Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 719–725

ISBN 978-83-60810-51-4

Analysis of Long-distance Word Dependencies and Pronunciation Variability at Conversational Russian Speech Recognition

Irina Kipyatkova, Alexey Karpov, Vasilisa Verkhodanova
St. Petersburg Institute
for Informatics and Automation of the Russian
Academy of Sciences (SPIIRAS),
St.Petersburg, Russia
Email: {kipyatkova, karpov}@iias.spb.su, interiora@gmail.com

Miloš Železný
The University of West Bohemia (UWB),
Pilsen, Czech Republic
Email: zelezny@kky.zcu.cz

***Abstract*—The key issues of conversational Russian speech processing at phonemic and language model levels are considered in the work. The multiple transcriptions for modeling word pronunciation variety and joint application of statistic and syntactic analysis of training text data for modeling long-distance grammatical relations between words in the phrase are proposed. The word error rate of the developed speech recognition system was 33% for the collected conversational speech corpus.**

I. Introduction

THE majority of state-of-the-art automatic speech recognition systems can efficiently analyze isolated pronounced words or read phrases. Recognition of conversational speech is difficult owing to its variability: different speakers may pronounce the same word differently, besides the pronunciation of the same speaker can vary depending on the context and the speech rate. Therefore it is necessary to take into account variability of word pronunciation when developing speech recognition system.

Any speech recognition system uses a phonemic vocabulary of the words. Generally such vocabulary is created with the use of the phonetic transcription rules. In spontaneous speech some phonemes can be assimilated and reduced up to complete disappearance [1], [2]. Therefore the transcriptions of the pronounced words often mismatch with the transcriptions made by the phonetic rules. The problem of appearance of reduction and assimilation phenomena could be solved by addition of alternative transcriptions to canonical transcriptions into the vocabulary of the recognition system. The accuracy of variation modeling of spontaneous speech depends on the way of alternative transcription creation.

There are two main approaches to the problem of pronunciation variability modeling [3]: knowledge-based and data-driven methods. In the knowledge-based methods variability of pronunciation is defined by the analysis of existing phonetic and linguistic knowledge formulated by experimental phonetics at analysis of speech data, acoustic and articulation characteristics of phonemes. In data-driven methods alternative transcriptions are found when analyzing a spontaneous speech corpus. Derived real transcriptions of words can describe only variants that are occurred in the given database therefore the fullness of alternative transcriptions directly depends on the speech corpus size. Unlike the knowledge-based methods in the data-driven methods it is possible to compute the probability of every alternative transcription appearance using the training speech corpus.

Direct and indirect approaches to creation of alternative variants of word pronunciation are applied for both methods. In knowledge-based methods direct modeling is made manually by an expert. In indirect modeling some reduction and assimilation rules are applied. In this case alternative transcriptions are made by applying these rules to the list of basic transcriptions. In the data-driven methods when applying the direct modeling only pronunciation variants that frequently occur in the training corpus are chosen as alternative transcriptions. When applying the indirect modeling the most typical changes in the pronunciation of the same phoneme sequences in different words are revealed, i.e. the rules of the most typical changes on the phoneme level are defined by the speech corpus.

Thus the mentioned above approaches to pronunciation variety modeling have its own advantages and disadvantages connected with manual data processing and creation of huge list of alternative transcriptions created automatically. So, a trade-off is required during development of speech recognition vocabulary.

The next stage after word recognition is generation of grammatically correct and sensible hypothesis of the pronounced phrase by a language model. Methods of language model creation, which increase accuracy of speech recognition, have been already developed for many natural languages. However, these methods cannot be directly applied to the

This research is supported by the Ministry of Education and Science of Russia (contract No.11.519.11.4020), the Russian Foundation for Basic Research (project No. 12-08-01265) and by the grant of the President of Russia (project No. MK 1880.2012.8).

Russian language owing to free order of words in sentences and existence of a large amount of word-forms for every lexical unit because of inflective nature of the language.

One of the most efficient natural language models is a statistical model based on word *n*-grams aimed to estimate a probability of word sequence *W=(*w_1,w_2,...,w_m*)* in some text. *n*-gram is a sequence of *n* elements (for example, words), and the *n*-gram language model is used for prediction of an element in a sequence containing *n*-1 predecessors [4]. This model is based on an assumption that a probability of any *n*-gram, which presents in an input text, can be estimated by information about its frequency of appearance in some large training text.

There are several types of *n*-gram models, which are described in the surveys [4], [5]. Class-based models use a function that maps every word w_i into a class c_i*: f:* $w_i \rightarrow$ *f(*w_i*)=*c_i. If any class contains more than one word, then this mapping results in less distinct classes than there are words.

Distance models describe a longer context than the *n*-gram model. In these models, a distance bigram is defined as a bigram, which predicts a word w_i based on the preceding word w_{i-d}, where *d* is the distance between the considered words.

Another type of models that determinates a correlation between word pairs in a longer context is trigger models. The appearance of a trigger word in a history increases the probability of another word referred to as a target word.

The simplified version of trigger pairs is a cache model. The cache model increases the probability of word appearance in accordance with frequency of appearance of this word in a history. If a speaker used a certain word, then he/she tends to use this word once more because this word is specific for the particular topic or because the speaker tends to use this word.

Particle-based models are used for inflected languages. In this case, a word is divided into some number of parts, and language model is created for these word parts.

There are models that do not restrict sequences of words to a certain *n* and store sequences of different lengths. These models are varigrams [4]. Varigrams can be considered as *n*-gram models with a large *n* and methods of *n*-gram pruning that store only a small subset of all long sequences.

In the paper [6], the class of compound language models has been proposed. For every word in a vocabulary, 15 attributes that determine grammatical features of a word-form are assigned. Every word in a sentence is considered as its initial form and a morphological class. As the result, the grammar is divided into 2 parts: a variable part based on the morphology and a constant part based on initial forms of words constructed in the form of the *n*-gram language model.

Free order of words in sentences permitted by Russian constrains implementation of the referred language models. Therefore some approaches to modeling long-distance dependencies between words are required.

In this work the software complex for conversational Russian speech processing is presented. The complex allows generating multiple transcription variants that take into account variability of pronunciation in conversational speech, and creating a stochastic Russian language model that is distinctive by joint application of statistic and syntactic analysis of training text data and uses long-distance grammatical relations between words in the phrase. In the two next sections the processes of creation of the vocabulary with words and multiple transcriptions, and creation of *n*-gram language model is considered. A software complex for Russian speech recognition is described in Section IV. Section V presents experimental results.

II. PHONEMIC VOCABULARY CREATION

One of the important challenges in development of spoken Russian ASR systems is grapheme-to-phoneme conversion or orthographic-to-phonemic transcription of a recognition lexicon. There are several issues: grapheme-to-phoneme mapping is not one-to-one, stress position(s) in word-forms is floating, substitution of grapheme Ё (always stressed) with E in the most of printed and electronic text data, phoneme reductions and assimilations in continuous and spontaneous speech, many homographs, etc.

According to the SAMPA phonetic alphabet, there are 42 phonemes in the Russian language (for 33 Cyrillic letters): 6 vowels and 36 consonants including plain and palatalized versions of some consonants. Russian consonants are: voiced-unvoiced pairs /p/ (Cyrillic grapheme П) and /b/ (Б), /t/ (Т) and /d/ (Д), /k/ (К) and /g/ (Г), /f/ (Ф) and /v/ (В), /s/ (С) and /z/ (З) (they have palatalized versions as well), /S/ (Ш) and /Z/ (Ж); sonorants /l/ (Л), /r/ (Р), /m/ (М), /n/ (Н) (these consonants are not paired, but have palatalized versions) and /j/ (Й), plus velar /x/ (and a soft version /x'/, grapheme Х), /ts/ (Ц), /tS'/ (Ч), /S':/ (Щ). However, according to the International Phonetic Alphabet (IPA), there are 17 vowels in Russian with different levels of reduction between stressed and unstressed vowels up to complete disappearance. Recent experiments showed [7], that distinction between models for stressed and unstressed vowels allows decreasing WER at ASR. Thus, six stressed (/a!/, /e!/, /o!/, /u!/, /i!/ and /1!/ in SAMPA format) and four unstressed vowels are used (/o!/ and /e!/ may have only stressed versions in the standard Russian with a few exceptions).

At grapheme-to-phoneme conversion the following positional changes of sounds are made: (1) changes of vowels in pre-stressed syllables, which are presented in Table I; (2) changes of vowels in post-stressed syllables, which are shown in Table II; (3) positional changes of consonants can happen in the following cases [8]:

- At the end of a word or before an unvoiced fricative consonant, voiced fricatives are devoiced.
- Before voiced fricatives (excluding /v/ and /v'/) unvoiced fricatives become voiced.
- Before the palatalized dentals /t'/ and /d'/ the phonemes /s/, /z/ become palatalized, as well as before /s'/ and /z'/, the consonants /s/, /z/ are disappeared (merged into one phoneme).
- Before the palatalized dentals /t'/, /d'/, /s'/ /z'/ or /tS'/, /S':/ the hard consonant /n/ becomes palatalized.

– Before / tS'/ the consonant /t/ (both for the graphemes Т and Д) is disappeared.
– Before /S/ or /Z/ the dental consonants /s/, /z/ are disappeared (merged).
– Two identical consonants following each other are merged into one.
– Some frequent combinations of consonants are changed: /l n ts/→ /n ts/, /s t n/→ /s n/, /z d n/→ /z n/, /v s t v/→ /s t v/, /f s t v/→ /s t v/, /n t g/→ /n g/, /n d g/→ /n g/, /d s t/→ /ts t/, /t s/→ /ts/, /h g/→ /g/, /s S':/→ /S':/, etc.

The developed algorithm for automatic grapheme-to-phoneme conversion of word-forms operates in two cycles, consisting of the following steps:

1) Stress positions are identified using a morphological database.

2) Hard consonants before graphemes И, Е, Ё, Ю, Я become palatalized (if possible) and these graphemes are converted into phonemes /i/, /e/, /jo!/, /ju/, /ja/ in the case if they are located in the beginning of a word or after any vowel, otherwise they are transformed into /i/, /e/, /o!/, /u/, /a/, correspondingly.

3) A consonant before grapheme Ь gets palatalization and the grapheme is deleted (it has no corresponding phoneme).

4) Transcription rules for positional changes of consonants (presented above) are applied.

5) Transcription rules for positional changes of vowels in pre-stressed and post-stressed syllables (presented above) are applied.

6) Steps (4)-(6) are repeated one time again, some changes may result in some other changes in preceding phonemes.

7) Grapheme Ъ is deleted (it has no corresponding phoneme), this letter just shows that the preceding consonant is hard.

For the grapheme-to-phoneme conversion, we employ an extended morphological database of more than 2.3M word-forms with a mark ("!") for stressed vowels/syllables. This database is a fusion of two different morphological databases: AOT (www.aot.ru) and Starling (starling.rinet.ru/morpho.php). The former one is larger and has above 2M items, but the latter one contains information about the secondary stress for many compound words as well as words with grapheme Ё, which is always stressed at pronunciation, but it is usually replaced to Е in official texts that results in loosing information on stress. Additionally, some alternative phonemic transcriptions can be generated for word-forms in order to model the effects of phonemes' reduction and assimilation in spontaneous speech using a set of cross- and within-word phonetic rules [9]. These rules can be divided into three groups [10]:

1) The rules of within-word reduction (for instance, unstressed vowels are reduced up to complete disappearance if they are located between the same consonants: *bal̲a̲laika* /balala!jka/ → /balla!jka/ (*balalaika* in English);

2) The rules of cross-word reduction (for instance, phoneme /j/ located at the word end is completely reduced if it follows an unstressed vowel: *dragotsennyj kamen'* /dragatse!nyj ka!m'in'/ → /dragatse!ny ka!m'in'/ (*precious stone* in English).

3) The rules of cross-word assimilation (for instance, the first vowel /i/ in a word, located after a hard consonant, transforms to the phoneme /y/: *fil'm i̲nteresnyj* /f'i!l'm ynt'ir'e!snyj/ (*interesting film* in English).

The set of all possible alternative pronunciation variants are produced by applying this rules to the basic word transcriptions. Forced alignment is performed for selection of the best transcription from multiple alternative transcriptions. At forced alignment a recognizer chooses the most appropriate transcription for speech signal from the alternative transcriptions list. In this case the selection of the transcription is carried out only between alternative transcriptions of the same word or phrase. For every alignment the Viterbi algorithm computes the probability that the phonemic transcription and speech signal match with each other [11]. The optimal transcriptions variants are chosen based on the highest probabilities [12]. As a result of the forced alignment a transcription that matches with a certain part of speech signal is chosen. The analysis of how often every transcription was chosen during training is performed. Only transcriptions with relative appearance frequency higher than a certain threshold are added to the resulting extended vocabulary. As a result the extended vocabulary containing the best transcriptions for every word is obtained.

Table I
POSITIONAL VOWEL CHANGES IN PRE-STRESSED SYLLABLES

Original vowel (for grapheme)	**Resulting phoneme depending on position**				
	At the beginning of a word	**After velar consonants**	**After paired hard consonants**	**After paired palatalized consonants**	**After fricatives /S/, /Z/, /ts/**
/e/ (Э,Е)	/1/	/i/	/1/	/i/	/1/
/i/ (И)	/i/	/i/	-	/i/	-
/1/ (Ы)	-	-	/1/	-	/1/
/a/ (А,Я)	/a/	/a/	/a/	/i/	/a/
/o/ (О,Ё)	/a/	/a/	/a/	/i/	/a/
/u/ (У,Ю)	/u/	/u/	/u/	/u/	/u/

Table II
POSITIONAL VOWEL CHANGES IN POST-STRESSED SYLLABLES

Original vowel (for grapheme)	**Resulting phoneme depending on position**		
	After velar consonants	**After paired hard consonants and /S/, /Z/, /ts/**	**After paired palatalized consonants and /tS'/, /S':/**
/e/ (Э,Е)	/1/	/i/	/1/
/i/ (И)	/i/	/i/	-
/1/ (Ы)	-	-	/1/
/a/ (А,Я)	/a/	/a/	/a/
/o/ (О,Ё)	/a/	/a/	/a/
/u/ (У,Ю)	/u/	/u/	/u/

III. LANGUAGE MODEL CREATION

We have proposed an integral language model (LM) that takes advantages of statistical and syntactic text analysis. Figure 1 presents the scheme of creation of the LM using some elements of the syntactic analysis. A training text corpus is processed in parallel identifying *n*-grams and syntactic dependencies in sentences and then the results of both analyzers are fused in the integral stochastic model that takes into account frequencies of the detected word pairs. These analyzers complement each other very well: syntactic one is used to find long-distance dependencies between words (potential *n*-grams not appeared in training data), but not relations between the adjacent words, which are covered by the statistical analyzer. For the statistical text analysis we employ the CMU SLM Toolkit v2, while VisualSynan v1.0 [13] from the AOT release is used for the syntactic analysis. The latter parses input sentences and produces a graph of syntactical dependencies between pairs of lexical items. There are 32 different types of syntactic groups in the analyzer in total, but we extract 9 of them, which can describe long-distance (over one word at least) relations between pairs of words. The following types of syntactic groups are selected:

1) subject - predicate, e.g., "мы её не знали" (English: "we did not know her");

2) adjective - noun: "ежегодный вокальный конкурс" ("an annual vocal competition");

3) direct object: "решить эту сложную проблему" ("to solve this complicated problem");

4) adverb - verb: "иногда такое бывает" ("sometimes this happens");

5) genitive pair: "темой текущего и следующего номера" ("a topic of the present and next issues");

6) comparative adjective - noun: "моё слово сильнее любого контракта" ("my word is stronger than any contract");

7) participle - noun: "дом, аккуратно построенный" ("house, carefully constructed");

8) noun - dangling adjective in a postposition: "цель, достаточно благородная" ("the aim is rather noble");

9) verb - infinitive: "мы хотим это потом изменить" ("we want to change it later").

Moreover, words of the syntactic groups (2)-(3), (7)-(9) and (1), but without subordinate attributive clauses starting with "which", "who", etc., are commutative in Russian and each such syntactic dependence produces two bigrams with direct and inverse word order. Figure 2 shows an example of the syntactic analysis of the phrase ("In the very expensive show, both military and civilian aircrafts, which arrived yesterday and today the airport of our little town, are involved") taken from the corpus. It demonstrates some types of long-distance dependences, whereas all the adjacent word pairs are modeled by the statistical bigrams. The commutative groups are denoted by the double-sided arrows. Thus, syntactic parsing of this sentence produces 6 long-distance word pairs additionally to the statistic processing, which gives 9 bigrams. *n*-gram likelihoods in the integral stochastic LM are calculated after merging the results (the count files) of both analyzers based on their frequency in the training text data. *n*-grams with words occurred once in the corpus are deleted from the LM, because they likely contain typos or these words are extremely rare.

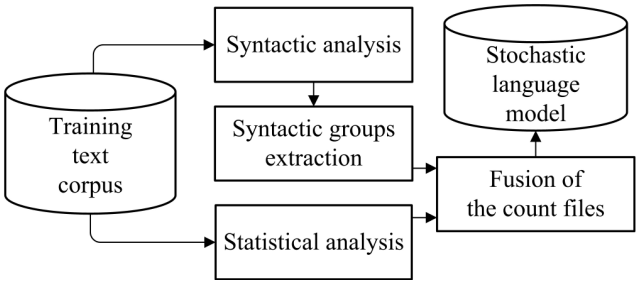

Figure 1. Integral syntactic-statistical LM generation

At present, there are some text corpora for Russian, for example, Russian National Corpus (www.ruscorpora.ru) of 140M words and Corpus of Standard Written Russian (www.narusco.ru) of 20M items. However, they contain a few of shorthand reports of the spoken language. For LM creation, we have collected and processed a text corpus consisting of the following on-line newspapers for the last five years: "Новая газета" (www.ng.ru), "СМИ" (www.smi.ru), "Лента.ру" (www.lenta.ru), "Газета.ру" (www.gazeta.ru). The news corpus contains texts that mirror state-of-the-art Russian including the spoken language. The volume of the corpus after text normalization and deletion of doubling or short (<5 words) sentences is over 110M words, and it has about 937K unique word-forms. As the results of the statistical analysis we have obtained almost 6M unique bigrams (*n*-gram cutoff is 1) and the syntactic analysis extended the integral LM to 6.9M items, i.e. 15% increase with respect to the baseline model.

IV. CONVERSATIONAL RUSSIAN SPEECH RECOGNITION SYSTEM

The architecture of software complex of conversational Russian speech recognition system is presented on Figure 3. The software modules are developed on programming language C++ and Perl, also some modules of software complexes of the CMU-Cambridge Statistical Language Modeling Toolkit (CMU SLM) [14], HTK (Hidden Markov Model Toolkit) [15], AOT [13] are used.

The system work in two modes: training and recognition. In this section the training mode of the system will be described in particular. In the training mode, acoustic models of speech units, language model, and phonemic vocabulary of word-forms that will be used by recognizer are created. For acoustic model's training manually segmented corpus of Russian speech is used; the language model is created based on a text corpus. Thus, the following stages of the training process can be distinguished:

– preliminary processing of the text material for creation of the language model;

Figure 2. An example of the syntactic phrase analysis (numbers of types of long-distance syntactic dependencies are shown)

- creation of transcriptions for words from the collected text corpus;
- selection of the best transcriptions from the multiple variants;
- creation of the stochastic language model;
- training of the acoustic models of speech units.

The block of preliminary text material processing carries out the following operations. At first, texts are divided into sentences, which must begin from an uppercase letter or a digit before which inverted commas may be situated. A sentence ends by the point, exclamation, question mark or dots. It takes into account that initials and/or a surname can be placed within the sentence. Formally, it is similar to a boundary between two sentences, therefore, if the point is after a single uppercase letter, the point is not considered as the end of the sentence. Sentences containing direct and indirect speech are divided into separate sentences. These sentences can be of the following types: (1) direct speech is placed after indirect speech; (2) direct speech is before indirect speech; (3) indirect speech is within direct speech. In the first case, a formal sign for distinguishing direct and indirect speech is presence of the colon mark followed by inverted commas. In the second case, the division is made if the comma follows the inverted commas and followed by the dash. In the third case, the initial sentence is divided into three sentences: (1) from inverted commas to the corresponding comma; (2) between the first comma with dash to the second comma with dash; (3) from comma with dash to the end of the sentence.

Then, a text written in any brackets is deleted, and sentences consisting of less than six words are also deleted. Then punctuation marks are deleted, symbols "N^o" and "#" are replaced by the word "number". All numbers and digits are combined in a single class that is denoted by the symbol "N^o" in the resulting text. A group of digits, which can be divided by point, comma, space or dash sign is denoted as a single number. Also the symbol "N^o" denotes Roman numbers that are a combination of Latin letters I, V, X, L, C, D, M, which can be divided by space or dash. Internet links and E-mails are distinguished in single classes and denoted by the symbols "<>" and "<@>", respectively. Uppercase letters are replaced by lowercase letters, if a word begins from an uppercase letter. If a whole word is written by the uppercase letters, then such change is made, when the word exists in a vocabulary only. Also at this stage of training the vocabulary of words occurred in the training corpus is created.

The word transcription creation block can generate both basic and alternative transcriptions for the words from the vocabulary obtained by the block of preliminary text processing. Basic transcriptions are made with help of canonical transcribing rules that describe standard pronunciation of a spoken isolated word. Alternative transcriptions take into account within-word and cross-word reduction and assimilation phenomena that are specific for conversational speech.

The block of best transcription selection is used only if mode of alternative transcription creation is chosen. In this block the most commonly used transcription variants are chosen as alternative variants of basic transcription. As a result of work of the block of phonemic transcription creation and best alternative transcriptions selection, the list of phonemic representations of words from the text corpus is created. This list contains basic transcriptions and the best transcriptions for words appeared in the training text corpus. This list of words with its canonical and alternative transcriptions is phonemic vocabulary of the speech recognition system.

The block of *n*-gram model creation performs statistic and syntactic analysis of text corpus and builds an integral stochastic language model. This model reflects connections between neighboring words as well as syntactically connected pairs separated by the other words in the training text material.

Training of acoustic models of speech units is carried out with use of Russian speech corpus. Speech databases with records of large number of speakers are needed to provide speaker-independent speech recognition. Recording is performed in soundproofing room. The phrases to be pronounced are sequentially shown to a speaker. Each phrase is recorded in separate sound wav file. Then semiautomatic labeling of acoustic signal on phrases, words, and phonemes is carried out.

Hidden Markov models (HMM) are used for acoustic modeling, and each phoneme (speech sound) is modeled by one continuous HMM. A phoneme model has three states: the first state describes phoneme's start, the second state present the middle part, and the third state is phoneme's end. HMM of a word is obtained by connection of phoneme's models from corresponding phonemic alphabet. Similarly the models of words are connected with each other, generating the models of phrases. The aim of training of the acoustic models based on HMM is to determine such model's parameters that would lead to maximum value of probability of appearance of this sequence by training sequence of observations [16].

In the speech recognition mode, an input speech signal is transformed into the sequence of feature vectors, and then search of the most probable hypothesis is performed with the help of preliminary trained acoustic and language models.

V. EXPERIMENTAL RESULTS

To test the system we used a speech corpus containing 100 continuously pronounced phrases consisting of 1068 words

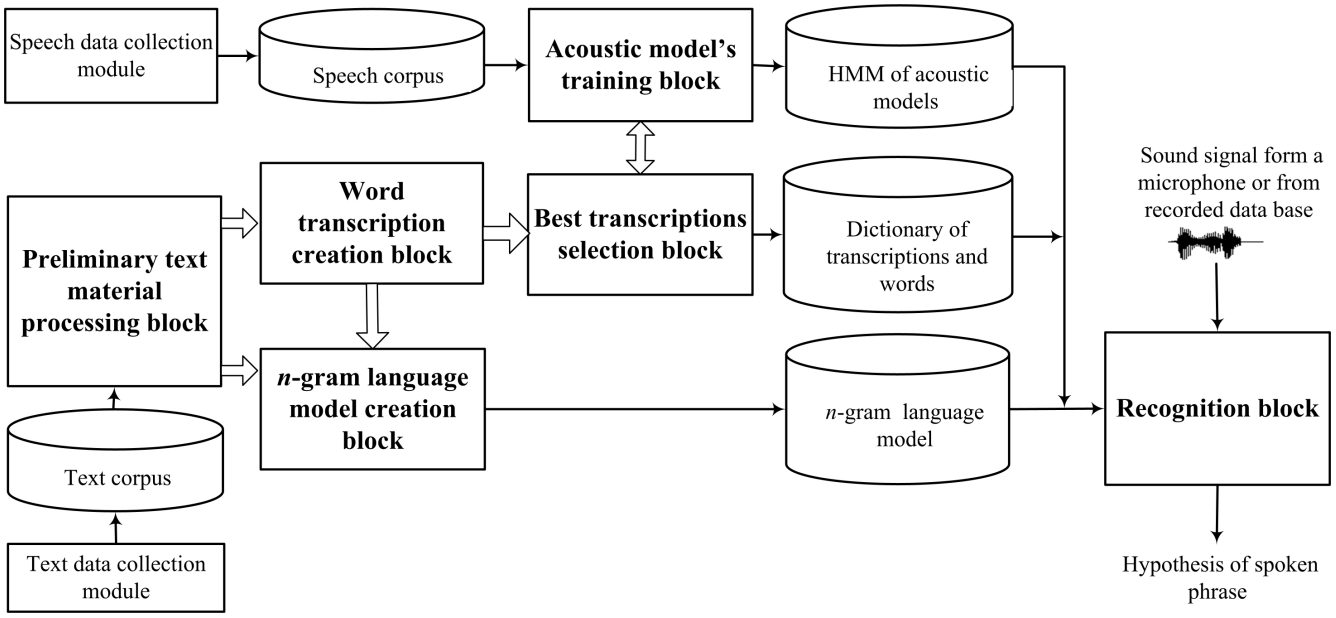

Figure 3. The architecture of software complex of conversational Russian speech recognition system

(7191 letters). The phrases were taken from the materials of the on-line newspaper "Фонтанка.ru" (www.fontanka.ru). The speech data were recorded with 44.1 KHz sampling rate (for ASR downsampled to 16 KHz), 16 bits per sample, SNR was 35dB at least, by a stereo pair of Oktava MK-012 stationary microphones (close talking ≈20 cm and far-field ≈100 cm microphone setup) connected to PC via Presonus Firepod sound board.

As for acoustic features, we used 13-dimentional Mel-Frequency Cepstral Coefficients (MFCC) with the 1^{st} and 2^{nd} order derivatives calculated from the 26-channel filter bank analysis of 20 ms long frames with 10 ms overlap. Cepstral mean subtraction is applied to audio feature vectors. Continuous density HMMs with 16 Gaussians per state model Russian context-dependent phones.

Table III summarizes language model parameters (LM informational entropy and perplexity, amount of out-of-vocabulary OOV words, bigram hit) and recognition results for the given corpora in terms of the word error rate (WER) and grapheme (or letter that is the same) error rate (LER). The pronunciation vocabulary contains almost 210.1K word-forms and the integral syntactic-statistical bigram LM is used.

Relatively high speech error rates can be explained by the inflective nature of the given Slavic language, where each stem corresponds with tens/hundreds of endings, which are usually pronounced in continuous speech not so clearly as the beginning parts of the words and often different orthographic word-forms have identical phonemic representations. We have also applied inflectional word error rate (IWER) measure [17], [18], which assigns a weight k_{inf_1} to all "hard" substitutions S_1, where lemma of the word-form is wrong, and a weight k_{inf_2} to all "weak" substitutions S_2, when lemma of the recognized word-form is right, but ending of the word-form is wrong:

$IWER = \frac{I+D+k_{inf\_1}S_1+k_{inf\_2}S_2}{N} * 100\%$

In our experiments, the IWER measure with k_{inf_1}=1.0 and k_{inf_2}=0.5 was 29.85%, so in total about 10% of the errors were caused by misrecognized word endings.

Table III
POSITIONAL VOWEL CHANGES IN POST-STRESSED SYLLABLES

Entropy, bit/-word	**Perplex**	**OOV words,%**	***n*-gram hit, %**	**WER,%**	**LER,%**
9.6	772	0.75	84.1	33.43	11.83

VI. CONCLUSION

The pronunciation variety is one of the main problem during the development of conversational speech recognition system. The inflective nature of Russian and the free word order are additional issues. The developed software complex generates multiple transcription variants that take into account the variability of pronunciation in conversational speech.Also this complex creates a stochastic Russian language model that is distinctive by joint application of statistic and syntactic analysis of training text data and that takes into account long-distance grammatical relations between words in the phrase. Further research will be devoted to other aspects of conversational speech like speech disfluency.

REFERENCES

[1] E. A. Zemskaya, "Conversational Russian speech", *E. A. Zemskaya (Edit.), Moscow: Nauka*, 1973, 485 p. (in Russian).
[2] C. P. Browman, L. Goldstein, "Articulatory phonology: An overview", *Phonetica*, 49, 1992, pp. 155-180.

[3] I. Amdal, “Learning pronunciation variation. A data-driven approach to rule-based lexicon adaptation for automatic speech recognition”, PhD thesis. *Department of Telecommunications Norwegian University of Science and Technology*, Norway, 2002.
[4] G. L. Moore, “Adaptive Statistical Class-based Language Modelling”. PhD thesis, Cambridge University, 2001.
[5] A. Vaičiūnas, “Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition”, PhD thesis, Vytautas Magnus University, Kaunas, 2006.
[6] A. B. Kholodenko, “On Construction of Statistical Language Models for Russian Speech Recognition Systems”. *J. Intelligent Systems*. vol. 6 (1-4), Moscow, 2002, pp. 381-394.
[7] D. Vazhenina, K. Markov, “Phoneme Set Selection for Russian Speech Recognition”, *in Proc. 7th Int. Conf. on NLP and Knowledge Engineering NLP-KE’11*, Japan, 2011, pp. 475-478.
[8] N. Shvedova, et al, “Russian Grammar”. Vol. 1, *Moscow: Nauka*, 1980, 783 p. (in Russian).
[9] B. M Lobanov, L. I. Tsirulnik, “Modeling of within-word and cross-word phonetic-acoustical phenomena of the complete and conversational speech style in the system of speech synthesis by a text”, *Proc. of First Interdisciplinary Workshop "Conversational Russian Speech Analysis"*. - SPb.: SUAI, 2007, pp. 57-71 (in Russian).
[10] I.S. Kipyatkova, A.A. Karpov, “The module of phonemic transcription for conversational Russian speech recognition system”, *Artificial intelligence*, Donetsk, Ukraine, Vol. 4, 2008, pp. 747-757 (in Russian).
[11] M. Saraclar, “Pronunciation Modeling for Conversational Speech Recognition”, PhD thesis. Baltimore, USA, 2000.
[12] J. M. Kessens, M. Wester, H. Strik, “Improving the performance of Dutch CSR by modeling within-word and cross-word pronunciation variation”, *Speech Communication*, vol. 29, 1999, pp. 193-207.
[13] A. Sokirko, “Morphological modules on the website www.aot.ru”, *in Proc. 10th International Conference "Dialog-2004"*, Protvino, Russia, pp. 559-564, 2004.
[14] P. Clarkson, R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge toolkit”, *in Proc. of EUROSPEECH. Rhodes*. Greece. 1997. pp. 2707-2710.
[15] S. Young et al., “The HTK Book (for HTK Version 3.4)”. Cambridge. UK, 2009, 375 p.
[16] L. Rabiner, B.-H. Juang, “Fundamentals of Speech Recognition”. *Prentice Hall*, 1993, 507 p.
[17] K. Bhanuprasad, M. Svenson. “Errgrams - a way to improving ASR for highly inflective dravidian languages”, *in Proc. 3rd International Joint Conf. on Natural Language Processing IJCNLP’08*, India, 2008, pp. 805-810.
[18] A. Karpov, I. Kipyatkova, A. Ronzhin, “Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis”, *in Proc. Interspeech’11, Florence*, Italy, 2011, pp. 3161-3164.