Automatic Phonetic Transcription for Russian: Speech Variability Modeling

Vera Evdokimova(✉), Pavel Skrelin, and Tatiana Chukaeva

Saint Petersburg State University, 7/9 Universitetskaya nab.,
St. Petersburg 199034, Russia
{postmaster,skrelin,chukaeva}@phonetics.pu.ru
http://www.phonetics.spbu.ru

Abstract. At the moment more advanced approaches to phonetic transcription are required for different speech technology tasks such as TTS or ASR. All subtle differences in phonetic characteristics of sound sequences inside the words and in the word boundaries need more accurate and variable transcription rules. Moreover, there is a need to take into account not only the normal rules of phonetic transcription. it is important to include the information about speech variability in regional and social dialects, popular speech and colloquial variants of the high frequency lexis. In this paper a reliable method for automatic phonetic transcription of Russian text is presented. The system is used for making not only an ideal transcription for the Russian text but also takes into account the complex processes of sound change and variation within the Russian standard pronunciation. Our transcribing system is reliable and could be used not only for the TTS systems but also in ASR tasks that require more flexible approach to phonetic transcription of the text.

Keywords: Automatic phonetic transcription · Russian · Phonetics · Speech processing · Speech transcription · Speech variability modeling

1 Introduction

For the last 30 years various large speech corpora have been developed through the world [1]. Well-known examples are TIMIT [2], Switchboard [3], Verbmobil [4], the Spoken Dutch Corpus [5] and the Corpus of Spontaneous Japanese [6]. At the moment a number of medium and large size Russian speech corpora are available. The largest published corpus of the Russian speech is ORD (One Day of Speech) corpus that is still under development [7]. It contains more than 1000 h of everyday speech. It has partial annotation and transcription. However, this corpus is not publicly available. The most annotated publicly available corpus nowadays is PrACS-Russ (Prosodically Annotated Corpus of Spoken Russian) that contains over 4 h of monologue speech [8]. It is available as part of Russian National Corpus [9]. The corpora containing well-annotated high-quality recordings are not publicly available. One of them is Corpus of Professionally Read Speech (CORPRES) contains over 30 h of speech recorded in a professional studio [10].

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 192–199, 2017.
DOI: 10.1007/978-3-319-66429-3_18

The corpus of monologues RuSpeech contains about 50 h of transcribed recordings produced by 220 speakers [11]. CoRuSS (Corpus of Russian Spontaneous Speech) is designed as a publicly available resource containing high-quality recordings of spontaneous speech with detailed prosodic transcription [12]. The recordings include dialogues between native Russian speakers, with a part of it - at least 14 h of speech from 60 speakers - annotated by expert linguists at lexical and prosodic levels.

One of the main reasons that provide the usability of large speech corpora is the availability and accuracy of annotations. For example, the TIMIT corpus is very popular for the phonetic and speech technology studies because of the very accurate phonetic transcriptions. The broad phonetic transcriptions are often used and sometimes even required for different tasks such as lexical pronunciation variation modelling for automatic speech recognition, unit selection for speech synthesis [10,11,13], automatic pronunciation training and assessment in Computer Assisted Language Learning [14] and general research on pronunciation variation [15]. Contemporary speech corpora are usually provided with a broad phonetic transcription of at least part of their material. In addition, time and money permitting, contemporary speech corpora are at least partially enriched with broad phonetic transcription with the help of expert phoneticians in order to ensure a more accurate representation of the material. The employment of experts is known to be exceedingly time-consuming and expensive when they have to transcribe speech from scratch. That is why, it is common practice to provide people with an example transcription they have to verify on the basis of their own perception of the speech signal [1].

Among the numerous approaches to providing text-to-speech transcription, the simplest is to use a small set of letter-to-sound rules to guess the pronunciation of any word. Each rule specifies a phonetic correspondence of sounds and letters. In some cases the letter's context is used to determine which rule should be applied. However, any language has great variation in the pronunciation. The transcription made for the TTS systems usually have one ideal variant for the text. It could be predicted and changed according to the acoustic and phonetic quality of the sounds, speaker characteristics and so on. In the speech recognition tasks it is more important to have the correct information not only about the phonemes but also about the exact acoustic characteristics and their variation. Those characteristics that can be predicted by the context beforehand. The grapheme-to-phoneme transcriber can use a dictionary-lookup approach but it tells nothing about the sound changes between the words and phrase boundaries. Therefore the rules of transcribing should use all the knowledge about the context variations of the sounds in the standard pronunciation, the phonetic changes and their frequencies of occurency in speech.

In this paper we present a reliable method for automatic phonetic transcription of Russian text into phonetic symbols. The system was used for modelling phonetic transcription for the Speech Corpus of spontaneous speech CoRuSS for Russian Language [12].

This paper is organised as follows. In Sect. 2, we introduce the automatic transcriber design and main principles. Section 3 sketches the problems of rules extensions. Section 4 presents the inclusion of the speech variability rules. In Sect. 5 we formulate our conclusions.

2 Design of the Automatic Phonetic Transcriber

The program was developed in java jdk 1.8. Each rule specifies a phonetic correspondence of phonetic symbols to letters. The letter's context is used to determine which rule should be applied. We implemented these processes as context-dependent rule modelling both within-word and cross-word contexts in which phones could be deleted, inserted or substituted with other phones.

The set of phonological and phonetic rules that differs according to conditions has been based on the phonetic knowledge obtained in experimental study of the great amount of the Russian speech corpora since the beginning of the previous century. There are 6 vowel phonemes and 36 consonant phonemes in the Russian literary speech [16–18]. The transcriber has been developed following the principles proposed by S. Stepanova [19] and K. Shalonova [20,21]. Besides, the coarticulation and sound change processes for Russian standard language (as for any other language) constantly modify. In order to include all the variation we decided to work not with separate letter-to-phoneme assosiations but use the characteristics of sound classes and the processes of assimilation, dissimilation, insertion or deletion of sounds. It gives us opportunity to model different allophone variations that are not usually provided by other phonetic transcription systems. Besides, all the exclusion are taken into account.

For example, the Russian phoneme "č" has no voiced pair in the system. Among the allophones of "č" there are voiced and unvoiced variants. Therefore it is important for the transcriber to model correctly the exact variant which should be used in the transcription using the preceding and following letters.

The quality of the vowel phonemes in Russian varies according to the word stress, position in a phrase and the quality of the neighboring sounds consonants before and after the vowel. For the correct result the transcriber needs information about the place of the word stress. It could process the words with primary and secondary stress. The signs for these are "1" for primary word stress and "0" for secondary stress. The numbers should be put after the vowel in the orthographic text. Our transcriber does not include the automatic stress detection in the orthographic text.

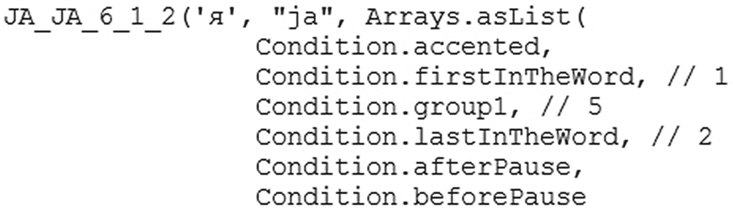
There are more than 200 rules for the vowel transformations that include all this information. Also the exclusions are taken into account for vowel transformation by inserting them into the rules (Fig. 1).

The consonant variation depends upon the quality of the neighboring sounds. There are different kinds of consonant assimilation in Russian which is usually regressive one. The consonants became similar or different in the palatalization, voiced/unvoiced characteristic, place of articulation, manner of articulation. The consonant insertions and deletion processes are also taken into account.

There more than 200 rules for consonant transformation including the consonant special sequences inside words (Fig. 2).

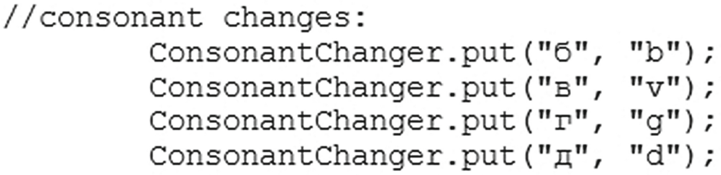
The resulting rule set comprised phonological and phonetic rules describing progressive and regressive voice assimilation, palatalisation and more specific rules modelling pronunciation variation in high-frequency words. We tried to take into account all the possible modifications and sound change that can happen within the word and on the word borders. Besides, the transcriber processes the pause signs and modifies the resulted transcription according to the place of the pause in the text and the pause type. There are several types of pauses: the end of phrase, the inhale sign, the sudden speech hesitation etc. According to the sound type the transcriber decides if the last consonant should be voiced or unvoiced for noise consonants (Fig. 3).

The processes in the word boundaries in the connected speech and the sound transformations in the end of the phrase are also included in the program. If the processed text has the phrase boundary markers and information about the pauses, speech breaks and intakes of breath it will process them automatically and decide about the phonetic quality of the sounds in the borders according to the Russian pronunciation (Figs. 4 and 5).



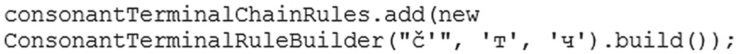
```
JA_JA_6_1_2('я', "ja", Arrays.asList(
                    Condition.accented,
                    Condition.firstInTheWord, // 1
                    Condition.group1, // 5
                    Condition.lastInTheWord, // 2
                    Condition.afterPause,
                    Condition.beforePause
```

Fig. 1. Example of the grapheme-to phoneme rules for vowels



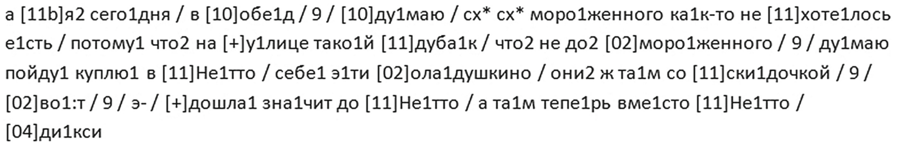
```
//consonant changes:
        ConsonantChanger.put("б", "b");
        ConsonantChanger.put("в", "v");
        ConsonantChanger.put("г", "g");
        ConsonantChanger.put("д", "d");
```

Fig. 2. Example of the grapheme-to phoneme rules for consonants



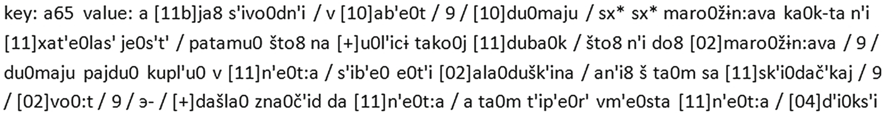
```
consonantTerminalChainRules.add(new
ConsonantTerminalRuleBuilder("č'", 'т', 'ч').build());
```

Fig. 3. Example of the grapheme-to phoneme rules for consonants sequencies



а [11b]я2 сего1дня / в [10]обе1д / 9 / [10]ду1маю / сх* сх* моро1женного ка1к-то не [11]хоте1лось
е1сть / потому1 что2 на [+]у1лице тако1й [11]дуба1к / что2 не до2 [02]моро1женного / 9 / ду1маю
пойду1 куплю1 в [11]Не1тто / себе1 э1ти [02]ола1душкино / они2 ж та1м со [11]ски1дочкой / 9 /
[02]во1:т / 9 / э- / [+]дошла1 зна1чит до [11]Не1тто / а та1м тепе1рь вме1сто [11]Не1тто /
[04]ди1кси

Fig. 4. Example of the Russian orthographic text for processing. ‘1’ is put after vowels to show the primary stress, ‘2’ is written after the vowels to show the secondary stress. The intonation markers are also included in the orthographic text. They show the intonation phrase borders and type of intonation



key: a65 value: a [11b]ja8 s'ivo0dn'i / v [10]ab'e0t / 9 / [10]du0maju / sx* sx* maro0žɨn:ava ka0k-ta n'i
[11]xat'e0las' je0s't' / patamu0 što8 na [+]u0l'icɨ tako0j [11]duba0k / što8 n'i do8 [02]maro0žɨn:ava / 9 /
du0maju pajdu0 kupl'u0 v [11]n'e0t:a / s'ib'e0 e0t'i [02]ala0dušk'ina / an'i8 š ta0m sa [11]sk'i0dač'kaj / 9
/ [02]vo0:t / 9 / э- / [+]dašla0 zna0č'id da [11]n'e0t:a / a ta0m t'ip'e0r' vm'e0sta [11]n'e0t:a / [04]d'i0ks'i

Fig. 5. Example of transcription. ‘0’ is put after vowels to show the primary stress, ‘8’ is written after the vowels to show the secondary stress

3 Rules Extensions and Refinements

At first we aimed at approximating transcription that were made with a limited rules and symbol set. Then we included the rules for pronunciation exclusions from the dictionary. The transcriber was developed to make transcriptions for the corpus *CoRuSS* [12] containing 30 h of high quality recorded spontaneous Russian speech. The recordings consist of dialogues between two speakers, monologues (speakers self-presentations) and reading of a short phonetically balanced text. Since the corpus is labeled for a wide range of linguistic-phonetic and prosodic information, it provides basis for empirical studies of various spontaneous speech phenomena. Besides, it allows comparing those phenomena with the ones we observe in prepared read speech. The corpus has orthographic and prosodic annotation for the part of the material. The orthographic decoding of the recording was made using no capital letters or punctuation marks; the only exception was a question mark to denote question phrases. Each word was written using standard spelling no matter whether it was pronounced in a proper way, mispronounced, or produced in a contracted form. Orthographic annotation also contained information about lexical stress: strong (primary) stress was marked with 1 after the vowel. Symbol 2 was used for vowels carrying secondary or weak stress, for vowels /o/, /e/ with no qualitative reduction. The Russian grapheme ‘ё’ in this corpus was never replaced by ‘e’.

The transcriber was properly tested manually. At first different texts from the CoRuSS corpus [12] were processed and checked by expert phoneticians. The manually verified phonetic transcriptions were required to tune the transcription procedures and to evaluate their performance. We took into account very special cases of Russian pronunciation that occur in the connected speech and cannot be known from the orthographic dictionary containing only word transcriptions.

In order to ensure the applicability of the transcription procedures in contexts we optimised our procedures with limited resources and minimal human effort using the statistics of the sound change in standard pronunciation from the real speech corpus CORPRES. Further additions and refinements to the rules could reduce the error rate still further.

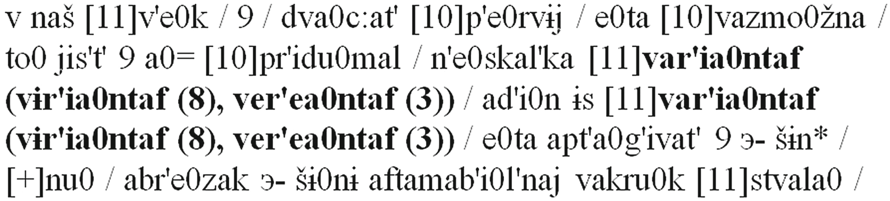
4 Modeling Speech Variation

The resulting transcription were updated using the results of the manual real speech segmentation and labelling that was made by expert phoneticians for the CORPRES speech corpus [10]. The material contains two types of transcription: manual phonetic transcription (the sounds actually pronounced by the speakers) and the level of rule-based phonetic transcription (automatically generated by another text transcriber for TTS and partially corrected by the experts). The ideal transcription in the CORPRES corpus did not contain phonetic variants within pronunciation standard.

We counted the occurrence rate of different phonetic sequences in the same contexts for ideal transcriptions in CORPRES corpus and improved the rules using several variants of transcription or the most frequent one.

For example in Russian the word /pagul'a0j/ has different variants of phonetic transcriptions that could be met in standard pronunciation (Fig. 6):

[**pəgul'a0i**] - that variant was met 0 times in corpus (the dictionary standard).
[**pogul'ai**] - that variant was met 3 times in corpus.
[**pugul'ai**] - that variant was met 5 times in corpus.



v naš [11]v'e0k / 9 / dva0c:at' [10]p'e0rvɨj / e0ta [10]vazmo0žna / to0 jis't' 9 a0= [10]pr'idu0mal / n'e0skal'ka [11]**var'ia0ntaf (vɨr'ia0ntaf (8), ver'ea0ntaf (3))** / ad'i0n ɨs [11]**var'ia0ntaf (vɨr'ia0ntaf (8), ver'ea0ntaf (3))** / e0ta apt'a0g'ivat' 9 э- šɨn* / [+]nu0 / abr'e0zak э- šɨ0nɨ aftamab'i0l'naj vakru0k [11]stvala0 /

Fig. 6. Example of transcription including the results of speech variability from the CORPRES. '0' is put after vowels to show the primary stress, '8' is written after the vowels to show the secondary stress

The example shows the variants of standard pronunciation and their frequency of occurrence in the phonetic transcription.

5 Conclusions

The results have shown that our transcriber is reliable and it could be used for the speech technology tasks that require the phonetic transcriptions of the text for speech segmentation, text-to-speech systems, and automatic speech recognition systems.

The transcriber could be adapted to the speaker as long as we know his/her speech peculiarities.

The automatic transcription can serve as an example for the human transcribers.

The ASR system and speech alignment system can be provided by a precise phonetic transcription if it has the text that has to be recognised.

Acknowledgments. The authors would like to thank the Saint Petersburg State University. This work has been carried out in the framework of SPbSU project n. 31.37.353.2015.

References

1. Van Bael, Ch., Boves, L., van den Heuvel, H., Strik, H.: Automatic transcription of large corpora. Comput. Speech Lang. **21**, 652–668 (2007)
2. TIMIT, Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065 (1990)
3. Godfrey, J., Holliman, E., McDaniel, J.: SWITCHBOARD: telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), San Francisco, USA, pp. 737–740 (1992)
4. Hess, W., Kohler, K.J., Tillman, H.-G.: The Phondat-Verbmobil speech corpus. In: Proceedings of Eurospeech, Madrid, Spain, pp. 863–866 (1995)
5. Oostdijk, N.: The design of the spoken Dutch corpus. In: Peters, P., Collins, P., Smith, A. (eds.) New Frontiers of Corpus Research, pp. 105–112. Rodopi, Amsterdam (2002)
6. Maekawa, K.: Corpus of spontaneous Japanese: its design and evaluation. In: Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR), Tokyo, Japan (2003)
7. Bogdanova-Beglarian, N., Martynenko, G., Sherstinova, T.: The "One Day of Speech" corpus: phonetic and syntactic studies of everyday spoken Russian. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS, vol. 9319, pp. 429–437. Springer, Cham (2015). doi:10.1007/978-3-319-23132-7_53
8. Kibrik, A., et al. (eds.): Rasskazy o snovidenijakh. Korpusnoe issledovanie ustnogo russkogo diskursa. Jazyki slavyanskoj kultury (2009)
9. Apresjan, J., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., Sizov, V.: A syntactically and semantically tagged corpus of Russian: state of the art and prospects 1. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, pp. 1378–1381 (2006)
10. Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., Evdokimova, V.: CORPRES: corpus of Russian professionally read speech. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 392–399. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15760-8_50

11. Krivnova, O.: Russkij rechevoj korpus ruspeech. In: Proceedings of the VII International Scientific Conference Fonetika Segodnia, pp. 54–56 (2013)
12. Kachkovskaia, T., Kocharov, D., Skrelin, P., Volskaya, N.: CoRuSS a new prosodically annotated corpus of Russian spontaneous speech. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portoro, Slovenia (2016)
13. Mizutani, T., Kagoshima, T.: Concatenative speech synthesis based on the plural unit selection and fusion method. IEICE Trans. Inf. Syst. **E88–D**(11), 2565–2572 (2005)
14. Neri, A., Cucchiarini, C., Strik, H.: Selecting segmental errors in non-native Dutch for optimal pronunciation training. Int. Rev. Appl. Linguist. **44**(4), 357–404 (2006)
15. Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G.: Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Commun. **29**, 209–224 (1999)
16. Avanesov, R.: Russian Standard Pronunciation [Russkoe literaturnoe proiznoshenie]. Prosveschenije (1984)
17. Bondarko, L.V.: Phonetics of Russian modern language, SPbSU (1998). (in Russian)
18. Kodzasov, S.V., Krivnova, O.F.: General Phonetics, Moscow (2001)
19. Stepanova, S.B.: The phonetic properties of Russian speech: realisation and transcription. Ph.D. dissertation. Leningrad (1988)
20. Shalonova, K.: Flexible transcriber for Russian continuous speech. In: 2nd International Conference on Speech and Computer, SPECOM, pp. 171–175 (1997)
21. Shalonova, K.B.: Automatic modelling of regional pronunciation variation for Russian. In: Matousek, V., Mautner, P., Ocelíková, J., Sojka, P. (eds.) TSD 1999. LNCS, vol. 1692, pp. 329–332. Springer, Heidelberg (1999). doi:10.1007/3-540-48239-3_60