Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

MULTI-PRONUNCIATION LEXICON FOR RUSSIAN AUTOMATIC SPEECH RECOGNITION (PILOT STUDY)

Shirokova A. (anna_a@stel.ru),
Telesnin B. (telesnin_ba@stel.ru),
Rogozhina V. (mind_your_own_business@rambler.ru)

Stel CS, MSLU, Moscow, Russia

Our pilot study is aimed at building a lexicon of effective pronunciation variants on the basis of canonical pronunciations, for implementing it into the automatic speech recognition system for Russian. We focus on phonetic changes in word pronunciation caused by different factors operating in spontaneous speech. Our speech data includes three different corpora of the conversational type. Manual expert processing and analysis of the audio data are used. The lexicon construction procedure is given. Some statistics for pronunciation variation in Russian, obtained from the speech data, is presented. A description of frequent types of this phenomenon is given. Parallel and sequential pronunciation variants are discussed. Ways of formulating general phonetic variation rules and predicting potential contexts, in which pronunciation variation is likely to appear, are considered. Test data, phoneset used, and automatic speech recognition (ASR) parameters are described. Preliminary results for ASR and key word spotting (KWS) are shown. The appropriateness of using multi-pronunciation lexicon is discussed.

Keywords: Russian spontaneous speech, pronunciation variants, Russian pronunciation, spontaneous speech, pronunciation lexicon, reduction, Russian ASR

1. Introduction

Variations in word pronunciation have multiple sources. First, it is a common phenomenon across languages that words share the same written form but have different pronunciations (homographs). Also there are orthoepic ambiguities when a word has

multiple pronunciations which are orthoepically acceptable. Specific pronunciations reveal and are due to individual manner or regional accent of a speaker. It is generally known that speech genres and speaking styles determine pronunciation peculiarities. In particular, relaxed or condensed pronunciation typical for rapid fluent (especially informal) spontaneous speech is characterized by various forms of contractions, reductions, elisions, deletions, etc. The above processes drastically affect articulatory and acoustic parameters of phones and cause grave changes in sound image of a word.

Pronunciation disambiguation is essentially important in speech synthesis, speech recognition and other fields of automatic natural language processing. Most state-of-the-art ASR systems use phone-based representations for acoustic modeling. As stated in (Schultz, Kirchhoff 2006), explicitly specified pronunciations allow spoken language to be modeled more accurately. A pronunciation-based approach includes the potential for reducing the ambiguity of a given language writing system. If different acoustic realizations of a word are unlikely to be covered properly by the acoustic models, a given lexical entry may be assigned multiple pronunciations to represent these significant differences. When adding variants, one has to consider the types of speech that will be processed in order to add pronunciation variants relevant for the actual genre and style.

Our work is aimed at constructing a lexicon of effective pronunciation variants on the basis of the canonical pronunciations and implementing it into the ASR system for Russian (Zulkarneev&al 2013). We take preliminary ASR and KWS experiments to roughly assess a potential profit of the explicit adding of phonetic variants for reduced tokens. Furthermore, our study is intended to assess the very appropriateness of taking into account multiple pronunciations in our ASR projects. It is essential to analyze whether there exist trends towards ASR performance gain achieved by using such an enhanced lexicon.

2. Related work

Our project has been inspired by a series of researches that deals with pronunciation variation phenomena and its influence on automatic speech recognition. The work (Adda-Decker, Lamel 1998) is aimed at evaluating the use of pronunciation variants across different system configurations, languages (English and French) and speaking styles (spontaneous and read speech). A correlation between the word frequency and the number of productive variants is outlined.

In (Adda-Decker&al 1999) authors focus on well-known pronunciation variants in French: the so-called mute *e* and liaisons. Their frequencies of occurrence in read speech and spontaneous speech are computed and compared regarding these types of speech.

Formal phonetic rules are recently developed for Austrian German conversational speech (Schuppler&al 2014).

For the Russian language the issue of pronunciation variety has been studied in theoretical and applied aspects.

The monograph (Bondarko&al 1988) describes the phonetic system of spontaneous speech. It is based mainly on evidences of the Russian oral speech, but also

considers some English and German data. It covers the problems of the pronunciation norm and acceptable variation, the allowable phonetic realization of phonemic units and the pronunciation types. As a supplement it includes a phonetic lexicon of 80 Russian high frequency words which gives a number of different phonetic representations for each word.

In (Lobanov, Tsirul'nik 2007) it is claimed that it is possible to predict potential contexts in which pronunciation variation is likely to appear in conversational speech. Moreover, there are deduced systematic phonetic changes in word pronunciation caused by the above factors which are generalized and formulated as strict phonetic rules.

An algorithm for automatic generation of pronunciation variants for Russian based on the results of the above-mentioned research (Lobanov, Tsirul'nik 2007) is proposed in (Kipyatkova, Karpov 2009) and is reported to be implemented into Russian ASR system (Kipyatkova&al 2013).

The pronunciation variety and peculiarities of reduced word forms in the ORD speech corpus of Russian everyday communication are analysed in (Bogdanova, Palshina 2010).

3. Pronunciation variants

3.1. Speech Data

Our speech data involve three separate speech corpora. The first corpus of retrieval queries contains short utterances of more than 4000 speakers of different gender and age. Speech material includes mostly exact address requests, geographic objects requests and proper names requests. Another corpus of professional telephone speech contains recordings of power engineers professional conversations. There are 30 adult male speakers. It is characterized by high portion of professional lexis, proper names and toponyms. The third corpus contains telephone speech recordings of the general conversational type. In all the corpora there is a portion of speakers with more or less distinct regional accent features. Table 1 summarizes the corpora characteristics. The given conversational data represents rapid fluent spontaneous speech. All the corpora are not publically available and are the property of our customers.

Table 1. Corpora characteristics

Corpus	Queries	Professional	Conversational
Number of speakers (multiple records per speaker available)	4,000+	30	1,000+
Duration average (per record)	10 sec	3 min	5 min
Duration total	30 h+	10 h+	50 h+
Gender	all	male	all
Age	all	adult	all

3.2. Building Canonical Pronunciations

Although there are many grapheme-to-phoneme conversion techniques (Bisani, Ney 2008), in our work we use a rule-based automatic transcription system to build the canonical pronunciations of words in Russian. This multifunctional transcribing tool (Krivnova&al 2001) has different representation levels: phonemes, phones, etc. Context-dependent rules cover major and slight conversion patterns. The system uses a number of exception lists. The phoneme error rate is 2% (mostly in proper names and loanwords) when testing on 500 Kb of texts.

3.3. Building Pronunciation Variants

The annotations of the speech data were created by linguists according to the adopted guidelines (Glavatskih&al 2015). A multi-pronunciation lexicon has been created on the basis of the orthographic transcripts which includes both the canonical phonetic representations of words and their pronunciation variants. The latter were created manually by expert phoneticians relying on the results of the perceptive and acoustic analyses. The process of building a lexicon had the following steps:

- when creating orthographic transcripts of the speech data expert phoneticians were asked to mark words with gravely reduced pronunciations (contracted forms, phone deletion) and incorrect or non-standard pronunciations (stress position, etc);
- marked words were ranked according to their frequency of occurrence in each data set; lists of the most frequent words were taken into account;
- up to four most common variants of actual pronunciations were added to each lexicon (after speech fragments corresponding to the marked words had been listened to by experts).

Thus, three separate lexicons have been formed. It should be noted that since those were not simultaneous projects, the data of the previously built lexicon was taken into account when creating a new one. It is obvious that they must share some lexical entries, but due to the corpora specifics their set of variants can still be partly different. Conversational lexicon has been chosen for the further processing, since the others contain a high portion of specific lexical data such as proper names, toponyms, etc.

3.4. Pronunciation Lexicon for Conversational Corpus

The following Table 2 and Table 3 highlight the main tendency in word reduction and its usage within 50 hours of spontaneous speech. The ratio of total amount of reduced words to total amount of words used in the database equals 5.53 % (see Table 2). It should be noted, however, that only word tokens with evident reduction (phone deletion, syllable contraction) are treated as reduced variants, other segmental changes are supposed to be covered by the acoustic models. Although, as an observation, such small reduction ratio in total corpora implicitly testifies, as we see it,

that the impact of adding reduced tokens for the total lexicon is not significant and, therefore, can be disregarded.

Table 2. Conversational corpus statistics

Duration of spontaneous speech	50 hours
Total amount of pronounced words	228,209
Total amount of reduced words	12,611
Ratio of total amount of reduced words to total amount of pronounced words, %	5.53

The list of the most frequent words affected by reduction is given in Table 3.

Table 3. List of the most frequent reduced words in conversational corpus

Reduced words	Frequency of reduced realizations	Overall frequency of the word	Ratio of reduced realizations to overall frequency of the word, %
что (what)	3,386	5,780	58.58
сейчас (now)	1,822	2,005	90.87
тогда (then)	561	987	56.84
сегодня (today)	511	779	65.60
говорить (to say) (all the verb forms are taken into account)	943	2,112	44.65
ничего (nothing)	427	632	67.56
чтоб (in order to)	480	777	61.78
только (just)	365	550	66.36
тебе (for you)	306	1,052	29.09
алло (hello)	322	845	38.11
сколько (how many)	232	357	64.99
когда (when)	231	491	47.05
тебя (you)	203	646	31.42
наверное (perhaps)	212	264	80.30
здравствуйте (greetings)	103	230	44.78

3.5. Phonetic evidence

As it has been suggested in (Adda-Decker, Lamel 1998) we use parallel (equipollent) and sequential (derived) pronunciation variants. Parallel are predominantly

used to cover homographs and orthoepically acceptable variants, while sequential represent different stages of reduction.

The examples given in Table 5 below and rules formulated in (Lobanov, Tsirul'nik 2007), (Kipyatkova, Karpov 2009) show that reduction can be viewed as a categorical phenomenon which can lead to the change of one phonological feature into another or to the total deletion of different segments.

According to (Hoole&al 2012) a lot of Russian words in spontaneous speech tend to syncope, (i.e. the strategy to make trisyllabic word bisyllabic), for example as in word 'сегодня' ('today') shown in Figure 1 in its full pronunciation. Figure 2 displays almost the deletion of the first syllable vowel [i], whereas there are some [i]-traces in the formant curve. There is no surprise in deletion of the phone [v] due to its intervocal position, while the optional presence of the phone [d^j] is due to the potential total regressive assimilation to [n^j], which has the same place of articulation. Tokens can be contracted even more, and this illustrates the gradient character of reduction process operating to some sequence of segments.

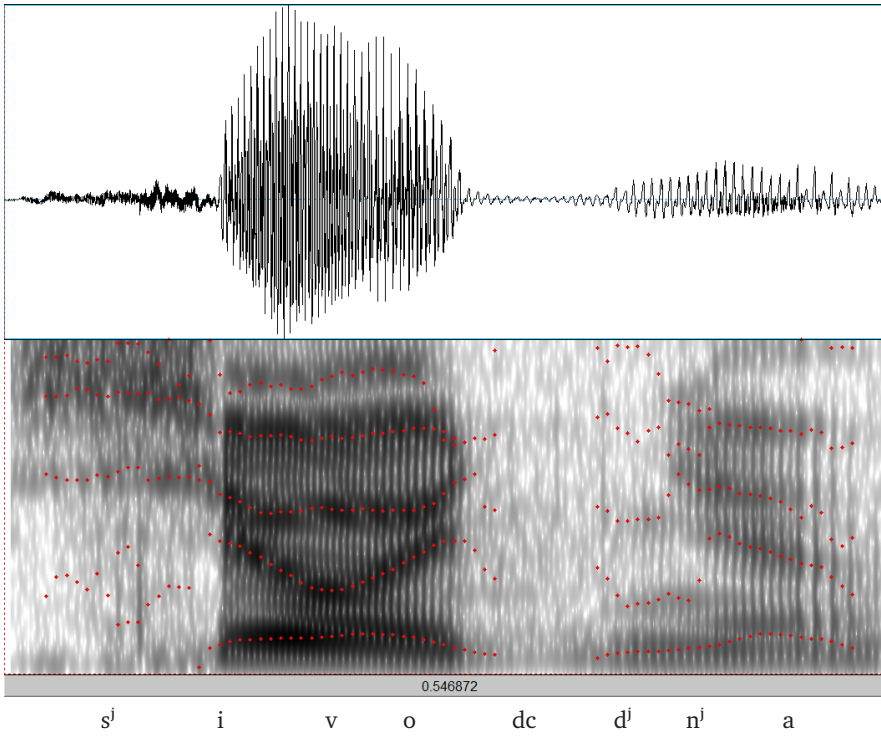

Fig. 1. Oscillogram and spectrogram for a pronunciation of the word 'сегодня' [s^j i v o d^j n^j a]. Figures are captured within our software annotation tool described in (Glavatskih&al 2015)

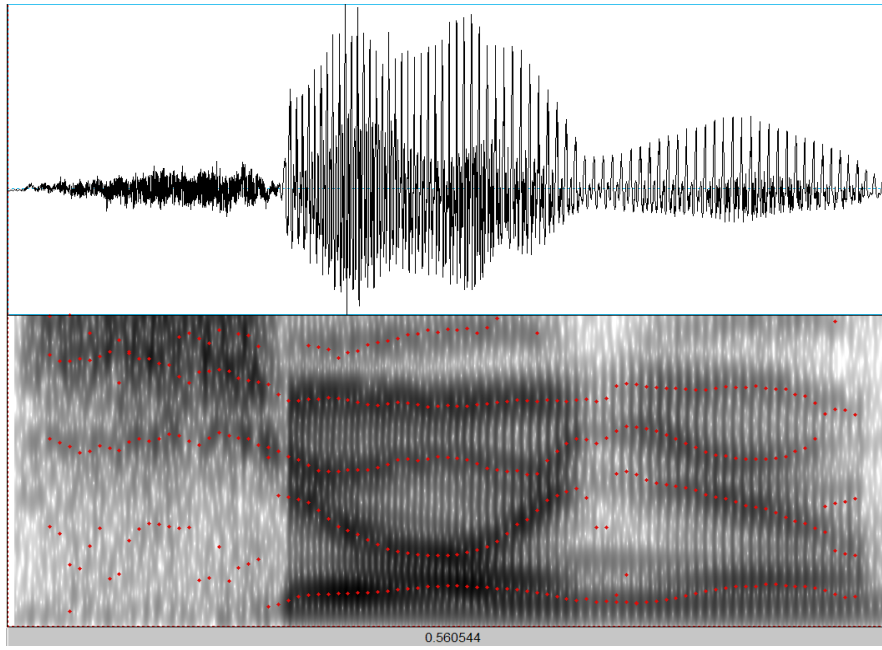

Fig. 2. Oscillogram and spectrogram for a pronunciation of the word 'сегодня' [sʲ o nʲ: ə]

3.6. General variation types

Our general observations verify that vowels are more robust to duration and quality reduction as well as to deletion than consonants. Experts have marked phonetic changes that are reported in (Bondarko&al 1988) to be systematic and typical for spontaneous speech. Those referred to consonants include:

- deletion of /j/ in word initial, word final, intervocal positions and in V/j/C contexts;
- deletion of /v, vʲ, bʲ, dʲ/ in intervocal position;
- deletion of one of double consonants;
- deletion of word-final plosives;
- consonant cluster reduction including phonetic changes across word boundaries (strong assimilation or total deletion of phones and phone sequences);

For vowels the following observations are made:

- stronger duration reduction even in stressed syllables;
- quality reduction in unlike position to the stressed syllable;
- delabialization of /u, o/ in weak positions;
- quality reduction of /u, y/ in weak positions;
- centralization in weak position;
- vowel deletion in unstressed syllables.

It is verified by all our data that numerals, common words and notional word fillers are the first to suffer compression in fluent spontaneous speech.

4. ASR & KWS Experiments

4.1. Baseline & Training Data

The experiments were performed using a speech recognition system based on Kaldi (Povey&al 2011). Recognition was carried out in two stages, and it can be described as hybrid HMM-DNN approach.

The training data (68.2 hours) is based on multiple sources: the major part of the data listed above, our Broadcast Russian speech data (Glavatskih&al 2015) and Russian Voxforge open speech corpus.

At first stage recognition system based on HMM was used to build the adaptation of MLLR matrix, then the MLLR transformation was applied to feature vectors. At the second stage recognition was performed using DNN, consisting of 4 hidden layers, each hidden layer composed of 1,024 elements.

As the language model, 3-gram model was used in speech recognition, trained on the text data of 772,365 words within 114,423 phrases, its pronunciation lexicon equals 44,446 tokens. Phrases, that are heavily distorted by artifacts due to the channel and/or other technical issues, or do not contain any intelligible word, are filtered and not used for the further analysis. Thus only 98,331 utterances (640,242 pronounced words and 38,737 word tokens), which equals approximately 54 hours of spontaneous speech, are used for building the acoustic models. Acoustic model adaptation set is composed of 38 hours (34,861 utterances, 322,861 pronounced words and 24,582 word tokens) of spontaneous speech. Speaker-independent model is applied.

4.2. Test Data

In our system the phonetic alphabet Worldbet is used, though our set of symbols for the Russian language slightly differs from the one suggested in (Hieronymus 1993).Thus, the property “dental” is not marked in plosives, nasals and the affricate. Furthermore, the symbol “I” is added to the vowel inventory. It represents the reduced high front vowel (the second stage of reduction) which is not considered in (Hieronymus 1993). The symbol “ax” represents the reduced mid central vowel (the second stage of reduction) and corresponds to “&” in the Worldbet list. It should be noted, that in Russian phonetic transcription palatalization is only marked when there is a corresponding nonpalatalized consonant. In unpaired consonants this property is not marked, since it is supposed to be implied in the symbol itself. “Ix” represents the reduced diphthongoid that has a higher and narrower beginning and a wider and lower ending. It is regarded as a prototypic realization of post-stressed

combinations of /j/ or /i/ and a wider and lower vowel. The vowel reduction in the terminal open post-stressed syllables is supposed to be only slight and for that reason is not taken into account. In continuous speech, however, terminal post-stressed vowels are reduced according to general rules. The above phoneset is mapped to IPA in Table 4.

Table 4. Phoneset mapping to IPA

IPA	Our Phoneset
ɪə	Ix
ʂ	S
ɕ	Sj
ʐ	Z
a	a
ə	ax
b	b
bʲ	bj
d	d
voiced closure	dc
dʲ	dj
e	e
f	f
fʲ	fj
g	g
gʲ	gj
i	i
ɨ	ix
j	j
k	k
kʲ	kj
ɬ	l
lʲ	lj

IPA	Our Phoneset
m	m
mʲ	mj
n	n
nʲ	nj
o	o
p	p
pʲ	pj
r	r
rʲ	rj
s	s
sʲ	sj
t	t
t͡s	ts
t͡ɕ	tSj
unvoiced closure	tc
tʲ	tj
u	u
v	v
vʲ	vj
x	x
xʲ	xj
z	z
zʲ	zj

We take preliminary speech recognition and key word spotting experiments to analyze the potential of the performance improvement when taking into account pronunciation variation of reduced words. Speech data for the test (not included in the training set) is 2 hours of spontaneous speech, i.e. 1,591 utterances with 15,836 words and 3,324 word tokens. In the course of experiment 33 most frequently used words were selected and their pronunciation variants were processed. Table 5 covers the majority of pronunciation variants built for several words included in the test set. The upper transcriptions for each word given in Table 5 represent these words as pronounced solely.

Table 5. Pronunciation variants for frequently used words

Frequent words	Pronunciation variants
сейчас (now)	sj i tc tSj a s Sj a s Sj a
что (what)	S tc t o S tc t ax S o tc tSj o
говорит (says)	dc g ax v a rj i tc t dc g ax a rj i tc t dc g a rj i tc t dc g rj i tc t dc g I tc t
тебе (for you)	tc tj i dc bj e tc tj i e tc tj e
когда (when)	tc k a dc g dc d a tc k a dc d a
сегодня (today)	sj i v o dc dj nj a sj i o dc dj nj ax sj o dc dj nj a sj o nj ax
будет (will)	dc b u dc dj I tc t dc b u I tc t dc b u I dc b u tc t
сказала (said)	s tc k a z a l a s tc k a a l ax s tc k a l ax
двадцать (twenty)	dc d v a tc tc ts ax tctj dc d v a tc tKs&
позвонишь (you'll call)	tc p ax z v a nj i S tc p a z v o nj I S

4.3. Preliminary Results & Discussion

Three tests based on the same acoustic and language models are carried out. The tests differ in pronunciation variants that have been used. For test_result_1tr only one canonical pronunciation variant is applied, so it represents baseline. For test_result_2tr only one additional variant is included, whereas in test_result_vartr four pronunciation variants are added (when given, otherwise less). Preliminary results are obtained (see Table 6), where FA and FR denote the false acceptance and the false rejection rates.

As a result of adding multi-variant pronunciation, the system manages to detect the word tokens that differ from their standard pronunciation. For that reason the

correctness rate (CORR) increases. On the other hand, when all the pronunciation variants are taken into account, as test_result_vartr shows, it leads to the increase of word error rate (WER) which is due to a higher number of insertions.

Table 6. System performance for ASR and KWS

tests	CORR%	WER%	FA	FR%
Baseline	64.01	40.44	2.005	26.98
Test_result_2tr	64.16	40.31	2.33	25.57
Test_result_vartr	64.31	40.62	3.52	22.67

It is supposed that strongly reduced variants tend to appear when an acoustic observation is unlikely to be recognized adequately. To reduce the number of insertions and, as a consequence, WER, the shortest transcriptions should be eliminated from the pronunciation lexicon and another set of experiments needs to be taken. For such purpose a special technique should be applied, which would enable to track the actual system choice of a pronunciation variant in the recognition process.

5. Conclusion

Our research verifies, expands and specifies the experimental results shown at (Bondarko&al 1988). It has been observed that the most likely words to be affected by reduction and adjacent mechanisms are numerals, common words and notional word fillers, which seems reasonable in the context that frequent words carrying little information are drastically affected by articulation relaxation. Obviously, the most robust segments of the words are stressed vowels, while in weak especially post-tonic syllables phone and phone sequences deletion is likely to appear. At the current stage of work there can be outlined the major factors that account for the evidences of pronunciation variation in the speech corpora, however, the potential contexts and the actual conditions can hardly be described in terms of patterns and summarized as a set of phonetic rules.

As it has been noted the pronunciation variants in our lexicon were created manually. Moreover, experts had to listen to audio samples in order to specify exact pronunciation and then to select the most common ones. Unfortunately, manual processing did not allow us to assign a unique acoustic form to its specific phonetic representation. This partly explains the lack of statistical data and estimations for the phone deletion and other segmental changes.

It is worth mentioning that experts are limited by the phone set of the speech recognition system so they have to approximate their actual observation and refer it to some phonetic unit in the set, therefore missing some slight phonetic differences.

Our further research involves learning potential contexts of the phonetic changes localization to be able to predict their occurrence. Alongside with it we plan to investigate the character of phonetic changes and to systematize their types regarding contexts of appearance.

As stated in (Adda-Decker, Lamel 1998), one should be aware that different words are likely to share the same pronunciation variants, when applying such multi-lexicon to the ASR system. It inevitably leads to a higher rate of word confusion.

In the further research the following strategies can be implemented to effectively introduce multi-pronunciation lexicon to ASR:

- to train acoustic models within added pronunciation variants to provide model compatibility;
- to take into account less variants and to select the most effective ones;
- to assign weight to variants.

Still, the question is open about the optimal strategy of such lexicon building, whether it is worth deducing general phonetic modification rules and applying them totally to all the words in lexicon (to a part of it) as suggested in (Kipyatkova, Karpov 2009), or to manually provide selected words with required pronunciations. The latter is not always the slowest strategy (if a list is not large), but enables including specific actual pronunciations and controlling their confusion potential. For the same reason the number of pronunciation variants should be limited. In any case the appropriateness of implementing an enhanced lexicon should be studied more thoroughly and the supposed performance profit should be estimated more accurately.

References

1. *Adda-Decker M., Lamel L.* (1998), Pronunciation variants across systems, languages and speaking style in Proc. ESCA Conf., May 1998, pp. 1–6.
2. *Adda-Decker M., Boula de Mareüil P., Lamel L.* (1999), "Pronunciation variants in French: schwa & liaison", in Proc. ICPhS Conf., 1999, pp. 2239–2242.
3. *Bisani M., Ney H.* (2008), Joint-sequence models for grapheme-to-phoneme conversion, Speech Communication, vol.50, May 2008, pp. 434–451.
4. *Bogdanova N. V., Palshina D. A.* (2010), The reduced forms in Russian (a lexicographical description) in Proc. Word. Lexicon. Philology: Lexicon Text and Lexicographical Content Conf., Nov. 2010, pp. 491–497.
5. *Bondarko L. V., Verbitskaja L. A., and N. I. Geilman* (1988) Spontaneous Speech Phonetics. [Fonetika spontannoy rechi] St. Petersburg: 1988.
6. *Glavatskih I. A., Platonova T. S., Rogozhina V. S., Shirokova A. M., Smolina A. A., Kotov M. A., Ovsyannikova A. S., Repalov S. A., Zulkarneev M. Ju.* (2015), The multi-level approach to speech corpora annotation for automatic speech recognition, in Proc. SPECOM Conf., Sept. 2015, pp. 438–445.
7. *Hieronymus J. L.* (1993) "ASCII phonetic symbols for the world's languages: Worldbet, Journal of International Phonetic association, Vol. 23.
8. *Hoole P., Bombien L., Pouplier M., Mooshammer C., Kuhnert B.* (2012), Consonant Clusters and Structural Complexity. Munich: Mouton de Gruyter.
9. *Kipyatkova I., Karpov A.* (2009), Creation of multiple word transcriptions for conversational Russian speech recognition, in Proc. SPECOM Conf., Sept. 2009, pp. 71–75.

10. *Kipyatkova I., Karpov A., Verkhodanova V., Zelezny M.* (2013), Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition, International Journal of Computer Science and Applications, vol. 10, №1, 2013, pp. 11–30.
11. *Krivnova O. F., Zakharov L. M., Strokin G. S.* (2001), Automatic transcriber of Russian texts: problems, structure and application in Proc. SPECOM Conf., Sept. 2001, pp. 408–409.
12. *Lobanov B. M., Tsirul'nik L.I.* (2007), Modelling of in-word and word-boundary phonetic-acoustic phenomena in full and conversational speaking styles for TTS synthesizer MULTIFON [Modelirovanie vnutrislovnykh i mezhslovnykh fonetiko-akusticheskikh yavleniy polnogo i razgovornogo stiley rechi v sisteme sinteza rechi po tekstu MULTIFON] in Proc. The first interdisciplinary seminar "Russian conversational speech analysis" RCSA [Trudy pervogo mezhdistsiplinarnogo seminara "Analiz russkoy razgovornoy rechi"], St-Peterburg, — Spb.: GUAP, Aug. 2007, c. 57–71.
13. *Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K.* (2011), The Kaldi speech recognition toolkit in Proc. IEEE Conf., 2011, iEEE Catalog No.: CFP11SRW-USB.
14. *Schultz T., Kirchhoff K.* (2006), Multilingual Speech Processing. Elsevier.
15. *Zulkarneev M., Grigoryan R., Shamraev N.* (2013), Acoustic modeling with deep belief network for Russian speech recognition, in Proc. SPECOM Conf., Sept. 2013, pp. 17–24.
16. *Schuppler B., Adda-Decker M., Morales-Cordovilla J. A.* (2014), Pronunciation variants in read and conversational Austrian German in Proc. INTERSPEECH Conf., Sept. 2014, pp. 1453–1457.