INTERSPEECH 2011

Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis

Alexey Karpov, Irina Kipyatkova, Andrey Ronzhin

Speech and Multimodal Interfaces Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia

{karpov,kipyatkova,ronzhin}@iias.spb.su

Abstract

In this paper, we present a word-based very large vocabulary automatic speech recognition system for Russian. Some novel methods are proposed for organization of the lexicon and the language model. Two-level morpho-phonemic prefix graph that uses some information on morphemic structure of lexical units is suggested for a compact representation of the pronunciation vocabulary and search space. Such model is more compact than the lexical tree or the linearly-based vocabulary and provides speeding up the recognition process. The syntactic analysis of a training text corpus in a combination with the statistical analysis is suggested for generation of N-gram language models. The syntax-based Russian language model allows taking into account long-distance syntactic dependencies between word pairs. The results have proved that the syntactic-statistic language model gives 5% relative improvement on the word and letter error rates with respect to the baseline models.

Index Terms: speech recognition, very large vocabulary, morphemic analysis, syntactic-based language modeling

1. Introduction

Russian, one of the major East Slavic languages, is the mother language for citizens of many Eastern European (CIS) countries even not being an official national language. However, being geographically vast, it has many various dialects, which differ both lexically and phonetically. For example, the so called okanye (existence of unstressed /o/ vowel, i.e. /golov"a/ ("head") in place of /galav"a/, while most of the Russian speakers have a tendency to merge unstressed vowels /a/ and /o/, but distinguish stressed versions) is typical of Northern Russian dialects; continuous using consonant /ɣ/ (in IPA notation) instead of /g/ is a norm for the Russian speakers in Ukraine; strong accent of Caucasian nations, etc.

It is often considered that only native speakers from Moscow and St. Petersburg regions talk "classical" Russian. However, there are some differences in these versions as well, such as vowels /i/ and /ɨ/ are distinguished by Petersburg phonology only, whereas only in Moscow region, there is a consonant /ʑː/ (soft version of hard /ʒ/). The present paper deals with the language version used mainly in the North-West area of the Russian Federation.

The Russian language is characterized by a combination (synthesis) of a lexical morpheme(s) and one or several grammatical morphemes (affixes) in one word-form. Its highly inflective nature results in essential extension of the lexicon for automatic speech recognition (ASR); it is required to apply a vocabulary in several times/orders larger than ones for English or French ASR due to numerous grammatical affixes that deteriorates both accuracy and performance of ASR. For example, the classical grammatical dictionary of A. Zaliznjak [1] contains over 110K lexemes and application of special word formation rules to all the word paradigms synthesizes the full lexicon of more than 2M diverse word-forms. Some verbs (e.g., "пить" - "to drink") produce up to two hundred grammatically correct word-forms due to inflectional or derivational affixes as well as alterations in stems. Moreover, many word-forms of different lexemes are distinguished in endings only, which are pronounced in spontaneous speech not as clearly as beginning word parts.

Another problem is that the word order in Russian sentences is rather flexible and not restricted by hard grammatical constructions, like that in English or German. It complicates creation of statistical language models and substantially decreases their predictive performance. N-gram language models (LM) for Russian have 3-4 times higher perplexity with respect to English LMs, as well as the vocabulary coverage is much worse. In [2], it is reported on 7.6% out-of-vocabulary (OOV) words for a large vocabulary of 65K Russian words and 1.1% OOV in an analogous English lexicon. These parameters for Russian correspond to ones for other morphologically rich languages, like agglutinative Finnish, Hungarian, Estonian or Turkish.

The conventional SAMPA phonetic alphabet for Russian includes 42 phonemes (covering 33 Russian graphemes): 36 consonants and 6 vowels; thus, consonant ambiguity is quite high in Russian. Moreover, a non-trivial automatic phonetic transcriber is required for creation of the pronunciation vocabulary. There are a lot of orthographic-to-phonemic transformation rules for Russian; however, the main problem at this step is to find a position of stressed vowel(s) for word-forms, which directly influences on pronunciation. There are no any common rules to determine stress (") positions in word-forms (we keep in mind it for each word), besides, compound words may have several stress positions. Thus, only knowledge-based approaches can handle with this problem.

Perhaps, one of the first attempts to develop a large vocabulary (up to 30K words) continuous ASR model for Russian was made by IBM in early 1990s [3]. Since that time there were some other projects on Russian ASR; however, there is still no system able to work with an acceptable quality.

Any modern ASR engine requires at least two types of models to recognize speech: acoustic and language models. The acoustic/pronunciation model, which contains probability representations of lexical items of the lexicon and uses a morphemic analysis, is proposed in Section 2; while the stochastic language model, which gives probabilities to word sequences and uses elements of a syntactic analysis, is described in Section 3, finally, Section 4 presents ASR results.

2. Pronunciation lexicon modeling with morphemic analysis

In order to recognize spoken Russian speech (and especially spontaneous speech) one has to utilize the recognition vocabulary of hundred thousand word-forms and such vocabulary size is now considered as a very large (>100K words) [2]. For some languages, it is efficient to split whole-words into sub-lexical units (morphemes or morphs as representation of abstract morphemes in text) and use them as tokens in the vocabulary and LM. There are such successful studies for Turkish [4], Finnish [5], Slovenian [6], Czech or Hungarian. In our previous research [7], we tried to implement a morpheme-based ASR for Russian as well, splitting words into prefixes, roots and suffixes+endings. Nevertheless, this approach did not bring any improvement on the word recognition accuracy (only on the real-time factor) because of the following reasons: (1) orthographic-to-phonemic transcription is not one-to-one (in contrast to the abovementioned agglutinative languages) and is context-dependent, which results in multiple phonemic representations of real morphs; (2) high variability of pronunciation of the same morphemes, depending on actual stress position and alternation of stems in many lexemes; (3) due to a complicated morphology one morph can be a prefix, root, suffix or ending (e.g., short morphs “a” or “o”) in different word-forms that makes troubles at word synthesis of recognized morph chains. Thus, it was found out as the result of our experiments that in spite of better OOV words coverage and lexicon size reduction with the morph-based ASR, the word error rate (WER) increases because of many grammatically incorrect words synthesized from sub-lexical units on the output of ASR. Moreover, morphs are much shorter and therefore more confusable acoustically than words. An elaboration of a stem-based ASR system was also attempted for Russian by another team, but without any improvement on the WER as well [8].

In practice, decomposition of word-forms into morphs can be performed by two different approaches: grammatical (knowledge-based) and unsupervised methods based on the statistical analysis of a large text corpus [9]. In our research, we used the former way, splitting up each word from the lexicon into stem (may consist of prefixes+roots) and ending (may consist of grammatical suffixes+inflexion+postfix; zero ending is allowed as well). For example, one Russian word-form ”переключающиеся” (”switching”), consisting of prefix “пере”, root “ключ” (“switch”), suffixes “а+ющ”, inflexion “ие” and reflexivity postfix “ся” is decomposed by the applied morphemic analyzer into the stem “переключ” and the ending “ающиеся”. For the morphemic analysis, we utilize a free software distributed with the LGPL license developed in the framework of AOT project [10], which is based on a morphological database containing >170K word paradigms including widespread names.

Considering the morphological structure of Russian word-forms we propose to apply a two-level morpho-phonemic prefix graph (TMPG) for a compact lexicon and search space organization. The straightforward approach to organize the vocabulary, as commonly used in ASR systems, is a linearly-organized pronunciation lexicon, where each word is represented as a linear sequence of phonemes independently of other words. However, another organization of the pronunciation lexicon in the form of a lexical (prefix) tree [11] gives significant improvement in its compactness. Taking into account similarity of phonemic transcriptions of many word-forms, the lexical tree shares common prefixes of their transcription and makes decoding more time efficient due to fewer local likelihood calculations in model states. Using the advantages of the lexical tree-based vocabulary organization and the knowledge-based morphemic analysis of Russian words we have implemented the TMPG in order to represent the recognition lexicon more compactly in the PC memory and to speed up the speech decoding process (Figure 1). The pronunciation vocabulary is organized as the lexical tree for stems linking with the linearly-organized list of unique endings. This structure has several times less nodes and arcs than the lexical tree structure [12]. TMPG represents all the word-forms of the ASR vocabulary and their phonemic transcriptions. Each phoneme is represented by a continuous left-to-right Hidden Markov Model (HMM) consisting of three emitting states. The first level of TMPG is the lexical tree of stem transcriptions, terminal nodes of which are orthographic stems. The second level is the linearly-organized list of ending transcriptions linked to the stems. The number of terminal nodes on the first level equals the number of different stem transcriptions in the vocabulary and the numbers of input and terminal nodes on the second level correspond to the number of unique ending transcriptions. Such lexicon structure allows generating only grammatically correct word-forms in contrast to the morph-based vocabulary.

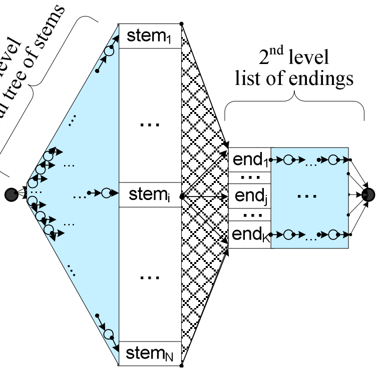

Figure 1: *TMPG for pronunciation vocabulary organization.*

The orthographic-to-phonemic translation of word-forms is made by an automatic transcriber, which employs a set of context-dependent and independent phonetic rules (almost 100 rules) for spoken Russian [13]. Additionally, alternative phonemic transcriptions were created for some vocabulary items in order to model the effects of phonemes’ reduction and assimilation in spontaneous speech using a set of cross- and intra-word phonetic rules (over 20), for example:

- vowel /i/, located in the first position of a word and after a hard consonant ending the previous word, is transformed into the phoneme /1/ (in SAMPA notation for the letter “ы”): /m"oZ1t izm'in'"it'/ → /m"oZ1t 1zm'in'"it'/ (“can change”);
- unstressed vowels are reduced up to complete disappearance if they are located between identical consonants: /aktr"isa s1gr"ala/ → /aktr"i s1gr"ala/ (“actress has played”);
- phonemes /t/,/t'/ and /d/,/d'/ located after /s/,/s'/ and /z/,/z'/, correspondingly, are reduced: /v"aZnas't' prabl'"em1/ → /v"aZnas' prabl'"em1/ (“importance of a problem”);
- phoneme /j/ located at the word end is completely reduced if it follows an unstressed vowel and the next word starts not from a stressed vowel: /dragats"en1j k"am'in'/ → /dragats"en1 ka"m'in'/ (“a precious stone”);
- if two identical consonants are situated on the break of two words, the last consonant of the former word is disappeared: /l'"es sasn"ov1j/ → / l'"e sasn"ov1/ (“pine forest”).

In the developed transcriber, we employ an extended dictionary of more than 2.3M word-forms with marks of the stressed vowels. This dictionary is a fusion of two different morphological databases: AOT (www.aot.ru) and Starling (starling.rinet.ru/morpho.php). The former one is larger and has above 2M items, but the latter one contains information about the secondary stress for many compound words as well as words with the Russian letter “ё” (which is always stressed in Russian original words, but often replaced to “е” in texts loosing accent information). The extended pronunciation vocabulary of 2.3M words is further used to generate a real TMPG-based ASR lexicon and LM.

3. Language model with syntactic analysis

Stochastic language models based on purely statistical analysis of some training text data are efficient for many natural languages with rather strict grammatical structure, like English or German. But for languages with more freedom in sentence construction, such models are much less perspective. There exist some other kinds of LMs: sub-word-, trigger-, cache-, class-based LMs, etc. In recent systems, a syntactic analysis is often embedded into various ASR levels, including LM. In [14], it is proposed to incorporate some syntactic information (part-of-speech tag, grammatical features, head-to-head dependency relations) into the discriminative morph-based LM for Turkish. In [15], it is 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. Another work [16] introduces a stochastic morpho-syntactic language model for agglutinative Hungarian ASR as well.

For Russian ASR, we have implemented an integral language model that takes advantages of both statistic and syntactic text analysis. Figure 2 presents the scheme of creation of the LM using some elements of the syntactic analysis. A training text corpus is processed in parallel identifying N-grams and syntactic dependencies in sentences and then the results of both analyzers are fused in the integral stochastic model that takes into account frequencies of the detected word pairs. These analyzers complement each other very well: syntactic one is used to find long-distance dependencies between words (potential N-grams not appeared in training data), but not relations between the adjacent words, which are covered by the statistical analyzer. For the statistical text analysis we employ the CMU SLM Toolkit v2, while VisualSynan v1.0 [10] from the AOT release is used for the syntactic analysis. The latter parses input sentences and produces a graph of syntactical dependencies between pairs of lexical items. There are 32 different types of syntactic groups in the analyzer in total, but we extract 10 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: "они уже полностью распределены" ("they are already fully allocated");

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

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

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

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

9) noun – verb in the subordinate attributive clause: "картины, которые вчера выставлялись" ("paintings that were exhibited yesterday").

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

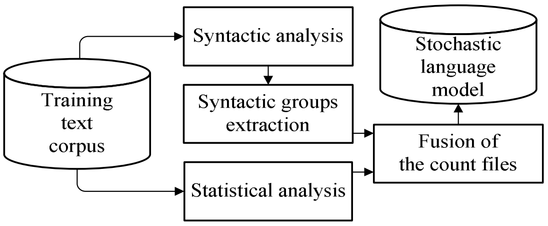
Syntactic analysis
Stochastic language model
Training text corpus
Syntactic groups extraction
Fusion of the count files
Statistical analysis

Figure 2: *Integral syntactic-statistical LM generation.*

Moreover, words of the syntactic groups (2)-(3), (7)-(10) and (1), but without subordinate attributive clauses starting "which", "who", etc.) are commutative in Russian and each such syntactic dependence produces two bigrams with direct and inverse word order. Figure 3 shows an example of the syntactic analysis of the phrase ("In the very expensive show, both military and civilian aircrafts, which arrived yesterday and today the airport of our little town, are involved") taken from the corpus. It demonstrates some types of long-distance dependences, whereas all the adjacent word pairs are modeled by the statistical bigrams. Commutative groups are denoted by the blue double-sided arrows. Thus, syntactic parsing of this sentence produces 13 long-distance word pairs additionally to the statistic processing, which gives 18 bigrams. N-gram likelihoods in the integral stochastic LM are calculated after merging the results (the count files) of both analyzers based on their frequency in the training text data. N-grams with words occurred once in the corpus are deleted from the LM, because they likely contain typos or these words are extremely rare.

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

4. Experimental evaluation

There are some medium-scale commercial databases of read Russian speech, e.g., RuSpeech, SPEECON, telephone speech ISABASE, SpeechDat(E), and even corpora of children's speech INFANTRU/CHILDRU [17]. In our research, we have used own corpus of spoken Russian speech Euronounce-SPIIRAS, created in 2008-2009 in the framework of the EURONOUNCE project. This corpus contains ~22 hours (+30 min for ASR testing) of continuous speech and dialogs of 52 native speakers from St. Petersburg (male and female voices are fifty-fifty) recorded in an acoustic studio with 44.1 KHz, 16 bits, SNR≈35dB, a stereo pair of Oktava MK-012 cardioid microphones, Presonus Firepod sound board.

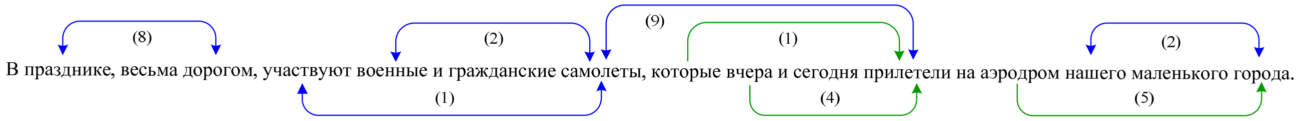
(8)
(2)
(9)
(1)
(2)
В празднике, весьма дорогом, участвуют военные и гражданские самолеты, которые вчера и сегодня прилетели на аэродром нашего маленького города.
(1)
(4)
(5)

Figure 3: *An example of the syntactic phrase analysis (numbers of types of long-distance syntactic dependencies are shown).*

As for acoustic features, we used 13-dimentional Mel-Frequency Cepstral Coefficients (MFCC) with the 1st and 2nd order derivatives calculated from the 26-channel filter bank analysis of 20 ms long frames with 10 ms overlap. Cepstral mean subtraction is applied to audio feature vectors. Continuous density HMMs with 16 Gaussians per state model Russian context-dependent phones. ASR system SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) [12] is used for the speech decoding. It integrates the Viterbi-based token passing algorithm with optimization of the beam pruning width and the lexicon as TMPG that allow us to recognize continuous speech with the real-time factor close to one.

We have evaluated 1-, 2-, 3-gram word-based LMs (in the ARPA format) with large pronunciation lexicons of 76-8K words and very large vocabularies of 208-9K items (with various cutoffs providing very close vocabulary sizes). Table 1 summarizes the recognition results in terms of the word error rate (WER) and letter/grapheme (includes all the letters and the white-space between words) error rate (LER). Lexicon size for each LM, rate of OOV words, perplexity (PP) and N-grams hits are also shown. The bigram+synt model demonstrates the advantage of the syntactic-based LM. The results with various lexicons have proved that the large vocabulary <100K in not enough for spoken Russian ASR, but the very large one is more adequate. Further enlargement of the vocabulary size does not reduce the WER because it covers words of the test data well enough. Higher order LMs are relatively weak for Russian ASR, because over 90% trigrams occurred once in the corpus and trigrams hit is very low for the test data.

Table 1. *Speech recognition results with various LMs.*

N-gram (N)	Lexicon size, K	OOV, %	PP	N-gram hit, %	WER, %	LER, %
1	208	0.8	5493	99.2	74.3	37.0
2	78	3.5	851	70.4	68.8	30.5
2	208	0.8	777	82.9	58.4	26.5
2+synt	209	0.8	772	84.1	56.1	25.3
3	76	4.9	452	35.8	69.9	35.5

Rather low speech recognition rates could be explained by the inflective nature of Russian. Each stem corresponds with tens of endings, which are usually pronounced in continuous speech not as clearly as the beginning parts of words and often word-forms have identical phonemic transcriptions. In [18], it was suggested to use the inflectional word error rate (IWER) metrics, which gives 1.0 weight to all hard substitutions confusing lemmas, while 0.5 weight is applied to all weak substitutions, when lemma of a recognized word-form is correct, but its ending is incorrect. IWER for the 2gram+synt LM, using an automatic Russian lemmatizer [10] to all the reference phrases and recognition hypotheses, was 48.2%.

5. Conclusion and future work

The paper presented SIRIUS ASR system for spoken Russian. This system uses the proposed two-level morph-phonemic prefix graph for compact organization of the pronunciation lexicon and search space as well as the integral syntactic-statistic language model for Russian. Application of this LM has provided the reduction of the LER from 26.5% to 25.3% and the WER from 58.4% to 56.1% in comparison with the baseline LM. The inflectional word error rate was 48.2%.

The proposed methods could also be efficient in ASR for other Slavic languages, like Belarusian or Ukrainian. Our future work will be dedicated to implementation of other types of language models: with embedded morphological characteristics (part-of-speech, grammatical features, etc.) in order to decrease the perplexity, as well as lemma-based and LMs with a partial semantic analysis of training text corpora.

6. Acknowledgements

This research is supported by the grant of the President of Russia (project MK-64898.2010.8) and by the Ministry of Education and Science of Russia (contracts П2579 and П2360).

7. References

[1] Zaliznjak, A.A., “Grammatical Dictionary of the Russian Language: Accidence”, 4th Edition. Moscow: “Russian dictionaries”, 800 p., 2003.

[2] Whittaker, E.W.D. and Woodland, P.C. “Efficient class-based language modelling for very large vocabularies”, in Proc. ICASSP’01 Conference, Salt Lake City, USA, 545-548, 2001.

[3] Kanevsky, D., Monkowski, M., Sedivy, J., “Large Vocabulary Speaker-Independent Continuous Speech recognition in Russian Language”, in Proc. 1st International Conference on Speech and Computer SPECOM’96, St. Petersburg. Russia, 117-121, 1996.

[4] Arisoy, E., Dutagaci, H. and Arslan, L., “A unified language model for large vocabulary continuous speech recognition of Turkish”, Signal Processing, Elsevier, 86(10):2844-2862, 2006.

[5] Hirsimäki, T., Pylkkönen, J. and Kurimo, M., “Importance of High-Order N-Gram Models in Morph-Based Speech Recognition”, IEEE Trans. on Audio, Speech and Language Processing, 17(4):724-732, 2009.

[6] Rotovnik, T., Maucec, M.S. and Kacix, Z., “Large vocabulary continuous speech recognition of an inflected language using stems and endings”, Speech Communication, Elsevier, 49(6): 437-452, 2007.

[7] Karpov, A. and Ronzhin, A., “Russian Speech Recognition Model with Morphemic Analysis and Synthesis”, in Proc. 19th International Congress on Acoustics ICA, Madrid, Spain, 2007.

[8] Oparin, I. and Talanov, A., “Stem-based approach to pronunciation vocabulary construction and language modeling for Russian”, in Proc. SPECOM, Patras, Greece, 575-578, 2005.

[9] Kurimo, M., et al., “Unsupervised segmentation of words into morphemes - Morpho Challenge. Application to automatic speech recognition”, in Proc. Interspeech’06, Pittsburgh, PA, USA, 1021-1024, 2006.

[10] Sokirko, A., “Morphological modules on the website www.aot.ru”, in Proc. 10th International Conference “Dialog-2004”, Protvino, Russia, 559-564, 2004.

[11] Ortmanns, S., Eiden, A. and Ney, H., “Improved Lexical Tree Search for Large Vocabulary Recognition”, in Proc. ICASSP’98, Seattle, USA, 817-820, 1998.

[12] Ronzhin, A., Leontieva, A., Kagirov, I. and Karpov, A., “Morpho-Phonetic Tree Decoder for Russian”, in Proc. 12th International Conference on Speech and Computer SPECOM’07, Moscow, Russia, 491-498, 2007.

[13] Kipyatkova, I. and Karpov, A., “Creation of Multiple Word Transcriptions for Conversational Russian Speech Recognition”, in Proc. 13th International Conference on Speech and Computer SPECOM’09, St. Petersburg, Russia, 71-75, 2009.

[14] Arisoy, E., Saraçlar, M., Roark, B. and Shafran, I., “Syntactic and sub-lexical features for Turkish discriminative language models”, in Proc. ICASSP’10, Dallas, USA, 5538-5541, 2010.

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

[16] Szarvas, M. and Furui, S., “Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR”, in Proc. ICASSP, HongKong, China, 368-371, 2003.

[17] Lyakso, E., Frolova, O., Kurazhova, A. and Gaikova, J., “Russian Infants and Children’s Sounds and Speech Corpuses for Language Acquisition Studies”, in Proc. Interspeech’10, Makuhari, Japan, 1878-1881, 2010.

[18] Svenson, M. and Bhanuprasad, K., “Errgrams - A Way to Improving ASR for Highly Inflective Dravidian Languages”, in Proc. 3rd International Joint Conference on Natural Language Processing IJCNLP’08, India, 805-810, 2008.