Preparing Audio Recordings of Everyday Speech for Prosody Research: The Case of the ORD Corpus

Tatiana Sherstinova(✉)

Saint Petersburg State University,
Universitetskaya nab. 11, St. Petersburg 199034, Russia
t.sherstinova@spbu.ru

Abstract. Studying prosody is important for understanding many linguistic, pragmatic, and discourse phenomena, as well as for solution of many applied tasks (in particular, in speech technologies). Prosody of everyday speech is extremely diverse, demonstrating high interpersonal and intrapersonal variations. Furthermore, natural everyday speech produces a multitude of effects which are hardly possible to obtain in speech laboratories. Because of this fact, it is very important to create resources containing representative collections of everyday speech data. The ORD corpus is a large resource aimed at studying everyday Russian speech. The paper describes the main stages of speech processing in the ORD corpus starting from segmentation of original files into macroepisodes and up to compiling prosody information into the database. This prosody database will be further used for building empirical prosody models.

Keywords: Russian · Everyday speech · Phonetics · Prosody · Duration · Pitch · Sociolinguistics · Communication settings · Pragmatics · Speech corpus

1 The ORD Corpus and Its Data

Studying prosody is important for understanding many linguistic, pragmatic, and discourse phenomena [1–8, etc.], as well as for solution of many applied tasks (in particular, in speech technologies) [9–13]. Prosody of everyday speech is extremely diverse, demonstrating high interpersonal and intrapersonal variations. At the same time prosody may be considered to be central in the interpretation of everyday spoken language [14], as it can completely change the meaning of utterances.

Natural everyday speech produces a multitude of effects which are hardly possible to obtain in speech laboratories [15]. Because of this fact, it is very important to create resources containing representative collections of everyday speech data.

The ORD corpus is a large resource aimed at studying everyday Russian speech. For collecting speech data for the ORD corpus the methodology of longitudinal recordings is used [16–18], for which the participants-volunteers have to spend a whole day with turned-on voice recorders that record all their audible communications. This methodology can be compared with a daily cardio monitoring, which is widely practiced in medicine.

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 623–631, 2017.
DOI: 10.1007/978-3-319-66429-3_62

The ORD corpus was started in 2007 [19]. Recently it was expanded significantly due to the support of the Russian Science Foundation in the framework of the project 'Everyday Russian Language in Different Social Groups' [20]. Nowadays, the corpus contains more than 1250 h of recordings which refer to about 2800 communicative episodes. Those are the recordings of 128 respondents and more than 1000 of their interlocutors, representing different social strata and different gender, age and professional groups of residents of a big Russian city.

The recordings were made in St. Petersburg, Russia in 2007–2016. Speech was recorded in diverse communication settings: the recordings were made at home, in the offices, outdoors, in service centers, in universities and colleges, in coffee bars and restaurants, in transport, in shops, in parks, etc. [19]. Text transcripts are made for 480 communicative episodes (17% audio recordings of corpus) and number 1 million of word usages [21].

All ORD recordings are supplied by sociological information concerning more than 1000 people recorded for the corpus. It allows to make search queries for speech of people with diverse social characteristics.

The ORD collection provides valuable research data for many other interdisciplinary studies like anthropological linguistics, behavioral and communication studies, studies in pragmatics, discourse analysis, psycholinguistics, and forensic phonetics.

Since ORD recordings are not "laboratory speech", only a part of gathered audio data is suitable for phonetic and prosody research. On average, there is only about 1/10 of all macroepisodes, the quality of which allows to conduct phonetic analysis of speech. The paper describes the main stages of speech processing in the ORD corpus: segmentation into macroepisodes, audio conversion, transcribing, segmentation onto words and syllables, obtaining prosody information and its implementation into the database.

2 The Main Stages of Speech Processing in the ORD Corpus

2.1 Segmentation into Macroepisodes

First of all, having received 8–14 h of recordings from each respondent, we are faced with the task to segment it into fragments, which are homogeneous in terms of communication settings (united by setting/scene of communication, social roles of participants and their general activity). We call such fragments "macroepisodes" [22].

Before segmentation, all files are subjected to audio conversion to the format adopted in the corpus: PCM, 22050 Hz, 16 bit, mono. The original recordings are kept in the archive.

The task of segmentation of audio recordings into macroepisodes is performed manually by linguists, who listen all gathered files, defining at the same time the boundaries between episodes. Further, the researchers save each macroepisode into a separate file, make a standardized description for each file in the database, and cut out all "pauses" (i.e., segments not containing speech which are longer than 5 min) from each audio file.

The methodology of macroepisode annotation was described in [22]. Thus, each macroepisode gets both verbal and standardized descriptions in three aspects: (1) Where does the situation take place? (2) What are the participants doing? (3) Who is (are) the main interlocutor(s). In addition, a concise description of the episode may be given in an auxiliary database field called *SceneName*. The duration of each file is indicated in the database, too.

The phonetic quality of each macroepisode is evaluated and measured in a 4-grade scale: 1 – the best quality, suitable for precise phonetic/prosody analysis, 2 – rather good quality, which is partially suitable for phonetic analysis, 3 – noisy recordings of intermediate and low quality, which are not suitable for phonetic analysis but are suitable enough for other aspects of research, and 4 – unintelligible conversations or remarks in extreme noise, which could not be understood without noise reduction techniques [23].

At this stage, macroepisodes, which are to be transcribed, are selected with a priority indication of their ranks in the database. When choosing files for transcribing, phonetic quality is usually considered, however, it is not the only factor that is taken into account (the other important causes may be linguistic, pragmatic or discourse peculiarities of the recorded data, as well as anthropological issues).

2.2 Speech Transcribing and Primarily Annotation

Selected macroepisodes are further subjected to transcribing and primarily multilevel annotation both of which are made in ELAN [24]. The main principles for transcribing and annotating are described in [19].

Besides speech transcripts and the correspondent anonymized codes of speakers, primarily annotation contains the following information: (1) voice quality (e.g., hoarse, whisper, scanning, irritated, imitating, ironical, dramatic, etc.); (2) non-language audio events (dog barking, squeak of a door, phone ring, etc.); and (3) "miniepisodes", which are minor communicational units homogeneous either by the topic of conversation or by its main pragmatic task [25]. Other linguistic, pragmatic or discourse comments are to be written on layers *FraseComment* and *Notes*.

Here, it should be mentioned that in the first transcripts of the corpus, there was only one level reserved for speech transcription in the annotation template. The multiple cases of overlapping speech were marked by special symbols # and @ in linearized transcript [ibid.]. This form of transcript is convenient enough for further linguistic annotation, however it does not reflect the audio reality in fragments with overlapping speech.

Because of that fact, since 2014 we practice multilevel speech transcribing similar to that used in Conversational Analysis, when each participant of the recorded conversation has his own level for transcription. In order to maintain compatibility with previous transcripts of the corpus, currently we practice both versions of transcribing: being initially made in a linear form, speech transcription is later converted into its multilevel variant.

Transcripts are made manually by linguists in ELAN, each transcript being then checked and approved by two or three experts. After that, the files are subjected to automatic processing.

2.3 Automatic Processing of Transcripts

Further, all annotation files are processed by means of *Corrector* software utility, specially developed for the ORD corpus. It was designed to automatically fix possible technical drawbacks in transcripts (e.g., to remove extra spaces), and to reveal possible mismatch between the levels of speech and speakers that may occur in cases of overlapping speech. Such a situation is often encountered in everyday conversations, making it very difficult to analyze speech. In cases where such discrepancies were detected, manual expert correction of the corresponding fragments is made followed by another launch of *Corrector* utility. This is a necessary step for further processing of annotation files.

After that, annotation files are processed by another ORD utility – *Eafer* program – with the help of which the linear one-level transcript is converted into several layers, each of which referring to one participant of the conversation. This approach allows to separate speech from different speakers, no matter how many people are participating in the conversation and to which social groups they belong.

At the next stage, the boundaries of annotation boxes are to be manually adjusted on fragments with overlapping speech. This procedure is made directly in ELAN. After that, the annotation files are ready for phonetic transcribing and segmentation.

2.4 Phonetic Transcribing and Segmentation

Phonetic transcribing of ORD transcripts is made automatically with the use of software specially designed for this purpose by Speech Technology Center [26].

The following set of allophones is used:
[a0], [a1], [a2], [a4], [o0], [o1], [o4], [e0], [e1], [y0], [y1], [y4], [u0], [u1], [u4], [i0], [i1], [i4], [b], [b'], [p], [p'], [d], [d'], [t], [t'], [g], [g'], [k], [k'], [c], [ch], [v], [v'], [f], [f'], [z], [z'], [s], [s'], [zh], [sh], [sc], [h], [h'], [m], [m'], [n], [n'], [l], [l'], [r], [r'], [j].

The numbers after vowels have the following meanings: 0 – stressed, 1 – pre-stressed, 4 – post-stressed. For /a/, in addition, the second pre-stressed position [a2] is distinguished.

For transcribing, the software uses the typical algorithm of conversion of text into sequence of allophones. Besides, it can distinguish different variants of word pronunciation, which are described in the Lexicon of exceptions.

For example, for the frequent Russian word "*sejchas*" ("*now*"), the transcription based on standard rules will be [s'i1jcha0s], but this full form rarely occurs in spontaneous speech. Instead, two other variants are usually used: [s'i1cha0s] and [sca0s]. Because of that, all non-standard forms should be listed in the Lexicon. When a program comes across any word from this list, its decision on its pronunciation is based on statistical variability of each variant which is calculated on the base of comparison of audio data from the corresponding wave segment with the variants described in the Lexicon.

The other important function of this software is to segment audio file into words and allophones. Actually, it means to define segment boundaries on these two levels. Technologically, the algorithm is also based on the usage of statistical probabilities [27], which takes into account three following aspects: acoustic data, speech transcript, and the Lexicon.

The program has two files as input: (1) audio file, and (2) ELAN-annotation file with the level of speech transcript, on which each utterance is referred to correspondent time segment. The result of the program is the updated ELAN-annotation file, which has two additional levels – for words and allophone segments (see Fig. 1).

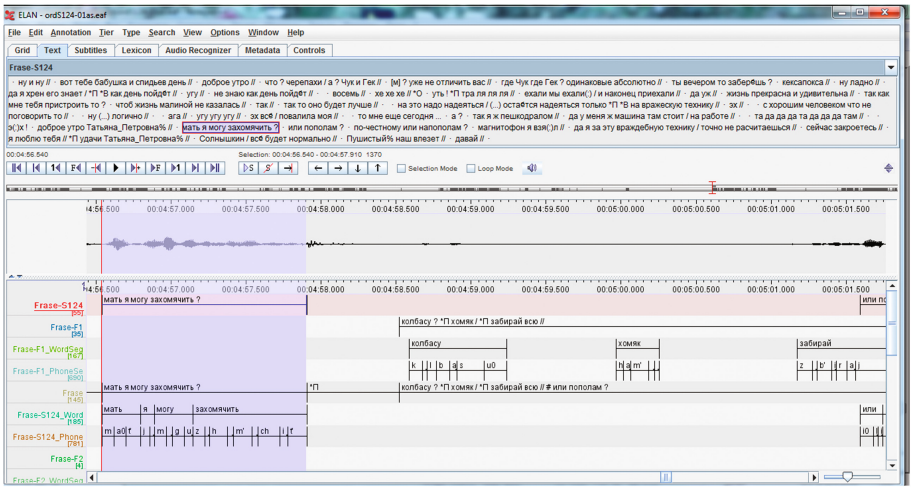

Fig. 1. Multilevel speech annotation in ELAN with its segmentation into words and allophones

The efficiency of this software depends to a large extent on the phonetic quality of the recorded signal. Thus, low level of recording, background noise or overlapping speech significantly worsen the results. As for the accuracy of segment boundaries, it is typically better on neutral speech fragments rather than on emotional speech, which is frequently characterized by a significant prolongation of sounds, unforeseen by the model, and therefore requiring expert correction. Generally, the use of this software allows to significantly reduce the labor costs for manual speech segmentation.

2.5 Duration, Pitch, and Intensity

The information on the duration of speech segments is easily obtained from segmentation data.

Recently, the new version of the utility described in the previous section has been developed by Speech Technology Center. Beside graphic interface, it has got new facilities allowing to automatically get the information concerning the mean values of F0, F1 and F2 measured in Hz.

Therefore, two prosody parameters – allophone duration and its average pitch – may be easily calculated for any allophone and further exported into the database.

For illustrative purposes, the example of such information for one phrase – *Vit'ka mne rasskazal vsjo pro jelektronnye sigarety [Vitka told me everything about electronic cigarettes]* – is presented in Table 1.

Table 1. The fragment of the table ALLOPHONES from the ORD Database

Macro-episode	SC	Phrase	Word	Allophone	Dur (ms)	F0	F1	F2
ordS33–15	S33	F32	*Vit'ka*	v'	70	122	384	1693
ordS33–15	S33	F32	*Vit'ka*	i0	140	185	446	1614
ordS33–15	S33	F32	*Vit'ka*	t'	40			
ordS33–15	S33	F32	*Vit'ka*	k	70			
ordS33–15	S33	F32	*Vit'ka*	a4	30	158	446	1401
ordS33–15	S33	F32	*mne*	m	40	145	443	1522
ordS33–15	S33	F32	*mne*	n'	60	139	458	1755
ordS33–15	S33	F32	*mne*	e0	40	137	493	1579
ordS33–15	S33	F32	*rasskazal*	r	50	131	501	1415
ordS33–15	S33	F32	*rasskazal*	a2	30	128	510	1321
ordS33–15	S33	F32	*rasskazal*	s	60			
ordS33–15	S33	F32	*rasskazal*	k	80			
ordS33–15	S33	F32	*rasskazal*	a1	50	124	512	1308
ordS33–15	S33	F32	*rasskazal*	z	90			
ordS33–15	S33	F32	*rasskazal*	a0	40	117	515	1443
ordS33–15	S33	F32	*rasskazal*	l	70			
ordS33–15	S33	F32	*vsjo*	f	60			
ordS33–15	S33	F32	*vsjo*	s'	160			
ordS33–15	S33	F32	*vsjo*	o0	120	181	519	1093
ordS33–15	S33	F32	*pro*	p	100			
ordS33–15	S33	F32	*pro*	r	50	117	435	1314
ordS33–15	S33	F32	*pro*	a2	30	117	454	1435
ordS33–15	S33	F32	*jelektronnye*	y1	30	113	448	1687
ordS33–15	S33	F32	*jelektronnye*	l'	70	109	387	1621
ordS33–15	S33	F32	*jelektronnye*	i1	30	107	415	1584
ordS33–15	S33	F32	*jelektronnye*	k	40			
ordS33–15	S33	F32	*jelektronnye*	t	70			
ordS33–15	S33	F32	*jelektronnye*	r	60	117	471	1168
ordS33–15	S33	F32	*jelektronnye*	o0	80	119	473	1069
ordS33–15	S33	F32	*jelektronnye*	n	60	120	323	1115
ordS33–15	S33	F32	*jelektronnye*	y4	45	117	404	1273
ordS33–15	S33	F32	*jelektronnye*	i4	45	111	472	1462
ordS33–15	S33	F32	*sigarety*	s'	110			
ordS33–15	S33	F32	*sigarety*	i1	30			
ordS33–15	S33	F32	*sigarety*	g	77			
ordS33–15	S33	F32	*sigarety*	a1	83	97	558	1652
ordS33–15	S33	F32	*sigarety*	r'	60	91	436	1772
ordS33–15	S33	F32	*sigarety*	e0	80	83	439	1744
ordS33–15	S33	F32	*sigarety*	t	120			
ordS33–15	S33	F32	*sigarety*	a4	110			

In particular, it contains data referring to (1) macroepisode (i.e., sound file), which is a link to the information on communication settings; (2) speaker's code (SC), which is a link to sociolinguistic information about speakers; (3) the phrase itself; (4) word; (5) allophone; (6) correspondent boundaries (not shown in Table 1); (7) allophone duration; (8) average pitch; (9) average F1; and (10) average F2.

As for the detailed dynamics of pitch and the intensity, they may be analyzed in Praat [28] after exporting annotation data from ELAN to TextGrid.

3 Conclusion

It this concise review, we have described the main points of preparation of the ORD audio data to prosody research. In the result of such processing, the prosodic data are accumulated in the corpus database, where they can be linked with other relevant information (linguistic, pragmatic and discourse). Therefore, it will be possible to analyze speech with specified parameters (e.g., recorded in a specific place, under specific circumstances, by a speaker of specific characteristics, etc.). The compiled prosody database will be further used for building empirical prosody models. Besides, it seems particularly perspective to combine prosody information with pragmatic annotation of speech acts [29].

Acknowledgements. The creation of the ORD speech corpus was supported by several grants: Russian Foundation for Humanities projects No. 07–04–94515e/Ya (Speech Corpus of Russian Everyday Communication "One Speaker's Day") and No. 12–04–12017 (Information System of Communication Scenarios of Russian Spontaneous Speech), the Russian Ministry of Education project "Sound Form of Russian Grammar System in Communicative and Informational Approach". Significant extension of the corpus and the software development was achieved in the framework the project "Everyday Russian Language in Different Social Groups" supported by the Russian Science Foundation, project No. 14–18–02070.

References

1. Couper-Kuhlen, E.: English Speech Rhythm: Form and Function in Everyday Verbal Interaction. John Benjamins Publications, Amsterdam (1993)
2. Couper-Kuhlen, E., Selting, M. (eds.): Prosody in conversation: Interactional studies. Cambridge University Press, Cambridge (1996)
3. Wells, B., Macfarlane, S.: Prosody as an interactional resource: turn-projection and overlap. Lang. Speech **41**, 265–294 (1998)
4. Klatt, D.H.: Linguistic uses of segmental duration in English: acoustic and perceptual evidence. J. Acoust. Soc. Am. **59**, 1208–1221 (1976)
5. Kello, C.T.: Patterns of timing in the acquisition, perception, and production of speech. J. Phonetics **31**(3–4), 619–626 (2003)
6. Campbell, N.: Timing in speech. A Multi-Level Process. In: Horne, M. (ed.) Prosody: Theory and Experiment, pp. 281–334. Kluwer Academic Publishers (2000)
7. O'Connell, D.C.: Communicating with One Another: Toward a Psychology of Spontaneous Spoken Discourse. Springer New York, New York (2008)

8. Barth-Weingarten, D., Reber, E., Selting, M.: Prosody in interaction. John Benjamins, Amsterdam, Philadelphia (2010)
9. Benesty, J., Sondhi, M., Huang, Y. (eds.): Handbook of Speech Processing, Springer (2008)
10. Harrington, J.: The Phonetic Analysis of Speech Corpora. Wiley-Blackwell, Chichester (2010)
11. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Pearson Prentice Hall, Englewood Cliffs (2001)
12. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall, Englewood Cliffs (2008)
13. Potapova, R.K., Potapov, V.V., Lebedeva, N.N., Agibalova. T.V.: Interdisciplinarity in the study of speech polyinformativity. Languages of Slavic Culture (2015)
14. Wennerstrom, A.K.: The Music of Everyday Speech: Prosody and discourse analysis. Oxford University Press, New York (2001)
15. Cummins, F.: Probing the dynamics of speech production. In: Sudhoff, S. et al. (ed.) Methods in Empirical Prosody Research. Language, Context and Cognition. W. De Gruyter, Berlin–New York, pp. 211–228 (2006)
16. Sibata, T.: Sociolinguistics in Japanese contexts. In: Kunihiro, T., Inoue, F., Long, D. (eds.) Mouton de Gruyter. Berlin-New York (1999)
17. Campbell, N.: Speech & expression; the value of a longitudinal corpus. LREC **2004**, 183–186 (2004)
18. Burnard, L. (ed.): Reference guide for the British National Corpus (XML edition). Published for the British National Corpus Consortium by Oxford University Computing Services (2007). http://www.natcorp.ox.ac.uk/docs/URG/. Accessed 2 June 2017
19. Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., Sherstinova, T.: The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 250–257. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04208-9_36
20. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Ermolova, O., Baeva, E., Martynenko, G., Ryko, A.: Sociolinguistic extension of the ORD corpus of Russian everyday speech. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS, vol. 9811, pp. 659–666. Springer, Cham (2016). doi:10.1007/978-3-319-43958-7_80
21. Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Ermolova, O., Baeva, E., Martynenko, G., Ryko, A.: Everyday Russian language in different social groups. Commun. Res. **2**(8), 81–92 (2016)
22. Sherstinova, T.: Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 268–276. Springer, Cham (2015). doi:10.1007/978-3-319-23132-7_33
23. Sherstinova, T.: The structure of the ORD speech corpus of Russian everyday communication. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 258–265. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04208-9_37
24. Hellwig, B., Van Uytvanck, D., Hulsbosch, M., et al.: ELAN – Linguistic Annotator. Version 5.0.0-alfa [in:]. http://www.mpi.nl/corpus/html/elan/. Accessed 28 Mar 2017
25. Sherstinova, T.: Pragmaticheskoe annotirovanie konnunicativnykh jedinic v korpuse ORD: mikroepisody i rechevye akty (Approaches to Pragmatic Annotation in the ORD Corpus: Microepisodes and Speech Acts). In: Proceedings of the International Conference on "Corpus linguistics-2015", pp. 436–446 (2015)
26. Speech Technology Center. http://speechpro.com

27. Prodan, A., Chistikov, P., Talanov, A.: The system of preparation of a new voice for the speech synthesis system "VITALVOICE". Komp'juternaja lingvistika i intellektual'nye tehnologii **9**(16), 394–399 (2010)
28. Praat: Doing Phonetics by computer. http://www.praat.org
29. Sherstinova, T.: Speech acts annotation of everyday conversations in the ORD corpus of spoken Russian. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) Speech and Computer (SPECOM 2016). LNAI. Springer, Switzerland (2016)