# Speech recognition algorithm for natural language management systems under variety of accents

*Irina* Gurtueva[1,*], *Olga* Nagoeva[1] , and *Inna* Pshenokova[1]

[1]Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, Institute for Computer Science and Problems of Regional Management, 37a, I. Armand Str., 360000, Nalchik, Russia

**Abstract.** This paper proposes a concept of a new approach to the development of speech recognition systems using multi-agent neurocognitive modeling. The fundamental foundations of these developments are based on the theory of cognitive psychology and neuroscience, and advances in computer science. The purpose of this work is the development of general theoretical principles of sound image recognition by an intelligent robot and, as the sequence, the development of a universal system of automatic speech recognition, resistant to speech variability, not only with respect to the individual characteristics of the speaker, but also with respect to the diversity of accents. Based on the analysis of experimental data obtained from behavioral studies, as well as theoretical model ideas about the mechanisms of speech recognition from the point of view of psycholinguistic knowledge, an algorithm resistant to variety of accents for machine learning with imitation of the formation of a person's phonemic hearing has been developed.

## 1 Introduction

The acoustic-phonetic approach, which was developed by Hemdal and Hughes in 1967, was the first attempt to formalize the processes of speech perception [1]. Its main postulate was in the following, each characteristic meaning-distinguishing unit of speech corresponds to certain acoustic properties. That is, decoding a message is possible based on research data on phonetics and linguistics. The recognition procedure includes three stages. At the first stage, spectral analysis of speech and detection are carried out, which convert spectral measurements into a set of properties that describe phonetic units. The next step is segmentation and layout. The speech signal is segmented into stable acoustic regions to which tags are attached. The result is a characteristic speech grid. At the final step, an attempt is made to determine a significant word (or sequence) from a sequence of phonetic labels obtained as a result of segmentation and markup. In the decoding process, linguistic restrictions (according to the dictionary, syntactic or semantic rules) corresponding to the task are used to evaluate the meaning of the word. The effectiveness of this approach is low due to the fact that formalizing the laws that determine the wide acoustic variability of

* Corresponding author: gurtueva-i@yandex.ru

speech is not a trivial task. The acoustic-phonetic approach was not used in commercial applications, but a deep analysis of the subject area certainly stimulated the development of this field of knowledge.

The approach called pattern recognition supposes that the identification of an unknown utterance is possible on the basis of its direct comparison with the previously prepared vocabulary standard of the system's internal library [1, 2]. Itakura proposed this approach in 1975 and later Rabiner summarized it. The speech pattern can be stored as a speech standard (template approach) or statistical model (stochastic approach) and applied to a word, sound or phrase. A collection of prototypical speech patterns is stored as standards, representing a system codebook. An unknown speech utterance is compared with each of the standards and the category of the best match is selected. Typically, a template is constructed for each word in the dictionary. The idea of the described approach attracts with its clarity, but taking into account the variability of the speech signal corresponding to the environment, speaker, type of microphone and recording devices, there is a need to create several standards for each word. Since preparation and approval of standards is an expensive procedure, and the size of the codebook becomes too voluminous for storage, these systems are impractical for creating systems with large dictionaries. In addition, the sequence of template patterns does not take into account pauses and the effect of co-articulation between words.

The stochastic approach uses probabilistic models to remove uncertainties or incomplete information. The system selects the sequence of words that will be pronounced with the greatest probability (hidden Markov models). The Markov process [1, 2] is an automaton with a finite number of states for which a matrix of probabilities of transition from one state to another is defined. Probabilities depend on a pair of numbers of consecutive states. For each pair of successive states, a function is also defined that evaluates the probability of observing a reference element $x$ from the codebook of language $X$ in state $j$ if the previous state was state $i$. If we neglect the effect of coarticulation, then we can assume that the probability functions do not depend on the previous state, but only on its number. The Markov model was developed for the analysis of written texts. In the late 70s, attempts began to adapt it to speech. The main drawback of Markov models is the fact that they make a priori assumptions that are not accurate and, therefore, reduce the efficiency of the system.

The connectionist approach [1-3] initially proceeded from the ideas of neurobiology, but later this approach became more interdisciplinary, covering computer science, engineering, mathematics, physics, psychology and linguistics. There is a wide variety of connectionist models with different architectures, but all of them used general principles [4, 5]. Artificial neural networks consist of a potentially large number of simple processing elements (called units, nodes, neurons) that interfere with their behavior through exciting and inhibiting network connections. The weakness of artificial neural networks is that it is a discriminatory model. This model successfully contrasts phonemes, notes their differential properties [5], but does not identify integral ones (allophones). In particular, for the same reason, the same sound in laboratory conditions and in noisy conditions the neural network is marked as different.

Summarizing the above, we have to admit that existing approaches have a number of fundamental limitations that do not allow speech recognition with high reliability. This is confirmed by the fact that although there are currently a large number of speech applications on the market with high recognition accuracy, when transferring from laboratory to real operating conditions, almost all existing systems are unstable to noise and useless in the case of a "cocktail party" [6], as well in a conference or meeting setting.

Thus, there is a need to develop a fundamentally new approach based on other mathematical methods for solving speech recognition problems.

Since a person recognizes speech more accurately than any modern computer, we believe that the most promising approach is modeling a person's cognitive abilities. This will allow to build a speech recognition system in an environment with several speakers based on a simulation of the attention mechanism. That is, this approach will allow to create a speech recognition system that automatically focuses on a specific speaker or topic of interest.

## 2 Materials and methods

Every speech message has information superfluity [7] and can be studied from the standpoint of various scientific disciplines. The results of psychoacoustic studies confirm the fact that a person actively uses supporting non-speech information when decoding a signal [8]. That is why in order to increase the reliability of recognition or to correct results when designing speech systems, video cameras and photo sensors have been used to analyze information on articulation, sensors that record the electrical activity of the skin, to enter information about the emotional and personal state of the speaker, etc. [1, 3]. Thus, to solve the problem of speech recognition successfully, it is necessary to analyze speech as a whole of all these aspects.

Cognitive modeling considers speech recognition as a task of multi-level pattern recognition. The fundamental foundations of these developments and the method of training multiagent neuro-like systems based on ontoneuromorphogenesis are described in detail in [9-15]. This approach is based on the theoretical foundation of cognitive psychology and cognitive neuroscience [16-20], as well as on modern achievements of computer science [21, 22].

The cognitive architecture of the automatic speech recognition system [23-25] considers the function of the auditory analyzer in the form of sequential processing of audio information at the following levels: preliminary recognition (level I), subconscious recognition (II level), conscious recognition (III level), situation level (IV level). The structure and functions of each level are determined by the sets of agents, actors and the systems of their contractual relationships.

The first level of architecture - preliminary recognition - is a system of software agents that record the acoustics of a signal, like human auditory receptors, and analyze multi-aspect sound information. Each layer forms its own signatures in order to describe a set of acoustic properties of fixed sound images specific for a given functional differentiation.

At the next level - subconscious recognition - the signatures of the previous layer are grouped around significant objects and actions. We believe that for reliable speech recognition it is necessary to establish a connection between the spectral characteristics of the signal and the "articulation event" underlying it. For each primitive that does not occur before, an agent is allocated that identifies the events and actions that became the source of sound.

At the third level - the level of conscious recognition - significant events of the current priority are identified, determined on the basis of the work of the so-called. cognitive emotion assessment [9] - a functional unit of a multi-agent recursive cognitive architecture containing a priori information acquired on the basis of training on the degree of significance of events for the implementation of the target function of an intelligent agent. Agents of the following levels of multi-agent recursive intelligence - goal-setting cognition and synthesis of action plans - form control commands for the effectors of fine-tuning the filtering system and amplifying the acoustic parameters of integration into the afferent path control of the observing apparatus and adjustment of afferent paths.

At the fourth level, where a situation is formed, the sound element is linked to the general context, including extra-linguistic connections.

Thus, the proposed cognitive architecture allows us to include in the signal analysis procedure all aspects of the speech message, including the extralinguistic component [23-25], expressed in this approach in terms of events and situations. Each subsequent level of signal structuring into a hierarchy of word elements, words, phrases, etc. has additional time limits, for example, known pronunciation of words or allowed sequences of words that can compensate for errors and uncertainties at lower levels. A hierarchy of constraints is used to organize the interaction between the decision-making level and on the basis of contractual relations.

## 3 Model and results

The speech stream is recorded by a system of microphones. Then its spectral composition is revealed using the short-term Fourier transform [1]. Then, the YIN method [26], similar to the autocorrelation function [2], is used to estimate the pitch frequency, since this method is most effective for extracting the pitch frequency of monophonic musical instruments and speech. The composition of harmonics is determined by the two-way mismatch method [27]. Thus, at the preprocessing stage, the audio signal is converted to the following set of signatures:

$$\langle F_0, F_1, F_2, \Delta t \rangle, \tag{1}$$

where $F_0$ is the pitch frequency, $F_1, F_2$ - the first and second formants, $\Delta t$ is the duration of the sound of the investigated phoneme.

This vector of features is fed to the input of multi-agent recursive cognitive architecture [9], in which the developers preformed a set of so-called neuro-factories - agents of a special type that dynamically create agents upon request, determine their type and place them in the corresponding space area of the multi-agent system. So, the neurofactory creates an agent responsible for some phoneme. To create such an agent, a special program is used. It reads the agent's genome - the starting set of production rules in the agent's knowledge base [9]. Then, the agent, realizing the behavior determined by the rules recorded in its own genome, in order to find the class to which it belongs, questions the expert. Based on the expert's response, a contract is concluded between the actor and the agent characterizing this class, that is, controlled machine learning is implemented [28]. The training process involves supplying a full set of training material to the input of the system, which is selected taking into account the characteristics of motherese [29]. As psycholinguistic studies show, one of the universal signs of speech addressed to children is the duration of vowel phonemes. Vowel phonemes last more than 3.5 times longer than vowels in speech addressed to adults [29, 30]. The multi-agent system, based on a comparative assessment of the duration of the analyzed phoneme with the existing statistical estimates of the longitude of the sound of baby talk [29] phoneme, assigns an additional feature to the agent - emotional coloring, and also identifies it as a prototype/non-prototype. The prototype is a binary feature, and emotional coloring is determined in the range from 0 to 1.

Thus, at the first level of cognitive architecture, the knowledge base of an agent characterizing a phoneme is composed of the following set of attributes:

$$\Phi_{vowel} = (\langle F_0, F_1, F_2, \Delta t \rangle, \mathrm{K}_{letter}, \varepsilon, P), \tag{2}$$

where $F_0$ is the frequency of the fundamental pitch, $F_1 F_2$ is the first and the second harmonics, $\Delta t$ is the duration of phoneme sound, $\mathrm{K}_{letter}$ – classifying contractual relationship with an agent-letter, $\varepsilon\epsilon[0,1]$ – emotional coloring, $P(0,1)$ – prototype/non-prototype, εϵ [0,1] - emotional coloring.

It is important to note that the lifetime of an agent in the system is determined by the step function of memory, that is, a long period of inactivity of the agent leads to its death, expulsion from the system. The settings of the agent's life expectancy parameter will allow to study the problems of mastering native language, as well as the problems of perceiving foreign speech [8].

As Figure 1 shows, the result of the functioning of the first level of the architecture under development is the creation of agent-actors that record the acoustic characteristics of the signal, like human auditory receptors do, and the formation of sets of agents corresponding to each minimal speech unit of a language. The proposed algorithm allows to reduce the severity of the problem of high variability of speech with respect to the variety of accents. Selecting training material with the high linguistic diversity characteristic of mountain enclaves, one can track the mechanism of formation of human auditory patterns and take into account speech features associated not only with individual physiology, but also with a wide variety of accents.

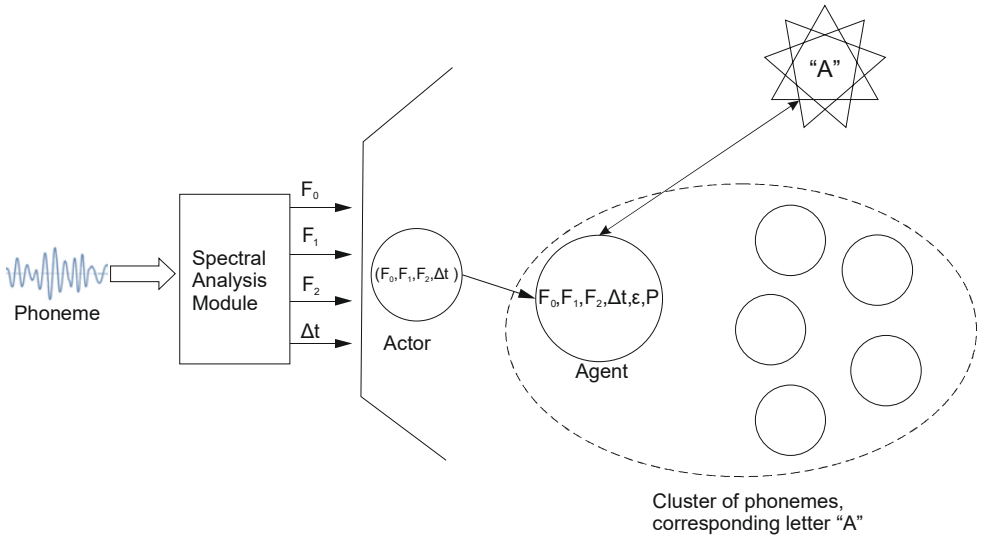


**Fig. 1.** The first level for the multi-agent system (MAS) of speech recognition.

At subsequent levels, to speed up the recognition process, it is planned to apply phonological, grammatical restrictions [7]. It is also planned to introduce feedbacks for correction and refinement of decoding results.

Thus, based on an analysis of experimental data from behavioral studies and model ideas about speech recognition mechanisms from the point of view of psycholinguistic knowledge, a machine learning method with imitation of the formation of a person's phonemic hearing was developed.

## 4 Conclusion and Discussion

Elements of a speech recognition system [25] based on a multi-agent recursive cognitive architecture have been developed, allowing to take into account the linguistic and extra-linguistic components of speech communication, which fundamentally distinguishes the agent-based approach from artificial neural networks, which are discriminative models.

An algorithm for machine learning with an imitation of the mechanism of the formation of the phonemic hearing of a person is proposed, which allows you to create speech systems that are resistant to the variety of accents that arose in mountain enclaves as a

result of using the Russian language as a means of interethnic communication. Which, undoubtedly, contributes to the accelerated development of the agro-industrial complex of the Caucasus.

The research was supported by the Russian Foundation of Basic Research, grants No. 18-01-00658, 19-01-00648.

## References

1. A. Waibel, K.-F. Lee, *Readings in Speech Recognition* (Morgan Kaufman, Berlington, 1990)
2. D. Jurafsky, J. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Prentice Hall, Boston, 2008)
3. V. Gupta, IJCSET **5(1)**, 14-16 (2014) http://www.ijcset.com/docs/IJCSET14-05-01-029.pdf
4. W. De Mulder, S. Bethard, M.-F. Moens, CS&L **30**, 61-98 (2015) https://doi.org/10.1016/j.csl.2014.09.005
5. L. Deng, X. Li, IEEE TASLP **21**, 1060-1089 (2013) DOI: 10.1109/TASL.2013.2244083
6. E.M. Zion Golumbic, N. Ding, S. Bickel, Neuron **77(5)**, 980-991 (2013) doi: 10.1016/j.neuron.2012.12.037
7. L.R. Zinder, *General Phonetics* (Higher School, Moscow,1979)
8. V.P. Morozov, I.A. Vartanyan, V.I. Galunov, *Perception of speech: Questions of functional asymmetry of the brain* (Nauka, Leningrad, 1988)
9. Z.V. Nagoev, *Intelligence, or thinking in living and artificial systems* (Publishing House of KBSC RAS, 2013)
10. Z.V. Nagoev, O.V. Nagoeva, *Materials of the XII All-Russian Scientific and Practical Conference "Perspective Systems and Control Problems"* (2017)
11. Z.V. Nagoev, O.V. Nagoeva, Izvestiya KBSC RAS **6**, 73-85 (2015); https://www.elibrary.ru/download/elibrary_25028776_80518166.pdf
12. Z. Nagoev, I. Pshenokova, I. Gurtueva, K. Bzhikhatlov, Springer's AISC **948** (2019) https://doi.org/10.1007/978-3-030-25719-4_48
13. Z. Nagoev, O. Nagoeva, I. Gurtueva , V. Denisenko, Springer's AISC **948** (2019) https://doi.org/10.1007/978-3-030-25719-4_49
14. Z. Nagoev, O. Nagoeva, I. Gurtueva, CSR **59**, 91-102 (2020) https://doi.org/10.1016/j.cogsys.2019.09.015
15. Z. Nagoev, O. Nagoeva, I. Pshenokova, I. Gurtueva, ICR, Springer's LNCS **116592019** (2019) https://doi.org/10.1007/978-3-030-26118-4_24
16. N.A. Chomsky, *A Review of Skinner's Verbal Behavior* (Prentice-Hall, 1967)
17. A. Newell, *Unified Theories of Cognition* (Harvard University Press, Cambridge, Massachusetts, 1990)
18. M.S. Gazzaniga, *Conversations in the Cognitive Neuroscience* (The MIT Press, Cambridge, 1996)
19. P. Haikonen, *The Cognitive Approach to Conscious Machines* (Exeter imprint Academic, UK, 2003)

20. D. H. Schunk, *Learning Theories: An Educational Perspective* (Pearson Merrill, Prentice Hall, Boston, 2011)
21. M. Wooldridge, *An Introduction to Multi-Agent Systems* (Wiley, Hoboken, 2009)
22. I. Kotseruba, J.K. Tsotsos, arxiv.org/abs/1610.08602 (2016)
23. Z. Nagoev, L. Lyutikova, I. Gurtueva, PCS **145**, 386-392 (2018) http://doi.org/10.1016/j.procs.2018.11.089
24. Z. Nagoev, I. Gurtueva, D. Malyshev, Z. Sundukov, Springer's AISC volume **948**, (2019) https://doi.org/10.1007/978-3-030-25719-4_47
25. Z.V. Nagoyev, I. Gurtueva, Izvestiya KBSC RAS **3**, 3-14 (2019) DOI: 10.35330/1991-6639-2019-3-89-3-14
26. A. De Cheveigne, H. Kawahara, JASA **111**, 1917 (2002) https://doi.org/10.1121/1.1458024
27. R.C. Maher, J.W. Beauchamp, JASA **95**, 2254 (1994) https://doi.org/10.1121/1.408685
28. A. Coates, Y. Andrew, LNCS **7700** (2012) https://doi.org/10.1007/978-3-642-35289-8_30
29. W. Strange, *Speech Perception & Linguistic Experience*: *Issues in Cross-Language* (York Press, 1995)
30. S.N. Zeitlin, *Language and child: Linguistics of children's speech: Textbook. allowance for students. higher study, institutions* (Humanity VLADOS Center, Moscow, 2000)