Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 213–228

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Large vocabulary Russian speech recognition using syntactico-statistical language modeling

Alexey Karpov [a], Konstantin Markov [b], Irina Kipyatkova [a,*], Daria Vazhenina [b], Andrey Ronzhin [a]

[a] *St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia*
[b] *Human Interface Laboratory, The University of Aizu, Fukushima, Japan*

Available online 24 July 2013

Abstract

Speech is the most natural way of human communication and in order to achieve convenient and efficient human–computer interaction implementation of state-of-the-art spoken language technology is necessary. Research in this area has been traditionally focused on several main languages, such as English, French, Spanish, Chinese or Japanese, but some other languages, particularly Eastern European languages, have received much less attention. However, recently, research activities on speech technologies for Czech, Polish, Serbo-Croatian, Russian languages have been steadily increasing.

In this paper, we describe our efforts to build an automatic speech recognition (ASR) system for the Russian language with a large vocabulary. Russian is a synthetic and highly inflected language with lots of roots and affixes. This greatly reduces the performance of the ASR systems designed using traditional approaches. In our work, we have taken special attention to the specifics of the Russian language when developing the acoustic, lexical and language models. A special software tool for pronunciation lexicon creation was developed. For the acoustic model, we investigated a combination of knowledge-based and statistical approaches to create several different phoneme sets, the best of which was determined experimentally. For the language model (LM), we introduced a new method that combines syntactical and statistical analysis of the training text data in order to build better *n*-gram models.

Evaluation experiments were performed using two different Russian speech databases and an internally collected text corpus. Among the several phoneme sets we created, the one which achieved the fewest word level recognition errors was the set with 47 phonemes and thus we used it in the following language modeling evaluations. Experiments with 204 thousand words vocabulary ASR were performed to compare the standard statistical *n*-gram LMs and the language models created using our syntactico-statistical method. The results demonstrated that the proposed language modeling approach is capable of reducing the word recognition errors.

Keywords: Automatic speech recognition; Slavic languages; Russian speech; Language modeling; Syntactical analysis

1. Introduction

First automatic speech recognition systems built in early 1950s were capable of recognizing just vowels and consonants. Later their capabilities advanced to syllables and isolated words. Nowadays, ASR systems are able to recognize continuous, speaker-independent spontaneous speech. Nevertheless, there are still a lot of possibilities to increase their accuracy, speed, robustness, vocabulary and usefulness for the end-user. Research in the ASR field have been traditionally focused on several main languages, such as English, French, Spanish, Chinese, and Japanese, while many other languages, particularly African, South Asian, and Eastern European, received much less attention. One of the reasons is probably the lack of resources in terms of appropriate speech and text corpora. Recently, research activities for such under-resourced European languages, including inflective Slavic languages like Czech, Polish,

[*] Corresponding author.
E-mail address: kipyatkova@iias.spb.su (I. Kipyatkova).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.07.004

Serbo-Croatian, Russian, and Ukrainian have been steadily increasing (Karpov et al., 2012).

Russian, Belarusian and Ukrainian are the East Slavic languages of the Balto-Slavic subgroup of the Indo-European family of languages. There are certain common features for all East Slavic languages, such as stress and word-formation. Stress is moveable and can occur on any syllable, there are no strict rules to identify stressed syllable in a word-form (people keep it in mind and use analogies; proper stressing is a big problem for learners of the East Slavic languages). These languages are synthetic and highly inflected with lots of roots and affixes including prefixes, interfixes, suffixes, postfixes and endings; moreover, these sets of affixes are overlapping, in particular one-letter grammatical morphemes set ('a', 'o', 'y'). Nouns and pronouns are identified with certain gender classes (masculine, feminine, common, neuter), which are distinguished by different inflections and trigger in syntactically associated words. Nouns, pronouns, their modifiers, and verbs have different forms for each number class and must be inflected to match the number of the nouns/pronouns to which they refer. Another grammatical category that is characteristic for these languages is case: Russian and Belorussian have 6 cases and Ukrainian has 7 (Ukrainian is the only modern East Slavic language, which preserves the vocative case). Conjugation of verbs is affected by person, number, gender, tense, aspect, mood, and voice. One of the Slavic language features is a high redundancy, e.g. the same gender and number information can be repeated several times in one sentence. Russian is characterized by a high degree of grammatical freedom (but not completely free grammar; random permutation of words in sentences is not allowable); however, the inflectional system takes care of keeping the syntax clear; semantic and pragmatic information is crucial for determining word order.

Russian is not the only language on the territory of the Russian Federation. There are up to 150 other languages (e.g., Tatar, Bashkir, Chechen, Chuvash, Avar, Kabardian, Dargin, Yakut, etc.) spoken by different peoples (Potapova, 2011). In addition, the Russian language has many dialects and accents because of the multi-national culture of the country. There exist essential phonetic and lexical differences in Russian spoken by Caucasian or Ukrainian people caused by the influence of their national languages. Major inner dialects of the standard Russian are North, Central and South Russian in the European part. The North Russian dialect is characterized, for example, by clear pronunciation of unstressed syllables with vowel /o/ (without typical reduction to /ɐ /), the so-called "okanye", some words from the Old Russian are used as well. On the contrary, the South Russian dialect has more distinctions including so called "akanye" (no difference between unstressed vowels /o/ and /a/) and "yakanye" (unstressed /o/, /e/, /a/ after soft consonants are pronounced as /æ/ instead of /i/ as usually) (Smirnova, 2011), voiced velar fricative /γ/ is used instead of the standard /g/ (like in Belarusian and Ukrainian), semivowel /w~u̯/ is often used in the place of /v/ or final /l/, etc. The Central Russian (including the Moscow region) is a mixture of the North and South dialects. It is usually considered that the standard Russian originates from this group.

In order to create a large recognition vocabulary, a rule-based automatic phonetic transcriber is usually used. Transformation rules for orthographic to phonemic text representation are not so complicated for the Russian language. The main problem, however, is to find the position of the stress (accent) in the word-forms. There exist no common rules to determine the stress positions; moreover, compound words may have several stressed vowels.

The International Phonetic Alphabet (IPA) is often used as practical standard phoneme set for many languages. For Russian it includes 55 phonemes: 38 consonants and 17 vowels. In comparison, the IPA set for American English includes 49 phonemes: 24 consonants and 25 vowels and diphthongs. The large number of consonants in Russian is caused by the specific palatalization. All but 8 of the consonants occur in two varieties: plain and palatalized, e.g. hard /b/ and soft /b'/, as in 'собор' (cathedral) and 'бобёр' (beaver). This is caused by the letter following the consonant and appears as secondary articulation by which the body of the tongue is raised toward the hard palate and the alveolar ridge during the articulation of the consonant. Such pairs make the speech recognition task more difficult, because they increase consonant confusability in addition to the fact that they are less stable than vowels and have smaller duration. Russian IPA vowel set includes also phoneme variants with reduced duration and articulation. There are 6 base vowels in Russian phonology (in the SAMPA phonetic alphabet), which are often used as stressed and unstressed pair, for example /a!/ and /a/. All unstressed syllables are subject to vowel reduction in duration and all but /u/ to articulatory vowel reduction tending to be centralized and becoming schwa-like (Padgett and Tabain, 2005). Thus, unstressed /e/ may become more central and closer to unstressed /i/, and unstressed vowel /o/ is always pronounced as /a/ except in the case of a few foreign words such as 'радио' (radio) or 'какао' (cacao).

Although, the underlying speech technology is mostly language-independent, differences between languages with respect to their structure and grammar have substantial effect on the recognition system's performance. The ASR for all East Slavic languages is quite difficult because they are synthetic inflective languages with a complex mechanism of word-formation, which is characterized by a combination of a lexical morpheme (or several lexical morphemes) and one or several grammatical morphemes in one word-form. For large vocabulary continuous speech recognition (LVCSR) of Russian it is necessary to use a lexicon several times larger than for English or French ASR because of the existence of many types of prefixes, suffixes and endings, that in turn decreases both the accuracy and speed of recognition. Baseline grammatical dictionary of Russian (Zaliznjak, 2003) contains more than 150 thousand lemmas. Applying word formation rules to this dictionary leads to a lexicon of over 2 million correct word-forms. For instance, a verb can generate up to two hundred word-forms, which have to be taken into account in the ASR process. Besides, most word-forms of the same lexeme differ only in endings, which when pronounced spontaneously are not as clear as the beginning parts of the words. This often results in misrecognition and reduces the ASR system performance.

Because of the complicated word-formation mechanism and multiple inflection rules, in practice, the size of the vocabulary increases a lot, which results in large number of out-of-vocabulary (OOV) words. In terms of OOV rates, Russian is comparable to some other morphologically rich European languages, such as Finnish, Hungarian, Lithuanian or Turkish (Ircing et al., 2006; Kurimo et al., 2006; Vaiciunas, 2006). But compared to some analytical languages like English, the OOV percentage can be up to 10 times higher. As can be seen from Table 1, using the same size text corpora of about 100 million words, OOV rate for the Russian 400 K vocabulary LM is still higher than for the English 65 K one.

Word order in Russian sentences is not restricted by hard grammatical constructions, like in modern English or German. While the English sentence has a strict structure: subject–verb–object, in Russian, words can change their place without significantly influencing the sentence meaning. This complicates the creation of statistical LMs and substantially decreases their predictive power. *N*-gram LMs for Russian are by orders of magnitude larger than English ones. As shown in Table 1, changing the vocabulary size from 100 K to 400 K words increases the English model perplexity by 5.8% relatively, while the Russian model perplexity increases by as much as 39.5%. This suggests that more sophisticated pronunciation and language models are needed to reach the performance level of the English ASR.

In this paper, we propose several approaches to adapt the general speech recognition process for the Russian language. The pronunciation and acoustic modeling are improved by including Russian language specifics into the pronunciation vocabulary and phoneme set creation processes. We designed several phoneme sets based on the analysis of the phoneme confusion matrix in combination with phonological knowledge. In addition, to some extent, the flexibility of the word order in Russian is taken into account by combining syntactic and statistical information in the LM training. Syntactical analysis of the training data, which takes into account the long-distance grammatical relations between words, is performed and new word bi-grams are created and pooled together with the regular bi-grams obtained the usual way.

This paper is organized as follows: Section 2 reviews the state-of-the-art technologies focused mainly on the Slavic languages. Section 3 presents our acoustic modeling approach. In Section 4, the pronunciation modeling is briefly described. Section 5 explains our language modeling method and the creation of the syntactico-statistical LM for the Russian language. Section 6 presents the experimental setup, obtained results and discussion. Conclusions and directions for further research are given in Section 7.

2. Related work

2.1. Language modeling using syntactic analysis

One of the most efficient natural LMs is the statistical word *n*-gram model aimed to estimate the probability of any word sequence $W = (w_1, w_2, \ldots. w_m)$ The *n*-gram is a sequence of n elements (for example, words), and the *n*-gram LM is used for prediction of an element in a sequence containing $n-1$ predecessors (Bellegarda, 2004; Moore, 2001). Stochastic LMs based on purely statistical analysis of some training text data are efficient for many languages with rather strict grammatical structure, but for languages, which have more freedom in the sentence formation (like Russian and most of the Slavic languages), such models are less efficient.

One way to account for the long-span word dependencies is to use syntactical text analysis. In recent ASR systems, it is sometimes embedded into various processing levels: language modeling, on-line speech decoding, N-best list re-scoring, post-processing of the ASR output for spoken language understanding tasks, etc.

In (Szarvas and Furui, 2003), a stochastic morpho-syntactical LM for Hungarian ASR is introduced. This model describes the valid word-forms (morpheme combinations) of the language. The stochastic morpho-syntactic LM decreased the morpheme error rate by 17.9% compared to the baseline tri-gram system. The morpheme error rate of the best configuration was 14.75% in a 1350 morpheme dictation task. For Turkish, it was proposed to incorporate some syntactic information (POS, grammatical features, head-to-head dependency relations) into discriminative morph-based LM (Arisoy et al., 2010). It is a feature-based

Table 1
Perplexity and OOV rates comparison for Russian and English tri-gram LMs (Whittaker, 2000).

Vocabulary size (# of words)	Russian		English	
	Perplexity	OOV, %	Perplexity	OOV, %
65 K	413	7.6	216	1.1
100 K	481	5.3	224	0.65
200 K	587	2.6	232	0.31
400 K	671	1.2	237	0.17

approach that re-ranks the ASR output with discriminatively trained feature parameters. Sub-lexical units are first utilized as LM units in the baseline recognizer; then, sub-lexical features were used to re-rank the sub-lexical hypotheses. In order to obtain those features all the words in the N-best lists were analyzed with a morphological analyzer. POS tags were utilized in order to obtain class-based generalizations that can capture well-formed tendencies. Head-to-head dependency relations were used since presence of a word or morpheme may depend on the presence of another word or morpheme in the same sentence and this information is represented in the dependency relations. The usage of morpho-syntactical features resulted in additional 0.4% word error rate (WER) reduction over the sub-lexical *n*-gram features.

Syntactically enhanced latent semantic analysis (SELSA) has been investigated in Kanejiya et al. (2003). It integrates both the semantic and syntactic information (part-of-speech – POS). A model, which characterizes a word's behavior across various syntactic and semantic contexts, is proposed. This model assumes that a word with different preceding syntax occurs in different semantic contexts. The preliminary results on WSJ corpus showed that SELSA decreases the bi-gram perplexity by 32%.

Recently, syntactic *n*-grams (sn-grams) were proposed (Sidorov et al., 2012). In this work, the neighbors are taken by following the syntactic relations in syntactic trees, and not by their position in the text. In the paper, sn-grams were applied for authorship attribution only, but it is suggested to apply sn-grams for speech recognition and machine translation tasks as well.

Syntactic parsing for N-best list rescoring is studied in Kuo et al. (2009). The syntactic features include exposed head words and their non-terminal labels both before and after the predicted word. Artificial neural networks are used for the language modeling. Experiments performed on Arabic broadcast news and conversational speech show 5.5% WER improvement compared to the baseline system (WER was 6.2–11.5% depending on the test corpus).

On the other hand, in Roark (2002), N-best list rescoring is performed by an incremental statistical parsing algorithm which uses a Markov assumption. Results presented in the paper show that a low order Markov assumption leads to the same accuracy as parsing with no Markov assumption, but with large efficiency gain in terms of computational complexity.

In (Chelba and Jelinek, 2000), a structured LM based on syntactic structure analysis capable of extracting meaningful information from the word history is developed. In this model, syntactical analysis is applied during speech decoding for partial parsing (and tagging) of the recognition output. Syntactical tree is constructed dynamically as the recognition progresses. The model enables the use of long-distance dependencies between words and allows predicting next word based not only on the previous few lexical tokens, but also on the exposed head words. An improvement over the standard tri-gram modeling for English was reported. This model is suitable for analytical languages with a strict grammatical structure such as English. However, for languages with more free grammar, only full utterance parsing is reasonable, not just part of the recognition output, because grammatically connected words can be located in the sentence quite far from each other. Syntactical analysis is also applied for post-processing of the recognition hypotheses in Rastrow et al. (2012), where a sub-structure sharing, which saves duplicate work in processing sets with redundant hypothesis structures, is proposed. The syntactic discriminative LM was trained using dependency parser and part-of-speech (POS) tagger, which results in significant decoding speedup and WER reduction. Also, in Huet et al. (2010), a syntactic analyzer was used for N-best list rescoring and selection of the best hypothesis. Morphologic and syntactic post-processing of the N-best lists in French ASR system is applied according to grammatical correctness criteria. This method relies on POS and morphological information; the N-best list is automatically tagged and each hypothesized word is referred to its morpho-syntactic class. Then morpho-syntactic scores are computed and combined with acoustic and language scores. New score including morpho-syntactical one is used to re-order the N-best lists.

A joint decoding approach to speech recognition with syntactic and semantic tagging is proposed in Deoras et al. (2012). In this work, two types of discriminative models sure as the Maximum Entropy model and Conditional Random Fields are used. In (Bechet and Nasr, 2009), a syntactic parser for spontaneous speech recognition outputs is used for identification of verbal sub-categorization frames for dialogue systems and spoken language understanding tasks. The processing is performed in two steps: in the first step, generic syntactic resources are used to parse the output of an ASR system represented as a lattice of words; the second step is re-ranking of all possible dependency analyses produced by the parser for each word in the lattice. In (Oparin et al., 2008), some morphological features (part-of-speech, word stem, morphological tag, inflection), which are used during syntactical analysis, were also applied for construction of a morphological decision tree for modeling inflectional languages (Czech).

In our work, we use a syntactic text analysis in order to detect grammatically connected word-pairs in the training data. The syntactic parser allows us to find long-span dependent words, which did not appear as regular bi-grams in the training corpus because of data sparseness. The inclusion of such new potential bi-grams into an *n*-gram LM results in increased *n*-gram coverage with respect to the same training data and hence leads to WER reduction. However, since the structure of the *n*-gram LM stays the same a standard speech decoder can be used without any modifications.

At present, there exist several syntactic analyzers for Russian: Treeton (Starostin and Mal'kovskiy, 2007), analyzer of the 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 analyzer (Sokirko, 2004). The latter one is used in this work because it has several advantages: it is open source and its databases permit modifications; it provides relatively high speed of text data processing, and uses constantly updated grammatical database based on the baseline grammatical dictionary (Zaliznjak, 2003).

2.2. Existing Russian speech recognition systems

Many small vocabulary ASR systems for Russian have been developed for voice command, incoming phone calls routing, and other similar applications (Vazhenina et al., 2012). For instance, one of the first ASR systems was developed in 1960s by T. Vintsyuk, who is considered as one of the Russian speech recognition pioneers (Vintsyuk, 1968). He proposed the use of dynamic programming methods for time alignment of speech utterances.

However, there are only a few systems for large vocabulary tasks. In order to dictate and recognize spoken Russian (and especially spontaneous speech) one has to utilize recognition vocabulary of several hundred thousand words (>100 K word-forms). Until recently such vocabularies were considered as very large (Whittaker and Woodland, 2001).

Researchers at IBM have developed one of the first Russian LVCSR systems (Kanevsky et al., 1996). The training and testing were carried out on the TASS news texts. Vocabulary size of the system was 36 thousand (36 K) words. About 30 K phrases pronounced by 40 speakers were used for acoustic model training as with 47 phoneme lexicon. A tri-gram LM was trained from texts consisting of 40 million (40 M) words. The WER was about 5%, but just for short read sentences. In this work, LM with word segmentation into stem and ending was also investigated.

Detailed comparison of LMs for the English and Russian is presented in Whittaker (2000). It has been shown that a 430 K vocabulary is needed for Russian to achieve the same vocabulary coverage as a 65 K English vocabulary. Two types of class-based LMs (two-sided and one-sided) as well as particle-based LM were considered for the modeling of both languages. In the two-sided symmetric class model, the same word classification function is used to map both the current and the predecessor words. In the one-sided model, mapping to a class is used for the predecessor words only, and the current word is not mapped. It was shown that combined word, one-sided class and particle-based models provide 19% improvement of the perplexity for Russian.

In (Viktorov et al., 2009), a Russian ASR system for broadcast news recognition is described. A text corpus consisting of 129 M words for LM training was collected from the Internet. Three frequency vocabularies were created: general frequency vocabulary, frequency vocabulary of proper names, and frequency vocabulary of common names. A vocabulary of 213 K words was selected to cover 98% of the training text data, but the coverage of the test data was not reported. Standard statistical *n*-grams ($n = 1 \div 3$) were used as LMs. The amount of the speech data for the acoustic model training exceeded 200 h; 3280 speakers have taken part in the corpus collection. Recognition accuracy was 60–70% depending on the sound files quality.

Another large vocabulary Russian speech recognition system is presented in Oparin and Talanov (2005). In this work, a model based on separating words into stems and endings is used. The resulting vocabulary consists of 85 K stems and 4 K endings that cover 1300 K word-forms; however, the morpheme-based LM and lexicon did not bring any reduction of the WER. In (Pylypenko, 2007), a two-pass algorithm for Extra Large Vocabulary Continuous Speech recognition based on Information Retrieval (ELVIRCOS) was studied for Ukrainian and Russian. This algorithm decomposes the recognition process into two passes, where the first pass builds a word subset for the second pass recognition.

Since Russian is a language with an alphabetic script and close grapheme-to-phoneme relation, another suitable approach is to use grapheme-based acoustic modeling (Stüker and Schultz, 2004). Graphemes have the advantage over phonemes that they make the creation of the pronunciation vocabulary easier. However, it complicates the creation of the acoustic model, because graphemes are generally less related to pronunciation, than phonemes. To solve this problem, enhanced tree clustering approach is used, which allows flexible parameter sharing across the acoustic units. The system is evaluated using the GlobalPhone corpus. For training, 17 h of speech are used and two sets of 1.6 h and 1.3 h are selected for testing. The system achieves WER of 32.8% using trigraphemes with enhanced tree clustering. In comparison, WER of 33% was obtained for triphone recognizer and 36.4% for trigrapheme recognizer with CART (classification and regression tree) clustering. It was hypothesized that the rich morphology of the Russian language is one major source of errors. Later, this grapheme-based method was used for multilingual speech recognition system (Stüker, 2008). In this work, multilingual ASR for 4 languages (English, German, Spanish and Russian) was implemented. Graphemes that are common to one language share the same model and are treated as identical in the other systems. All information about which language a grapheme belongs to, is discarded in the system and the data from all languages for this grapheme are used for training. The WER for Russian was 41.9% on the GlobalPhone evaluation test data.

Most recently, a large vocabulary continuous speech recognizer that uses syllable-based LM was described in Zablotskiy et al. (2012). A method for concatenation of the recognized syllables and error correction is proposed. The syllable lexicon has about 12 K entries. The final sentence is constructed from the recognized syllables by the designed co-evolutionary asymptotic probabilistic genetic algorithm (CAPGA).

Another recent work (Lamel et al., 2012) studies the conversational telephone Russian speech transcription task. For the system training about 8 h of conversational Russian speech were used, and about 1 h was used for development. The pronunciation lexicon was created using grapheme-to-phoneme conversion rules. Texts for the LM training were taken from the Web, and transcriptions of broadcast and conversational telephone speech data. The total size of the training text corpus was 297 M words. Vocabulary of the system contained up to 500 K words and the WER was 50–60% (for real conversational speech data) depending on the used acoustic models and the type of LM training corpus. Another collaborative work (Lamel et al., 2011) of the same authors summarizes the experimental results on speech transcription for 9 European languages (including Russian). Those results were obtained by ASR systems used for the Quaero 2010 and 2011 evaluations campaigns (test data contained various broadcast data including news and conversations). The WER of Russian ASR was 19.2% in 2010 and 18.3% in Quaero 2011 (among 9 European languages only Portuguese ASR system had higher WER).

Finally, for automatic voice search in the Internet, Google Inc. has developed the on-line Voice Search service (Schalkwyk et al., 2010), which uses speech recognition technology. This service allows users to find necessary information in the Internet pronouncing a word or a phrase. For the LM creation, written queries to Google search engine were used. This technology is also applied to other Google services, for example, Google maps, where it is possible to do voice request for searching a place on the map. For short and common sentences it works pretty well, but it fails for conversational Russian speech.

Any modern ASR system requires at least three types of models in order to recognize speech: acoustic, pronunciation and language model. These models for our LVCSR of spoken Russian are presented in the following sections.

3. Acoustic modeling

The first step, we are faced with in developing any speech recognition system, is choosing appropriate units for acoustic modeling. Phoneme set size determines the number of context-independent models and also influences the number of context-dependent models and the amount of data needed for training. If it is too large, the complexity of the phoneme hypotheses lattice will increase significantly, making the decoding process computationally more expensive. If it is too small, recognition performance may degrade because of low acoustic space resolution.

3.1. Development of Russian phonetic decision tree

Depending on the task and the availability of training data, an inventory of sub-word units has to be defined, which could consist of word particles, syllables, phonemes, or phonemes in context. The advantage of introducing context for phonemes is its potential to model the effects of co-articulation between adjacent speech units. For this reason, in the modern speech recognition systems, context-independent (CI) models, also called monophones, are most often replaced with context-dependent (CD) models (triphones). If two phones have the same identity, but different left and/or right contexts, they are considered as different triphones.

Introducing uncontrolled context-dependency causes training data sparsity, because for each model we need enough observations of each phone type to train and not all possible triphones can be seen in the training data. The complexity of the phoneme hypotheses lattice will increase significantly, making the decoding process computationally more expensive. The most common way to reduce data sparsity is by clustering some of the contexts and by tying together states whose context falls in the same cluster.

Contexts are usually clustered using phonetic decision trees (Young et al., 1994; Young et al., 2009). Questions used in such trees ask whether the context phoneme to the left or right has a certain phonetic feature (e.g. whether the left/right context phoneme is voiced). Using multiple questions set allows us to increase context dependency degree. We used English phonetic decision tree (Odell, 1995) as basis for the development of a new Russian phonetic decision tree. In addition to the general questions like place of articulation during pronunciation, specifics of the Russian phonology were also included in the question set such as whether the left/right context phoneme is soft (palatalized), or whether the left/right context phoneme is stressed. Some questions specific to English were removed from the set. Table 2 summarizes the main differences between the English and Russian phonetic decision tree question sets.

3.2. Best phoneme set selection

Phoneme set size determines the number of context-independent models, and also influences the number of context-dependent models. There are two main approaches to determine phoneme set for acoustic modeling: knowledge-based and statistical.

Table 2
Differences between the English and Russian phonetic decision tree question sets.

Removed questions	New questions
Is left/right vowel long?	Is left/right vowel stressed?
Is left/right vowel short?	Is left/right vowel unstressed?
Is left/right vowel diphthong?	
Is left/right vowel reduced?	
Is left/right consonant syllabic?	Is left/right consonant soft?
Is left/right consonant continuant?	Is left/right consonant hard?
	Is left/right consonant trill?

For the Russian language, knowledge-based phoneme sets (alphabets) are often used. They are manually designed by human experts according to some linguistic and phonological rules (Cubberley, 2002). The rules for transformation from orthographic text to phonemic representation are not very complicated for Russian. In (Psutka et al., 2005), 43 phoneme set was used, which consists of the standard Russian SAMPA phoneme set plus additional consonant /ɣ/ because of the data specificity. On the other hand, direct spelling conversion produces 49 phoneme set, which was developed for comparison with grapheme recognizer introduced in Stüker and Schultz (2004). For Russian LVCSR, a 59 phoneme set was proposed in Tatarnikova et al. (2006), but no results were reported. In most cases, researchers use extended sets of vowels including their stressed and unstressed variants (Stüker and Schultz, 2004; Tatarnikova et al., 2006; Ronzhin and Karpov, 2004).

For other languages, however, there are studies, where statistical information is utilized for the phoneme set derivation, such as data-driven sub-word unit sets selection (Singh et al., 2002) or usage of the mutual information between the word tokens and their phoneme transcriptions in the training text (Zhang et al., 2008).

In our work, we combine the two conventional approaches described above: knowledge-based and statistical. Information from phonological knowledge and statistics from the phoneme confusion matrix is combined in our method. Phonological knowledge includes pronunciation rules and phonological alternation information. This allows us to determine acoustically close phonemes. Confusion matrix allows us to determine most frequently mismatched phonemes. Combining these information sources, we can make a decision how to select the phonemes in the set. First, we begin with largest phoneme set available and then gradually reduce its size by deleting or merging some phonemes.

Phoneme selection workflow, shown on Fig. 1, includes the following steps:

(1) First, from the IPA set, we select phoneme set P0 applying phonological pronunciation rules.
(2) For further merging, we define phoneme pair candidates according to the language phonology specifics. Those include both soft/hard consonant and stressed/unstressed vowel pairs.
(3) Using P0 set, we perform phoneme recognition and obtain phoneme confusion matrix. For the selected pairs, we calculate the confusion rate (CR), which is defined as follows:

$$CR = \frac{M_1 + M_2}{H_1 + M_1 + H_2 + M_2} \times 100\% \qquad (1)$$

where H_1 is the number of correctly recognized occurrences of the first phoneme in the pair, e.g. /a/ recognized as /a/, H_2 is the number of correctly recognized occurrences of the second phoneme in the pair (e.g., /a!/ recognized as /a!/), M_1 is the number of misrecognized occurrences of the first phoneme in the pair, e.g. /a/ recognized as /a!/, and M_2 the is number of misrecognized occurrences of the second phoneme in the pair, e.g. /a!/ recognized as /a/. The higher confusion rate, the more phonemes are mismatched within the pair, which makes it a likely candidate for merging.
(4) Phoneme pairs are sorted by decreasing confusion rates.
(5) Finally, we select top N phoneme pairs and by merging them obtain a new phoneme set.

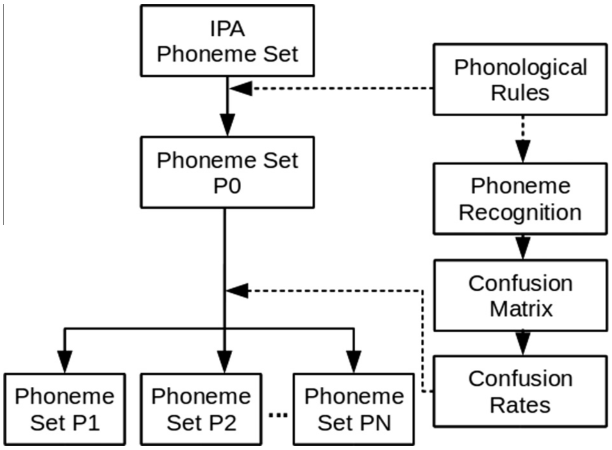

Fig. 1. Flowchart of the phoneme selection method for Russian ASR.

Different choices of N produce different phoneme sets. The best phoneme set can be found by evaluating its speech recognition performance in terms of recognition accuracy and speed. The speech recognition results and discussion on the best phoneme set selection are reported in Section 6.2.

4. Pronunciation modeling

One important challenge in the development of spoken Russian ASR is the grapheme-to-phoneme conversion (or orthographic-to-phonemic transcription) of all the words in the recognition vocabulary. The aim of the grapheme-to-phoneme conversion is to automatically generate pronunciation lexicon from the vocabulary represented in the orthographic form. There are some problems at this step: grapheme-to-phoneme mapping is not one-to-one: stress position in word-forms is floating, substitution of grapheme ‘ё’ (always stressed) by ‘е’ in the most of printed and electronic text data, phoneme reductions and assimilations in continuous and spontaneous speech, and a lot of homographs.

The main common feature of all the East Slavic languages is the Cyrillic alphabet, which is shared by Russian, Ukrainian, Belarusian and some other Slavic languages (Serbian, Bulgarian and Macedonian) as well as some non-Slavic languages of the former Soviet Union (Kazakh, Kyrgyz, Tajik, etc.).

In the grapheme-to-phoneme conversion for Russian, the following positional changes of speech sounds are possible (Shvedova et al., 1980): (1) changes of vowels in pre-stressed syllables; (2) changes of vowels in post-stressed syllables; (3) positional changes of consonants:

- At the end of a word or before 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/and /z/ become palatalized, as well as before /s'/ and /z'/, the consonants /s/and /z/ disappear (merged into one phoneme).
- Before the palatalized dentals /t'/, /d'/, /s'/, /z'/ or / ch/, /sch/ the hard consonant /n/ becomes palatalized.
- Before /ch/ the consonant /t/ (both for the graphemes 'т' and 'д') disappears.
- Before /sh/ or /zh/ the dental consonants /s/ and /z/ disappear (merged).
- Two identical consonants following each other are merged into one.
- Some frequent combinations of consonants are changed: /l n c/ → /n c/, /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/ → /c t/, /t s/ → /c/, /h g/ → /g/, /s sch / → /sch/, etc.

The method for automatic grapheme-to-phoneme conversion operates in two cycles, consisting of the following steps (Kipyatkova et al., 2012):

(1) Stress positions are identified using the morphological database.
(2) Hard consonants before graphemes 'и', 'е', 'ё', 'ю', 'я' become palatalized (if possible) and these graphemes are converted into phonemes /i/, /j e/, /j o!/, /j u/, /j a/ in the case, when they are located at the beginning of a word or after a vowel, otherwise they are transformed into /i/, /e/, /o!/, /u/, and /a/ correspondingly.
(3) A consonant before grapheme 'ь' gets palatalization (getting soft) and this grapheme is deleted (it has no corresponding phoneme).
(4) Transcription rules for positional changes of consonants (Karpov et al., 2012) are applied.
(5) Transcription rules for positional changes of vowels in pre-stressed and post-stressed syllables (Karpov et al., 2012) are applied.
(6) Steps (4)–(6) are repeated once more, because some changes may lead to other changes in the preceding phonemes.
(7) Grapheme 'ъ' is deleted (it has no corresponding phoneme; it only shows that the preceding consonant is hard).

For the grapheme-to-phoneme conversion, we apply an extended morphological database, consisting of more than 2.3 M word-forms with the symbol '!' indicating stressed vowels. This database is a fusion of two different morphological databases: AOT[1] and Starling.[2] The former one is larger and has more than 2 M entries, but the latter one contains information about secondary stress for many compound words, as well as words with grapheme 'ё', which is always stressed, but usually replaced with 'е' in official texts that leads to loosing required information on the stress position.

5. Language modeling

5.1. Collection and analysis of the training text corpus

At present, there are several large commercial text corpora of Russian, for instance, the Russian National Corpus[3] and the Corpus of Standard Written Russian,[4] 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 LM 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 four on-line Russian newspapers for the years 2006–2011.[5] The database contains text data that reflect contemporary Russian including some spoken language.

Automatic pre-processing of the collected text data is carried out as follows. At first, texts are divided into sentences. Then, text written in any brackets is deleted, and sentences consisting of less than five words are excluded. Further, punctuation marks are deleted and the symbols "№" and "#" are replaced by the word 'номер' (number). All numbers and digits are combined in a single class that is denoted by the symbol "№" in the resulting text. A group of digits, which can be divided by point, comma, space or dash sign is considered as a single number. The symbol "№" also denotes Roman numbers, which are a combination of the Latin letters I, V, X, L, C, D, and M, which can be divided by space or dash. Internet links and E-mails are replaced by the symbols "⟨⟩" and "⟨@⟩" respectively. If a word begins with an uppercase letter, it is changed to lowercase. When the whole word is written with uppercase letters, such change is made only if the word exists in the vocabulary.

The size of the corpus after the text normalization and the deletion of double and short (less than 5 words) sentences is over 110 M words, and it has about 937 K unique word-forms. The frequency distribution of these word-forms is shown in Fig. 2. This plot shows that there are a lot of rare words in the corpus. More than 350 K words occurred only once in the training corpus; these words are mainly personal names and some misprints as well.

[1] http://www.aot.ru.

[2] http://starling.rinet.ru

[3] http://ruscorpora.ru.

[4] http://www.narusco.ru.

[5] www.ng.ru, www.smi.ru, www.lenta.ru, www.gazeta.ru.

Based on this text corpus, stochastic LMs were created using both statistical and syntactic automatic analyzers.

5.2. Language modeling with syntactic analysis

Our syntactico-statistical language modeling approach takes advantage of both the statistic and syntactic text analyses. Fig. 3 illustrates the process of stochastic LM creation for Russian using some syntactic analysis elements. The training text corpus is processed in parallel discovering regular *n*-grams and syntactic word dependencies in sentences. Then, results of both analyzers are processed to obtain count files and calculate two stochastic LMs, at the final step these LMs are linearly interpolated to create an LM of the required order. So, our method allows building integral *n*-gram language model as the result of joint statistical and syntactical text analysis at the training stage without the need to do syntactical parsing during decoding. Such simultaneous statistical and syntactical processing may discover more bi-grams from the same training data (that is important in the case of data sparseness for under-resourced languages and languages with nearly free word ordering like Russian) and produce an enhanced LM with improved coverage.

Both analyzers complement each other very well: the syntactic one is used to find long-distance grammatical dependencies between words (potential bi-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 SRI Language Modeling Toolkit (SRILM) Stolcke et al., 2011, while the software "VisualSynan" from the AOT project (Sokirko, 2004) is used for the syntactic analysis. The latter parses the input sentences and produces a graph of syntactical dependencies between the pairs of lexical units.

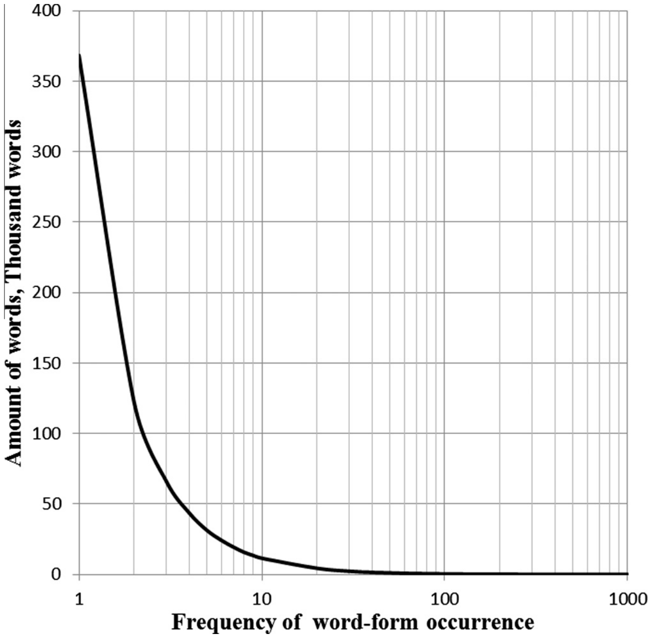

Fig. 2. Word frequency distribution for the text corpus.

The main goal of the syntactical analysis is to extract syntactical groups in a sentence (Nozhov, 2003; Sokirko, 2004). Syntactical groups are defined as the following sequence: a group type, a pair of syntactically connected words, and grammemes. Group type is a string constant, for example: "ПРИЛ_СУЩ" (adjective-noun), "ПГ" (prepositional phrase), etc. Group grammemes are the morphological characteristics of words, which determine behavior and compatibility of the 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 (more than one word) relations between pairs of words. The types of syntactic groups that are selected during the syntactic analysis are presented in Table 3.

Moreover, words of the syntactic groups (2), (3), (7), (8), (9), and group (1), but without subordinate attributive clauses starting with words 'which', 'who', etc., are commutative in Russian and each such syntactic dependence produces two bi-grams with direct and inverse word order. Fig. 4 shows an example of the syntactic analysis of one Russian phrase ('In the very expensive show, military and civilian aircrafts, which arrived yesterday and today in the airport of our little town, are involved') from the training corpus. It demonstrates some types of long-distance dependences, whereas all the adjacent word pairs are modeled by statistical bi-grams. Commutative groups are denoted by the dark red double-sided arrows. Thus, syntactic parsing of this sentence produces 13 long-distance word pairs additionally to the statistic processing, which gives 18 bi-grams. N-gram likelihoods are calculated after merging the results (the counts) of both analyzers based on their frequency in the training data.

After the normalization of the text corpus, statistical processing is carried out and a list of bi-grams with their frequency of occurrence is created. Then, syntactic analysis of the text corpus is performed and grammatically connected pairs of words (syntactic groups), which were separated in the text by other words, are detected. Then the list of bi-grams obtained by the statistical analysis and the list of grammatically connected pairs of words, which were extracted during the syntactical analysis, are merged (bi-gram counts are added). Finally, this model is linearly interpolated with the conventional statistical *n*-gram LM serving as a baseline model. The optimal interpolation weight in syntactico-statistical LM was 0.27 for syntactically derived bi-grams.

During LMs creation we used the Kneser–Ney discounting method, and did not apply any *n*-gram cutoff. After the statistical analysis of the collected text corpus we obtained 19.1 M bi-grams. As a result of the syntactical analysis we added more than 2.7 M new bi-grams. Thus, the total number of bi-grams in the extended LM is 21.8 M, and consequently the size of the syntactico-statistical LM increased by 14% compared with the statistical model. The inclusion

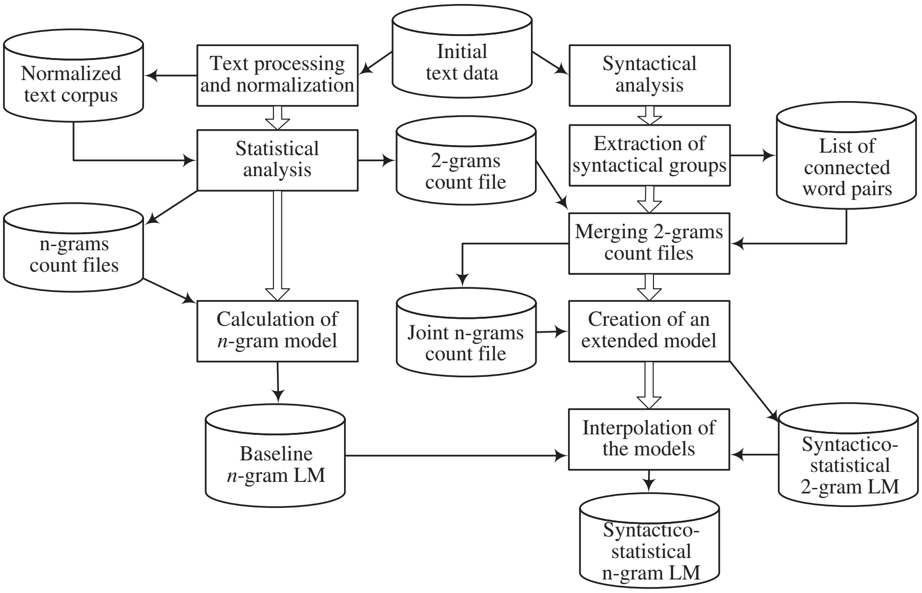
Normalized text corpus
Text processing and normalization
Initial text data
Syntactical analysis
Statistical analysis
2-grams count file
Extraction of syntactical groups
List of connected word pairs
n-grams count files
Merging 2-grams count files
Calculation of n-gram model
Joint n-grams count file
Creation of an extended model
Baseline n-gram LM
Interpolation of the models
Syntactico-statistical 2-gram LM
Syntactico-statistical n-gram LM

Fig. 3. Flowchart of the syntactico-statistical n-gram LM creation for Russian.

Table 3
Types of extracted syntactic groups.

#	Syntactic group	An example in Russian	An example in English
1	Subject–predicate	мы её не знали	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 issues
5	Genitive pair	темой текущего и следующего номера	A topic of the present and next happens
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

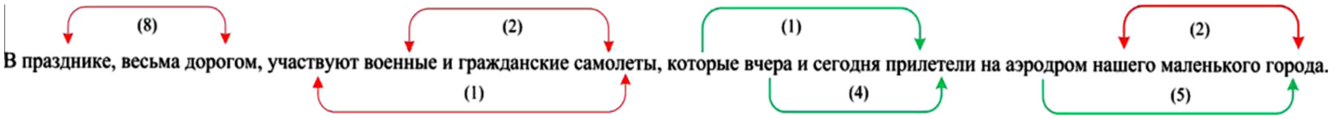
(8)
(2)
(1)
(2)
В празднике, весьма дорогом, участвуют военные и гражданские самолеты, которые вчера и сегодня прилетели на аэродром нашего маленького города.
(1)
(4)
(5)

Fig. 4. An example of syntactic phrase analysis (long-span syntactic dependencies are shown by arrows).

of such new bi-grams into the *n*-gram LMs results in increased *n*-gram coverage with respect to the same training data and consequently, in WER reduction. Also, in the Russian language, each word-form itself yields unique grammatical information (POS, gender, tense, etc.) and there is no necessity to additionally incorporate any syntactic tags into LM.

In order to compare LMs we created, their perplexity and *n*-gram hit rates (summarized in Table 4) were calculated using text data consisting of phrases (33 M words in total) from another online newspaper 'Фонтанка.ru'.[6] Perplexity is given with two different normalizations: counting all input tokens (PPL) and excluding end-of-sentence tags (PPL1). In Table 4, we present the *n*-gram hit rates in the form of SRILM output, for 3-gram LMs it shows both 3-gram hit rate and 2-gram hit rate independently. It means that 3-gram hit rate is calculated at first, and then 2-gram hit rate is estimated for text data not covered by 3-grams.

We have tried several vocabulary sizes and the best one in terms of performance/OOV trade-off has 204 K words; its OOV rate for the given test set is 4.1%. These parameters have quite large values for Russian LMs, which is a great challenge for the LVCSR.

6. ASR experimental results and discussion

6.1. Russian speech databases

There are several Russian speech corpora collected in quiet office environment without any background noise: Global-

[6] www.fontanka.ru.

Phone (Schultz and Waibel, 1998), STEL (Shirokova, 2007), CORPRES (Skrelin et al., 2010), SpeechOcean Russian corpora and others. There are also some telephone speech corpora: ISABASE (Arlazarov et al., 2004), RuSpeech (Kouznetsov et al., 1999), SpeechDat(E), TeCoRus, LDC RuSTeN Russian speech corpora, etc. Other speech corpora include various background environments, which represent typical surroundings such as: home, car, public place, etc.

In our speech recognition experiments, we combined the speech data from two databases. The first one is the GlobalPhone Russian speech corpus, which was developed in the framework of the multilingual GlobalPhone project (Schultz and Waibel, 1998). The speech data were collected in ordinary, but quiet rooms. The data is recorded with 16 kHz sampling rate, 16-bit audio quality. The surrounding noise level ranges from quiet to loud. The database contains records of Russian newspaper articles (12,321 meaningful utterances) read by 115 speakers (61 men and 54 women). The number of utterances per speaker varies from 19 to 228. In total, the recordings' duration is about 26 h. All utterances with disfluencies and loud noise level are excluded. The rest of the recordings are split into training set (15 h 25 min) and test set (1 h 40 min), which contains recordings from 5 male and 5 female speakers.

The second database is the SPIIRAS Russian speech database, which was developed in the framework of the EuroNounce project (Jokisch et al., 2009). The speech data were collected in clean acoustic conditions, with 44.1 kHz sampling rate, 16-bit audio quality. The signal-to-noise ratio (SNR) is about 35 dB. For the recognition experiments, all recordings were down-sampled to 16 kHz. The database consists of 16,350 utterances pronounced by 50 native Russian speakers (25 men and 25 women). Each speaker pronounced the same set of 327 phonetically balanced and meaningful phrases and texts. The total duration of the SPIIRAS speech corpus is more than 21 h. Recordings from 5 male and 5 female speakers are selected for testing and the rest are used for acoustic model training. Additionally, 30 min of continuous Russian speech were recorded from 1 male and 1 female speaker for the task of Russian LVCSR system evaluation; these data contain meaningful phrases of 6–20 words each.

6.2. Best phoneme set selection experiments

6.2.1. Experimental setup

We used the HTK toolkit (Young et al., 2009) for our acoustic models training. The acoustic signal was coded with energy and 12 MFCCs (Mel Frequency Cepstral Coefficients), calculated from 26-channel filter bank analysis of 20 ms long frames with 10 ms overlap, and their first and second order derivatives, which results in a 39-dimension feature vector. Phone models were represented by 3 state HMMs with left-to-right topology except the silence model, which also has transition from the third to the first state. Probability distribution function in each state was modeled with 16 component Gaussian mixture. Triphones were clustered by phonetic decision tree state tying using the question set described earlier in Section 3.1.

For each phoneme set we created, we built one monophone based phoneme recognition system and three different triphone based word recognition systems, which share the same acoustic model, but have different vocabulary size and LM. In all cases, the acoustic model was trained using both the GlobalPhone and SPIIRAS training data.

The first system uses closed form bi-gram LM trained on all the SPIIRAS speech transcriptions. The vocabulary size is 1,146 words and the text perplexity is 241. The SPIIRAS test data are used for this system evaluation. Next system is similar, but uses the GlobalPhone transcripts for the bi-gram LM training. The vocabulary consists of about 20 thousand words and the perplexity of the text data is 137. The test set for this system is the GlobalPhone test set. The third system uses the baseline bi-gram LM with vocabulary size of 204 K words trained for the LVCSR experiments. This system was also tested with GlobalPhone test set, for which the LM has perplexity of 844 and the OOV rate is 3.41%.

6.2.2. Phoneme sets creation process

Based on the phonological specifics of the IPA phonemes, set P0 (Table 5) was obtained by the following steps (Vazhenina and Markov, 2011):

- Stressed /a/, /æ/ and /ɑ/ were merged because they are very close acoustically and differ only depending on place within the word.
- Unstressed /ɐ / and /ə/ were merged because their pronunciations differ just slightly depending on the distance from the stressed syllable. The same is the reason for merging /ʉ/ and /ʊ/.
- /ɵ/ is very similar to the combination of sounds /j/ and /o/ and was split accordingly.
- Consonants /ʑ/ and /ɣ/ were excluded, because they are used only in some dialects.

In total, it includes 47 phonemes: 6 stressed and 5 unstressed vowels except unstressed /o/, because of its rare

Table 4
Characteristics of the LMs.

LM type	# *n*-grams, M	PPL	PPL1	*n*-gram hit rate, %
Statistical bi-gram model	19.1	938	1541	79.4
Syntactico-statistical bi-gram model	21.8	946	1555	79.8
Statistical tri-gram model	49.6 (+19.1 bi-grams)	672	1078	38.0 (+41.3 bi-grams)
Statistical tri-gram model interpolated with syntactico-statistical bi-gram	49.6 (+21.8 bi-grams)	652	1043	38.0 (+41.8 bi-grams)

occurrences, and 36 consonants. As candidates for merging, the hard and soft consonants pairs were selected, because of their small differences. In addition, stressed and unstressed vowel pairs were chosen because their main difference is in the duration. Vowel reduction is already embedded in the pronunciation lexicon word forms.

Using the set P0, we built a phoneme recognition system and performed phoneme recognition experiment. From the resulting confusion matrix we made a list of all candidate phoneme pairs sorted by their confusion rates. These pairs, sorted by confusion rates, are shown in Fig. 5. As can be seen, the most confusable pairs are the vowel pairs. All of them are in the top five. Therefore, we first merge all vowel pairs and obtain phoneme set P1.

Next we look at the consonant pairs and see that the confusion rates of the pairs /n/–/n’/, /z/–/z’/, /s/–/s’/, /r/–/r’/ and /t/–/t’/ are much higher than for the other pairs. In addition, the difference in confusion rates between /s/–/s’/ and /r/–/r’/ pairs is quite big. Therefore, we make two new sets P2 by merging pairs /n/–/n’/, /z/–/z’/, /r/–/r’/ and P3 by additionally merging pairs /s/–/s’/ and /t/–/t’/. Finally, we merged all candidate pairs in P4. Summary of all the phoneme sets is given in Table 6.

6.2.3. Phoneme set performance evaluation results

In addition to the system based on P0, we trained 4 more systems using phoneme sets P1-P4. Phoneme recognition rates (accuracy) of these systems using monophone models are shown in the top of the Table 7. The LM is simple phoneme bi-gram trained on the phoneme transcriptions of both the databases. The test sets were pooled together as well. The best accuracy was achieved by the P2 set. The big difference between the results of P0 and all others can be explained by the absence of the highly confusable vowel pairs. Slightly worse performance of P4 suggests that too many consonants have been merged resulting in decreased phonetic space resolution. Phoneme recognition experiments showed that candidate pairs were chosen correctly and their merging gives better results.

Next, word recognition experiments were performed using cross-word triphone models. In order to reduce as much as possible the errors due to the LM and focus on the acoustic differences between the different phoneme sets, bi-gram LMs were trained on both the training and test transcriptions. Thus, our system uses a closed set LM. Acoustically, however, the training and the test sets are different. Word recognition results are presented in the bottom of Table 7. As can be seen, the performance of both systems is quite similar with slight advantage of the set P0. Higher accuracies for the SPIIRAS test set are due to the same lexical content of each speaker’s utterances. The number of CD triphones for each phoneme set is shown in Table 8. It also presents the number of tied states of the corresponding acoustic model.

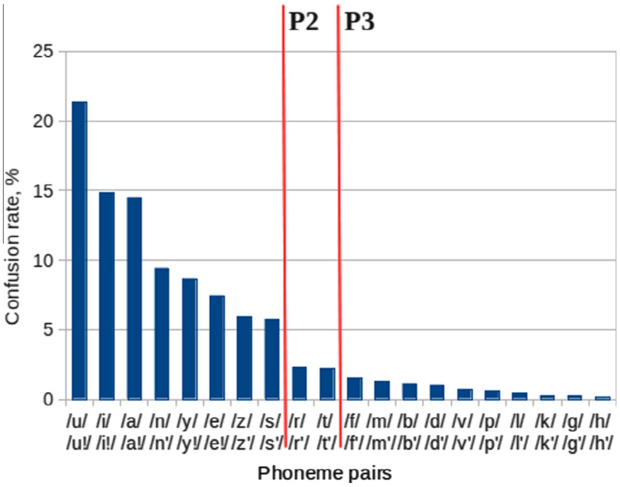

Fig. 5. Confusion rates for all phoneme pairs sorted in descending order. P2 and P3 show the N-best threshold for appropriate phoneme sets.

Since the difference in word recognition accuracy of the systems with various phoneme sets for SPIIRAS test set was quite small (less than 1% in absolute), an additional experiment to measure the real-time factor (RTF) was conducted and the results are presented in Fig. 6. For RTF of 1.2 and less, P3 phoneme set shows better recognition accuracy, while for larger RTF values, P0 phoneme set is more effective.

The high word accuracy results in Table 7 are biased due to the closed set bi-gram LM usage. To obtain unbiased performance we used the large vocabulary statistical bi-gram model (see Section 5.2) with the GlobalPhone test set. Speech recognition results for CD triphone models presented in Table 9. Although much lower, they show similar tendency in the word accuracy differences, as in the previous experiments.

Thus, for our next large vocabulary language model evaluation with triphone acoustic models, we selected the P0 phonemic alphabet consisting of 47 phonemes.

6.3. Syntactico-statistical language model performance evaluation results

In this experiment, we used the tied state triphone acoustic model built for the phoneme set selection investigations.

Table 5
Phoneme set P0. Symbol ‘!’ means that vowel is stressed, while symbol ‘’’ denotes the soft (palatalized) version of a consonant.

Class of phonemes		List of phonemes
Consonants	Hard	/b/, /v/, /g/, /d/, /zh/, /z/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /f/, /h/, /c/, /sh/
	Soft	/b’/, /v’/, /g’/, /d’/, /z’/, /j/, /k’/, /l’/, /m’/, /n’/, /p’/, /r’/, /s’/, /t’/, /f’/, /h’/, /ch/, /sch/
Vowels	Stressed	/a!/, /e!/, /i!/, /o!/, /u!/, /y!/
	Unstressed	/a/, /e/, /i/, /u/, /y/

In this case, however, the recognition engine was the open-source large vocabulary decoder Julius ver. 4.2 (Lee and Kawahara, 2009).

For the language scale and word insertion penalty optimization, we selected 4 speakers' data (40 min, 318 utterances) from the GlobalPhone test set and used them as a development set. The rest of the test data were combined with the additional SPIIRAS test set. Thus, in total, for this test we used speech from 8 speakers (4 male, 4 female) with total duration of 1 h 30 min.

Table 10 summarizes the speech recognition results in terms of word error rate (WER), letter (includes all the letters and the white-space between words) error rate (LER) as well as perplexity (PPL) and *n*-gram hit ratios for speech test set transcriptions. All LMs were built using the same vocabulary of 204 K words. The OOV rate for the combined test set is 2.5%. The best WER and LER results were obtained with the statistical tri-gram model linearly interpolated with the proposed syntactico-statistical bi-gram LM with the optimal (in terms of minimal WER using the development set) interpolation coefficient of 0.27. RTF was less than 2.0 for the speech decoder installed on a desktop PC with multi-core Intel Core i7-3770K 3.5 GHz processor.

Table 6
Desciption of the created phoneme sets.

Phoneme set	Number of phonemes	Description
P0	47	See Table 5
P1	42	P0 without /a!/, /e!/, /i!/, /u!/, /y!/
P2	39	P1 without /n'/, /s'/, /z'/
P3	37	P2 without /r'/, /t'/
P4	27	P3 without /b'/, /g'/, /d'/, /k'/, /l'/, /m'/, /p'/, /f'/, /h'/

Table 7
Speech recognition performance of the obtained phoneme sets using closed set LMs.

Test set	P0	P1	P2	P3	P4
	Phoneme recognition accuracy (%)				
SPIIRAS + GlobalPhone	48.92	52.43	53.2	53.06	52.93
	Word recognition accuracy (%)				
SPIIRAS	96.64	96.55	96.41	96.62	96.4
GlobalPhone	80.92	80.35	80.67	79.82	78.52

Table 8
Number of triphones and states pool sizes for different phoneme sets.

Phoneme set	Number of triphones	Number of states
P0	112849	5342
P1	81314	5356
P2	65562	5359
P3	56279	5335
P4	22709	5342

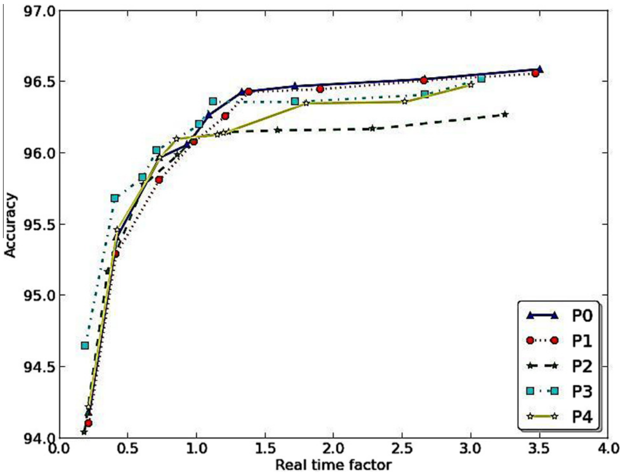

Fig. 6. Word recognition accuracy depending on RTF for the SPIIRAS test set.

Table 9
Word recognition accuracy (%) of the systems built using obtained phoneme sets and open set language model.

Test set	P0	P1	P2	P3	P4
GlobalPhone	67.07	66.78	66.53	65.93	65.28

6.4. Discussion of the experimental results

Various phoneme sets performance evaluation was conducted separately for two databases, and the obtained results were consistently similar. This suggests that they can be accepted with high confidence. The differences in the word recognition accuracy between the different phoneme sets are not significant except for the highly reduced set P4. For an off-line ASR system, where recognition accuracy is the most important feature, phoneme set P0 could be recommended to achieve the best possible result. In some real-time systems, however, where the trade-off between the recognition accuracy and the decoding speed is crucial, phoneme set P3 may provide some advantage. Word recognition results using both the small (closed set) and large vocabulary bi-gram LMs showed similar trends of performance change depending on the phoneme set.

In the large vocabulary Russian speech recognition task, the proposed syntactico-statistical language modeling approach demonstrated an advantage in comparison with the baseline *n*-gram models. Depending on the training data, the amount of bi-grams newly discovered by the syntactical analysis may vary, but we expect it to grow with the

Table 10
Summary of the results on Russian LVCSR using various LMs.

LM type	PPL	*n*-gram hit rate, %	WER, %	LER, %
Statistical bi-gram model	750	82.7	30.5	9.5
Syntactico-statistical bi-gram model	758	83.2	30.0	9.4
Statistical tri-gram model	569	39.2 (+43.5 bi-grams)	27.5	8.6
Statistical tri-gram model interpolated with syntactico-statistical bi-gram model	549	39.2 (+44.0 bi-grams)	26.9	8.5

size of the corpus. In addition, the counts for part of the regular bi-grams increase as well. This means that our method will always produce more robust models, which will have better coverage.

Relatively high WERs can be explained by the inflective nature of the Slavic language, where each stem corresponds to tens or even hundreds of endings. In continuous speech, they are usually pronounced not as clearly as the beginning parts of the words. Additionally, some different orthographic word-forms have identical phonemic representations. In order to account for the recognition errors resulting from such cases, we also applied the inflectional word error rate (IWER) measure Bhanuprasad and Svenson, 2008; Karpov et al., 2011. It assigns weight k_{inf_1} to all "hard" substitution errors S_1, where the lemma of the word-form is wrong, and weight k_{inf_2} to all "weak" substitution errors S_2, when the lemma of the recognized word-form is right, but its ending is wrong:

$$IWER = \frac{I + D + k_{inf\_1}S_1 + k_{inf\_2}S_2}{N} \times 100\% \qquad (2)$$

where I is the number of word insertions, D is the number of word deletions, and N is the total number of words in the reference utterances.

In our experiments, the IWER measure with $k_{inf\_1} = 1.0$ and $k_{inf\_2} = 0.5$ was 24.34% for the best tri-gram LM interpolated with the syntactico-statistical bi-gram LM. It turns out that in total, about 37% of the substitution errors were caused by misrecognized word endings. An automatic lemmatizer from the AOT linguistic software (Sokirko, 2004) was used to get the lemma for each word-form in the hypotheses. In practice, some of the recognition errors are due to reference transcription mistakes and misses in the pronunciation lexicon. However, the most of the errors, which can be attributed to LM, occur because of the rather high perplexity, low *n*-gram hit values, and high number of out-of-vocabulary words, all of which are consequence of the Russian language specifics.

7. Conclusions

In this paper, we presented several approaches to improve the efficiency of the ASR for inflected Slavic languages with practically free order of words in sentences and studied these approaches in application to the Russian language. During the acoustic model development, we developed new phonetic decision tree for triphone state tying taking into account the specifics of the Russian phonology. A combination of knowledge-based and statistical information approaches was used to create several different phoneme sets. Experimentally we found that the best in terms of word recognition performance is the set P0, which consists of 47 phonemes. To some extend, this finding is in accordance with the common ASR experience in acoustic models development for various other languages. We also proposed a method for syntactico-statistical language modeling. Syntactical parsing of the training text data allows us to reveal grammatically connected word pairs, which are not neighbor words in the sentence. Combining these word pairs with the statistically derived *n*-grams results in an increased number of *n*-grams, better *n*-gram coverage and hence an improvement of the speech recognition accuracy. The experiments with Russian LVCSR (>200 K word-forms) have demonstrated the advantage of the syntactico-statistical LM with respect to the standard *n*-gram language modeling achieved without any noticeable change in the decoding speed and without change in the speech decoder. Using our original syntactico-statistical LM method, we obtained 26.9% WER, having 0.6% absolute and 2.2% relative improvement with respect to the standard *n*-gram LMs.

With a few modifications, the proposed approaches can be applied to the other East Slavic languages (Belarusian and Ukrainian) and further to some more synthetic languages with high freedom of language grammar.

Our future work will be focused on the implementation and evaluation of some other types of LMs: with embedded morphological characteristics (part-of-speech, grammatical features, etc.) in order to decrease the perplexity, as well as lemma-based models and LMs with partial semantic analysis of the training text data.

Acknowledgements

This research is supported by the Ministry of Education and Science of Russia (contract No. 07.514.11.4139), by the grant of the President of Russia (project No. MK-1880.2012.8), by the Russian Foundation for Basic Research (project No. 12-08-01265) and by the Russian Humanitarian Scientific Foundation (project No. 12-04-12062).

References

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.

Arisoy, E., Saraclar, M., Roark, B., Shafran, I., 2010. Syntactic and sublexical features for Turkish discriminative language models. In: Proc. Int. Conf. ICASSP'2010, Dallas, USA, pp. 5538–5541.

Arlazarov, V., Bogdanov, D., Krivnova, O., Podrabinovich, A., 2004. Creation of Russian speech databases: design, processing, development tools. In: Proc. Int. Conf. SPECOM'2004, St. Petersburg, Russia, pp. 650–656.

Bechet, F., Nasr, A., 2009. Robust dependency parsing for spoken language understanding of spontaneous speech. In: Proc. Interspeech'2009, Brighton, UK, pp. 1039–1042.

Bellegarda, J.R., 2004. Statistical language model adaptation: review and perspectives. Speech Commun. 42, 93–108.

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.

Chelba, C., Jelinek, F., 2000. Structured language model. Comput. Speech Lang. 10, 283–332.

Cubberley, P., 2002. Russian: A Linguistic Introduction. Cambridge University Press.

Deoras, A., Sarikaya, R., Tur, G., Hakkani-Tur, D., 2012. Joint decoding for speech recognition and semantic tagging. In: Proc. Interspeech'2012, Portland, Oregon, USA.

Huet, S., Gravier, G., Sebillot, P., 2010. Morpho-syntactic postprocessing of N-best lists for improved French automatic speech recognition. Comput. Speech Lang. 24 (4), 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.

Ircing, P., Hoidekr, J., Psutka, J., 2006. Exploiting linguistic knowledge in language modeling of Czech spontaneous speech. In: Proc. Int. Conf. on Language Resources and Evaluation LREC'2006, Genoa, Italy, pp. 2600–2603.

Jokisch, O., Wagner, A., Sabo, R., Jaeckel, R., Cylwik, N., Rusko, M., Ronzhin, A., Hoffmann, R., 2009. Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system. In: Proc. SPECOM'2009, St. Petersburg, Russia, pp. 515–520.

Kanejiya, D.P., Kumar, A., Prasad, S., 2003. Statistical language modeling using syntactically enhanced LSA. In: Proc. TIFR Workshop on Spoken Language Processing, Mumbai, India, pp. 93–100.

Kanevsky, D., Monkowski, M., Sedivy, J., 1996. Large vocabulary speaker-independent continuous speech recognition in Russian language. In: Proc. SPECOM'1996, St. Petersburg, Russia, pp. 117–121.

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.

Karpov, A., Kipyatkova, I., Ronzhin, A., 2012. Speech recognition for East Slavic languages: the case of Russian. In: Proc. 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages SLTU'2012, Cape Town, RSA, 2012, pp. 84–89.

Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M., 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.

Kouznetsov, V., Chuchupal, V., Makovkin, K. Chichagov, A., 1999. Design and implementation of a Russian telephone speech database. In: Proc. SPECOM'1999, Moscow, Russia, pp. 179–181.

Kuo, H.-K.J., Mangu, L., Emami, A., Zitouni, I., Lee, Y.-S., 2009. Syntactic features for Arabic speech recognition. In: Proc. International Workshop ASRU'2009, Merano, Italy, pp. 327–332.

Kurimo, M. et al., 2006. Unlimited vocabulary speech recognition for agglutinative languages. In: Proc. Human Language Technology Conference of the North American Chapter of the ACL, New York, USA, pp. 487–494.

Lamel, L. et al., 2011. Speech recognition for machine translation in Quaero. In: Proc. International Workshop on Spoken Language Translation IWSLT'2011, San Francisco, USA, pp. 121–128.

Lamel, L., Courcinous, S., Gauvain, J.-L., Josse, Y., Le, V.B., 2012. Transcription of Russian conversational speech. Proc SLTU'2012. Cape Town, RSA, pp. 156–161.

Lee, A., Kawahara, T., 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.

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

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

Odell, J., 1995. The use of context in large vocabulary speech recognition, PhD thesis, Cambridge Univ.

Oparin, I., Talanov, A., 2005. Stem-based approach to pronunciation vocabulary construction and language modeling for Russian. In: Proc. SPECOM'2005, Patras, Greece, pp. 575–578.

Oparin, I., Glembek, O., Burget, L., Cernocky, J., 2008. Morphological random forests for language modeling of inflectional languages. In: Proc. IEEE Spoken Language Technology Workshop SLT'2008, Goa, India, pp. 189–192.

Padgett, J., Tabain, M., 2005. Adaptive dispersion theory and phonological vowel reduction in Russian. Phonetica 62 (1), 14–54.

Potapova, R., 2011. Multilingual spoken language databases in Russia. In: Proc. International Conference Speech and Computer SPECOM'2011, Kazan, Russia, pp. 13–17.

Psutka, J., Ircing, P., Psutka, J.V., Hajic, J., Byrne, W.J., Mirovsky, J., 2005. Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In: Proc. Interspeech'2005, Lisbon, Portugal, pp. 1349–1352.

Pylypenko, V., 2007. Extra large vocabulary continuous speech recognition algorithm based on information retrieval. In: Proc. Interspeech'2007, Antwerp, Belgium, pp. 1809–1812.

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.

Roark, B., 2002. Markov parsing: lattice rescoring with a statistical parser. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics ACL'2002, Philadelphia, USA, pp. 287–294.

Ronzhin, A., Karpov, A., 2004. Automatic system for Russian speech recognition SIRIUS. In: Proc. SPECOM'2004, St. Petersburg, Russia, pp. 291–296.

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B., 2010. Google search by voice: a case study. In: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, pp. 61–90.

Schultz, T., Waibel, A., 1998. Development of multilingual acoustic models in the GlobalPhone project. In: Proc. TSD'1998, Brno, Czech Republic, pp. 311–316.

Shirokova, A., 2007. STEL speech database for speaker recognition and multispeaker segmentation. In: Proc. SPECOM'2007, Moscow, Russia, pp. 877–881.

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

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L., 2012. Syntactic dependency-based n-grams as classification features. In: LNAI, 7630. Springer, Mexico, pp. 1–11.

Singh, R., Raj, B., Stern, R., 2002. Automatic generation of subword units for speech recognition systems. IEEE Trans. Acoust. Speech Signal Process. 10 (2), 89–99.

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.

Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., Evdokimova, V., 2010. CORPRES – Corpus of Russian professionally read speech. In: Proc. TSD'2010, Brno, Czech Republic, pp. 392–399.

Smirnova, J., 2011. Compound systems of pretonic vocalism after palatalized consonants in Russian dialects: a synchronic and diachronic analysis. In: Proc. 17th Int. Cong. of Phonetic Sciences ICPhS'2011, Hong Kong, pp. 1870–1873.

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

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 Russian).

Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011. SRILM at sixteen: update and outlook. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop ASRU'2011, Waikoloa, Hawaii, USA.

Stüker, S., 2008. Modified polyphone decision tree specialization for porting multilingual grapheme based ASR systems to new languages. In: Proc. ICASSP'2008, Las Vegas, Nevada, USA, pp. 4249–4252.

Stüker, S., Schultz, T., 2004. A grapheme based speech recognition system for Russian. In: Proc. Int. Conf. SPECOM'2004, St. Petersburg, Russia, pp. 297–303.

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.

Tatarnikova, M., Tampel, I., Oparin, I., Khokhlov, Y., 2006. Building acoustic models for a large vocabulary continuous speech recognizer for Russian. In: Proc. SPECOM'2006, St. Petersburg, Russia, pp. 83–87.

Vaiciunas, 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. Int. Conf. on Natural Language Processing and Knowledge Engineering NLP-KE, Tokushima, Japan, pp. 475–478.

Vazhenina, D., Kipyatkova, I., Markov, K., Karpov, A., 2012. State-of-the-art speech recognition technologies for Russian language. In: Proc. Joint Int. Conf. on Human-Centered Computer Environments HCCE'2012, ACM, Aizu, Japan, pp. 59–63.

Viktorov, A., Gramnitskiy, S., Gordeev, S., Eskevich, M., Klimina, E., 2009. Universal technique for preparing components for training of a speech recognition system. Speech Technol. 2, 39–55 (in Russian.).

Vintsyuk, T., 1968. Speech discrimination by dynamic programming. Kibernetica 1, 15–22 (in Russian).

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

Whittaker, E.W.D., Woodland, P.C., 2001. Efficient class-based language modelling for very large vocabularies. In: Proc. ICASSP'2001, Salt Lake City, USA, pp. 545–548.

Young, S., Odell, J., Woodland, P., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proc. Int. Workshop on Human Language Technology HLT'1994, Stroudsburg, PA, USA, pp. 307–312.

Young, S. et al., 2009. The HTK Book. Cambridge Univ. Press, p. 384.

Zablotskiy, S., Shvets, A., Sidorov, M., Semenkin, E., Minker, W., 2012. Speech and language recources for LVCSR of Russian. In: Proc. LREC'2012, Istanbul, Turkey, pp. 3374–3377.

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

Zhang, J.S., Hu, X.H., Nakamura, S., 2008. Using mutual information criterion to design an efficient phoneme set for Chinese speech recognition. IEICE Trans. Inform. Syst. E91-D (3), 508–513.