Survey of Russian Speech Recognition Systems

Andrey L. Ronzhin, Rafael M. Yusupov, Izolda V. Li, Anastasia B. Leontieva

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

Abstract

An idea of the natural speech interaction was arrived since electronic calculation machines were created. The command interface of the first computers did not provide acceptable speed and naturalness of interaction. By many years of investigations large number of methods and software, solved the speech recognition problem, were developed. Significant results were achieved in speaker depended recognition of isolated speech and today the attention of researches is focused on problems of spontaneous speech, speaker independency, robustness in noisy conditions. In this paper the authors present the survey of Russian language specifics, methods and applied models for Russian speech recognition developed in some organizations in Russia and abroad. The survey is prepared by proceedings of the last conferences and internet webs of developers.

1. Introduction

Today one of the main direction in informatics become development of the natural means of human computer interaction. Permanent advancing possibilities of computers and network technologies already are not used in full measure owing to of unnatural dialog between a human and a computer. Lack of a decision of this problem suppresses the development of various applied systems in telecommunication, medicine, edutainment and everyday life, since all modern techniques and network services use automate devices for control and information processing. Already more than half a century specialists of different scientific directions are researching problems of the automatic speech recognition. It is connected with complexity and interdisciplinary character of the task. Besides, the task of human-computer interaction became to be considered as important scientific one, where it is necessary to take into account various natural modalities (text, speech, image, video, 3D, touch, and gestures).

The paper consists of three parts. Fist part describes specifics of Russian language and troubles aroused at automatic speech recognition. Second part is devoted to survey of speech recognition methods and applications developed in Russian organizations and abroad. In the last part Russian speech corpora intended for training of speech recognition engines are described.

2. Specifics of Russian language

Russian belongs to Slavic Branch of Indo-European family of languages, which are characterized by tendency to combination (synthesizing) of the lexical morpheme (or several lexical morphemes) and one or several grammatical morphemes in one word-form. So Russian is a synthetic language and has a complex mechanism of the word formation. Let's consider basic meaning units, which are used at the word formation [1]. One of basic units of language is a word. It exists in language as a system of word-forms. Word-form is one of representations of the word. The word-form is extracted of the speech stream as the meaning unit, characterized by two properties: 1) relative freedom of replacement and 2) impermeability, i.e. inability to include inside any meaning units of speech which possess such freedom of replacement. The mininal meaning part extracted in structure of the word form is a morpheme. The morpheme is an elementary indicator of grammar and influences to morphological characteristics of words. There are two types of the morphemes: root and affix. The affix morphemes is divided on: prefixes, suffixes, postfixes, interfixes and inflections. For the speech recognition task we use simplified language models, in which three types of morpheme (root, affix and inflection) should be marked.

The root obligatory exists in every word forms and contains a basic lexical meaning of the word. The basis of each word is a stem. The root morpheme can completely coincide with a stem. If the word form consists of one morpheme then this morpheme is a root. The root morpeheme is determinated by a notion of the stem, which is presented a part of the word form without inflection and postfix. At the failing of these morphemes the stem coincides with the word form. The affix allows to denote an additional meaning of the word formation or morphology. The affix morpheme can absent in the word form and never completely coincides with the stem. The stems are divided on simple (containing one root morpheme) and complex (containing more of one morpheme). Since morphemes are elementary supersegment parameters of grammatical meaning it is possible to select classes of segment and supersegment morphemes. Such classification of morphemes is presented Table 1 [1].

Table 1. Morpheme classification on grammatical meaning

<table>
<tr><td rowspan="11">Morphemes</td><td colspan="5">Morphemes –words</td></tr>
<tr><td rowspan="10">Morphemes – parts of words</td><td colspan="4">Roots</td></tr>
<tr><td rowspan="9">Affixs</td><td colspan="3">Morphemes-operations</td></tr>
<tr><td colspan="3">Empty</td></tr>
<tr><td rowspan="7">Segment</td><td colspan="2">Prefixes</td></tr>
<tr><td colspan="2">Infixes</td></tr>
<tr><td colspan="2">Circumfixes</td></tr>
<tr><td colspan="2">Transfixes</td></tr>
<tr><td rowspan="2">Postfixes</td><td>Inflexion</td></tr>
<tr><td>Suffixes</td></tr>
</table>

Inflection morpheme responses for morphological values of the word form (gender, number, case and person). In word forms postfix morphemes can be only used after inflection.

The inflection morpheme is used by strict grammar rules. Besides every part of speech has own rule of declension. Russian parts of speech can be inflective and uninflective. Five basic parts of speech - noun, verb, adjective, numeral and pronoun - are inflective. Their characteristics and examples of declension are presented in Table 2. The number of the word form for these examples was calculated by web tool [2]. Inflective parts of speech can change case, sort, number and other morphological features of word. Thus, the word form can be constructed by inflection as well as suffixes, prefixes, etc. Taking into account possible changeable forms of words, and various combinations with affixes the number of word forms sharply increases.

Table 2. Inflective forms of Russian parts of speech

Part of speech	Grammatical categories	Example
Noun	2 numbers 6 cases	The word «дерево» has 12 word forms (дерево, деревом, дереве, деревьям…)
Verb	2 voice (active/passive) 2 aspect (perfect, continue) 3 mode (indicative, subjunctive, imperative) Participle Verbal adverb 3 persons 2 numbers 3 tenses 3 genders 6 cases	The word «делать» has 180 word forms (делаю, делаем, делаешь, делаете, делает, делают, делал, делала, делало, делали, делаюсь, делаемся, делай, делайте, делающая, делающие, делаемого, делавши, делав, делая…)
Adjective	3 comparative degrees short form 2 grades (animate/inanimate) 3 genders 2 numbers 6 cases	The word «прекрасный» has 34 word forms (прекрасный, прекрасная, прекрасен, прекрасней…)
Numeral	Ordinal Cardinal 2 grades (animate/inanimate) 3 genders 6 cases	The word «третий» has 29 word forms (третий, третья, третье, третьи, третьей, третьем, третьих…)
Pronoun	2 numbers 6 cases 3 genders	The word «я» has 6 word forms (я, меня, мне, мной, мною, мне)

Thus, many of word-forms of the same words differ only by endings which are usually pronounced not so clearly as the beginnings of words and errors of inflections' recognition lead to errors of phrase recognition because of word discordance. In Russian the subject and object of sentence can only be determined by a word inflection and a verb congruence rather then order of words. Even the interchanging of words does not lead to the loss of the meaning. Also a punctuation play an important role in Russian. Accents and pauses should be used for a correct recognition of the inflective part of speech and senstence as a whole.

Let's consider some differences of Russian and English, as researches and system engineering of English speech recognition are most actively conducted. For comparison of language models two characteristics are often used: (1) perplexity factor and (2) number of missed words or number of Out Of Vocabulary words (OOV). The perplexity factor defines language model and equals to an average number of words which can be connected with the previous word in a phrase. In table 3 comparative characteristics of Russian and English languages according to these parameters are presented [3]. You can see that N-gram language models of Russian are characterized by increased data, conditioned by the inflectional nature of Russian and flexible restrictions for the word order. Such models are less predictive and have high rates of the perplexity and OOV.

Table 3. Comparison of statistical models of English and Russian languages

Vocabulary size, thousands of words	English		Russian	
	Perplexity	OOV, %	Perplexity	OOV, %
65	216.1	1.10	413.3	7.60
100	224.5	0.65	481.0	5.31
200	232.4	0.31	586.8	2.65
400	236.8	0.17	670.9	1.19
500	-	-	689.9	0.93
800	-	-	713.8	0.64
1000	-	-	718.8	0.53

Also an important role for speech recognition plays various phonetic alphabet of languages. In international phonetic alphabet SAMPA for Russian 44 phonemes are accepted: 38 phonemes of consonants and 6 phonemes of vowel sounds. In American variant of English language phonetic alphabet SAMPA there are 41 phoneme: 24 consonants and 17 vowels (including a lot of diphthongs). It is obvious, that recognition of consonants is more complex, than vowels, as they are less stable, than vowels and have much smaller duration.

Thus, the inflectional nature of Russian, very rich vocabulary and relaxed word order constraints lead to huge N-gram language models. Taking into account specifics of Russian language we shall review some developed systems and databases for Russian speech recognition.

3. Russian speech recognition systems

In this section the survey of existing models and methods for Russian speech recognition is proposed. Also modern systems and speech corpora developed in Russia and abroad are described. The first systems of automatic speech recognition were speaker-dependent, could recognize limited number of words and required preliminary adjustment for a speaker. Among most well known systems of Russian speech recognition it can mark series of devices "Rech", developed by laboratory of Professor Taras Vintsuk [4]. This model based on conception of consequent speech processing by

dynamic time warping. During the recognition a speech signal was segmented by phonetic units and words.

New direction in speech recognition area was started by Professor Valerian Trunin-Donskoy [5]. The especial attention was paid to acoustic speech processing, feature extraction (temporal, frequency, amplitude) and modeling of speech production. Developed methods have strong mathematical base. Besides, some successful models were introduced in serial production. The devices for automatic speech synthesis and recognition MARS-1, MARS-2 were based on a formant analysis and synthesis and worked with telephone lines [6]. The terminal IKAR for input of voice commands was developed in 1980 [7]. Other speaker dependent device developed by Prof. Victorov recognized up to 1000 words with accuracy 87% and time processing 0,1 sec. The 15 bands spectral processor and dynamic time warping were used there [8].

The device DIS-332 based on microprocessor K580IK80 could recognize 200 voice commands with accuracy 96% [9]. Features vectors were created by 16 spectrum bands and parameters of base frequency F0. Components of feature vector were normalized by a logarithm scale. A comparison of input signal with templates was accomplished by gradient method. The compact board version use also zero-crossing parameters and recognized up to 200 words with accuracy 98% and time processing 2 sec.

Thus, at the end of 20 century there are some available systems for speaker dependent recognition of isolated Russian words with a small vocabulary (less 1000 words). However, the communication based on input of separated words does not possess naturalness and high speed. As a result these systems did not find a wide application and the works were practically stopped. Also the economical situation in Russia played own role in those period. The development of high-end computers gave the possibility of real time processing of large databases. The processing of a large text material and the creation of a statistical language model were started. Recently some approaches to recognition of continuous Russian speech were developed.

Today intellectual systems for telecommunication and various informational services are actively developed. These systems provide the access to information more easy and faster, that attracts many users. By this reason, the most important characteristic of modern speech recognition systems is a speaker independency.

There are some developed models of speaker independent Russian speech recognition with a large vocabulary based on the statistical language model, where various units of the word are used (full word form, stem, morpheme, etc.) [10,11,12]. The most applied statistical methods are based on hidden Markov modeling (HMM). The arising of high-end computers allowed to apply these complex calculation models for speech recognition. However, the development of any speech recognition and understanding system mainly deal with study of language specifics, so the research should carry out immediately in several neighboring scientific areas: linguistics, phonetics, digital signal processing, information theory, computer science, etc.

Among Russian scientific groups, which deal with speech processing, it should be mark Institute of System Analysis RAS, Dorodnicyn Computing Centre RAS, Institute of Control Sciences RAS, SPIIRAS, speech groups of philological and mechanic-mathematical departments of Moscow State University, Moscow State Linguistic University, St. Petersburg State University, Taganrog State University of Radioengineering, Tomsk State University, Speech Technology Center, etc. The most significant results were achieved by scientific groups, which have possibility to acquire or development of large Russian vocabulary and speech corpora.

More than 30 years a laboratory of Queuing Systems of Institute of Control Sciences RAS conducts the research in a speech processing area. The main work of the laboratory is focused on introduction of continuous speech recognition technologies in the mass using services [13]. The system for Russian and other languages are applied. Mathematical models are developed for a description of the automatic speech recognition process. The special system for adjusting the speech recognition model for an application and the assessment of influence of different parameters of recognition quality are developed. As a result the voice interface for SIRENA system, dispatching office for taxi and WebMoney service were developed [14].

Dorodnicyn Computing Centre RAS focuses on the robustness of speech recognition methods, which save a high quality in real conditions of speech communications. Today there are many speech recognition systems, which created and tested in laboratory conditions, but at the real using these systems do not provide the declared accuracy. Therefore the task is to save the sufficiently high recognition accuracy in real environments at presence of noisy channels, background noise, variation and inaccuracies of voice, etc. A common approach is based on using multiple parallel acoustic-phonetic models of allophones and non-speech acoustical events. Every allophone and morpheme has several special acoustical models, which are jointly used in the lexical network at speech stream decoding. A selection of the models is accomplished automatically by the analysis of speech corpora and classification subject to a surrounding area and a speaker voice [15].

Moscow State Institute of electronics and mathematics develops automatic speech recognition systems based on the genetic machinery, fuzzy logic methods, Hidden Markov's models and other methods for construction speech understanding systems at the minimal constraints [16].

Institute of System Analysis RAS conducts a theoretical and applied research of real time speech processing based on neuron networks and models of human speech perception and production [17].

The chair of mathematical theory of intelligence systems and the laboratory of theoretical cybernetic problems of the faculty of mathematics and mechanics of Moscow State University after M.V. Lomonosov are studying problems, which prevent a creation of industrial continuous speech technologies for Russian language. A decomposition of the common language model into two components, such as a morphology model and a model based on the initial form of words, allows to use all advantages of the N-gram approach. Furthermore, the creation of the independent model, which contains the morphological information, allows to solve a problem of the acoustic similarity of different word-forms of the same words. This solution may be efficiently applied for many inflective languages, which are characterized a great variability of word-forms (e.g. Slavonic group languages). The result of theoretical research was a creation the package of programs for constructing different Russian language

models including composite models based on the category method. Also the software for the morphological analysis and synthesis with a vocabulary of 150 thousand words has been developed [18].

Speech Informatics Group of Saint-Petersburg Institute for Informatics and Automation the Russian Academy of Science developed a set of methods and program modules for training acoustic models of phonetic units of speaker-independent recognition systems for Russian speech as well as collected required training material to perfecting training methods for acoustic models of continuous Russian speech recognition systems [12]. Also a system for automatic transcription Russian texts and separate words was designed. The database of the different types of Russian morpheme was developed. The model based on the morpheme analysis for speaker-independent recognition of Russian continuous speech was developed. A division of word-forms into morphemes allowed significantly reducing the vocabulary of recognizable lexical units. Such processing provides invariance to the grammar deviation and increase a speed of Russian speech recognition. Besides, such approach may be use to recognition of other languages with the complex mechanism of word formation. For testing developed methods an experimental model with voice-activated access for rubrics' searching in electronic catalogue "Yellow Pages of Saint-Petersburg" was created. In this experiment the size of a vocabulary was 1850 words and the accuracy of recognition was over 90% [19].

The speech recognition group of Donetsk Institute of Artificial intelligence developed a program, which automatically recognizes up to 1000 separate pronounced words with a high reliability. It was applied for a voice-activated typesetting program of mathematical formulas in the system "Equation", a voice-activated control program for mobile robots. At present time the group studies a problem of phonemic recognition [20]. To solve this problem original methods of segmentation (automatic fragmentation the speech signal into several parts which correspond to separate phonemes) were developed. Moreover the software and libraries for automatic labeling Russian word-forms and the morphological analysis are actively developed [21].

Today existing speech recognition systems are mainly based on statistical methods, therefore it is necessary to solve a problem of collecting and processing the information about speech and language to design such systems. Thus a development of effective speech recognition methods and a database creation are interrelated and very important. It requires gross funds, which is able to provide just large corporations such as IBM, Intel, Microsoft, and AT&T developing the theoretical base for speech technology area for years. Many speaker-independent and speaker-dependent recognition systems for English language with acceptable quality level as well as standardized speech corpora were developed. At the end of 90s west companies, which invested in ARS research and achieved essential results at this area, made some attempts to adapt existing methods for Russian language.

One of early attempts to design a speaker-independent recognition system for Russian language was a trigram statistic model developed by researchers of IBM (Human Language Technologies, Computer Science Dept., IBM T.J. Watson Research Center) [22]. The system was trained by 30000 utterances (40 Russian speakers). The trigram language model was trained by a vocabulary of 40 million words. A system of Russian phonetic sub-groups and a set of rules for phonetic words' transcription were elaborated. Although during the testing of this model by 8 speakers, which pronounced 149 utterances, the error level was less than 5%, but these investigations did not obtain a further development. It turned out, that Russian language is inflected in contrast to English language, there is a more complicated word-formation mechanism, order of words in a utterance is not so rigid, and recognition vocabulary is richer. It is necessary to consider all these features during the development of Russian speech recognition systems.

During the cooperation project of VNIIEF-STL and Intel Corporation (1999-2001) a continuous speech recognition system with a large vocabulary SDT (Speech Developer Toolkit) was developed in Nizhniy Novgorod [23]. STD toolkit involves modules for the attribute vector calculation, construction and adaptation of acoustic models, prompt speech decoding by finite state and stochastic grammars, evaluation of decoding results. STD was applied to design recognition systems for English and Chinese languages as well as a prototype of Russian speech recognition system. Now VNIIEF-STL is developing Russian speech recognition system with a vocabulary of over 1 million words including development of the compact conception of Russian phonetic vocabulary, modified decoding algorithm and statistic language model for Russian language.

During last ten years attempts of companies such as Intel and Lernout&Hauspie to create Russian speech recognition systems are not so successful. These investigations were reduced by economic reasons. Among commercial systems, which achieved the end user, there is only a line of "Gorynich" systems developed by Russian company "VoiceLock" and firm "White computers" on the basis of Dragon system. "Gorynich" system was unsatisfactory in recognition (about 70%) since it did not consider Russian phonetics and linguistics features.

The automatic speech recognition task is interdisciplinary so specialists from various scientific fields should take part in speech technology designing (engineers, mathematicians, cognitive researchers, philologists, physician, educational specialist and others). Therefore it is necessary to integrate a potential in various scientific areas such as signal processing, pattern recognition, phonetics, computational linguistics. At the same time it is related with use of large financial and temporal resources. In the circumstances Russian research teams aspire to exchange of knowledge and cooperation, so they obtain a result in fundamental and applied research. In present-day economic conditions private companies prefer supporting short-term and safe projects. As a result Russian speech recognition software market is represented by single developments.

Speech Technology Center designed VoiceCom toolkit for speech command recognition [24]. The system allows a synchronous recognition of about 100-200 commands in speaker-dependent, about 30-50 commands in speaker-independent mode and speaker-independent recognition a vocabulary of 10-20 words by telephone. Also the possibility of vocabulary extension was provided. The system may be applied to a voice control, speech enquiry messages for databases (perhaps, by telephone), a keyword search in WAV-files, an embedding of voice functions in autonomous devices – DSP programming. Advantages of the toolkit are a quick

performance of algorithms, small memory requirements, a noise adaptation, and language and accent independence. Also the investigation of Russian language models, in which base units are stems and endings, is conducted. As a result of such approach a recognition unit's vocabulary is 16 times less in contrast to common used whole word form model [25].

Company "IstraSoft" engage in speech technology developments including speech recognition and synthesis as well as speech identification by voice. A real-time algorithm of phoneme separation from continuous speech was developed here. As a result the program "IstraSoft Voice Commander" for speaker-independent recognition Russian speech commands was designed [26]. The program is intended for a voice-activated control of different Windows applications. The addition of new voices and commands to existing models is provided in a simple and convenient mode. The maximal number of commands is 45. As follow from the user manual the program is speaker-dependent with a small recognition vocabulary.

Byelorussian company "Sakrament" developed a program package Sakrament ASR Engine to design speech recognition applications with various hardware and software: IVR-systems, mobile electronic devices, domestic techniques and so on [27]. Sakrament ASR Engine module may be easy transferred to any hardware or software platform as well as tuned in to any application configuration. A quality of recognition system depends on the vocabulary size, a quality of transcription, index of recognition words tie-up, background noise level, parameters of using communication channels and microphone features.

Among introduced speech recognition systems in Russia, which were developed by west companies, it should be mark SpeechPearl [28]. It is speech recognition toolkit for telephone applications, which support Russian language. The system "Auto secretary" of CTI company, the system "Rechevoy portal" of company "Svetec", the system "Smartphone" of company "Novavox", the system "Telepat" of Institute of control problems of the Russian Academy of Sciences and other are based on this technology [29,30,31].

Thus, a survey of existing approaches and modern speech recognition systems showed that methods of statistic modeling of speech and language processes are more popular. A design of speech applications requires large-scale speech corpora. A creation of acoustical speech corpora, which also called speech databases, requires a large bankroll and time processing. In spite of existing difficulties in Russia several Russian speech corpora including about a thousand speakers recorded in various environments were collected.

4. Commercial corpora of Russian speech

One of the first Russian speech databases with labeling by phoneme units was ISABASE [32]. It was developed during 1996-1998 years by Cognitive Technologies Company with the assistance of specialists of Institute of System Analysis RAS and phonetician-experts of a speech group of the philological faculty of Moscow State University after M.V. Lomonosov. It contains 4653 speech fragments which are marked into words and phonemes using semiautomatic marking system. Phonetic alphabet is based on Avanesov transcription system [33]. Isolated and continuous speech is pronounced by 36 speakers (20 male and 26 female voices) with total duration about 8 hours. These texts were recorded in different conditions: in the room with a background noise and in the closed studio.

Today the most full Russian speech corpus is RuSpeech database [34], which contains continuous Russian speech fragments with corresponding text, phonetic transcription and information about speakers (there are 237 speakers at the age from 18 to 65 years including 127 male and 110 female voices). This database was developed by Cognitive Technologies Company for Intel Corporation during 2000-2001 years, and it is a result of the investment project to a creation of Russian speech recognition system. As a result of the project the toolkit for designing recognition systems, which includes a large Russian speech corpus, were developed for the first time in Russia.

Since 1996 "STEL – Computer Systems" company with the assistance of leading experts of the philological faculty of Moscow State University after M.V. Lomonosov, Computing center of the Russian Academy of Sciences and others is realizing a project of a creation speaker-independent Russian speech recognition system. In the framework of this project Russian speech database for recognition and synthesis was developed [35]. The acoustic-phonetic database consists of some composite parts differed by type of reading text and corresponding transcription material (tables, figures, fictions, sections of related news items, other sentences). Speakers at the age from 18 to 65 years old took part in recording; over 90 percents of them were at the age of about 50 years old. Speakers belonged to different dialect groups such as north-russian, middle-russian and south-russian dialects as well as natives of Saint-Petersburg and Moscow. In total the database contains 14 CD ROM ISO9660 with WAV-files (continuous speech of 137 speakers with total duration about 3 hours). The database was recorded in laboratory conditions – the quiet room.

In 2001 in the framework of European projects SpeechDat(II) and SpeechDat(E) Russian speech databases were developed [36]. The goal of these projects was a creation of speech databases of most European languages for speech recognition systems and an identification/verification a speaker by telephone channel. Russian database SpeechDat(E) (Eastern European Speech Databases for Creation of Voice Driven Teleservices) involves 2500 records of Russian native speakers (1242 male and 1258 female voices) which were recorded via the stationary telephone network. A volume of the database is about 60 hours. Each record contains examples of speakers' speech which are accompanied by corresponding annotations [37]. SpeechDat(E) passed certification tests of SPEX (the Netherlands). SpeechDat(II) database contains records of 1000 Russian speakers (500 male and 500 female voices). A volume of the database is about 25 hours. The database contains a speech material, which satisfies special technical (record conditions, computer file format) and linguistic-demography requirements (the records were made in five Russian regions: Moscow, Saint-Petersburg, Northern Russia, Center Russia, Southern Russia, Ural and Siberia).

In 2002 Auditech Ltd Company (Saint-Petersburg) recorded a multispeech database for VNIIEF-STL company [38]. It contains records of 102 speakers; a volume of the database is about 40 hours. Each speaker pronounced 250 sentences. The records were made in different conditions: 52 speakers in office rooms with a noise level about 40-60dB; 25 speakers in public places with a noise level about 50-

80dB; 25 speakers in a car with a noise level about 40-80dB. Each phrase was recorded by 4 microphones. Each of microphones had 4 channels recording with 16 kHz frequency and 16- bit quantification. So a good quality was provided. The corpus allows to design acoustic models in normal conditions as well as in the high noise level.

Recently company Ectaco developed Russian speech corpus (Russian Acoustic-Phonetic Continuous Speech Corpus) for commercial purposes. 46 speakers were recorded. All possible phonetic combinations, which may be found in the natural speech, were considered: each sentence contains about 70 phonemes. The total record time is about 7.75 hours.

In 2003 Linguistic Data Consortium presented Russian speech corpus West Point [39]. The corpus was developed by Department of Foreign Languages of United States Military Academy at West Point. This corpus is used by cadets of United States Military Academy, which are taking part in Russian language program, for training and development speaker-independent recognition systems. The corpus consists of 4181 speech files in SPHERE format, a volume of the corpus is about 4 hours of speech. There are 2290 files with native speaker voices and 1891 files with non-native speaker voices.

GlobalPhone[40] is a high-quality database of reading speech and texts for different languages, which is suitable for development of a speech recognition system with large vocabulary for many languages. It is successfully used both for language-independent and adaptive speech recognition. At present GlobalPhone has 15 languages including French, Japanese, Russian and others. Russian database is being developed in Minsk and will contain records of 100 speakers; each of them should pronounce about 100 phrases. The reading text was chosen from newspapers available by internet and providing a vocabulary of 65 thousand words. During the record vary speaker features were considered such as age, gender, profession and others. GlobalPhone corpus contains over 300 hours of transcribed speech of over 1500 native speakers and in the near future it will be available via ELRA.

Computing center of the Russian Academy of Sciences developed a speech corpus TeCoRus to fundamental investigations. The speech corpus (2,7 Gb) contains acoustic-phonetic (600 Mb) and speech (2,1 Gb) parts. The acoustic-phonetic part is intended for creating of basic models of speech sounds. Texts of the acoustic-phonetic part contain 3050 sentences which were read by 6 speakers. All records are two-channel (telephone + microphone). The speech part is intended for training of speech sound models, parameter adjustment, and testing of algorithms and programs for speech signals processing. It consists of interactive tests - interview and includes both reading and spontaneous pronouncing material. The speech part involves records of 100 speakers. The speech corpus is recorded on 5 CD which contain the data corpus (audio and annotation files), the pronouncing vocabulary of the data corpus and required software. The corpus may be use for training parameters of context-dependent Hidden Markov Model, a creation and testing of algorithms and programs of continuous speech recognition and digital speech signal processing as well as systems of a key-words detection in continuous speech and others [41].

5. Conclusions

The interaction between a human and a computer, which is similar to the interaction between humans, is one of the most important and difficult problems of the artificial intelligence. Existing models of speech recognition yield to human speech capabilities yet; it evidences of their insufficient adequacy and limits the introduction of speech technologies in industry and everyday life. Besides, multimodal interfaces, which combine different ways of input (speech, lips articulation, gestures, look direction and so on) are actively investigated now. The multimodal interface is natural for the inter-humans communication. Here a person can choose himself what channel is more convenient to use for information transmission at the present moment [42]. Such interfaces provide more natural and effective interaction with different automatic controls and communication facilities. For example, during the audio-visual processing the lip movement information could essentially improve speech recognition accuracy under noisy conditions.

At present multimodal interfaces are used abroad in some applied areas: mapping systems, virtual reality systems, medical systems, robotics, web-applications and so on. Besides, the multimodal interface may be used in mobile devices where use of the ordinary keyboard is impossible. In Pocket Personal Computers handwriting recognition systems are used. The combining of such systems with speech input allows to exchange the information with users more effective. The use of multimodal interfaces in Smart Phones or electronic kiosk, where it is possible to use separate input by a voice, a non-ergonomic keyboard or a sensor screen, is also actually. The optimal combination of these communicative channels allows users to exchange the information between such devices more effective and robust.

In Russia research activities of this area have started recently, and their successful realization is complicated by necessity to combine efforts of various scientific laboratories dealing with speech, video, handwriting processing and so on in different research institutes. In 2003 Speech Informatics Group of Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences has started fundamental and applied works of investigation of multimodal interfaces in the network of European scientific community SIMILAR sponsored by FP6 program of ES.

In the conclusion it is necessary to mark that investigation of problems of automatic speech recognition and multimodal interfaces is an important fundamental area. The lack of decision of the problem holds in a development of different applied systems in telecommunications, medicine, education and everyday life. Practically all modern techniques and different services use automatic controls and information processing facilities, so the development of effective means for the interaction between humans and computers is a top-priority scientific problem.

6. Acknowledgements

This research is supported by the European Community in framework of the SIMILAR NoE FP6-IST-2002#507609, INTAS Project № 04-77-7404, INTAS Project №: 05-1000007-426 and by Human Capital Foundation.

7. References

[1] Russian Grammar (1980)*, 2 vols, Moscow: Nauka.*

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

[3] Whittaker, E. W. D. Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis, Cambridge University, Cambridge (2000).

[4] Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 264 с.

[5] Трунин-Донской В.Н. Опознавание набора слов с помощью цифровой вычислительной машины. // Работы по технической кибернетике. – М.: ВЦ АН СССР, 1967. – С. 37-51.

[6] Архитектура речевого телефонного терминала МАРС-2 "Электроника МС7602"/ В.П. Афанасьев, Н.П. Дегтярев, Л.Ю Карабаева и др. // Тр. Всесоюз. шк.-семинара АРСО-14, г. Каунас , 26-28 авг 1986г. - Каунас, 1986. ч.2. - с 77.

[7] Авирн С.Б. О характеристиках надежности распознавания устных команд устройством ИКАР// Тр. Всесоюз. шк.-семинара АРСО-13, г.Новосибирск, 23-28 июля 1984г. -Новосибирск, 1984. ч.1. - с 170-180.

[8] Викторов А.Б., Жаков М.Л.,Форш Б.Н. Система распознавания дискретной речи до 1000 слов для персонального компьютера// Тр. Всесоюз. шк.-семинара АРСО-15, г. Таллинн , 13-17 марта 1989г. - Таллинн, 1989. с 314-315.

[9] Система распознавания речи ДИС-332. 03 / А.А. Горловский, Н.А. Лендяшев, Н.А. Петров и др.// Тр. Всесоюз. шк.-семинара АРСО-13, г. Новосибирск , 23-28 июля 1984г. -Новосибирск, 1984. ч.2. - с 95-96.

[10] Холоденко А.Б. Использование лексических и синтаксических анализаторов в задачах распознавания для естественных языков. // Интеллектуальные системы. Т.4, вып. 1-2, 1999, с.185-193.

[11] Соколова Е.Н. Алгоритмы лемматизации для русского языка. // Рабочий проект многоязычного автоматического словаря на 60 тыс. словарных статей. Т.1. Лингвистическое обеспечение. -М. 1984. Стр. 45-62.

[12] A. Karpov, A. Ronzhin, I. Li. SIRIUS: A system for speaker-independent recognition of continuous Russian speech. TRTU Proceedings, № 10, 2005, pp. 44-53.

[13] http://www.ipu.ru

[14] V.A. Zhozhikashvili, M.P. Farkhadov, N.V. Petukhova, A.V. Zhozhikashvili. The first voice recognition applications in Russian language for use in the interactive information systems, 9-th International Conference SPECOM'2004, St. Petersburg: "Anatoliya", 2004, pp. 304-308.

[15] Чучупал В.Я., Маковкин К.А., Чичагов А.В. К вопросу об оптимальном выборе алфавита моделей звуков русской речи для распознавания речи. Искусственный интеллект, №2, стр 575-579, "Наука і освіта", 2002.

[16] http://www.eva.miem.edu.ru/index.php

[17] http://www.isa.ru

[18] Kholodenko, A.B., "O postroenii staiisticheskih yazykovyh modelei dlya system raspoznavaniya slitnoi russkoi rechi", Intellektual'nye Sistemy (Rus.), vol 6, 1-4, MSU, Moscow, 2001.

[19] A.A. Karpov, A.L. Ronzhin. "Speech Interface for Internet Service Yellow Pages". Intelligent Information Processing and Web Mining: Advances in Soft Computing, Proc. of the International IIS: IIPWM'05 Conference, Gdansk, Poland, Springer-Verlag, 2005, pp. 219-228.

[20] Шелепов В.Ю., Ниценко В.Ю. К проблеме пофонемного распознавания // Искусственный интеллект. - 2005. - № 4. - С. 662-668.

[21] Дорохина Г.В., Павлюкова А.П. Модуль морфологического анализа слов русского языка // Искусственный интеллект. - 2004. - № 3. - С. 636-642.

[22] D. Kanevsky, M. Monkowski, J.Sedivy. Large vocabulary speaker-independent continuous speech recognition in Russian language. Proc. International Workshop SPECOM'96, St.Petersburg, Russia, pp.117-121, 1996.

[23] Баранников В.А., Кибкало А.А. Пакет программ построения систем распознавания речи. Труды III Всероссийской конференции «Теория и практика речевых исследований» АРСО-2003. Москва, МГУ им. М.В. Ломоносова, Сентябрь 2003г., с.7-12.

[24] http://speechpro.com/production/?id=471&fid=44

[25] Ilya Oparin, Andre Talanov. Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling for Russian. In Proc. of 10-th International Conference "Speech and Computer" SPECOM'2005, Patras, Greece, pp. 575-578.

[26] http://www.istrasoft.ru

[27] http://www.sakrament.com/viewprod.php?TopId=30&ProdId=24

[28] http://scansoft.com

[29] http://www.ipsoft.ru

[30] http://www.svetets.ru/portal_0.html

[31] https://www.telepat.ru

[32] www.cognitive.ru/innovation/voice-recog.htm

[33] Avanesov R.I., "Russkoe literalurnoe proiznoshenie'", (Rus), Moscow: Prosvechshenie, 1972.

[34] Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я. Современный инструментаций для разработки речевых технологий // "Информационные технологии и вычислительные системы, 2, 2004.

[35] http://www.stel.ru

[36] Galunov V.I., van den Heuvel.H. , Kochanina J.L., Ostroukhov A.V., Tropf H., Vorontsova A.V. Speech Database for the Russian Launguage. In Proc. of 3 International Conference "Speech and Computer" SPECOM'98.

[37] Project SpeechDat(E) - Eastern European Telethone Speech Database http://www.speechdat.org/

[38] www.auditech.ru

[39] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003S05

[40] http://www.cs.cmu.edu/~tanja/GlobalPhone/russian.html

[41] В.Я. Чучупал, К.А. Маковкин, Д.В. Ковков, А.В. Чичагов. Распознавание речи и диктора в системе мультимедийной идентификации личности/ Сб. Трудов Конф. Математические Методы распознавания образов, ММРО-12, Москва, 2005.

[42] A.A. Karpov, A.L. Ronzhin. Multimodal interfaces in automated control systems. Scientific journal "Instrument-making", St. Petersburg, 2005, Vol. 48, № 7, pp. 9-14.