10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Prosodic phrasing in Russian spontaneous and read speech: evidence from large speech corpora

Tatiana Kachkovskaia, Pavel Skrelin

Department of Phonetics, Saint Petersburg State University, Russia
kachkovskaia@phonetics.pu.ru, skrelin@phonetics.pu.ru

Abstract

This paper explores the differences between spontaneous and read speech in terms of prosodic phrasing. For both types of speech material we analysed the following data: IP duration (in seconds) and IP length (in clitic groups); distribution of melodic types in the material; silent pause duration; the ratio between the number of silent pauses and the number of IP boundaries. The measurements were made over 30 hours of read and 15 hours of spontaneous Russian speech. Our data have revealed the following differences in prosodic phrasing between the two types of speech material. In spontaneous speech we observe (1) shorter IPs, (2) more IPs with non-final nucleus, (3) a different frequency distribution of intonation models. Silent pause duration and the ratio between the number of silent pauses and the number of IP boundaries are highly variable across speakers.

Index Terms: prosody, spontaneous speech, read speech, intonational phrase, pause

1. Introduction

Many studies have shown that listeners can often quite easily distinguish between read and spontaneous speech [1], and it is still true in case of de-lexicalized speech signal [2]. The success of perceiving this difference, however, depends hugely on the material selected for the experiment [1], as with very careful selection listeners are unable to hear the difference at all [3]. Therefore, despite the obvious overlap between the two speech styles, in most cases spontaneous speech must contain certain markers that enable us to tell one from the other.

A lot of differences between read and spontaneous speech are discussed in the literature, mostly, lexical and segmental differences [4], differences in marking the information structure [5] (for Russian [6]), and prosody.

As the amount of speech material is growing, we are becoming able to revise our previous conclusions based on larger speech corpora. For Russian the previous comparison of read and spontaneous speech was published in 2003 [7]; the material included recordings from 10 speakers, around 10 minutes in total for each speaker—which makes around 5 minutes per speaker per speech style. Nowadays two large annotated corpora are at our disposal: a 30-hour corpus of read speech [8] and a 15-hour corpus of spontaneous speech [9].

In the focus of this paper are prosodic features of spontaneous Russian speech. We analyzed duration of intonational phrases (IPs) and silent intervals (along with IP length measured in clitic groups), the use of certain melodic types and position of the nuclear stress within the IP; we compared the number of silent pauses with the the total number of IP boundaries.

2. Material and methods

Read speech. Corpus of Professionally Read Speech (CORPRES) [8] contains recordings of fictional texts read by 8 speakers aged 30–45. The speakers were selected based on their reading skills—most of them had much professional experience as newsreaders, narrators, or lecturers. The total amount of annotated speech material is 30 hours.

Spontaneous speech. Corpus of Russian Spontaneous Speech (CoRuSS) [9] contains free dialogues recorded from 60 speakers aged 16–77 in a studio. The speakers were not specifically selected—apart from some lecturers and school teachers, there were also engineers, accountants, entrepreneurs, housewives, young college students etc. The total amount of annotated speech material is 15 hours.

Thus, in total, this paper is based on 45 hours of speech material recorded from 68 speakers.

Both corpora are comparable in terms of prosodic phrasing because their authors used the same principles of prosodic annotation. The basic segmentation unit is an intonational phrase (IP), such that: each IP usually contains one main word (the nucleus); other words are joined by a single declination (or inclination) trend; certain prosodic phenomena (e.g., pre-boundary lengthening) occur at IP boundaries. Within an IP the annotators marked the position of the nucleus, type of pitch movement for the nucleus, additional prosodic prominence[1] [10]. Additionally, word stress is marked with a special symbol. For spontaneous speech the annotation also includes false-starts, hesitations and non-speech events (these do not occur in the corpus of read speech); the corpus has a special tier containing physical boundaries of filled pauses and non-speech events (noisy breathing, clicks, coughing etc.).

Prosodic annotation in both corpora was made manually by experts. The difference between the corpora is in segmentation. In CORPRES (read speech) IP boundaries are time-aligned to the corresponding .wav file, while in CoRuSS (spontaneous speech) only turn boundaries are time-aligned, but boundaries of IPs are only marked in the orthographic transcription.

Thus, we knew the exact physical boundaries of IPs only for our read material. This is why for the spontaneous material we had to estimate IP and silent interval duration in a different way. As the total number of IPs is known from the annotation, average IP duration requires the total duration of speech (excluding silent intervals). Silent intervals (SIs) were detected using Praat [11] with the minimal duration of 0.2 seconds, the threshold of –0.35 dB and minimum pitch of 50 Hz for males and 100 Hz for females (for all other parameters default values were used); SIs shorter than 0.2 seconds were considered not perceptually significant and automatically qualified as speech. Next, SI boundaries obtained with Praat were corrected with respect to physical boundaries of non-speech events—which were

[1] That is, the words bearing extra prominence within the pre-nucleus.

10.21437/SpeechProsody.2020-34

classified as unfilled pauses. Filled pauses were intentionally classified as speech.

Using the same boundaries, average SI duration was calculated. In order to make read and spontaneous data comparable, we corrected IP boundaries in CORPRES (read speech) by setting the threshold of 0.2 seconds for SI duration.

Traditionally, an IP can never contain an internal physical pause. However, in our experience with annotation of Russian speech there were a number of cases where all other requirements for an IP were met, except for an interruption or short break between some of the words. Therefore, our procedure for segmentation into IPs includes the following rule: when we have two fragments divided by a break (silence, click, cough etc.), we assign them to the same IP if one of the fragments does not have a clear nucleus but it occurs in the other fragment, and if the declination line continues after the break as if there were no break at all (both must be true). Technically, as the function of such breaks has not been formulated yet, these breaks can not be called pauses; this is why in this paper we use the term "silent interval".

When calculating the ratio between the number of SIs and the total number of IP boundaries, we intended to answer the question of "how often IP boundaries are followed by a physical pause". However, due to differences in segmentation of the corpora, only for read speech we could get the answer for this particular question. For spontaneous speech, among the SIs calculated as described above there are some that occur not at IP boundaries but inside IPs; unfortunately, the annotation does not provide reliable information on the number of such internal SIs. On the other hand, in the corpus of read speech there are almost no (less than 0.2 %) IP-internal silent intervals longer than 0.2 seconds. Therefore, in this case, the ratio between the number of SIs and the total number of IP boundaries should be interpreted as answering the question "how often does the speaker make a break in the speech flow".

IP length was measured in clitic groups [12]. A clitic group includes at least one lexically stressed word plus the surrounding unstressed clitics (e.g., не ходи́ли бы, "would not go"—a verb with one proclitic and one enclitic). The number of clitic groups was calculated as the number of stressed vowels. Filled pauses were not counted as clitic groups.

The number of intonational phrases with non-final nucleus were calculated as the number of those IPs where after the nucleus there was at least one clitic group (i.e. at least one stressed syllable).

3. Results and discussion

3.1. Intonational phrase and the nuclear stress

Figure 1 shows mean IP length measured in clitic groups for read and spontaneous speech. It is clear that in general, IPs tend to be shorter in spontaneous speech. We have also found that the data on read speech depend on the type of text. In our corpus of read speech the material includes mostly narrative texts (short novels), but 4 of 8 speakers also recorded a play—a genre which is designed to imitate spontaneous speech—and mean IP length for that material was as low as 2.13, ranging between 2.09 and 2.22 for different speakers. Thus, in terms of average IP length, spontaneous speech lies somewhere between read narratives and read drama. From now on we present the results for narrative texts (novels) *only*, and in figure 1, bottom, recordings of the play are not included.

In order to estimate statistical significance of the difference,

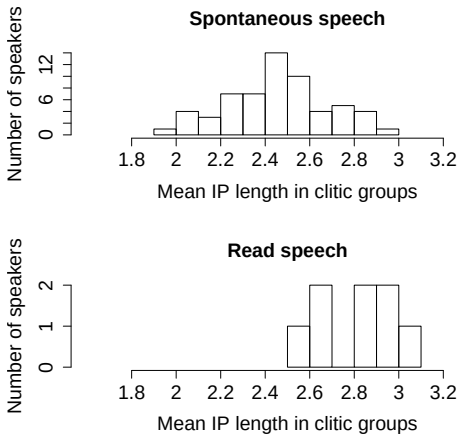

Figure 1: *Mean IP length (measured in clitic groups) for all speakers in read and spontaneous speech*

we performed pairwise t-tests between speakers from read and spontaneous corpora (two-tailed Welch t-tests). Table 1 presents p-values for such comparisons between speakers aged 30–45 in females and male groups separately, as by balancing age and gender we aimed to eliminate these factors and thus obtain more reliable results. Columns correspond to speakers from the read corpus, rows—to speakers from the spontaneous corpus. At the top and to the left we provide mean values for the respective speakers. These data prove that in most pairs of speakers IPs are longer in read speech. Still, a few speakers do show opposite tendencies. In the female group they are f24, f44, f49 from the spontaneous material and speaker K from the read material. At the intersection of the respective lines and rows (these p-values are underlined) we observe statistically significant differences—which means that here the opposite result is observed, i.e. IPs are significantly *shorter* in read speech.

In the male group a similar result is observed. Speakers m02, m19 and m43 differ from other speakers having greater IP length, and in a few cases the values are significantly greater than those in read speech (these p-values are underlined).

Thus, we observe a tendency that IP length is greater in read speech than in spontaneous speech. At the same time, around 30 % of speakers show individual strategies in this respect—a result that was also shown on other Russian material [7].

Another measurement which illustrated the difference even better is the percentage of one-word IPs (i.e. consisting of only one clitic group). The mean for spontaneous speech is 30 % (ranging from 20 % to 39 %). For read speech it is significantly lower—the mean is 16 % (ranging from 13 to 20 %). On the one hand, some of the one-word IPs in spontaneous speech are just pause fillers—similar to hesitations, but consisting of "real" lexemes. But on the other hand, we might hypothesize that speakers could avoid short IPs for the sake of rhythm and harmony; this could explain why in well-planned read speech the number of one-word IPs is lower. However, testing this hypothesis requires a separate series of experiments.

The main word within the IP—the one bearing nuclear stress—is often the last word within the IP. However, there are cases when the nucleus is not IP-final. For read speech the percentage of such IPs is quite low: on average, 8.5 % (5.7–11.7 %). In spontaneous speech this is observed more often: on average, 16 % (8–26 %). In Russian there are some types of

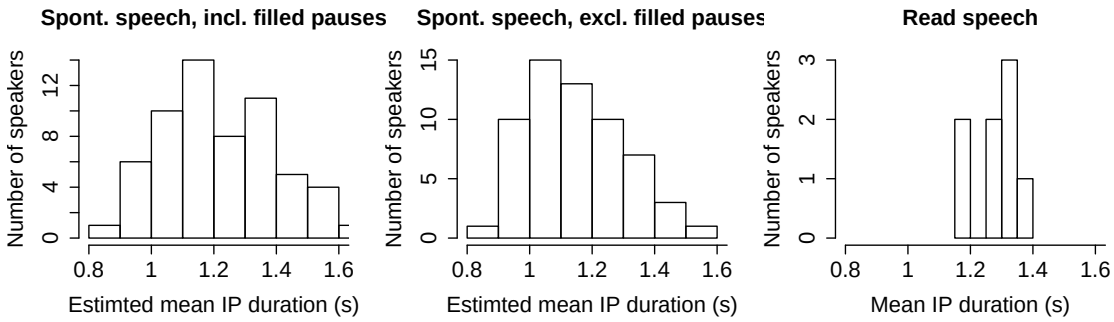

Figure 2: *Mean IP duration for all speakers in read and spontaneous speech*

Table 1: *P-values for differences in IP length (in clitic groups) calculated for speakers aged 30-45. Female speakers: C, K, O, S (read speech), f23-f59 (spontaneous speech); male speakers: A, M, G, U (read speech), m01-m56 (spontaneous speech). At the top and to the left, mean values for the speakers are provided*

females	mean	2.83	**2.56**	2.88	2.99
mean		C	K	O	S
2.52	f23	<0.001	0.49	<0.001	<0.001
2.73	f24	0.19	0.02	0.05	<0.001
2.28	f25	<0.001	<0.001	<0.001	<0.001
2.28	f26	<0.001	<0.001	<0.001	<0.001
2.53	f36	<0.001	0.53	<0.001	<0.001
2.83	f44	0.99	<0.001	0.52	0.04
2.46	f46	<0.001	0.14	<0.001	<0.001
2.79	f49	0.44	<0.001	0.10	<0.001
2.31	f57	<0.001	<0.001	<0.001	<0.001
2.56	f59	<0.001	0.93	<0.001	<0.001
males	mean	2.64	2.93	2.68	3.01
mean		A	M	G	U
2.15	m01	<0.001	<0.001	<0.001	<0.001
2.78	m02	0.049	0.037	0.15	0.002
2.42	m03	<0.001	<0.001	<0.001	<0.001
2.36	m04	<0.001	<0.001	<0.001	<0.001
2.81	m19	0.04	0.14	0.11	0.017
2.50	m21	0.09	<0.001	0.037	<0.001
2.26	m27	<0.001	<0.001	<0.001	<0.001
2.37	m28	<0.001	<0.001	<0.001	<0.001
2.91	m43	<0.001	0.73	<0.001	0.13
2.44	m56	<0.001	<0.001	<0.001	<0.001

Table 2: *P-values for differences in estimated IP duration (in seconds) and silent interval duration (SID, in seconds) calculated for all speakers and for speakers aged 30-45 only. IP duration for spontaneous speech was measured both with filled pauses (+ f.p.) and without (– f.p.); for read speech these values do not differ as there were no filled pauses. For each type of speech the mean value is provided*

	mean (s) reading	mean (s) spont.	p-value read vs spont.
age group 30–45			
mean IP dur + f.p.	1.281	1.221	0.25
mean IP dur – f.p.	1.281	1.136	**0.002**
SID mean	0.552	0.569	0.65
all speakers			
mean IP dur + f.p.	1.281	1.219	0.07
mean IP dur – f.p.	1.281	1.146	<**0.001**
SID mean	0.552	0.553	0.99

utterances that sometimes require a "shifted" nucleus—mainly, wh-questions and yes-no questions (in yes-no questions, e.g., the nucleus is by default on the verb, which is often in non-final position). A reasonable explanation of the difference could have been that spontaneous speech contains many questions. But this is not the case. In our spontaneous material, questions are very rare (except for the short "da?" (="yes?", or "right?") which usually do not function as questions). Rather, such "shift" of nucleus means that in spontaneous dialogues speakers move the most meaningful words towards the beginning of the IP, and this may serve as evidence for specific topic-comment organization. This is in accordance with the traditional understanding of information structure in colloquial speech (see, e.g., Kovtunova [6][p. 134]).

There are also some evidence that frequency of melodic types is different in read and spontaneous material. The annotation system used in these corpora is based in the following basic principles: each IP has one and only one nucleus [2], which is the main word within the IP; then the type of nucleus is defined—the "melodic type" is chosen from a limited set of values defined in terms of melodic movement, which itself is associated with certain function (e.g., general question vs. special question) and sometimes additional shades of meaning. The number of basic melodic types is 7 in the system suggested by E. A. Bryzgunova [13] and 30 in a more elaborate version suggested by N. Volskaya [10]; the latter is used here.

A striking difference is the frequency of low falls, which are used to signal end of utterance or end of paragraph (models 01a and 01 in [10]): their number is 11 % in read speech and less than 1 % in spontaneous speech. This is compensated by the number of falls to non-low (almost level tones), which is higher in spontaneous speech. This seems to lead to a more general observation: in spontaneous speech, compared with read speech, there are fewer finished utterances; this is in accordance with previous results for Russian [14]. Another difference is the frequency of "commenting" intonation (a whole IP produced in low register; models 09, 09a, 09b in [10]; used to signal parentheses, author's remarks etc.): 7.5 % in spontaneous speech and only 2.5 % in read speech. As switching between registers is

[2]Except for Bryzgunova's "intonation construction 5" (and the corresponding "model 05" in Nina Volskaya's notation), which has two nuclei, see [13]

a means to convey information structure—to separate more important information from less important—it follows that spontaneous speech uses this means more often. However, this difference may also be result of the textual material chosen for the read corpus.

Figure 2 presents histograms for mean duration of IP in spontaneous and read speech. For spontaneous speech, two variants of estimation are provided: including filled pauses and excluding filled pauses. The mean values are provided in table 2, as well as the results of statistical analysis (two-tailed Welch t-test). If we count filled pauses as parts of IPs, IP duration in read speech is on average higher (around 60 ms), but the difference is not statistically significant. A statistically significant difference is obtained only when we do *not* count filled pauses as parts of the IP: p-value 0.002 for the age group 30–45, and $<$0.001 for all speakers.

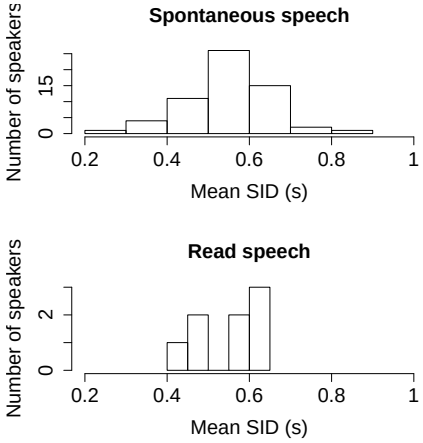

Figure 3: *Mean silent interval duration (in seconds) for all speakers in read and spontaneous speech*

3.2. Pauses

Mean silent interval duration (SID) for speakers in read speech lies within the range of 0.433 to 0.645 s (median values are between 0.325 and 0.575), for spontaneous speech—0.280 to 0.808 s (median values are between 0.240 and 0.656)—see figure 3. Thus, the difference between the two types of speech is at least in the range, which is higher for spontaneous speech. As for the means (see table 2), the difference is not statistically significant: p=0.65 for age group 30–45 and p=0.99 for all speakers.

That is, for many speakers from the spontaneous corpus, average SID values fall within the range typical for read speech. That is, in terms of SID, some speakers may talk just the way they would read; however, this cannot be proved or rejected by our data, as speakers are all different. A comparison where the same Russian speakers recorded similar read and spontaneous material [7] showed that for some speakers average pause duration is greater in read speech, for some speakers—in spontaneous speech. Therefore, speakers probably use different phrasing strategies, and SID is not a characteristic of either spontaneous or read speech.

The ratio between the number of silent pauses and the total number of IP boundaries (functional pauses) shows a similar pattern. The values for read speech range between 56 and

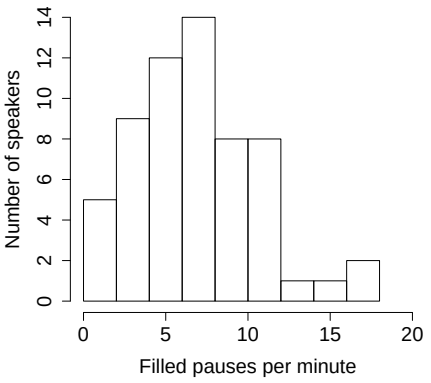

Figure 4: *Number of filled pauses per minute of pure speech (i.e. excluding silent intervals) in spontaneous speech*

69 %, for spontaneous speech – between 39 and 90 %. Thus, the speakers might be roughly divided into three groups: those who use pausing the way it is done in reading, those who use more silent breaks (and this is a sign of spontaneity), and those who use fewer silent breaks (maybe because they also use facial expression and gestures as additional boundary markers).

Figure 4 shows the number of filled pauses per minute of pure speech (i.e. excluding silent intervals). The graph shows that filled pauses—probably, the most outstanding trait of spontaneous speech—are used by all 60 speakers from the corpus of spontaneous speech, but the frequency differs across speakers. The mean across speakers is 6.9, and in most cases the value does not exceed 12 filled pauses per minute.

While analyzing the variability of the number of filled pauses per minute, we made an interesting observation: though the values vary across speakers within the range 0.8–17.8, those speakers that are close relatives tend to have very close values. In our spontaneous corpus there were 4 pairs of close relatives: two pairs of same-gender siblings (aged 30–34) and two married couples (aged above 54); the differences within the pairs did not exceed 1.5. This opens a direction of further research involving more pairs of close relatives. Another observation here is that students tend to have few filled pauses per minute compared with other speakers of greater age, as 4 of 5 students fall within the range 1.3–4.4. This tendency also requires further investigation.

4. Conclusion

Our data confirm that differences in speech planning often result in differences in prosodic phrasing. Generating speech "in real time" requires more resources, and at some point the speaker runs out of time, which leads to unplanned interruptions; speakers hurry up to say the most important words first, before the time they might run out of resources. However, reading is not as easy as it may seem, either. Looking at a printed sentence, a speaker must quickly interpret the meaning and generate the appropriate melody and phrasing, and this requires special skills or preparation. But in this case we might assume that one can prepare for spontaneous recording as well or just be skilled at spontaneous speaking. On the other hand, there might be examples of "unskilled" and unprepared *read* speech, which is rarely described in literature. All of this could explain the overlap between the values that we obtained here, as well as the fact that there is often no perceptual difference between the two styles.

5. References

[1] R. M. Haynes, L. White, and S. L. Mattys, "What do we expect spontaneous speech to sound like?" in *Proc. of ICPhS*, 2015.

[2] H. Levin, C. A. Schaffer, and C. Snow, "The prosodic and paralinguistic features of reading and telling stories," *Language and Speech*, vol. 25, no. 1, pp. 43–54, 1982.

[3] G. P. M. Laan, "Perceptual differences between spontaneous and read aloud speech," in *Proceedings of the Institute of Phonetic Sciences Amsterdam, 16*, 1992, pp. 65–79.

[4] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech and Language*, vol. 22, no. 2, pp. 171 – 184, 2008.

[5] L. E. de Ruiter, "Information status marking in spontaneous vs. read speech in story-telling tasks evidence from intonation analysis using GToBIs," *Journal of Phonetics*, vol. 48, pp. 29 – 44, 2015.

[6] I. Kovtunova, *Contemporary Russian. Word order and information structure [Sovremenniy russkiy jazyk. Poryadok slov i informatsionnaya strukturapredlozhenya]*. Russia: Editorial URSS, 2002.

[7] L. Bondarko, N. Volskaya, S. Tananaiko, and L. Vasilieva, "Phonetic properties of Russian spontaneous speech," in *Proc. of ICPhS*, 2003.

[8] P. Skrelin, N. Volskaya, D. Kocharov, K. Evgrafova, O. Glotova, and V. Evdokimova, "CORPRES—corpus of russian professionally read speech," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horak, I. Kopecek, and K. Pala, Eds. Springer Berlin Heidelberg, 2010, pp. 392–399.

[9] T. Kachkovskaia, D. Kocharov, P. Skrelin, and N. Volskaya, "CoRuSS—a new prosodically annotated corpus of Russian spontaneous speech," in *Proc. of LREC*, 2016, pp. 1949–1954.

[10] N. Volskaya and T. Kachkovskaia, "Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS," in *Proc. of Speech Prosody 8*, 2016, pp. 917–921.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. Version 6.0.37," http://www.praat.org/, 2018.

[12] M. Nespor and I. Vogel, *Prosodic Phonology: With a New Foreword*. Mouton de Gruyter, 2007.

[13] J. A. Bryzgunova, "Intonation [intonaciya]," in *Russian Grammar [Russkaya Grammatika]*, N. J. Shvedova *et al.*, Eds. Moscow, Nauka, 1980, vol. 1.

[14] N. Svetozarova and A. Kuosmanen, "Declination and finality in spontaneous and read speech in Russian," in *Proc. of ICPhS*, 2003.