International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 10, No. 1, pp. 11 – 30, 2013

MODELING OF PRONUNCIATION, LANGUAGE AND NONVERBAL UNITS AT CONVERSATIONAL RUSSIAN SPEECH RECOGNITION

IRINA KIPYATKOVA

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS),
St. Petersburg, Russia
kipyatkova@iias.spb.su

ALEXEY KARPOV

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS)
St. Petersburg, Russia
karpov@iias.spb.su

VASILISA VERKHODANOVA

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS)
St. Petersburg, Russia
verkhodanova@iias.spb.su

MILOŠ ŽELEZNÝ

The University of West Bohemia
Plzen, Czech Republic
zelezny@kky.zcu.cz

The main problems of a conversational Russian speech recognition system development are variability of pronunciation, free word-order in sentences and presence of speech disfluencies. In the paper, pronunciation variability is modeled by creation of multiple word transcriptions. A syntactic-statistical language model that takes into account long-distant word dependencies is proposed for Russian language modeling. Also in this paper the results of analysis of such speech disfluencies as artefacts and filled pauses, which were extracted during segmentation of the Russian speech corpus, are presented. The recognition accuracy of nonverbal elements in the collected corpus was 87%. The proposed methods of pronunciation variability modeling and syntactic-statistical language model creation were realized in the software complex for Russian speech recognition. The performed experiments with large vocabulary using syntactic-statistical language model showed that word error rate of the system was 33%.

Keywords: Conversational speech recognition, multiple transcriptions, Russian language modeling, speech disfluencies.

1. Introduction

The majority of state-of-the-art automatic speech recognition (ASR) systems can efficiently analyze words pronounced in isolation 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 while developing speech recognition system.

Any speech recognition system uses a phonemic vocabulary of the words. In general, 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 [Zemskaya (1973); Browman and Goldstein (1992)]. 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 modeling of spontaneous speech variation depends on the way of alternative transcription generating.

The next stage after word recognition is generation of grammatically correct and meaningful hypothesis of the pronounced phrase by a language model (LM). 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 Russian language owing to the free word order in sentences and the existence of a large amount of word-forms for every lexical unit caused by the inflective nature of the language.

Additional problem of automatic spontaneous (informal) speech recognition is presence of speech disfluencies such as filled pauses, nonverbal pauses, artefacts that occur within the flow of otherwise fluent speech. Such disfluencies are an obstacle for automatic processing of speech and its transcriptions. These elements may be recognized as key words and thus impair accuracy of speech recognition. Eliminating such uninformative elements from speech signal on initial processing stage and transmitting only useful information on the next level of processing will allow to avoid a lot of errors during speech recognition.

In this work, our approach to speech variability and language modeling and the software complex for conversational Russian speech processing are 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 Sections II, the state-of-the-art methods of creation of the vocabulary with words and multiple transcriptions, creation of *n*-gram language model, and speech disfluencies processing are considered. Section III is devoted to description of original method of multiple phonemic transcriptions generating, which helps to take into account speech pronunciation variety. An approach to creation of statistical language model with using syntactical rules is described in Section IV. The training corpus with speech disfuencies which was used for creation of acoustical models of nonverbal units is presented in Section V. A software complex for Russian speech recognition is described in Section VI. Section VII presents experimental results.

2. Related work

2.1. *Methods for pronunciation modeling*

There are two main approaches to the problem of pronunciation variability modeling [Amdal (2002)]: knowledge-based and data-driven methods. In the knowledge-based methods variability of pronunciation is specified 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. These 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 occurrence using the training speech corpus.

Also there are direct and indirect approaches to creation of alternative variants of word pronunciation, which could be applied for knowledge-based and data-driven 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 there are only pronunciation variants that frequently occur in the training corpus 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 tied with manual data processing and creation of a huge list of alternative transcriptions created automatically. So, a trade-off is required during development of speech recognition vocabulary. Combinations of these methods are often used. For example, using knowledge about reduction and assimilation phenomena a set of rules is constructed, and the dictionary with alternative transcriptions is generated using these rules. Then with a hand-labeled speech corpus it is checked which of the alternative transcriptions really exist and the probabilities of appearance of the pronunciation variants are estimated. A similar way of generating of alternative transcriptions was applied in [Byrne *et al.* (1997)].

2.2. *Methods for language modeling*

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 $W=(w_1, w_2,\ldots, w_n)$ 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 [Moore (2001)]. 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 [Moore (2001); Vaičiūnas (2006)]. 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 [Moore (2001)]. 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 [Kholodenko (2002)], 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 word order in sentences permitted by Russian constrains implementation of the referred language models. Therefore some approaches to modeling long-distance dependencies between words are required. One of the methods for long-distance word dependencies modeling is the syntactical text analysis.

In recent ASR systems, syntactical analysis is often embedded into various processing levels: language modeling, speech decoding, hypotheses processing, etc. In [Szarvas and Furui (2003)], a stochastic morpho-syntactical language model for agglutinative Hungarian ASR is introduced. This model describes the valid word-forms (morpheme combinations) of the language. The stochastic morpho-syntactic language model decreased the morpheme error rate by 17.9% compared to the baseline trigram system. The morpheme error rate of the best configuration was 14.75% in a 1350 morpheme dictation task.

Syntactical analysis is applied for post-processing recognition hypothesis in [Rastrow *et al.* (2012)]. The substructure sharing, which saves duplicate work in processing

hypothesis sets with redundant hypothesis structures, is proposed. The syntactic discriminative LM was trained using dependency parser and part of speech tagger that led to significant speedup and reductions in word error rate (WER).

A syntactic analyzer was used for rescoring N-best list and selection of the best hypothesis in [Huet *et al.* (2010)]. It was offered to apply a morphologic and syntactic post-processing of N-best lists in French ASR in order to parse and re-order the list of recognition hypothesis according to grammatical correctness criteria. This method relies on part-of-speech and morphological information. In this method, N-best list was automatically tagged and each hypothesized word was referred to its morpho-syntactic class. Then morpho-syntactic scores were computed and combined with acoustic and language model scores. New score including morpho-syntactical one is used to reorder the N-best lists.

In our work, a syntactic text analysis was employed in order to detect grammatically-connected word-pairs in the training data. Results of the syntactic analysis are combined with statistical *n*-grams to improve the *n*-gram language model. At present, there are several syntactic analyzers for Russian: Treeton [Starostin and Mal'kovskiy (2007)], analyzer of ETAP-3 machine translator [Iomdin *et al.* (2012)], dependency parser SyntAutom [Antonova and Misyurev (2012)], ABBYY syntactic and semantic parser [Anisimovich *et al.* (2012)], Dictum [Skatov *et al.* (2012)], syntactic analyzer SMART [Leontyeva and Kagirov (2008)], and AOT Synan syntactic analyzer [Sokirko (2004)]. The latter one has several advantages: it is open source and its databases permit modifications; provides high speed of text data processing, and uses constantly updated grammatical database based on a standard grammatical dictionary [Zaliznjak (2003)]. So, the last version of the AOT "VisualSynan" syntactical analyzer (www.aot.ru) was used in our work as well.

2.3. *Methods for speech disfluencies processing*

The phenomenon of speech disfluencies is known under a wide range of different terms including "non-fluency", "discontinuity", "flustered speech", "speech disturbance", "hesitation", "speechmanagement", "changes of mind", "self repair", "self correction", "self editing" etc. [Eklund (2003)]. The origin of speech disfluencies appearance may be of different nature: they may be caused by external influences as well as by failures in speech act planning [Podlesskaya and Kibrik (2004)]. Among failures in speech act planning there are filled pauses, self-repairs, slips of the tongue. Taking into account different causes of speech disfluencies it is possible to introduce the classification shown on Figure1.

Pause of hesitation (filled pause) is the break in the phonation, which is often filled with certain sounds. Usually such pauses are semantic lacunas and show that speaker needs an additional time to formulate next piece of utterance [Clark and Fox Tree (2002)]. There are different types hesitation fillers [Clark and Fox Tree (2002)]: unfilled pauses; sound prolongations in words; words-alike "pre-lexical" pause fillers; auxiliary discourse elements (words and phrases (so to say)).

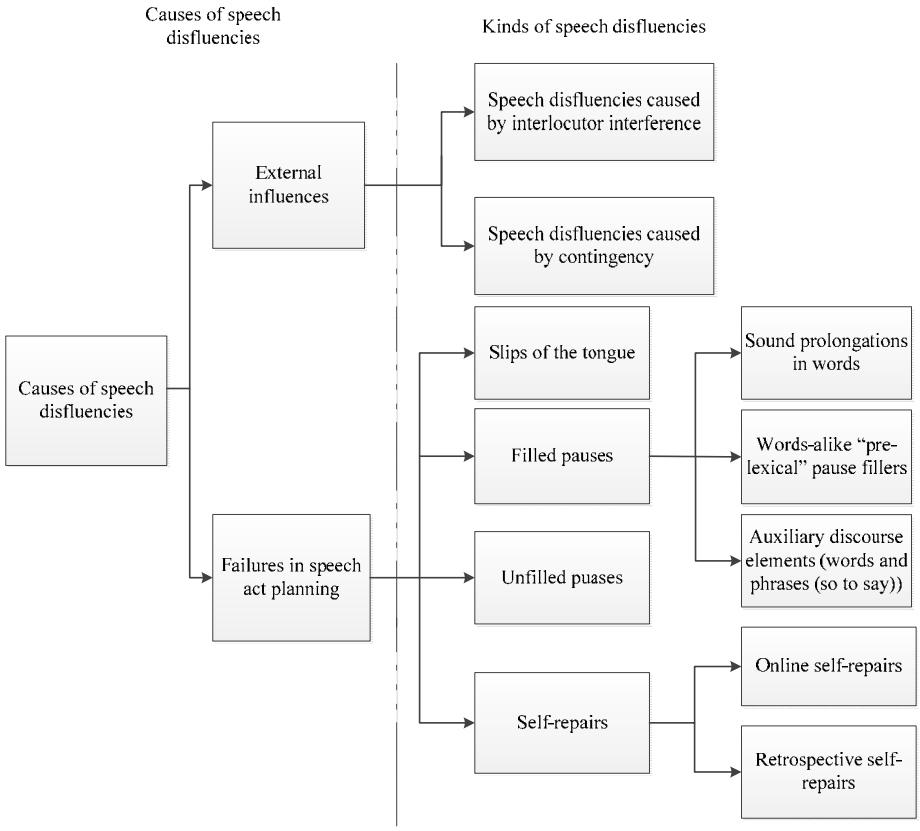

Fig. 1. Classification of speech disfluencies.

Self-repairs appear when at a particular moment in discourse speakers decide that certain part of their utterance does not correspond to their intentions, and replace it partly or wholly. There are online and retrospective self-repairs, that is made right after the mistake or post factum. Traditionally slips of the tongue are also considered among speech disfluencies.

There is a tradition of describing speech disfluencies in terms of temporal characteristics. Shriberg [Shriberg (1994)] uses such terms:

- reparandum (RM) - the part of signal, that corresponds to the whole deleted speech section;
- interruption point (IP) – the beginning of speech section, corresponding to the "moment of interruption" of fluent speech and the occurrence of speech disfluency;
- interregnum (IM) (other authors use "editing phase" [Levelt (1983)] and "disfluency interval" [Nakatani and Hirschberg (1994)]) – this term is used for describing the stretch from RM till the beginning of repair.
- repair (RR) – the stretch of speech, corresponding to the reparandum material.

There are a lot of publications, dedicated to the speech disfluencies modeling for ASR systems [Masataka *et al.* (1999); Liu *et al.* (2006)]. There is a group of method that

deals with the disfluencies phenomenon by means of parametrical signal processing [Masataka *et al.* (1999)]. Also there is a group of methods aimed to rise of the quality of spontaneous speech recognition by means of detecting and deleting speech disfluencies at the stage of speech signal preprocessing [Kaushik *et al.* (2010)] or by means of deleting speech disfluencies using speech transcripts [Liu *et al.* (2003); Snover *et al.* (2004)].

In [Kaushik *et al.* (2010)] an algorithm, which defines and eliminates filled pauses and repetitions from the speech signal, is proposed. For detection of boundaries of filled pauses the following characteristics were applied: duration, pitch, spectral and formant characteristics. For extraction and further elimination of repetitions the proposed algorithm used duration and frequency of the repeated segments as well as the Euclidian distance between the logarithms of the Linear Predictive Coding (LPC) spectra of each pair of the voiced sections around a long pause. Also the fact that repetitions are usually accompanied by a pause was taken into account.

There are number of publications aimed to rise speech disfluencies recognition quality by means of additional knowledge sources such as different language models. In [Liu *et al.* (2003)] three types of speech disfluencies are considered: (1) repetition, (2) revisions (content replacement), (3) restarts (or false starts). A part of Switchboard-I as well as its transcription (human transcriptions and ASR output) was taken for research. Normalized word and pause duration, pitch, jitter (undesirable phase and/or random frequency deviation of the transmitted signal), spectral tilt, and the ratio of the time, in which the vocal folds are open to the total length of the glottal cycle were taken as the prosodic features. Also three types of language models were used: (1) hidden-event word-based language model that describes joint appearance of the key words and speech disfluencies in spontaneous speech; (2) hidden-event POS-based language model that uses statistics on part-of-speech (POS) to capture syntactically generalized patterns, such as the tendency to repeat prepositions; (3) repetition pattern language model for detection of repetitions.

3. Multiple phonemic transcriptions for speech variety modeling

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 [Vazhenina and Markov (2011)], 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).

We perform grapheme-to-phoneme conversion by applying several phonetic rules to a list of word-forms. For detection of stressed vowels in words we employ an extended morphological database of more than 2.3M word-forms [Kipyatkova (2012)].

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 1; (2) changes of vowels in post-stressed syllables, which are shown in Table 2; (3) positional changes of consonants can happen in the following cases [Shvedova *et al.* (1980)]:

- 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.

Table 1. 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 2. 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*)	/i/	/1/	/i/
/i/ (*И*)	/i/	/1/	/i/
/1/ (*Ы*)	-	/1/	-
/a/ (*А*,*Я*)	/a/	/a/	/a/
/o/ (О,Ё)	/a/	/a/	/a/
/u/ (У,Ю)	/u/	/u/	/u/

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 once 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.

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 [Lobanov and Tsirulnik (2007)]. These rules can be divided into three groups [Kipyatkova and Karpov, 2008]:

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: *bal̲alaika* /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: *dragotsennyj̲ 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 [Saraclar (2000)]. The optimal transcriptions variants are chosen based on the highest probabilities [Kessens *et al.* (1999)]. 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.

4. Syntactic-statistical language model for Russian

At present, there are several large commercial text corpora of Russian, for instance, the Russian National Corpus (www.ruscorpora.ru) and the Corpus of Standard Written Russian (www.narusco.ru), which mainly contain text material of the end of the 20th century. These corpora include different types of texts: fiction, political essays, scientific, etc. They also contain a few shorthand reports in spoken language. For the language model creation, we collected and automatically processed a new Russian text corpus of on-line newspapers. This corpus was assembled from recent news published in freely available Internet sites of 4 on-line Russian newspapers for the years 2006-2011 (www.ng.ru, www.smi.ru, www.lenta.ru, www.gazeta.ru). The database contains text data that reflect contemporary Russian including some spoken language.

The size of the corpus after the text normalization and the deletion of doubling and short (less than 5 words) sentences is over 110M words, and it has about 937K unique word-forms. As a result of the statistical analysis, we obtained almost 6M unique bigrams (*n*-gram cutoff=1).

Based on this text corpus, stochastic language models were created using both statistical and syntactic automatic analyzers. Fig. 2 illustrates the process of creation of a stochastic LM for Russian using some syntactic analysis elements.

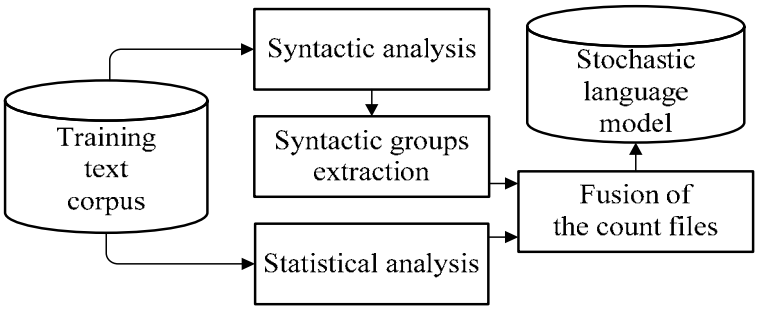

Fig. 2. Diagram of the process for creation of the integral syntactic-statistical language model for Russian.

The training text corpus is processed in parallel identifying *n*-grams and syntactic dependencies in sentences; 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: the syntactic one is used to find long-distance dependencies between words (potential *n*-grams unseen in the training data), but not the relations between the adjacent words, which are discovered by the statistical analyzer. For the statistical text analysis we use the CMU SLM Toolkit ver. 2 [Clarkson and Rosenfeld (1997)], while the software "VisualSynan" ver. 1 from the AOT project [Sokirko (2004)] (www.aot.ru) is used for the syntactic analysis. The latter parses input sentences and produces a graph of syntactical dependencies between the pairs of lexical units.

The main aim of the syntactical analysis is to extract syntactical groups in a sentence [Nozhov (2003); Sokirko (2004)]. Syntactical group is defined as follows: it is a group type, a pair of syntactically-connected words, and grammemes. Group type is a string constant, for example: "*ПРИЛ_СУЩ*" (adjective-noun), "*ПП*" (prepositional phrase), etc. Group grammemes are the morphological characteristics of words, which determine behavior and compatibility of elements in other groups. There are 32 different types of syntactic groups in the analyzer in total, but we extract only 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) verb – infinitive: "мы хотим это потом изменить" ("we want to change it later").
8) participle – noun: "дом, аккуратно построенный" ("house, carefully constructed");
9) noun – dangling adjective in a postposition: "цель, достаточно благородная" ("the aim is rather noble");

Moreover, words of the syntactic groups (2), (3), (7), (8), (9) and (1), but without subordinate attributive clauses starting with words 'which', 'who', etc., are commutative in Russian and each such syntactic dependence produces two bigrams with direct and inverse word order. Fig. 3 shows an example of the syntactic analysis of the phrase taken from the corpus: "In the very popular show, military and civilian aircrafts, which arrived yesterday and today in the airport of our city, are involved".

Fig. 3. An example of syntactic phrase analysis (long-distance syntactic dependencies are shown by arrows).

The example demonstrates some types of long-distance dependences, whereas all the adjacent word pairs are modeled by 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.

After the normalization of the text corpus, statistical processing is carried out and a list of bigrams with their frequency of occurrence is created. Then, syntactic analysis of the text corpus is performed, and results are processed. Grammatically-connected pairs of words (syntactic groups) which were separated in the text by other words are detected during the processing. Then the list of bigrams obtained by the statistical analysis and the list of grammatically-connected pairs of words, which were extracted during syntactical analysis, are merged.

We applied the Good-Turing discounting method, when creating the language models, and used *n*-gram cutoff =1, so only bigrams and syntactical groups with frequency of appearance more than 1 were added to the language model. After the statistical analysis of the collected text corpus we obtained 6M bigrams. As a result of the syntactical analysis we added more than 0.9M new bigrams. Thus, the total number of bigrams in the extended language model is 6.9M, and consequently the size of the syntactic-statistical language model increased by 15% compared with the statistical model. The size of the vocabulary of this model was 210K word-forms. For comparison the statistical bigram language model with a smaller recognition vocabulary of 79K entries was created (*n*-gram cutoff=8).

The entropy, perplexity, out-of-vocabulary words (OOV) and *n*-gram hit rates calculated for evaluation of the created language model are summarized in Table 3. For this purpose, 100 phrases from the on-line newspaper "Фонтанка.ru" (www.fontanka.ru) were used. The speech corpus that further was used for experiments on continuous speech recognition was recorded based on these phrases. The *n*-gram hit is the number of *n*-grams in test data that are present in the language model.

Table 3. Characteristics of the language models.

Language model type	Lexicon size, K words	Number of *n*-grams, M	Entropy, bit/word	Perplexity	OOV rate, %	*n*-gram hit, %
Statistical bigram model	79	1.0	9.7	851	3.5	72.6
Statistical bigram model	208	6.0	9.6	777	0.8	83.6
Syntactic-statistical model	210	6.9	9.6	772	0.8	84.2

These parameters have quite large values for Russian language models, which is a great challenge for the speech recognition. For comparison, relative percentage of the

OOV words in English texts is 0.31% for a text corpus of 1.12M words [Whittaker (2000)].

5. Acoustic models of nonverbal units

In order to separate speech disfluencies from the key words and exclude them from further processing one needs to create acoustic models for these phenomena. To train acoustic models of nonverbal units in the given study there was collected a corpus of Russian speech, which is composed of 6 conference reports (of three men and three women). The corpus is 70 minutes long. During segmentation there have been detected artefacts and filled hesitation pauses – features, that are characteristic for any spontaneous speech. To train and test the system only those nonverbal units that have occurred more than two times in the corpus were used. The list of such units is given in Table 4.

Table 4. Description of modeled nonverbal units in the spontaneous speech.

Group of the nonverbal units	Designation	Nonverbal unit
Artefact	ar.brth	Breath
	ar.clth	Clearing throat/coughing
	ar.smck	Smacking
Filled pauses	h.a	/a/
	h.au	/au/
	h.e	/e/
	h.em	/em/
	h.eu	/eu/
	h.m	/m/
	h.me	/me/
	h.mne	/mne/

As the result there were created acoustic models for three types of artefacts (breath, clearing throat/coughing, and smacking) and of eight types of filled pauses. Every model of nonverbal element is based on left-right Hidden Markov Model, which contains three basic states.

The distribution of frequency of different nonverbal units pronounced by different speakers and its average duration in the collected corpus is presented in Table 5. In all we have segmented 1052 nonverbal elements; their total duration is 7 minutes that is approximately 10% of duration of all records.

From the table one may see, that the majority of nonverbal elements consists of filled pause h.e (46.15% of total number of nonverbal units) and breath (31.91%), these units were found in the speech of all six speakers. Also ar.clth, h.em, h.m were present in the speech of the speakers majority.

Table 5. Description of the collected corpus of nonverbal units.

Speaker	Speech duration, min	Number of occurrences of nonverbal units											
		ar.brth	ar.clth	ar.smck	h.a	h.au	h.e	h.em	h.eu	h.m	h.me	h.mne	Overall
1	18	94	15	7	1	0	147	12	1	25	0	0	302
2	15	9	1	1	10	0	141	4	0	20	4	0	190
3	8	49	22	0	5	1	64	23	3	11	1	2	181
4	2	9	0	0	0	0	26	0	0	0	0	0	35
5	13	149	4	0	0	4	61	2	12	16	6	1	255
6	14	26	8	0	0	0	47	1	0	7	0	0	89
Total number of unit occurrence		336	50	8	16	5	486	42	16	79	11	3	1052
Relative number, %		31.94	4.75	0.76	1.52	0.48	46.20	3.99	1.52	7.51	1.05	0.29	100.0
Average duration (ms)		392	345	194	454	833	423	679	834	504	465	892	–

6. Conversational Russian Speech Recognition System

The architecture of software complex of conversational Russian speech recognition system is presented on Figure 4. 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) [Clarkson and Rosenfeld (1997)], HTK (Hidden Markov Model Toolkit) [Young *et al.* (2009)], AOT [Sokirko (2004)] are used.

The system works 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 and nonverbal 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;
- 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 and nonverbal units.

Training of acoustic models of speech units is carried out with use of the Russian speech corpus. Speech databases with records of large number of speakers are needed to provide speaker-independent speech recognition. In our research, we have used own corpus of spoken Russian speech Euronounce-SPIIRAS, created in 2008-2009 in the framework of the Euro-Nounce project [Jokisch (2009)]. The speech data were collected in clean acoustic conditions, with 44.1 kHz sampling rate, 16-bit audio quality. A signal-to-noise ratio (SNR) at least 35-40 dB was provided. The database consists of 16,350 utterances pronounced by 50 Russian native speakers (25 male and 25 female). Each speaker read 327 phonetically-balanced and meaningful sentences carefully, but fluently one time only. Total duration of speech data is about 21 hours.

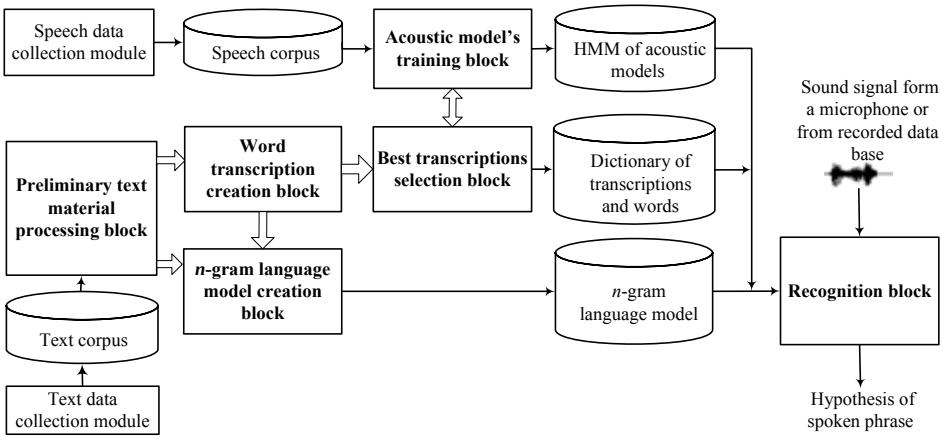

Fig. 4. The architecture of software complex of conversational Russian speech recognition system.

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 [Rabiner and Juang (1993)].

The block of preliminary text material processing performs text normalization and deletion of doubling and short sentences. 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 the 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 the 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.

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.

7. Experimental Results

At the beginning the experiments of recognition of retrieved nonverbal units were performed. The accuracy rate of all elements was 86.98% in average. The results of every nonverbal unit recognition are presented in Table 6.

Table 6. Analysis of the results of nonverbal units recognition.

Input unit	Recognition result, %										
	ar.brth	ar.clth	ar.smck	h.a	h.au	h.e	h.em	h.eu	h.m	h.me	h.mne
ar.brth	**96.73**	0.60	0.00	0.00	0.00	0.60	0.30	0.00	1.79	0.00	0.00
ar.clth	2.00	**94.00**	0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00
ar.smck	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h.a	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h.au	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00
h.e	1.23	0.82	0.00	3.29	0.00	**79.22**	3.91	5.56	5.35	0.62	0.00
h.em	0.00	0.00	0.00	2.38	0.00	4.76	**85.71**	0.00	7.14	0.00	0.00
h.eu	0.00	0.00	0.00	0.00	0.00	6.25	0.00	**93.75**	0.00	0.00	0.00
h.m	7.59	5.06	0.00	0.00	0.00	3.80	2.53	0.00	**81.01**	0.00	0.00
h.me	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00
h.mne	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**

So, one may see, that the accuracy rate of six units (ar.smck, h.a, h.au, h.me, h.mne) was 100%. The filled pause h.e was recognized worse, its recognition accuracy was 79.22 %. This unit got confused with units: ar.brth, ar.clth, h.a, h.em, h.eu, h.m, h.me. In the further work, probably, it will be necessary to check the accuracy of segmentation of the element h.e in the corpus, and to introduce additional variants of pronunciation of this hesitation type. Also units h.em, h.m had the recall rate lower than 90 %, during speech recognition they were mixed up with each other.

Next experiment was devoted to very large vocabulary Russian speech recognition. To test the speech recognition system we used a speech corpus containing 100 continuously pronounced phrases consisting of 1068 words (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.

Table 7 summarizes the obtained speech recognition results and performance in terms of word error rate (WER), letter/grapheme (includes all the letters and the white-space between words) error rate (LER) and real time-factor (RTF). The vocabulary size for each language model is given as well. Some more parameters of the created LMs were

presented above in Table 3. Test speech data of 30 minutes were used and the speech decoder was installed on a desktop PC with the Intel Core2Quad 2.66 GHz processor. The best WER and LER results were obtained with the integral syntactic-statistical language models applying the very large recognition vocabulary of 210K words.

Table 7. Summary of the results of very large vocabulary Russian speech recognition using various language models.

Language model	Vocabulary size, K words	WER, %	LER, %	RTF
Statistical bigram model	79	40.26	14.64	3.44
Statistical bigram model	208	35.72	12.38	3.96
Syntactic-statistical model	210	33.43	11.63	3.82

Relatively high word error rates can be explained by the inflective nature of the given Slavic language, where each stem corresponds to 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.

Table 8 presents some recognition examples; recognition mistakes made by the ASR system are given in a boldface. These examples show that WER decreases when the syntactico-statistical language model is used.

Table 8. Examples of recognized phrases with analysis of recognition errors.

An input sentence example		Statistical bigram language model			Syntactico-statistical language model		
Pronounced in Russian	Translation in English	Best recognition hypothesis	WER, %	LER, %	Best recognition hypothesis	WER, %	LER, %
Однако купить вакцину могут только организации	However, only organization can buy the vaccine	Однако купить **к активному каток** организаци**я**	66.7	32.6	Однако купить **картины** могут только организаци**я**	33.3	10.9
Я просто не мог пройти мимо такого события в моем родном городе	I just could not pass by a such event in my native town	Я **про снимок** пройти мимо такого события в моем родном городе	42.9	11.1	Я просто не мог пройти мимо такого события в моем родном городе	0	0
Слушай, это потрясающе, ты делаешь, что хочешь, такая возможность эксперименти-ровать	Listen, it is amazing, you do what you want, it such an opportunity for experimenting	Слушай это потрясающ**ий подделали** что хочешь такая возможность эксперименти-ровать	30.0	10.0	Слушай это потрясающе ты дела**й** что хочешь такая возможность эксперименти-ровать	10.0	3.8

We also applied inflectional word error rate (IWER) measure [Bhanuprasad and Svenson (2008); Karpov et al. (2011)], which assigns the weight kinf_1 to all "hard" substitution errors S1, where lemma of the word-form is wrong, and the weight kinf_2 to all "weak" substitution errors S2, when lemma of the recognized word-form is right, but ending of the word-form is wrong. In our experiments, the IWER measure with kinf_1=1.0 and kinf_2=0.5 was 29.85%, so in total above 11% of the errors were caused by misrecognized word endings. An automatic lemmatizer from the AOT linguistic software (www.aot.ru) was applied to get lemma for each word-form in the hypotheses. Some other reasons for the recognition errors are out-of-vocabulary words, a few transcription mistakes in tokens of the pronunciation vocabulary, rather high language model perplexity and low *n*-gram hit values, which are caused by high freedom of the Russian grammar.

8. Conclusion

The pronunciation variety is one of the main problems during the development of conversational speech recognition system. The inflective nature of Russian and free word order are additional issues. The developed software complex generates multiple transcription variants that take into account 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.

Another problem of conversational speech recognition is the presence of disfluencies that significantly impair automatic audio signal processing. Without explicit models of nonverbal elements such as hesitations, artefacts, the corresponding audio segments will be recognized as keywords from the system vocabulary. The extension of the vocabulary by lexical models for each type of the hesitations and artefacts allows the speech recognition system to detect these non-verbal elements and avoid false recognition of the keyword units. The performed experiments showed a quite high recognition accuracy of nonverbal units and large vocabulary speech recognition. As further work, we plan to increase the number of modeled nonverbal elements and to carry out experiments to test the correctness of separation of nonverbal units from the key words.

Acknowledgments

This research is supported by the Ministry of Education and Science of Russia (contract No.07.514.11.4139), the Russian Foundation for Basic Research (project No. 12-08-01265, 12-06-31203) and by the grant of the President of Russia (project No. MK 1880.2012.8).

References

Amdal, I. (2002): 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.

Anisimovich, K., Druzhkin, K., Minlos, F., Petrova, M., Selegey, V., Zuev, K. (2012): Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, in Proc. "Dialogue-2012", Moscow, Russia, Vol. 2, pp. 91-103.

Antonova, A., Misyurev, A. (2012): Russian dependency parser SyntAutom at the Dialogue-2012 parser evaluation task, in Proc. Int. Conf. "Dialogue-2012", Moscow, Russia, Vol. 2, pp. 104-118.

Bhanuprasad K., Svenson M (2008): Errgrams - a way to improving ASR for highly inflective Dravidian languages, in Proc. 3rd Int. Joint Conf. on Natural Language Processing IJCNLP'2008, India, pp. 805–810.

Browman, C. P., Goldstein, L. (1992): Articulatory phonology: An overview. Phonetica, 49, 1992, pp. 155-180.

Byrne, B., *et al*. (1997): Pronunciation Modelling for Conversational Speech Recognition, A Status Report from WS97 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, California.

Clarkson, P., Rosenfeld, R. (1997): Statistical language modeling using the CMU-Cambridge toolkit, in Proc. Eurospeech'1997, Rhodes, Greece, pp. 2707–2710.

Gelbukh, A., Sidorov, G. (2001): Zipf and Heaps Laws' Coefficients Depend on Language. In Proc. Int. Conf. on Intelligent Text Processing and Computational Linguistics CICLing-2001, LNCS 2004, Mexico City, pp. 332–335.

Clark, H. H., Fox Tree, J.E. (2002): Using uh and um in spontaneous speaking // Cognition. Vol. 84. pp. 73–111.

Huet, S., Gravier, G., Sebillot, P. (2010): Morpho-syntactic postprocessing of N-best lists for improved French automatic speech recognition, Computer Speech and Language, 24(4), pp. 663–684.

Iomdin, L., Petrochenkov, V., Sizov, V., Tsinman, L. (2012): ETAP parser: state of the art. In Proc. "Dialogue-2012", Moscow, Russia, Vol. 2, pp. 119-131.

Jokisch, O., *et. al.* (2009): Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system. In Proc. SPECOM'2009, St. Peterburg, Russia, pp. 515–520.

Karpov, A., Kipyatkova, I., Ronzhin, A. (2011): Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis, in Proc. Interspeech'2011, Florence, Italy, pp. 3161–3164.

Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A. (2010): Automatic Detection and Removal of Disfluencies from Spontaneous Speech // In Proc. of the Proceedings of the Thirteenth Australasian International Conference on Speech Science and Technology (SST). Melbourne, Australia, pp. 98–101.

Kessens, J. M., Wester, M., Strik, H. (1999): Improving the performance of Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication*, vol. 29, pp. 193-207.

Kipyatkova, I.S., Karpov, A.A. (2008): The module of phonemic transcription for conversational Russian speech recognition system. *Artificial intelligence*, Donetsk, Ukraine, Vol. 4, pp. 747-757 (in Russian).

Kipyatkova, I.S., *et al.* (2012): Analysis of Long-distance Word Dependencies and Pronunciation Variability at Conversational Russian Speech Recognition. In Proc. Federated Conference on Computer Science and Information Systems FedCSIS-2012, Wroclaw, Poland, pp. 719-725.

Kholodenko, A. B. (2002): On Construction of Statistical Language Models for Russian Speech Recognition Systems. J. *Intelligent Systems*. vol. 6 (1-4), Moscow, pp. 381-394.

Lee, A., Kawahara, T. (2009): Recent Development of Open-Source Speech Recognition Engine Julius, in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009), Sapporo, Japan, pp.131–137.

Leontyeva, A., Kagirov, I. (2008): The Module of Morphological and Syntactic Analysis SMART. In Proc. Int. Conf. on Text, Speech and Dialogue TSD'2008, LNAI 5246, Brno, Czech Republic, pp. 373–380.

Levelt, W.J.M. (1983): Monitoring and self-repair in speech // Cognition. Vol. 14. P. 41–104.

Liu, Y., Shriberg, E., Stolcke, A. (2003): Automatic Disfluency Identification in Conversational Speech Multiple Knowledge Sources // In Proc. of the EUROSPEECH 2003. Geneva, Switzerland, pp. 957–960.

Lobanov, B. M., Tsirulnik, L. I. (2007): 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, pp. 57-71 (in Russian).

Masataka, G., Katunobu, I., Satoru, H. (1999): A real-time filled pause detection system for spontaneous speech Recognition // In Proc. of the 6th European Conference on Speech Communication and Technology (Eurospeech '99). Budapest, Hungary, pp. 227–230.

Moore, G. L. (2001): Adaptive Statistical Class-based Language Modelling". PhD thesis, Cambridge University.

Nakatani, C. H., Hirschberg, J. (1994): A corpus-based study of repair cues in spontaneous speech // Journal of the Acoustical Society of America.

№ 95 (3). P. 1603–1616.

Nozhov, I. M. (2003): Realization of automatic syntactic segmentation of a Russian sentence. PhD thesis. 140 p. (in Rus.), http://www.aot.ru/docs/Nozhov/msot.pdf.

Podlesskaya, V.I., Kibrik, A.A. (2004): Speech disfluencies and their reflection in discourse transcription. In Proceedings of VII International Conference. Cognitive Modelling in Linguistics. Varna, Bulgaria, v.1, pp. 194-204.

Eklund, R. (2003): Preambulum**.** Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop // Gothenburg Papers in Theoretical Linguistics 90 / ed. by Robert Eklund.Sweden : Göteborg University, pp. 3–4.

Rabiner, L., Juang, B.-H. (1993): Fundamentals of Speech Recognition. Prentice Hall, 507 p.

Rastrow, A., Dredze, M., Khudanpur, S. (2012): Fast Syntactic Analysis for Statistical Language Modeling via Substructure Sharing and Uptraining. In Proc. 50th Annual Meeting of Association for Computational Linguistics ACL'2012, Jeju, Korea, pp. 175-183.

Saraclar, M. (2000): Pronunciation Modeling for Conversational Speech Recognition. PhD thesis. Baltimore, USA, 2000.

Shriberg, E.E. (1994): Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley, 225 p.

Shvedova, N., et al. (1980): Russian Grammar. Vol. 1, Moscow: Nauka, 783 p. (in Russian).

Skatov, D., Okat'ev, V., Patanova, T., Erekhinskaya, T. (2012): Dictascope Syntax: the Natural Language Syntax Parser, http://dialog-21.ru/digests/dialog2012/materials/pdf/Скатов.pdf.

Snover, M., Dorr, B., Schwartz, R. (2004): A lexically-driven algorithm for disfluency detection // In Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics 2004 (HLT-NAACL-Short '04). Boston, Massachusetts, USA, pp. 157–160.

Sokirko A. (2004): Morphological modules on the website www.aot.ru, in Proc. "Dialogue-2004", Protvino, Russia, 2004, pp. 559–564 (in Rus.).

Starostin, A., Mal'kovskiy, M. (2007): Algorithm of Syntax Analysis Employed by the "Treeton" Morpho-Syntactic Analysis System, in Proc Int. Conf. "Dialogue-2007", Moscow, Russia, pp. 516–524 (in Rus.).

Szarvas, M., Furui, S. (2003): Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR. In Proc. ICASSP'2003, Hong Kong, China, pp. 368–371.

Vaičiūnas, A. (2006): Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition. PhD thesis, Vytautas Magnus University, Kaunas.

Vazhenina, D., Markov, K. (2011): Phoneme Set Selection for Russian Speech Recognition. In Proc. 7th Int. Conf. on NLP and Knowledge Engineering NLP-KE'11, Japan, pp. 475-478.

Whittaker E.W.D. (2000): Statistical language modelling for automatic speech recognition of Russian and English, PhD thesis, Cambridge Univ., 140 p.

Young, S., et al. (2009): The HTK Book (for HTK Version 3.4). Cambridge. UK, 375 p.

Zaliznjak A.A. (2003): Grammatical Dictionary of the Russian Language, 4th Edition. Moscow: Russian dictionaries, 800 p.

Zemskaya, E. A. (Edit) (1973): Conversational Russian speech. Moscow: Nauka, 485 p. (in Russian).