A Fully Annotated Corpus of Russian Speech

Pavel Skrelin, Nina Volskaya, Daniil Kocharov,
Karina Evgrafova, Olga Glotova, Vera Evdokimova
Department of Phonetics, Saint-Petersburg State University
Universitetskaya Emb., 11, 199034, Saint-Petersburg, Russia
E-mail: skrelin@phonetics.pu.ru, volni@phonetics.pu.ru, kocharov@phonetics.pu.ru,
evgrafova@phonetics.pu.ru, oglotova@phonetics.pu.ru, postmaster@phonetics.pu.ru

Abstract

The paper introduces CORPRES – a fully annotated Russian speech corpus developed at the Department of Phonetics, St. Petersburg State University as a result of a three-year project. The corpus includes samples of different speaking styles produced by 4 male and 4 female speakers. Six levels of annotation cover all phonetic and prosodic information about the recorded speech data, including labels for pitch marks, phonetic events, narrow and wide phonetic transcription, orthographic and prosodic transcription. Precise phonetic transcription of the data provides an especially valuable resource for both research and development purposes. Overall corpus size is 528 458 running words and contains 60 hours of speech made up of 7.5 hours from each speaker. 40% of the corpus was manually segmented and fully annotated on all six levels. 60% of the corpus was partly annotated; there are labels for pitch period and phonetic event labels. Orthographic, prosodic and ideal phonetic transcription for this part was generated and stored as text files. The fully annotated part of the corpus covers all speaking styles included in the corpus and all speakers. The paper contains information about CORPRES design and annotation principles, overall data description and some speculation about possible use of the corpus.

1. Introduction

Contemporary research both in linguistic phonetics and speech technology is largely based on and can largely benefit from the use of large speech corpora. The corpus to be used for these purposes needs to meet the following requirements: it has to contain a large sample of speech data, to ensure a consistently high quality of the data, and to have annotation that enables researchers of a wide range of phonetic issues to search for and find specific data that is valid and reliable. Good examples of such a resource are the corpora developed for Dutch (Van Son et al., 2001). For the Russian language, the existing speech corpora tend to serve a narrow practical purpose (Arlazarov et. al., 2004). Therefore, the need for a fully annotated large corpus of Russian speech recorded at a consistently high quality is evident.

In this paper we present CORPRES – a fully annotated COrpus of Russian Professionally REad Speech developed at the Department of Phonetics, Saint-Petersburg State University as a result of a three-year project. The corpus meets all of the requirements to databases of this kind listed above and may be used both for the purposes of development and scientific research. It is large enough for statistical machine learning (60 hours of continuous speech) and has six annotation levels including prosodic annotation, rule-based canonical phonetic transcription and manual transcription reflecting the actual sounds pronounced by the speakers. In the paper, we describe the corpus design and data and discuss the principles and issues behind its development.

2. Corpus Design

The aim of the corpus was to provide a large sample of Standard Russian continuous speech. It was originally intended for use in unit-selection TTS synthesis, however, with the idea that it might be suitable for use in a wider range of phonetic research and development. Therefore, the corpus was designed along a number of principles.

Firstly, the sample was to represent a number of speaking styles. As the corpus included only read speech, different styles of texts were selected for recording with specific characteristics of those styles in mind:

- an action-oriented fiction narrative resembling conversational speech;
- a fiction narrative of a more descriptive nature containing longer sentences and very little direct speech;
- a play containing a high number of conversational remarks and emotionally expressive dialogues and monologues;
- purely informational neutral texts on IT, politics and economy containing terminology, geographical and proper names, numerals, acronyms and abbreviations.

The choice of diverse texts served our other goal of making the corpus phonetically and prosodically rich, i.e. to contain a large number of all Russian phonemes in all possible contexts and a wide range of diverse prosodic structures, and to provide for good lexical representation. The corpus is composed of 60 hours of speech recorded from 8 speakers (7.5 hours from each speaker).

Thirdly, the corpus was intended as a sample of Standard Russian (St. Petersburg pronunciation variant); dialect variation was not accounted for. However, records were made from eight speakers, four men and four women, in order to cover a certain degree of variation within the St. Petersburg pronunciation variant.

Fourthly, it was necessary to ensure consistently high quality of all data both in terms of technical characteristics and voice quality. The latter objective was achieved by recording professional speakers: some of them worked in radio broadcasting; some were professional actors or television newsmen. In addition to voice training, pleasantness of voice and clear articulation were considered.

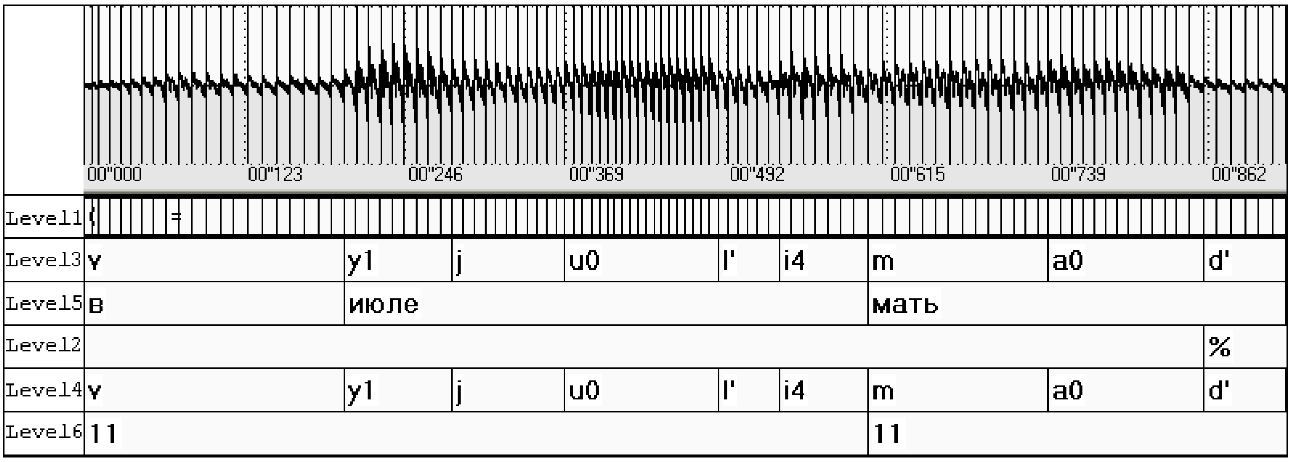

Figure 1: Annotation levels.

The recordings were made in the recording studio at the Department of Phonetics, University of St. Petersburg. Motu Traveler multi-channel recording system, an AKG capacitor microphone and WaveLab software were used. The recordings have a sample rate of 22050 Hz and a bitrate of 16 bits. Before the recording sessions, all texts were revised to detect and resolve ambiguities caused by nonstandard words, terminology etc. All transliterated foreign language words, terminology, acronyms and numbers were clarified in the prompts to avoid difficulties and mistakes. In case of doubt, speakers could ask for instructions from researchers present at the studio. Slips of the tongue were noted, and the speakers were asked to read the passages where they occurred once again.

The final, but the most crucial objective we had in mind was to ensure that the annotation of the corpus covers a wide range of information that may be of interest to those involved in most areas of phonetic research. There are six annotation levels that will be further discussed in greater detail.

3. Annotation

The annotation captures the maximum amount of phonetically and prosodically relevant data. The six annotation levels are as follows:

Level 1 – pitch marks;

Level 2 – phonetic events labeling;

Level 3 – real phonetic transcription (this is performed manually and reflects the sounds actually pronounced by the speakers);

Level 4 – ideal phonetic transcription (this level is automatically generated by a linguistic transcriber in accordance with a canonical set of rules);

Level 5 - orthographic transcription;

Level 6 – prosodic transcription.

Levels 1 and 2 contain information on various phonetic events: epenthetic vowels, voice onset time, voiced plosure, stationary parts of voiceless consonants, laryngalization, and glottalization. The phonetic events were annotated manually by expert phoneticians.

Level 5 also contains information on prosodically prominent words.

Prosodic transcription on Level 6 includes labels for different types of pauses, types of tone unit, and non-speech events such as laughter or breathing. Figure 1 shows the six levels at which the annotation is done. (Levels 1-6 are not in numerical order for the purpose of clearer visual design.)

3.1 Detecting and Labeling Periods of Fundamental Frequency

The fundamental frequency periods were detected automatically. A linear combination of the following methods was used for this purpose: autocorrelation, analysis-by-synthesis, spectral domain analysis, estimation of the energy of signal peaks and estimation of the ratio of lengths and correlation of neighboring periods. For a detailed description of the algorithm, see (Kocharov, 2008). The efficiency of automatic pitch detection and pitch periods labeling was about 98%. The results of the automatic procedure were checked and corrected manually.

3.2 Phonetic Transcription

Phonetic transcription is of fundamental importance in speech corpora as it reflects characteristic phonetic features of speech. The transcription system should be well-grounded linguistically and also comprehensible for corpus users. In CORPRES transcription is available at two levels. Level 3 contains narrow phonetic transcription. We called this transcription level ‘real’ phonetic transcription because it reflects the sounds actually pronounced by the speakers. The ‘ideal’ transcription found at Level 4 was generated in accordance with a set of phonological rules without reference to the actual sound. As a result, Level 4 contains a canonical phonetic transcription of the speech sample. The transcription symbols used were a version of SAMPA for the Russian language. To mark positional allophones of 6 Russian vowel phonemes /a/, /o/, /i/, /u/, /e/, /y/ 18 symbols were used. Each vowel symbol contained indication of the sound’s position regarding stress. Thus 0 was used to for a stressed accented vowel, 1 - for an unstressed vowel in a pretonic syllable, 4 – an unstressed one in a post-tonic syllable.

The set of consonant symbols included 41 symbols to cover 36 Russian consonant phonemes and 5 voiced allophones of voiceless consonants which occur frequently at word junctions.

To produce the real phonetic transcription, the speech signal was manually segmented, transcribed and peer-revised by expert phoneticians.

Ideal phonetic transcription was generated automatically by an automatic transcriber. The labels were placed automatically to coincide with the label positions produced manually on the real transcription level. Procedure of automatic labeling is based on calculating the Levenshtein distance. Automatic labeling is not perfect due to the mismatch of ideal and real phonetic transcriptions. Therefore, the results of the automatic procedure were further manually corrected.

3.3 Orthographic and Prosodic Transcription

Prosodic information was marked by expert phoneticians on the basis of perceptual and acoustic analysis of the speech data in a text file containing orthographic transcription. Labels were later automatically transferred from the text file to the annotation files to coincide with the phonetic transcription levels. Orthographic transcription was stored on Level 5, it contains the boundaries of words and word labels. Besides the prosodically prominent words are labeled with special symbols. Prosodic information was stored on Level 6, it contains the boundaries of tone units and pauses and their labels. The set of symbols to label pauses and tone units and the principles behind the labeling process are described in detail in (Volskaya & Skrelin, 2009).

4. Corpus Data Description

Overall corpus size is 528,458 running words. 40% of the corpus (24 hours of speech) was manually segmented and fully annotated on all six levels. 60% of the corpus was partly annotated; there are labels for pitch period and phonetic event labels. Orthographic and prosodic transcription, as well as the ideal phonetic transcription (see Section 3 for detail) for this part was generated and then stored as text files, but was not transferred to sound file labels. The fully annotated part of the corpus covers all speaking styles included in the corpus and all speakers. Table 1 shows general corpus statistics.

	Fully Annotated Data	Partly Annotated Data	Total Amount
Phonemes	1 048 867	–	–
Words	211 437	317 021	528 458
Tone Units	64 055	86 546	150 601
Hours	24	36	60

Table 1: General corpus statistics.

It is impossible to estimate the number of phonemes in the part of the corpus which was not annotated on phonetic transcription levels, therefore, two cells in the table remain empty.

5. Findings Based on the Corpus Data

As CORPRES contains a large sample of high quality speech data with detailed annotation, it enables researchers of a wide range of phonetic issues to search for and find specific data that is valid and reliable. The fact makes it suitable for use in a wide range of phonetic research. For the time being, the necessary information from the corpus (e.g. sound variants and their frequency distribution and etc.) is obtained by means of specially designed computer programs to suit a certain task.

For instance, consulting the corpus we can obtain important information about the changes in the Russian standard pronunciation (Bondarko, 2009). In Table 2 we compare the ideal phonetic transcription reflecting the way the speech sample is supposed to be pronounced according to the canonical transcription rules of the Russian language and the real phonetic transcription reflecting the way it actually was pronounced by the speakers recorded.

	Total	Correctly	Mispronounced	Elided
Count	1 118 833	947 508	101 292	70 033
Percents	100	84.7	9.05	6.25

Table 2: Ideal vs. real transcription.

Table 2 reveals that despite the fact that as many as 84.7% of the ideal transcription reflects the actual pronunciation, 9.05% of the expected sounds are replaced by other sounds, and 6.25% of the expected sounds are actually not pronounced at all.

Table 3 shows in percentage terms the ratio between vowel realizations according to ideal transcription (down) and real transcription (across). 0 is used to mark a stressed vowel, 1 – a pretonic vowel, and 4 – a post-tonic vowel. The column Total shows the whole number of corresponding allophones.

This data shows that there is a certain degree of variation even for stressed vowels that tend to be more stable than the unstressed ones, with approximately 1-3% of them pronounced as allophones of other phonemes. Some of the unstressed vowels are especially unstable, e.g. less than 50% of post-tonic /a/ vowels are pronounced as /a/, while a third of them is pronounced as /y/ allophones. The vowel variation findings support those obtained earlier on a smaller corpus of read and spontaneous speech (Bolotova 2003).

A closer look at vowel variation data provides insight into the changes in Standard Russian. The general phonotactic rule for unstressed vowels is that /e/ and /o/ do not generally occur in the unstressed position, but can be found in a small number of words, mostly loan words and foreign names, and contexts (post-tonic /e/ is mostly found in word-final open syllables) (e.g. *radio /r a0 d' i4* ***o4/***, *izvinite /i1 z v' i1 n' i0 t'* ***e4/***, *Hemingway /h e1 m' i1 n g u1 e0 j/*. Our data showed that unstressed /e/ is pronounced as /i/ or /y/ in 40-45% of the cases. The unstressed /o/ is pronounced in 77.4% and appears to be more stable. Therefore, we may assume that the phonotactics of Standard Russian is going through change in this respect.

	a	**e**	**i**	**o**	**u**	**y**	**Total**
a0	**98.3**	1.5		0.1		0.1	52 769
a1	**80.7**	3.9	0.1	1.6	0.5	**13.1**	76 992
a4	**46.3**	**13.2**	1.6	4.6	1.3	**33**	53 667
e0		**97.6**	1	0.4		0.9	30 861
e1	0.6	**61**	**13.2**	0.6	0.6	**23.9**	159
e4		**55.6**	**18.9**	1.1	2.2	**22.2**	90
i0		0.5	**98.9**		0.1	0.5	20 596
i1	0.1	6.2	**91**	0.2	0.8	1.8	47 840
i4	0.6	**19**	**77.4**	0.3	0.9	1.9	38 799
o0	0.1	0.2		**99.1**	0.2	0.3	43 875
o1	1.3	0.3	0.1	**93.4**	2.2	2.8	1 945
o4	7.1	3		**71.7**	5.1	**13.1**	99
u0				0.2	**99.7**	0.1	12 503
u1			0.2	0.9	**98.5**	0.4	12 729
u4	0.2	1.6	0.9	2.4	**92.8**	2.1	9 144
y0		0.4	0.6		1	**97.9**	9 355
y1	1.3	6.9	7.1	0.8	2	**81.9**	6 275
y4	1	9.2	0.3	0.8	2	**86.7**	14 337

Table 3: Ideal vs. real transcription: vowels.

As the annotated part of the corpus used for this analysis includes an even distribution of all of the represented speaking styles and speakers, we can expect that similar results could be obtained from the analysis of the rest of the corpus. This clearly shows that the ideal transcription alone does not yield data that would be sufficient or valid for any type of phonetic research or practical application. Therefore, despite the large amount of human and financial resources required, precise phonetic transcription seems to be an indispensible part of corpus annotation at the present moment. There appear to be two ways of overcoming the discrepancy between rule-based transcription and manual transcription. One possible solution is to bring the automatic transcriber up-to-date by using the obtained information about the actual sound pronunciation. In this respect, the present corpus and its two levels of phonetic transcription may be used as a database for revising the traditional view of Standard Russian pronunciation and introducing new phonetic transcription rules. The other solution is to avoid automatic rule-based transcription altogether and transcribe all of the data manually. The former course of action appears to be more preferable as the emergence of a set of rules reflecting the current state of the language would largely benefit both the development of speech technology applications and theoretical research in Russian phonetics.

6. Conclusion

The Department of Phonetics, SPSU developed a fully-annotated large corpus of Russian speech including samples of different speaking styles produced by 4 male and 4 female speakers. The six levels of annotation cover all phonetic and prosodic information about the recorded speech data. Precise phonetic transcription of the data provides an especially valuable resource for both research and development. The corpus may be used for unit-selection TTS synthesis purposes, as well as a bootstrapping corpus for speech recognition systems, or as data for research in Russian phonetics and inter- and intra-speaker variability.

7. References

Arlazarov V.L., Bogdanov D.S., Krivnova O.F., and Podrabinovitch A.Ya. (2004). Creation of Russian Speech Databases: Design, Processing, Development Tools. In *Proceedings of SPECOM'2004.* St. Petersburg, pp. 650--656.

Bolotova O. (2003). On some acoustic features of spontaneous speech and reading in Russian (quantitative and qualitative comparison methods). In: *Proceedings of the 15th International Congress of Phonetic Sciences,* Barcelona: Causal Productions Pty Ltd, pp. 913--916.

Bondarko L. (2009). Short Description of Russian Sound System. In: De Silva V., Ullakonoja R. (Eds.), *Phonetics of Russian and Finnish: General Description of Phonetic Systems. Experimental Studies on Spontaneous and Read-Aloud Speech.* Frankfurt am Main: Peter Lager, pp. 77--87.

Kocharov D. (2008). Avtomaticheskoe opredelenie chastity osnovnogo tona pri pomoschi linejnoj kombinatsii razlichnih metodov // In. *Materialy XXXVII mezhdunarodnoj filologicheskoj konferentsii*, St. Petersburg: SPbSU, pp. 7--11. (In Russian)

Van Son R.J.J.H., Binnenpoorte D., Van Den Heuvel H. and Pols L.C.W. (2001). The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. In *Proceedings of Eurospeech 2001.* Aalborg, pp. 2051--2054.

Volskaya N.B., Skrelin P.A. (2009). Prosodic model for Russian. In *Proceedings of Nordic Prosody X.* Frankfurt am Main: Peter Lager, pp. 249--260.