[18] 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.

[19] Kohler, K.J., “Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech”,4th International Conference on Spoken Language Proc., 3: 1938–1941, 1996.

[20] Wave Assistant, the speech analyzer program by Speech Technology Center. Online: http://www.phonetics.pu.ru/wa/WA_S.EXE, accessed 5 Sep 2013.

[21] Krivnova, O.F., Chadrin, I.S., ”Pausing in the Natural and Synthesized Speech”, Conference on Theory and Practice of Speech Investigations Proc, 1999 (in Rus).

[22] Verkhodanova V., Shapranov V., “Automatic Detection of Speech Disfluencies in the Spontaneous Russian Speech”, in M. Zelezny et al. [Eds.], SPECOM 2013, LNAI 8113, 2013, pp 70-77, Springer International Publishing Switzerland, 2013.