Development of a method and software system for dialogue in real time*

Andrey Tarasiev
Institute of Radiolectronics and Information Technologies – RTF
Ural Federal University named after the first President of Russia B.N.Yeltsin
Yekaterinburg, Russian Federation
andrew4800@mail.ru

Egor Talancev
Institute of Radiolectronics and Information Technologies – RTF
Ural Federal University named after the first President of Russia B.N.Yeltsin
Yekaterinburg, Russian Federation
i.spyric@gmail.com

Konstantin Aksyonov
Institute of Radiolectronics and Information Technologies – RTF
Ural Federal University named after the first President of Russia B.N.Yeltsin
Yekaterinburg, Russian Federation
bpsim.dss@gmail.com

Olga Aksyonova
Institute of Radiolectronics and Information Technologies – RTF
Ural Federal University named after the first President of Russia B.N.Yeltsin
Yekaterinburg, Russian Federation
wiper99@mail.ru

Igor Kalinin
LLC "UralInnovation
Yekaterinburg, Russian Federation
igor_kalinin@hotmail.com

Margarita Filippova
Institute of Radiolectronics and Information Technologies – RTF
Ural Federal University named after the first President of Russia B.N.Yeltsin
Yekaterinburg, Russian Federation
rituly_22@mail.ru

Abstract

In this paper, we propose a method for recognizing the audio stream in real time by cleaning the input signals from noise, as well as speech recognition using various third-party services. At the same time, the results of testing and analysis of the quality of speech recognition by these systems are presented. Based on the obtained test results, improvements and modifications to the recognition system are proposed.

1 Introduction

Automatic speech recognition is one of the key tasks in the construction of human-machine interaction systems based on the speech interface.

The development of theoretical foundations and applied developments of question-answer systems, as well as intelligent systems with a voice interface, is an urgent scientific and technical task. Theoretical and practical approaches used in question-answer systems are actively used in search engines and application software, as well as for tasks supporting the context of dialogue with the end user.

Various software systems for developing and operating voice robots (Akkulab, Zvonbot, CallOffice, Infobot, IVONA) are presented on the market. The main disadvantages of these systems include the lack of integration with the enterprise corporate system; the use of a static dialogue scenario, the high cost of maintenance and work.

The main goal of this project is to develop a flexible and adaptive platform for the development of voice agents and text, which allow you to use various proprietary and third-party services to solve the following tasks of automatic calling and information support of context-sensitive dialogs:

1. Collection, processing and storage of dialogs;
2. Recognition of Russian speech using Google and Yandex;
3. Constructing a dialogue script;
4. Dynamic learning based on imported history of dialogs based on a neural network;
5. Control the progress of the dialogue of the voice robot.

Also an urgent task of this project is to develop an integrated approach using various methods that improve the quality of work of question-answer systems: from the stages of speech synthesis recognition to conducting a flexible and context-sensitive dialogue.

At the same time, today the most popular and at the same time the most difficult to implement systems are recognition of spontaneous speech. The complexity of constructing such systems is caused by such features as significant variability in the rate of speech and the psychophysical state of the speaker (the manner of pronouncing phrases, emotional state, coughing, stuttering), the presence of accents, or a large number of word forms used.

The task is complicated by the presence of pauses, repetitions, non-lexical inserts, parasitic words, etc. To date, a large number of speech recognition methods have been developed taking into account the described limitations of spontaneous speech, and there are also a large number of open source and commercial speech engines that can serve as the basis for such systems.

However, existing speech recognition systems have disadvantages in recognizing both speech in general and individual language units.

Thus, when constructing the speech recognition module of the developed real-time voice dialogue system “TWIN”, decisions were made based on the idea of using, adapting and finalizing existing and well-established approaches, rather than creating conceptually new algorithms.

Today, Google and Yandex systems demonstrate the highest recognition accuracy of continuous Russian speech - about 85% [1-2]. This recognition quality is provided, firstly, by huge sets of acoustic data for training (thousands of hours of speech), and, secondly, by the presence of many requested phrases and word forms from text search queries on which language models were trained. By integrating the data of the two systems, our own speech recognition system was implemented.

The recognition module of the TWIN system consists of three main subsystems:

1. Virtual PBX - Implements the functionality of making a call and routing traffic to the speech recognition subsystem;
2. Speech Recognition Subsystem - a software package whose main task is to redirect traffic to the required recognition system;
3. The decision-making module is a software package consisting of copyright algorithms for processing text information. Equipped with a decision-making routine and a speech synthesis routine.

The choice of recognition system can be pre-configured, or determined dynamically using the decision support module.

For this it is necessary to carry out minimal preparation of incoming streaming audio from the point of view of noise purification.

2 Audio Stream Preparation

The use of telephone voice signals as a direct source for recognition leads to a deterioration in the quality of the speech recognition module, which significantly reduces the effectiveness of the dialogue system. These limitations include a small bandwidth, the presence of hospitals (for example, white and pink noise) and non-linear distortions, as well as loss of information as a result of encoding a speech signal. In addition, if the person receiving the call is on the street or in a moving car, then an enormous amount of extraneous noise may be present in the audio signal, which reduces the quality of replica recognition. Therefore, in order to reduce recognition errors, a noise cleaning system was introduced. To separate the useful signal in difficult acoustic conditions, we used instruments developed at the Center for Speech Technologies LLC (MDG) [3] and described in [4-5]. The main component of noise reduction is the VAD algorithm (modification of

the algorithm based on the statistics of the fundamental tone [4-5]), which distinguishes voiced portions of speech [6]. The main idea of highlighting these sections of speech is to use vowels and nasalized consonants. On the one hand, the disadvantage is the loss of some consonants, on the other hand, explosive consonants and affricates have less identification value. Then it can be assumed that the loss of some part of insignificant speech material will be compensated by the qualitative removal of non-speech sections.

This allows, for example, to reduce the dependence of speaker identification quality on channel distortions in pauses. The developed VAD algorithm is based on the spectral analysis of a speech signal. On each frame of the spectrogram, the positions of the maxima corresponding to the harmonics of the fundamental tone are searched for, according to which the value of its frequency is estimated. In this case, the signal may lack the lower harmonics of the fundamental tone, which is typical for a telephone channel with a frequency band of 300 ... 3400 Hz. As noted in [6], such a detector has the following advantages: the speech signal is extracted, including in relatively noisy areas (signal-to-noise ratio of 10 dB and below); the continuity of the value of the fundamental tone and the belonging of this value to the range of frequency values typical of speech.

To verify the module's operability, records of previously made conversations were used, on which the system gave an incorrect answer during recognition. And at the same time, it should be noted that in some cases of the functioning of the system (low signal quality, the presence of external noise or extraneous conversations) even such methods of dealing with interference may not provide acceptable speech recognition quality.

To further improve the quality of recognition, we can distinguish a number of methods that will be implemented in the system and, in our opinion, will be able to increase the quality of the developed product. These include:

1. Dividing the audio stream into segments depending on the speaker;
2. Recognition based on context (topic and history of conversation, emotional state of the speaker, etc.);
3. Accounting for semantic errors (the meaning of the spoken phrase as a whole), and not the number of mistakenly recognized words.

3 Recognition Module Description

The decision to use both popular speech recognition systems is caused by several factors.

1. These systems are closed, which makes it impossible to rely unambiguously on the quality of recognition of each.
2. The recognition quality of individual language structures varies for these systems.
3. These systems use various internal recognition mechanisms, as a result of which they can generate the final result in different ways, which can be used for different subject areas (in the case of explicitly setting up the recognition system at the stage of designing dialogue scenarios).
4. These systems offer varying additional functionality that can also be variably applied to different areas of use.
5. These systems vary in cost of use, which allows in some subject areas to use cheaper solutions with a simpler infrastructure.

Of the items listed, the most controversial and requiring attention is the assertion about the different quality of recognition of various language structures. As a result of this, it is necessary to test the recognition quality of some basic speech structures for both systems.

For this testing, the traditional approach of analyzing the quality of recognition of isolated phrases independent of the speaker cannot be used, since such analytics cannot be representative. This is primarily due to the peculiarities of language models, the presence of paronyms, variable pronunciation of words in various situations or by different people, the presence of noise, long, difficult phrases, the presence of emotional coloring, etc.

Thus, to solve the problem, it is necessary to use an integrated approach based on a large number of experiments using simulation.

In this case, real dialogue scenarios over time will be simulated.

4 Experiment setup

Based on the foregoing, the following criteria for the quality of speech recognition can be distinguished for tests:

• Percentage of recognized short emotional phrases.

• Percentage of recognized long phrases.
• Percentage of recognized domain-specific terms.
• Percentage of recognized proper names.
• Percentage of recognized simple numbers.
• Percentage of recognized complex numerals.
• Percentage of recognized dates, addresses, and other audio information containing numerals.
• Percentage of recognized speech in noise and other distortions.

The problematic issue in the context of this task is the way to build models based on the implementation of specific dialogue scenarios. Traditional modeling systems do not have such functionality in their implementation, due to the narrow focus of this problem.

The TWIN system has in its implementation a module for visual configuration of dialog scripts - scripts (Figure 1) [7-10].

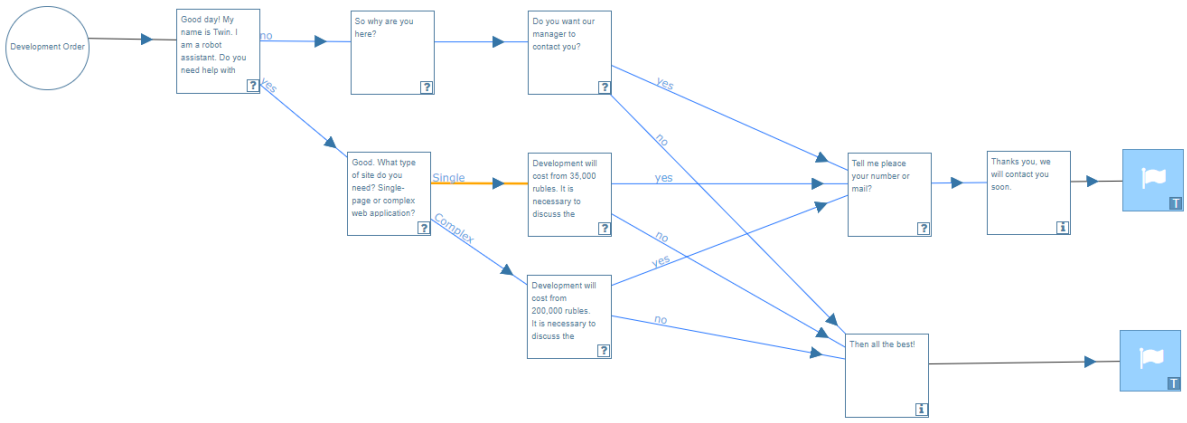

Figure 1: Example of dialogee script

To solve the problem of organizing simulation experiments, a specialized complex was developed for generating (reproducing dialogs) on the basis of technologies already existing in the system.

The idea was to organize automatic dialogs according to previously described possible scenarios between robots. That is, it is necessary to compose several pairs of scenarios for the study of each selected criterion.

Based on these technologies, several scripts have been developed for modeling and testing selected lexical units. Particular attention was paid to the quality of recognition of numerals, dates and addresses.

The various question-answer speech recognition systems under consideration were connected to this module.

The modeling process consisted of multiple automatic runs of pre-prepared audio files for the first script, which were pre-tested by experts on selected lexical categories. At the same time, recognition settings were changed. As the dialogue progressed according to the scenario based on the recognition of the second participant in the conversation, the dialogue went into the desired predicted scenario branch or not. Based on statistics at the end of the simulation, results were recorded.

5 The discussion of the results

Based on the information received, it can be concluded that the Yandex company system better recognizes short expressive phrases, as well as numerals, while the Google API better recognizes long phrases and terms.

At the same time, both systems have problems with recognition in noise, hectic speech rate and voice defects of the interlocutor.

This is due to the fact that both systems better recognize phonemes and phrases, and worse individual sounds, especially in the case of noise and other factors that distort the quality of the transmitted audio message. This observation is confirmed by the conclusions obtained by other independent researchers [11].

Based on the obtained data, an algorithm for the operation of the speech recognition system that is part of the TWIN complex was formed. The system also includes a module for the subsequent processing of recognized text - normalization, highlighting keywords, etc. The use of this module greatly simplifies the final perception by the robot of the phrase uttered by the interlocutor and the choice of subsequent actions (pronouncing the corresponding remarks) provided for by the given script [12].

The proposed method was implemented in the modules of preliminary and subsequent third-party recognition of processing. The resulting speech recognition module was also tested using previously used models.

The recognition quality in this case has increased in terms of the tested indicators. Table 1 shows the statistics for phrase recognition by the TWIN system of phrases by category.

Table 1: Phrase recognition statistics by the TWIN system by category

Compare the quality of recognition	Type of recognized phrase	Count of phrases tested by module	Count of phrases recognized correctly	Number of errors in recognition of sounds, syllables	Number of word shape recognition errors	Number of full word recognition errors	Percentage of recognition errors
1	Short emotional phrases	4200	3402	187	529	82	19
2	Long phrases	3699	3107	104	471	17	16,00432549
3	Terms that are specific to the subject area	1216	1143	31	35	7	6,003289474
4	Own names	2540	2159	249	46	86	15
5	Simple Numerals	5366	5205	25	62	74	3,000372717
6	Complex numerals	2200	1958	95	59	88	11
7	Dates, addresses and other audio information containing numerals	1289	1006	163	86	34	18,00254453
8	Phrases in noise and distortion conditions	1659	553	436	391	279	40

The recognition quality in this case increased relative to the best indicators of third-party services in the following indicators:

- The percentage of recognized short emotional phrases is 5%.
- The percentage of recognized long phrases is 3%.
- The percentage of recognized terms specific to the subject area is 1%.
- The percentage of recognized proper names is 1.5%.
- The percentage of recognized simple numerals is 3%.
- The percentage of recognized complex numerals is 2%.
- The percentage of recognized dates, addresses and other audio information containing numerals is 4%.
- The percentage of recognized phrases in noise conditions is 1%.

6 Conclusion

The speech recognition module in the TWIN system uses integration in its work by the two most developed currently existing solutions YandexSpeechKit and GoogleSpeech API.

Based on the use of simulation, testing of the used speech recognition systems of the speech recognition module was carried out. For this, an additional dialog playback module was implemented.

Based on the information received, we can conclude that the Yandex company system better recognizes short expressive phrases, as well as numerals. In contrast, the Google API recognizes longer phrases and terms better.

Based on the information received, pre-processing modules, dynamic selection of the recognition system, and subsequent processing of the recognized text were created. The recognition quality of the integrated solution - the speech recognition module of the TWIN system has increased significantly.

System development involves the development and implementation of additional functions, such as sending statistics and creating tips when compiling a script. Analysis of statistics will help identify priority areas for improving the interface.

The range of use of the system can be expanded due to the initial design flexibility.

References

1. “Speech Kit Cloud”, Speech Kit Cloud, 2019. [Online]. – URL: https://tech.yandex.ru/speechkit/cloud/ (Accessed: 03.11.2019).
2. “SpeechKi”, Tech.yandex.ru, 2018. [Online]. – URL: https://tech.yandex.ru/speechkit/ (Accessed: 03.11.2019).
3. T. Mikolov, K. Chen, G.S. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. { 2013.
4. I.P Medennikov. Methods, algorithms, and software for recognizing Russian telephone spontaneous speech: dissertation of a candidate of technical sciences: 05.13.11 / Medennikov Ivan Pavlovich {Place of defense: St. Petersburg State University, 2016
5. I. B Tampel., A. A. Karpov. Auto Speech Recognition {138, St. Petersburg: ITMO University, 2016.
6. E.D. Loseva, L.V. Lipinsky. Recognition of human emotions by spoken using intelligent data analysis methods { Actual problems of aviation and astronautics, 2016.
7. K. Aksyonov, D. Antipin, T. Afanaseva, I. Kalinin, I. Evdokimov, A. Shevchuk, A. Karavaev, O. Aksyonova, U. Chiryshe. Testing of the speech recognition systems using Russian language models {5th International Young Scientists Conference on Information Technologies, Telecommunications and Control Systems, ITTCS 2018. Yekaterinburg, Russian Federation, December 2018
8. K. Aksyonov, E. Bykov, O. Aksyonova, N. Goncharova, A. Nevolina. Extension of the multi-agent resource conversion processes model: Implementation of agent coalitions {5th International Conference on Advances in Computing, Communications and Informatics, 2016
9. K. Aksyonov, E. Bykov, E. Sysoletin, O. Aksyonova, A. Nevolina. Integration of the Real-time Simulation Systems with the Automated Control System of an Enterprise {International Conference on Social Science, Management and Economics, 2015
10. K. Aksyonov, I. Kalinin, E. Tabatchikova, U. Chiryshev, O. Aksyonova, E. Talancev, A. Tarasiev, V. Kanev. Development of decision making software agent for efficiency indicators system of IT-specialists {5th International Young Scientists Conference on Information Technologies, Telecommunications and Control Systems, ITTCS 2018. Yekaterinburg, Russian Federation, December 2018
11. D.V. Bobkin, K.Y. Zhigalov. The study of the reliability of speech recognition by the system Google Voice Search. {Volume 2 / Cloud of Science. 2015
12. A. Tarasiev, E. Talancev, K. Aksyonov, I. Kalinin, U. Chiryshev, O. Aksyonova. Development of an Intelligent Automated System for Dialogue and Decision-Making in Real Time {2nd European Conference on Electrical Engineering & Computer Science (EECS 2018). Bern, Switzerland, December 2018