## HYBRID ACOUSTIC MODELS FOR REMOTE AND MULTI-CHANNEL SPEECH RECOGNITION WITH A LARGE VOCABULARY

**Gorinsky Stepan Alexandrovich**

Undergraduate, International University of Information Technologies, Kazakhstan, Almaty

**Nurtas Marat**

научный руководитель, Associate professor, PhD, International University of Information Technologies, Kazakhstan, Almaty

# ГИБРИДНЫЕ АКУСТИЧЕСКИЕ МОДЕЛИ ДЛЯ ДИСТАНЦИОННОГО И МНОГОКАНАЛЬНОГО РАСПОЗНАВАНИЯ РЕЧИ С БОЛЬШИМ ЛЕКСИКОМ

***Горинский Степан Александрович***

*магистрант, Международный университет информационных технологий, РК, г. Алматы*

***Марат Нуртас***

*Международный университет информационных технологий, РК, г. Алматы*

The issues of human-machine interaction are among the most important when creating new computers. The most effective means of human-machine interaction would be those that are natural to him: through visual images and speech. The creation of speech interfaces could find application in systems of various purposes[1]: voice control for people with disabilities, reliable control of combat vehicles that "understand" only the commander's voice, answering machines that automatically process hundreds of thousands of calls per day (for example, in the air ticket sales system), etc. At the same time, the speech interface should include two components: an automatic speech recognition system for receiving a speech signal and converting it into text or command, and a speech synthesis system that performs the opposite function – converting a message from a machine into speech.

However, despite the rapidly increasing computing power, the creation of speech recognition systems remains an extremely difficult problem. This is due to both its interdisciplinary nature (it is necessary to have knowledge in philology, linguistics, digital signal processing, acoustics, statistics, pattern recognition, etc.) and the high computational complexity of the developed algorithms. The latter imposes significant restrictions on automatic speech recognition systems – on the volume of the dictionary being processed, the speed of receiving an answer and its accuracy. It is also impossible not to mention that the possibilities of further increasing the speed of the computer due to the improvement of integrated technology will sooner or later be exhausted, and the increasing difference between the speeds of memory and processor only exacerbates the problem.

There are areas of application of automatic speech recognition systems where the described problems are particularly acute due to severely limited computing resources, for example, on mobile devices[2]. Manufacturers of mobile phones and tablets have found a way out in transferring resource-intensive computing from users' devices to servers in the cloud, where, in fact, recognition is performed. The user application only sends speech requests there and accepts responses using an internet connection. Apple's Siri and Google* Voice Search systems work successfully according to this scheme [3,4]. However, for such implementation, certain conditions are necessary, for example, continuous access to the Internet, which in some cases are unattainable, and it is necessary to create a compact and reliable independent device that uses only available "on-site" computing power. The described difficulties arise when creating intelligent devices in both the military and civilian spheres. An example of such devices is the REX robot, developed by the Israeli concern Israel Aerospace Industries. REX is designed to transport ammunition, food and other ammunition, which allows you to unload a soldier. At the same time, the robot is able to follow the person leading it, and it is controlled entirely by voice commands. Another example of the active use of speech recognition technologies in combat complexes is the introduction of voice control modules (or direct voice input – Direct Voice Control) into the cockpits of modern fighters, such as Eurofighter Typhoon1, Dassault Rafale2, JAS 39 Gripen. This made it possible to significantly unload the pilot so that he could focus only on completing the task. In the non-military sphere, speech recognition is widely implemented in the automotive industry (for example, BMW, Ford), when part of the functionality of the machine for which a recognition error will not lead to emergency situations (climate control, navigation, multimedia, etc.) is controlled by voice 3. As with the use of voice control in military aircraft, this technology made it possible to remove part of the load from the driver so that he could focus only on the road. Finally, it is necessary to note the relevance of the implementation of the speech interface for people with disabilities, for example, in wheelchairs.

All the examples described above are united by the need to create a compact, reliable, independent and maximum-speed device. A lot of specialists are working on solving this problem. The following areas of research and development can be distinguished in the field of improving the performance and implementation of independent speech recognition modules:

1. Implementation of hardware support for algorithms for preprocessing and feature extraction (for example, implementation in programmable logic chips) of the block for finding small-scale coefficients);

2. Hardware implementation of recognition algorithms.

**The latter direction is represented by many works.** At the same time, there is a general trend in the development of hardware implementations of the recognition block: firstly, user-programmable logic is used as chips due to their availability and versatility, and secondly, they are all focused on introducing hardware support for algorithms of hidden Markov models - the forward running algorithm and the Viterbi algorithm. As a prerequisite for this, the high computational complexity of the designated algorithms is indicated. To solve this problem, in [10] it is proposed to construct a systolic matrix for performing calculations using forward running and Viterbi algorithms;

Thus, the search for new architectural solutions that are not based on von Neumann architecture is an urgent topic, especially in its application to solving artificial intelligence problems, which include speech recognition. One of the promising directions is the development and research of associative environments and the construction of heterogeneous cellular automata using them.

Associative access is carried out, as opposed to addressable, by the content of the information, and not by its address in the storage environment. This allows you to process information directly in the logical-storage environment, and the time of associative search practically does not depend on the storage capacity. At the same time, the associative oscillatory medium consists of simple cells, each of which has its own law of functioning, and together these cells produce streaming information processing.

**The aim of the work** is to develop methods for remote and multi-channel speech recognition with a large vocabulary.

To achieve this goal, the following **tasks are solved:**

- selection of the method of speech extraction and preprocessing, extraction of features;

- software implementation of speech extraction and its preprocessing;

- choice of speech recognition method;

- choosing an associative environment for implementing recognition in it;

- development of a recognition block based on elements of an associative environment;

- creation of a speech base for training and testing the system;

- creation of a software model of the developed speech recognition methods in the environment;

**The object of research** is speech recognition methods and methods of their hardware support.

**The subject of the research** is methods and algorithms of speech recognition and ways of their implementation.

**The scientific novelty of the work** consists in the following:

- a method has been developed for the allocation of speech sites based on the analysis of the distribution of local extremes;

- the recognition algorithm has been modified by switching to a simplified calculation of the logarithm of the probability value, which made it possible to replace multiplication operations with addition. On the basis of the cell ensemble "Differential", a new cell ensemble "Comparator" was built, which selects the flow of spikes with maximum intensity. Thanks to this, it was possible to fully implement it on the elements of the associative oscillatory environment and successfully apply it to the recognition of Russian words;

- a new recognition method has been developed based on the selection of the most suitable hidden Markov model without taking into account the order of sounds in the pronunciation of the word. This made it possible to simplify the hardware implementation performed on the elements of the associative oscillatory medium and increase the recognition speed. The method was successfully applied to recognize Russian words;

**The practical value of the work consists in the following:**

- a method has been developed for identifying areas with speech in the original signal;

- speech recognition methods have been developed;

- an experimental speech database of Russian words has been formed, which can be used both for training recognition systems and for testing them;

- a software package has been developed that includes tools for compiling a speech database and software models of the proposed speech recognition methods in the environment.

- the practical application of the proposed implementation of speech recognition on the example of recognition of Russian words is investigated;

- hardware implementation has been developed;

**During the work on the dissertation**, the following research methods were used: methods of designing and analyzing software tools, simulation modeling, probability theory and mathematical statistics.

The validity of the scientific results and conclusions presented in the work is determined by the correct application of the research methods used. The reliability of scientific statements, conclusions and practical recommendations formulated in the dissertation is confirmed by computational experiments and data obtained during simulation modeling.

**Approbation of the work.** The main results of the work were reported at international scientific and technical conferences.

**The structure of the work.** The master's thesis consists of an introduction, three chapters, a conclusion and a list of references.

Three steps are differentiated in the functioning of the automated speech recognition system (SARR): feature selection, training, and recognition. The original signal is used to create a feature vector, which is a compressed representation of the speech signal that only contains the information needed for recognition. Methods that work both in the frequency domain (mel-kepstral coefficients, linear prediction coefficients) and in the temporal domain (for example, on a short-term energy value) are used to accomplish this, while the problem of speech representation remains unsolved and research is ongoing, including by the authors of this paper [27, 28]. An acoustic or observable sequence O  (o1, o2, ... , oT).  A person sends a chain of words W  (w1, w2, ... , wN). The goal of speech recognition is to locate a string of words W that correlates to the auditory sequence O [26, 29, 30].

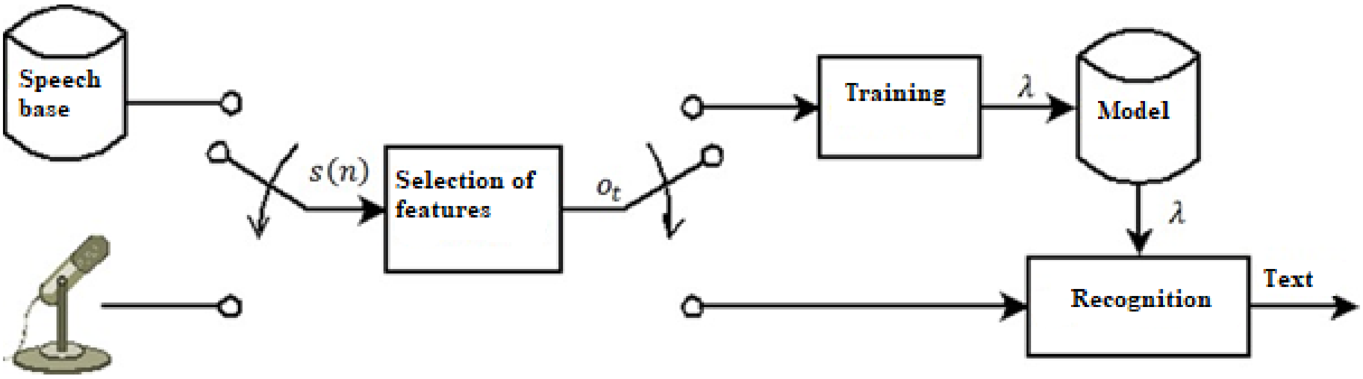

***Figure 1 .The general scheme of the automatic speech recognition system***

To overcome this challenge, at the training stage, a model is created that can generate all possible sequences O for every possibilities W. Allow the function (W, ) to return all potential O for a single W. The recognition will then consist of identifying a string of words that, according to the model, will produce the audio sequence that is closest to the one in question:

$$\boldsymbol{W}^{*} = ArgMin_{W\in}((\boldsymbol{W}, \lambda), \boldsymbol{O})$$

where $d(\boldsymbol{O'}, \boldsymbol{O})$ is the distance between $\boldsymbol{O'}$ and $\boldsymbol{O}$.

As a result, you should verify all the chains of words W in theory, which is obviously impossible in practice. To make this process easier, certain limitations are imposed using the language's grammar, or a narrower objective is handled, such as recognition of solely isolated words.

Furthermore, both SARR blocks – feature selection and recognition – are examined in greater depth.

*(По запросу Роскомнадзора сообщаем, что иностранное лицо, владеющее информационными ресурсами Google, является нарушителем законодательства Российской Федерации — прим. ред.)

**References:**

1. A. Ronzhin, A. Karpov, Russian Voice Interface // Pattern Recognition and Image Analysis, 2007. Vol. 17, No. 2, pp. 321–336.

2. A. Schmitt, D. Zaykovskiy, W. Minker, Speech recognition for mobile devices// International Journal of Speech Technology, Springer, 2008. Vol. 11, Pp. 63-72.

3. M. Pinola, Speech recognition through the decades: how we ended up with Siri, [Электронный ресурс] // PCWorld, 2011. URL: http://www.techhive.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html?page=0.

4. B. Johnson, How Siri works, [Электронный ресурс] // How Stuff Works. URL: http://electronics.howstuffworks.com/gadgets/high-tech- gadgets/siri2.htm.