Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 4062786, 9 pages
http://dx.doi.org/10.1155/2016/4062786

Research Article

A Russian Keyword Spotting System Based on Large Vocabulary Continuous Speech Recognition and Linguistic Knowledge

Valentin Smirnov,[1] Dmitry Ignatov,[1] Michael Gusev,[1] Mais Farkhadov,[2] Natalia Rumyantseva,[3] and Mukhabbat Farkhadova[3]

[1]*Speech Drive LLC, Saint Petersburg, Russia*
[2]*V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia*
[3]*RUDN University, Moscow, Russia*

Correspondence should be addressed to Mais Farkhadov; mais.farhadov@gmail.com

Received 11 July 2016; Revised 27 October 2016; Accepted 14 November 2016

Academic Editor: Alexey Karpov

The paper describes the key concepts of a word spotting system for Russian based on large vocabulary continuous speech recognition. Key algorithms and system settings are described, including the pronunciation variation algorithm, and the experimental results on the real-life telecom data are provided. The description of system architecture and the user interface is provided. The system is based on CMU Sphinx open-source speech recognition platform and on the linguistic models and algorithms developed by Speech Drive LLC. The effective combination of baseline statistic methods, real-world training data, and the intensive use of linguistic knowledge led to a quality result applicable to industrial use.

1. Introduction

The need to understand business trends, ensure public security, and improve the quality of customer service has caused a sustainable development of speech analytics systems which transform speech data into a measurable and searchable index of words, phrases, and paralinguistic markers. Keyword spotting technology makes a substantial part of such systems. Modern keyword spotting engines usually rely on either of three approaches, namely, phonetic lattice search [1, 2], word-based models [3, 4], and large vocabulary speech recognition [5]. While each of the approaches has got its pros and cons [6] the latter starts to be prominent due to public availability of baseline algorithms, cheaper hardware to run intensive calculations required in LVCSR and, most importantly, high-quality results.

Most recently a number of innovative approaches to spoken term detection were offered such as various recognition system combination and score normalization, reporting 20% increase in spoken term detection quality (measured as ATWV) [7, 8]. Deep neural networks application in LVCSR is starting to achieve wide adoption [9]. Thanks to the IARPA Babel program aimed at building systems that can be rapidly applied to any human language in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech [10] in recent years wide research has been held to develop technologies for spoken term detection systems for low-resource languages. For example, [11] describes an approach for keyword spotting in Cantonese based on large vocabulary speech recognition and shows positive results of applying neural networks to recognition lattice rescoring. Reference [12] provides an extensive description of modern methods used to build a keyword spotting system for 10 low-resource languages with primary focus on Assamese, Bengali, Haitian Creole, Lao, and Zulu. Deep neural network acoustic models are used both as feature extractor for a GMM-based HMM system and to compute state posteriors and convert them into scaled likelihoods by normalizing by the state priors. Data augmentation via using multilingual bottleneck features is offered (the topic is also covered in [13]). Finally language independent and unsupervised acoustic models are trained for languages with no training data. An average MTWV reported for these languages ranges from 0.22 for Zulu to 0.67 for Haitian Creole. In [14] the use of recurrent neural networks for example-based word spotting in real time for English is described. Compared to more widespread text-based systems, this approach makes use of spoken examples of a keyword to build up a word-based model and then do the search within speech data. As an alternative to hybrid ANN-HMM approaches authors in [15] offer a pure NN based keyword search system for conversational telephone speech in Vietnamese and Lao. For Vietnamese the "pure" NN system provides ATWV comparable with that reported for a baseline hybrid system while working significantly faster (real-time factor 3.4 opposed to 5.3 for a hybrid system).

As high-quality language modeling is an indispensable part of any modern keyword spotting system, a lot of effort is now aimed at improving LMs. One of the most recent trends is to use web data in training. The advent of the Internet has provided rich amount of data to be easily available for speech recognition community [16]. This is of particular interest for low-resource languages and among most recent improvements [17] suggests an approach to effectively deal with the challenge of normalizing and filtering the web data for keyword spotting. Two methods are offered, one using perplexity ranking and the other using out-of-vocabulary words detection. This resulted in more than 2% absolute improvement in ATWV across 5 low-resourced languages. Reference [18] covers the aspect of augmenting baseline LMs with carefully chosen web data, showing that blogs and movie subtitles are more relevant for language modeling of conversational telephone speech and help to obtain large reductions in out-of-vocabulary keywords.

Russian research in the domain of speech recognition falls in line with global scientific trends. It is noteworthy however that most frequently the research is conducted to meet a more general target of creating LVCSR systems per se with no specific focus on spoken term detection. The most well-known systems include Yandex SpeechKit [19] used to recognize spoken search queries via web and mobile applications, real-time speech recognition system by Speech Technology Center [20] used for transcribing speech in the broadcasting news, LVCSR system developed by SPIIRAS [21, 22] used for recognizing speech in multimodal environments, and speech recognition system by scientific institute Specvuzavtomatika [23] based on deep neural networks.

Current paper presents the results of the ongoing research underlying the commercial software for speech analytics. The software design follows the concept of a minimum viable product, which motivates incremental complication of the technology while the product evolves. Such approach motivated us to rely on generally available open-source toolkits and a number of readily available knowledge-based methods developed under our previous studies.

Sections 2 and 3 outline the overall setup of applying LVCSR technology to keyword spotting for Russian telephone conversational speech, including the key system parameters and the description of experiments run to assess the quality and performance of the system. Special focus is given to linguistic components used at the training and spotting stage. Section 4 describes the off-the-shelf speech analytics system developed using the ideas and results discussed in this paper.

2. Key System Parameters

The system described in the paper is intended to be used to perform keyword search in telephone conversational speech. The system is provided both as SDK to be integrated with speech recording systems and as a stand-alone MS Windows application. The system is created on top of CMU Sphinx [24]; this framework has been chosen due to its simplicity and licensing model which allows for freely using the code in commercial applications. Following the idea of minimum viable product we mostly use the standard settings across all system modules. 13 MFCCs with their derivatives and acceleration are used in the acoustic front-end; triphone continuous density acoustic models are trained on around 200 hours of telephone-quality Russian speech (8 kHz, 8 bit, Mono) recorded by 200 native speakers. 5-state HMMs are used with diagonal covariation matrix, and CART (classification and regression trees) algorithm is used to cluster acoustic models into 9000 senones, each senone being described by 10 to 30 Gaussians. Texts in the training database for language models are transcribed automatically with a morphological and contextual linguistic processor [25]. A set of transcription variation rules are applied. Unigram and bigram language models are trained on hundreds of thousands of modern Russian e-books generally available on the Internet. Decoder makes use of a standard CMU Sphinx token-passing algorithm with pruning methods widely employed in the system setup including maximum beam width, word insertion penalty, and acoustic likelihood penalty.

The core novelty of the system is granted by extensive use of linguistic knowledge on both the training and spoken term detection steps. The system uses a linguistic processor with built-in information on Russian morphology which helps to generate high-quality transcriptions for any word form and thus train more viable acoustic models. The same processor is used to generate various forms of words which ensures better spoken term detection on the spotting step. A rule-based transcription variation algorithm is applied to generate alternative phoneme sequences. Ultimately on the language modeling step the texts are automatically prefiltered by the type of text to let only dialogues stay in the training corpus.

3. Algorithms, Associated Problems, and Solution

3.1. Acoustic Front-End. While throughout the system standard MFCCs are used, an additional effort was required to make the front-end work for keyword spotting in a real-world application. First, audio files to be analyzed are chunked into 10-second long chunks in order to split the decoding process over multiple CPUs. An overlap of 1 second is used to guarantee that a keyword is not truncated between two subsequent speech segments. Further on, a parsing algorithm

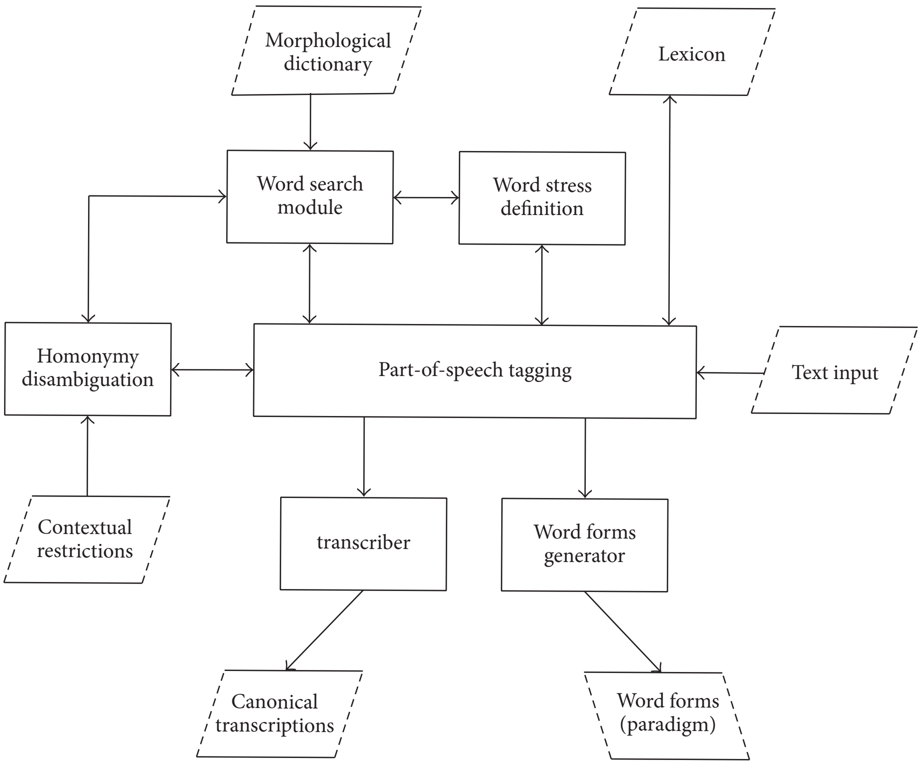


Figure 1: Linguistic processor.

is applied to combine partial decoding results into a single file in order to avoid redundant false alarms. The future plan is to use VAD to divide the audio stream into phrases which would better suit the LVCSR-based approach used in this paper; however, our current VAD implementation has shown worse results, hence the use of chunks of equal length.

3.2. Acoustic Modeling, Grapheme-to-Phoneme Conversion, and a Transcription Variation Algorithm. The system discussed in the paper is intended to be used in real-world telephone environment under low sound quality conditions. To cover this requirement the acoustic model is trained on real-world data encountering the telephone channel quality speech in Russian telephone networks. Continuous density HMMs are used, resulting in a representative set of 9000 senones each described with a GMM with 10–30 components.

Under our previous research [25] a linguistic processor has been developed which makes use of information on morphological characteristics of around 600 000 Russian words (see the structure on Figure 1) to transcribe words and generate forms of words. Processor parses the text and defines the part of speech for every word in the sentence; then the word stress is defined, and a set of preprogrammed contextual grapheme-to-phoneme rules is applied to derive a canonical (“ideal”) word transcription.

The current state of the art for transcribing words in speech recognition systems is to use statistical grapheme-to-phoneme converters [26, 27]. The research has been held on combining various techniques, for example, in [28] Conditional Random Fields and Joint-Multigram Model are used to bring an additional improvement in quality. Studies have been done [29, 30] to introduce weighted finite state transducers to grasp the probabilities of in-word phonetic sequences. Altogether these studies outline the key advantages of probabilistic approach compared to knowledge-based methods, namely, language independency (easily ported to a new language), ability to generalize and provide transcriptions to new (out-of-vocabulary) words, and the need of a smaller footprint of linguistic data (and hence effort) to train a grapheme-to-phoneme converter.

On the other hand, the majority of the results shared in cited studies relate to languages characterized with low number of word forms (e.g., English and French). Meanwhile Russian is a highly inflectional language with a word stress depending on the exact word form in the paradigm and a high frequency of homonymy also affecting word stress and thus being a source for potential transcription errors [31]. This means that one needs a much bigger hand-crafted lexicon to train a high-quality probabilistic grapheme-to-phoneme converter for Russian. This obstacle together with the concept of minimum viable product described above motivated us to set probabilistic grapheme-to-phoneme conversion as a target for our future research and to use a readily available high-quality knowledge-based linguistic processor instead. Another important factor which guided this choice is the ability to disambiguate homonymy and to generate word forms (to be discussed later on).

The key element of the acoustic model training process is transcription variation. Every phrase used to train the models receives several alternative transcriptions by applying

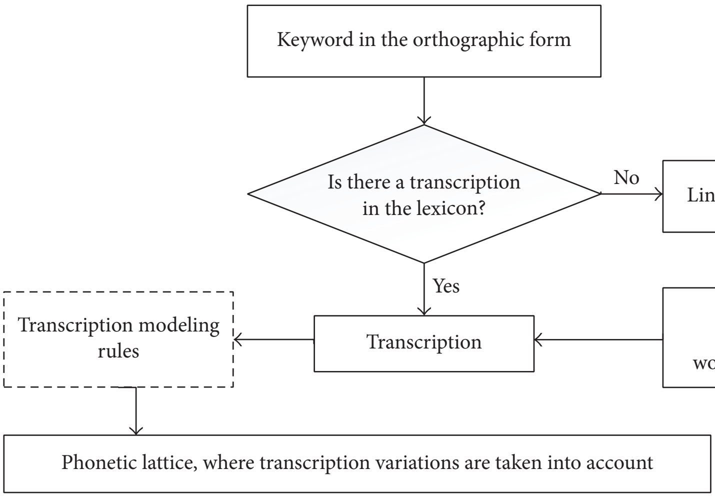

Keyword in the orthographic form
Is there a transcription in the lexicon?
No
Linguistic processor
Yes
Automated (canonical) word transcription
Transcription
Transcription modeling rules
Phonetic lattice, where transcription variations are taken into account

Figure 2: Transcription variation algorithm.

a set of predefined linguistic rules. Then on the training step CMU Sphinx train module chooses the best alternative which maximizes expectation. The experiments showed a 4% absolute increase in keyword detection rate achieved thanks to such implementation (please refer to Section 4 for more details on experiments). At the moment the rules are derived manually based on linguistic knowledge. The list of rules is then embedded in the recognizer which is run on a training dataset to define which rules provide for quality improvement and should be kept in the production system. As the next step of our research we plan to develop a sufficient corpus to train these rules automatically.

The ultimate list of transcription variation rules chosen on the training set contains 30 contextual phoneme-to-phoneme and grapheme-to-phoneme mappings based on both from the modern research on Russian spontaneous speech [32] and from the authors' proper experiments with real-life data audio analysis. The main steps of the transcription variation algorithm are outlined below (please also refer to Figure 2):

(1) A textual annotation of the trained database is loaded.

(2) If the word is not in the lexicon (used mainly for foreign words and named entities), automatic transcriber is launched which makes use of the digitized dictionary of the Russian language, containing 600 thousand words with morphological information and a part-of-speech (POS) tagger. As a result of this stage the word stress is assigned to the right syllable of every word.

(3) Automated, or canonical, transcription is generated by applying context-dependent letter-to-phone rules.

(4) Pronunciation variation is executed by iteratively applying a separate set of phoneme-to-phoneme and grapheme-to-phoneme transcription modeling rules to the canonical transcriptions.

It is well known that knowledge-based rules, being "laboratory" in origin, may happen to be inadequate when confronted with real-world data. However this was our intent to check this critical assumption on our test material. Moreover, during the past decades, Russian phonetics has undergone a general shift from laboratory speech to fully spontaneous [32, 33], and the rules we use are based on vast research on spontaneous speech traits.

The rules are divided into two main groups. The first contains substitution, deletion, and insertion rules, which apply to initial phonetic transcriptions. Here are some examples of such rules:

(i) [@] ("schwa"), followed by a consonant, is deleted in the unstressed position after stressed syllable.

(ii) [f] is deleted from consonant sequence [fs] + (any) unvoiced consonant.

(iii) Affricates [c] and [tʃ'] are substituted by fricatives [s] and [ʃ^j], respectively (sign j denotes that a consonant is palatalized).

(iv) Sonorant [j] is deleted before unstressed vowel at the beginning of words.

(v) Noise stops (e.g., [p], [t], [p^j], and [t^j]) are deleted in the final position after vowels due to implosive pronunciation (i.e., without burst following articulators closure).

The second group of rules makes use of both morphological and orthographical level of linguistic representation. Hence, this is not correction to initial transcriptions (phoneme-to-phoneme rules) but a separate set of grapheme-to-phoneme rules. Here are some examples:

(i) [@j@] and [uju] in unstressed inflections of adjectives "–ая" and "–ую" are changed to [@e] and [u], respectively.

(ii) [@v@], [ɨv@], and [iv@] in unstressed noun inflections "–ого" and "–его" are changed to [@@], [ɨ@], and [i@].

(iii) [@t] in verb inflections “–ат” is changed to [it].

For frequent words we also added another set of rules, which generate simplified pronunciation, which is common to informal spontaneous speech. These include [d^j] and [v] deletion in intervocalic position, [$s^j t^j$] changing to [s^j], and so forth.

3.3. Language Models and the Choice of Relevant Content to Train Them. Initially language models have been trained with a few gBs of user-generated content to be found on the Internet, including public forums, social networks, and chats. The idea behind this was that such content would better represent spontaneous speech and thus ensuring more sustainable keyword spotting results. However the experiments have shown that such linguistic material occurred to bear an intrinsic drawback, because it contains enormous number of spelling errors which led to a statistic bias and wrong lemmas to appear in the lexicon. Hence a decision was taken to rely on standard and error-free texts derived from a wide range of books of different genres available on the Internet. Only books by modern authors (1990s and later) were chosen to reflect current traits of Russian speech. However only the dialogues have been extracted from such books to guarantee the “live” style of communication, which is characteristic of real-world telephone speech. 2 gB of raw text data was used as a result to train a unigram and bigram language models containing 600 000 Russian lemmas. The LMs were trained using SRILM toolkit [34] with Good-Turing discounting algorithm applied.

Current research in the domain of language modeling is focused on applying deep neural networks and high-level LM rescoring [35]. In our case there is insufficient data to train such models, which motivated us to shift to much simpler models. As outlined in Section 3.4 we do not rely on the most probable word sequence in the recognition result to detect keywords; rather we want to generate as diverse and “rough” lattice on the indexing step to guarantee high probability for the spoken term detection. Simple bigram/unigram language modeling fits this aim quite well.

3.4. Decoding, Word Spotting, and Automated Word Form Generation. The main idea behind using LVCSR to find keywords is to transform speech signal into a plain text and then search for the word in this text. However due to diverse types of communication context in the telephone conversational speech it is not viable to use the top decoding result per se. Rather, it makes sense to parse the resulting recognition lattice to find every possible node with the keyword. Hence speech is first indexed into recognition lattices; the keyword search is performed on-demand at a later stage.

To improve spotting results we make intensive use of the linguistic processor described above. When a word is entered as a search query its forms are automatically generated by addressing the morphological dictionary (see Figure 1) and a set of variants are derived for the word which are then searched in the lattice and appear in the recognition results list. For example, when the word “кусок” is to be searched (Russian word for “a piece”) all the words containing this sequence will be searched within a recognition lattice; hence the user will be able to spot the words “куска” and “куском” and so forth. Since Russian is an inflectional language numerous forms are available for one word. Consequently low-order (unigram and bigram) language models used in our system cause the recognizer to make errors in the word endings. The simple idea described above helps to avoid errors and achieve much better results.

Table 1: Experimental results.

Parameter	Value
MTWV	0.37
RTF	2.0

4. Experimental Results

The system described hereby is intended to be used in real-world applications to analyze telephone-quality speech. To test it a 10-hour database including the recordings of dialogues of around 50 speakers has been recorded using the hardware and software of SpRecord LLC (http://www.sprecord.ru/). 1183 different keywords are searched within the database. The signal-to-noise ratio falls between 5 and 15 dB, reflecting an adverse real telephone channel environment.

Maximum Term-Weighted Value (MTWV) is a predicted metric corresponding to the best TWV for all values of the decision threshold; θ (see formula (1)) and real-time factor (RTF) metrics (formula (2)) are used to evaluate system performance; the former metric reflects the quality of word spotting, and the latter reflects its speed. RTF parameter is calculated on 1 CPU unit of 3 gHz. The results are shown in Table 1.

$$\text{TWV}(\theta) = 1 - \left[P_{\text{Miss}}(\theta) + \beta \cdot P_{\text{FA}}(\theta)\right]. \tag{1}$$

θ is the threshold used to determine a hit or a miss, and β is the weight that accounts for the presumed prior probability of a term and the relative costs of misses and false alarms are equal to 0.999 in our study.

$$\text{RTF} = \frac{T_{\text{proc}}}{T_{\text{set}}}. \tag{2}$$

T_{proc} is the time spent on processing the file, and T_{set} is the duration of the test set.

In order to understand whether these results correspond to the current state of the art we compared them to the result of another scientific group for spoken term detection in telephone conversational of another underresourced language (Cantonese) [11]. What we saw is that our results in terms of keyword search quality fall in between those reported for Cantonese when GMMs are used in the acoustic model and are slightly worse when deep neural networks are used (MTWV 0.335 and 0.441, resp.). As for the real-time factor our results outperform those reported in [14], which may be attributed to a relatively small number of Gaussians we use per senone.

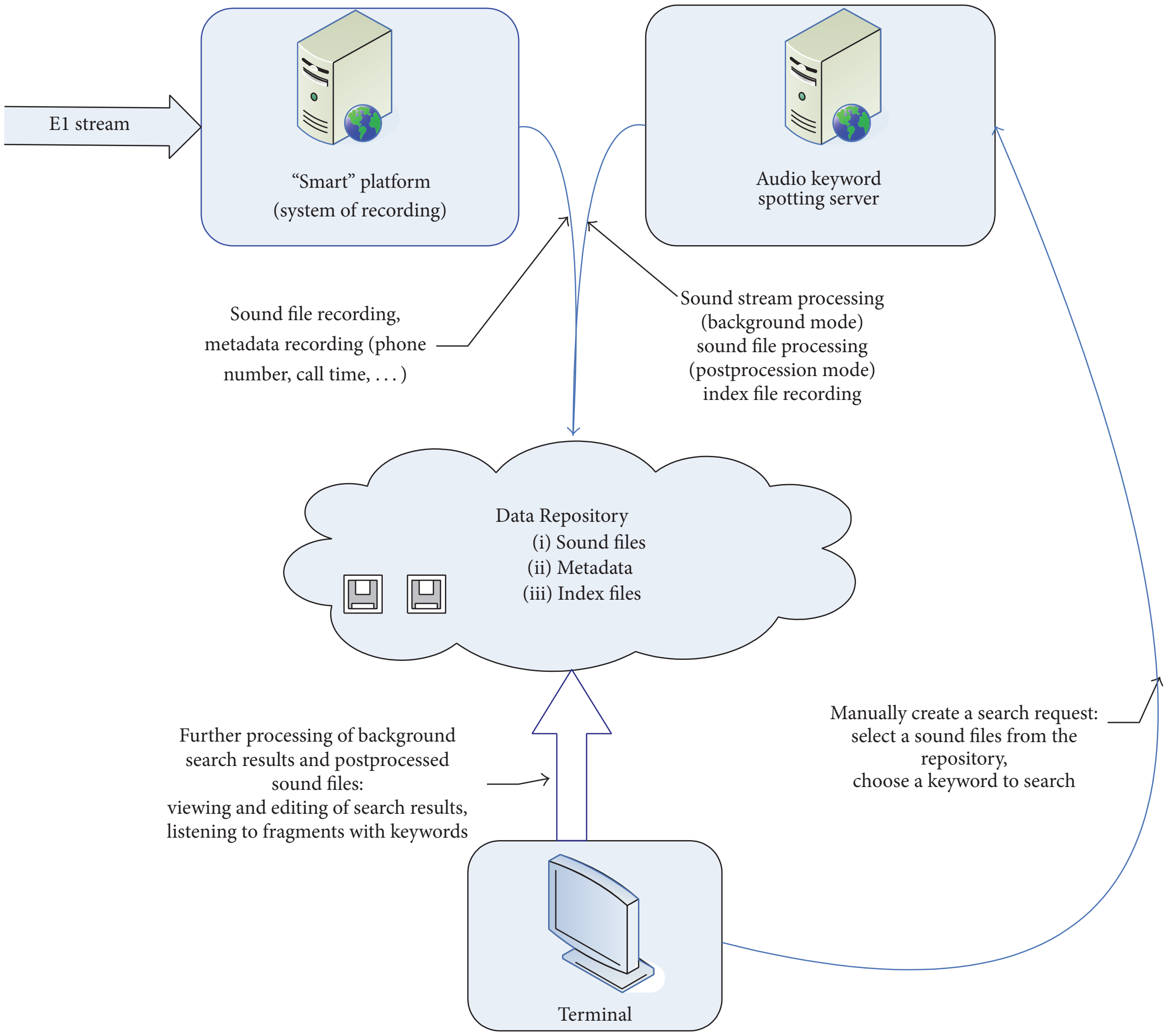


Figure 3: System architecture.

5. System Architecture and User Interface

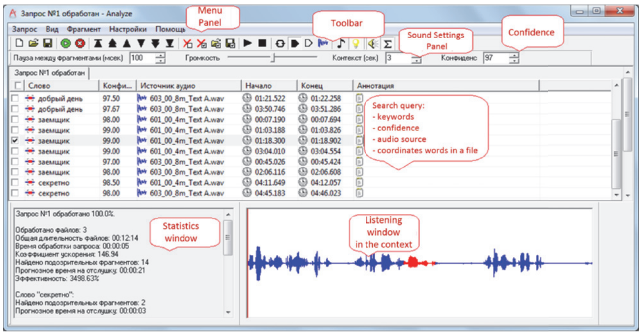
5.1. Principal Components. The algorithms described in Section 2 were used in creating "ANALYZE" software—an industrial speech analytics system. Figure 3 outlines the key system components: word spotting server, terminal, and data repository. Word spotting server processes speech data and saves index with positions of searched keywords into the database. The terminal is used to schedule or launch immediate search queries and to view the search results. The search is performed in two steps: first, the lattice with speech recognition results is generated for each wave file; second, the keyword is found via a substring search within this lattice. The data repository contains both speech files and corresponding indices.

5.2. User Interface. The key problems of human-machine interaction within speech analytics systems, including accurate treatment of the keyword spotting results, and the role of in the optimization of workflows in modern organizations are reflected in [36–39]. Figure 4 outlines the user interface of the ANALYZE software which has been developed based on use-cases validated with the end-users. Usability and use-case integrity were tested in the real-world environment. All settings are available in 1-2 clicks; real-time reporting is shown on the screen; navigation panel provides access to all needed functions. Table 1 with search results provides easy filtering and listening modes. Figure 3 presents the main board of the system's user interface.

An essential benefit of the software is the ability to work in real-time mode on the workstations with limited resources, which makes it worthy for small organizations with a fraction of telephone lines in use.

6. Conclusion and Further Plans

A keyword spotting system for Russian based on LVCSR has been described in this paper. General availability of



Запрос №1 обработан - Analyze

Пауза между фрагментами (мсек) 100 Громкость Контекст (сек) 3 Конфиденс 97

Запрос №1 обработан

☐	Слово	Конфи...	Источник аудио	Начало	Конец	Аннотация
☐	добрый день	97.50	603_00_8m_Text A.wav	01:21.522	01:22.258	
☐	добрый день	97.67	603_00_8m_Text A.wav	03:50.746	03:51.286	
☐	заемщик	98.00	601_00_4m_Text A.wav	00:07.190	00:07.694	
☐	заемщик	99.00	601_00_4m_Text A.wav	01:03.188	01:03.826	
☑	заемщик	99.00	601_00_4m_Text A.wav	01:18.300	01:18.902	
☐	заемщик	99.00	601_00_4m_Text A.wav	03:04.010	03:04.554	
☐	заемщик	97.00	603_00_8m_Text A.wav	00:45.026	00:45.424	
☐	заемщик	98.00	603_00_8m_Text A.wav	02:06.116	02:06.608	
☐	секретно	98.50	601_00_4m_Text A.wav	04:11.649	04:12.057	
☐	секретно	98.00	603_00_8m_Text A.wav	04:45.183	04:46.023	

```
Запрос №1 обработано 100.0%.

Обработано файлов: 3
Общая длительность файлов: 00:12:14
Время обработки запроса: 00:00:05
Коэффициент ускорения: 146.94
Найдено подозрительных фрагментов: 14
Прогнозное время на отслушку: 00:00:21
Эффективность: 3498.63%

Слово "секретно":
Найдено подозрительных фрагментов: 2
Прогнозное время на отслушку: 00:00:03
```

Figure 4: The user interface of the "ANALYZE" software.

open-source software made it easy to be implemented and linguistic modules helped to improve the system quality, while representative training and test data ensured the applicability of the system to real-world problems.

The ongoing research is aimed at further tuning the acoustic and language models, trying probabilistic frameworks for grapheme-to-phoneme conversion, data-driven transcription variation, introducing noise compensation and pause detection into the front-end and at creating specific confidence measures to minimize false alarms which are caused by frequent words in the language model.

In building our automated keyword spotting system based on large vocabulary continuous speech recognition we relied on the results of the scientific community, namely, the open-source software CMU Sphinx for acoustic modeling and decoding and SRILM for language modeling. At the same time the system has several technological advantages: the use of linguistic knowledge in training and decoding, namely, a morphological parser of texts and transcription variation to generate word transcriptions, transcription variation rules, and automated generation of word forms on the spotting step; real-world industrial data used to train acoustic models; accurate language modeling achieved via cautious choice of training data; real-time operation mode on limited computer resources.

We believe that high-quality automated keyword spotting system based on large vocabulary continuous speech recognition for online speech data analysis can be used both as a technological platform to create effective integrated systems for monitoring and as a ready-to-use solution to monitor global information space.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank SpRecord LLC authorities for providing real-world telephone-quality data used in training and testing of the keyword spotting system described in this paper.

References

[1] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, pp. 659–664, Kyoto, Japan, December 2007.

[2] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent wordspotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 377–380, Adelaide, Australia, 1994.

[3] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12–15, 2005. Proceedings*, vol. 3658 of *Lecture Notes in Computer Science*, pp. 302–309, Springer, Berlin, Germany, 2005.

[4] M. Yamada, M. Naito, T. Kato, and H. Kawai, "Improvement of rejection performance of keyword spotting using anti-keywords derived from large vocabulary considering acoustical similarity to keywords," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[5] M. Matsushita, H. Nishizaki, H. Nishizaki, S. Nakagawa et al., "Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 1205–1208, Geneva, Switzerland, September 2003.

[6] I. Szoke et al., "Comparison of keyword spotting approaches for informal continuous speech," in *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (INTERSPEECH '05)*, pp. 633–636, Edinburgh, UK, 2005.

[7] D. Karakos, R. Schwartz, S. Tsakalidis et al., "Score normalization and system combination for improved keyword spotting," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '13)*, pp. 210–215, Olomouc, Czech Republic, December 2013.

[8] J. Mamou, J. Cui, X. Cui et al., "System combination and score normalization for spoken term detection," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 8272–8276, Vancouver, Canada, May 2013.

[9] G. Hinton, L. Deng, D. Yu et al., “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] M. Harper, “IARPA Babel Program,” https://www.iarpa.gov/index.php/research-programs/babel?highlight=WyJiYWJlbCJd.

[11] J. Cui, X. Cui, B. Ramabhadran et al., “Developing speech recognition systems for corpus indexing under the IARPA Babel program,” in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’13)*, pp. 6753–6757, IEEE, Vancouver, Canada, May 2013.

[12] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, “Speech recognition and keyword spotting for low resource languages: babel project research at CUED,” in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 16–23, Petersburg, Russia, 2014.

[13] F. Grézl, M. Karafiát, and M. Janda, “Study of probabilistic and Bottle-Neck features in multilingual environment,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU ’11)*, pp. 359–364, December 2011.

[14] P. Baljekar, J. F. Lehman, and R. Singh, “Online word-spotting in continuous speech with recurrent neural networks,” in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT ’14)*, pp. 536–541, South Lake Tahoe, Nev, USA, December 2014.

[15] K. Kilgour and A. Waibel, “A neural network keyword search system for telephone speech,” in *Speech and Computer: 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5–9, 2014. Proceedings*, A. Ronzhin, R. Potapova, and V. Delic, Eds., pp. 58–65, Springer, Berlin, Germany, 2014.

[16] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and Ö. Çetin, “Web resources for language modeling in conversational speech recognition,” *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, article 1, 2007.

[17] A. Gandhe, L. Qin, F. Metze, A. Rudnicky, I. Lane, and M. Eck, “Using web text to improve keyword spotting in speech,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU ’13)*, pp. 428–433, Olomouc, Czech Republic, December 2013.

[18] G. Mendels, E. Cooper, V. Soto et al., “Improving speech recognition and keyword search for low resource languages using web data,” in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech ’15)*, pp. 829–833, Dresden, Germany, September 2015.

[19] SpeechKit API, http://api.yandex.ru/speechkit/.

[20] K. Levin, I. Ponomareva, A. Bulusheva et al., “Automated closed captioning for Russian live broadcasting,” in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH ’14)*, pp. 1438–1442, Singapore, September 2014.

[21] A. Karpov, I. Kipyatkova, and A. Ronzhin, “Speech recognition for east slavic languages: the case of russian,” in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU ’12)*, pp. 84–89, Cape Town, South Africa, 2012.

[22] A. Karpov, K. Markov, I. Kipyatkova, D. Vazhenina, and A. Ronzhin, “Large vocabulary Russian speech recognition using syntactico-statistical language modeling,” *Speech Communication*, vol. 56, no. 1, pp. 213–228, 2014.

[23] M. Zulkarneev, R. Grigoryan, and N. Shamraev, “Acoustic modeling with deep belief networks for Russian Speech,” in *Speech and Computer: 15th International Conference, SPECOM 2013, Pilsen, Czech Republic, September 1–5, 2013. Proceedings*, vol. 8113 of *Lecture Notes in Computer Science*, pp. 17–23, Springer, Berlin, Germany, 2013.

[24] K.-F. Lee, H.-W. Hon, and R. Reddy, “An overview of the SPHINX speech recognition system,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[25] V. A. Smirnov, M. N. Gusev, and M. P. Farkhadov, “The function of linguistic processor in the system for automated analysis of unstructured speech data,” *Automation and Modern Technologies*, no. 8, pp. 22–28, 2013.

[26] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[27] T. Hain, “Implicit modelling of pronunciation variation in automatic speech recognition,” *Speech Communication*, vol. 46, no. 2, pp. 171–188, 2005.

[28] D. Jouvet, D. Fohr, and I. Illina, “Evaluating grapheme-to-phoneme converters in automatic speech recognition context,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’12)*, pp. 4821–4824, IEEE, Kyoto, Japan, March 2012.

[29] L. Lu, A. Ghoshal, and St. Renals, “Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU ’13)*, pp. 374–379, Olomouc, Czech Republic, December 2013.

[30] S. G. Paulo and L. C. Oliveira, “Generation of word alternative pronunciations using weighted finite state transducers,” in *Proceedings of the Interspeech 2005*, pp. 1157–1160, Lisbon, Portugal, September 2005.

[31] I. Kipyatkova, A. Karpov, V. Verkhodanova, and M. Železný, “Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition,” *International Journal of Computer Science and Applications*, vol. 10, no. 1, pp. 11–30, 2013.

[32] L. V. Bondarko, A. Iivonen, L. C. W. Pols, and V. de Silva, “Common and language dependent phonetic differences between read and spontaneous speech in russian, finnish and dutch,” in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS ’03)*, pp. 2977–2980, Barcelona, Spain, 2003.

[33] L. V. Bondarko, N. B. Volskaya, S. O. Tananaiko, and L. A. Vasilieva, “Phonetic properties of russian spontaneous speech,” in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS ’03)*, p. 2973, Barcelona, Spain, 2003.

[34] A. Stolcke, “SRILM-an extensible language modeling toolkit,” in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colo, USA, September 2002.

[35] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” *Interspeech*, vol. 2, no. 3, 2010.

[36] R. V. Bilik, V. A. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, “Analysis of the oral interface in the interactive servicing systems. II,” *Automation and Remote Control*, vol. 70, no. 3, pp. 434–448, 2009.

[37] V. A. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, “Computerized queuing systems and speech technologies,” *Control Sciences*, no. 2, pp. 3–7, 2006.

[38] V. A. Zhozhikashvili, R. V. Bilik, V. A. Vertlib, A. V. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, "Open queuing system with speech recognition," *Control Sciences*, no. 4, pp. 55–62, 2003.

[39] N. V. Petukhova, S. V. Vas'kovskii, M. P. Farkhadov, and V. A. Smirnov, "Architecture and features of speech recognition systems," *Neurocomputers: Development, Applications*, no. 12, pp. 22–30, 2013.

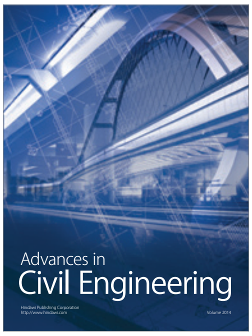
Journal of
Engineering

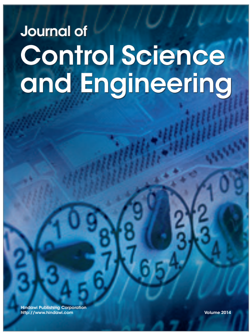
The Scientific World Journal

International Journal of
Rotating Machinery

Journal of
Sensors

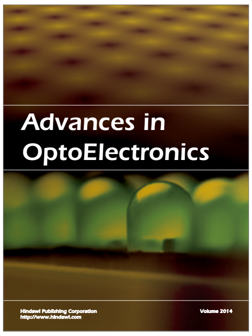
International Journal of
Distributed Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science and Engineering

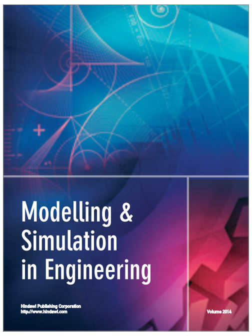
Journal of
Robotics

Journal of
Electrical and Computer Engineering

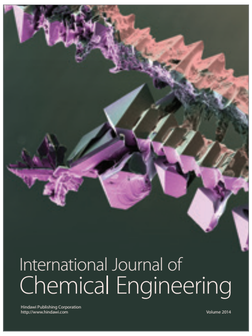

Advances in
OptoElectronics

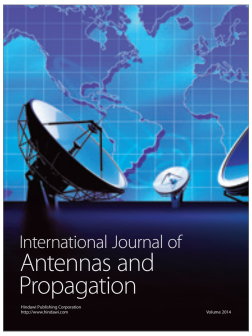
VLSI Design

International Journal of
Navigation and Observation

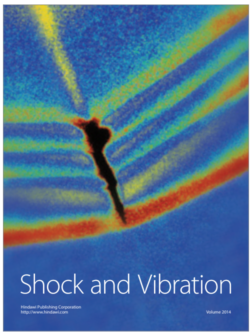

Modelling & Simulation in Engineering

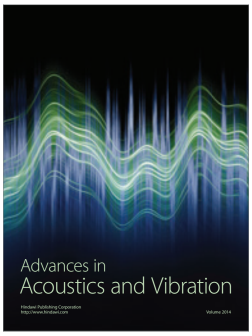
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration