Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 145 (2018) 386–392

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society)

Model for Automatic Speech Recognition Using Multi-Agent Recursive Cognitive Architecture

Zalimhan Nagoev[a], Larisa Lyutikova[a], Irina Gurtueva[a]*

[a]*The Federal State Institution of Science Federal Scientific Center Kabardino-Balkarian Scientific Center of Russian Academy of Sciences, I. Armand Street, 37-a, Nalchik, 360000, Russia*

Abstract

A concept of a fundamentally new approach to the development of speech recognition systems is proposed, as applications built based on existing approaches are not effective enough when used in noisy conditions and cocktail party situations. The architecture of the speech recognition system in an environment with several speakers based on multi-agent recursive cognitive models with imitation of the attention mechanism is constructed. The speech recognition system allows to model selectivity of perception in speech peculiarities for a speaker using multi-agent self-organization. Principles for selective signature processing inside sound modality are defined. They allow to tune on a speaker. Articulatory primitives were chosen as minimal functional pattern in the speech recognition problem. Due to multi-agent nature, use of space-time characteristics and self-learning this approach allow us to separate from each other and analyze sounds of different nature. Screenshots of the cognitive architecture of the speech recognition system based on multi-agent models of semantics are presented.

Keywords: Speech Recognition; Artificial Intelligence; Artificial Neuron Networks; Multi-Agent Systems.

* Corresponding author.
E-mail address: gurtueva-i@yandex.ru

10.1016/j.procs.2018.11.089

1. Introduction

Automatic speech recognition systems are program complexes that transform sound wave of spoken message from acoustic form into graphic, i. e. text [1]. Speech technologies are applied in the different spheres [1-6]. Modern speech applications use different learning techniques, algorithms and methods [1, 2, 7], but all of them with some deviations are based on the same procedure of sound message analysis. In particular, at the initial steps, previous processing (noise removal and allocation of a useful signal) and feature extraction (input speech signal transforms into a set of acoustic parameters) are implemented. The third step is the reduction of the acoustic waveform to the internal alphabet of the reference elements and decoding in final. Differences in approaches to sound signals recognition are discovered at the final stage consequently three main approaches are formed to solve the problem of decoding.

Acoustic phonetic approach postulates that speech is based on a limited set of characteristic phonetic units and the corresponding acoustic properties [8]. To assess word meaning in the process of decoding linguistic constraints are applied. As formalization of laws that govern wide acoustic variability in speech is not a trivial problem, the efficiency of the acoustic phonetic approach is not high.

In the frames of the pattern matching approach identification of unknown utterance is in its comparison with previously prepared reference of code book [1, 7]. Speech pattern can be stored in the form of speech template (template-based approach) or statistic model (stochastic approach) applied to sound, word or phrase.

Connectionist models with different architectures are based on general principles of artificial neural networks, which consist of a potentially large number of simple processing elements [9-14]. The weakness of artificial neural networks is that it is a discriminative model. This model successfully detects the difference between phonemes in this fragment of sound, but does not demonstrate the difference between allophones.

Existing approaches have a number of fundamental limitations that do not allow to recognize speech with high reliability. Speech applications based on them are unstable in real operation conditions or in the case of cocktail party [15, 16]. Thus, there is a need to develop fundamentally new 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 human cognitive abilities. This will allow us to build a speech recognition system in an environment with several speakers based on the simulation of the attention mechanism. In other words, the indicated approach will create a speech recognition system that automatically focuses on a particular speaker or topic of interest.

Formalization for the semantics of rational thinking, the fundamental problem in artificial intelligence, is supposed to be solved using cognitive modeling based on concept of recursive cognitive architecture and hypothesis on invariant organizational and functional structure of process of intellectual decision-making by dint of cognitive functions [17]. The developments and method for multi-agent neural systems learning based on ontoneuromorphogenes is have an aim to build self-organizing multi-agent emergent systems that are capable to emulate psyche functions and adapted purposeful behavior, to set goals using the semantification of reality and the construction social affairs. The present article shows possibilities for speech recognition in an environment with several sound sources (speakers). The architecture for automatic continuous speech recognition on the base of multi-agent cognitive models for semantics that allow to focus automatically on a specific speaker.

2. The Main Principles for Modeling on the Base of Multi-Agent Cognitive Architectures

In this paper, we proposed the development of an automatic speech recognition system using cognitive modeling and the hypothesis of an invariant of organizational and functional structure of the intellectual decision-making process based on cognitive functions. The fundamental foundations of this development and method of learning multi-agent neural-like systems based on ontoneuromorphogenesis are described in details in [17]. This approach is based on the theoretical foundation of cognitive psychology and cognitive neurology [18, 19] as well as modern advances in computer science [20-23].

In the context of the proposed approach, the elementary agent (or agent of zero rank) is the system [17, 24, 25]:

$$\overset{0}{\aleph}_i = S_i\{R_i, F_i, G_i\},$$

that consists of the genome of the agent G_i, the set of receptors $R_i\{r_{i1}, \dots, r_{ik}\}$, the set of effectors $F_i\{f_{i1}, \dots, f_{ik}\}$.

The agent functions in the outer continuum. Its communication with the environment is carried out through systems of direct and feedback. Rational agents are also interconnected, exchanging energy and information, and are then combined into *cognitons* - systems that focus on the realization of individual cognitive functions.

Agents, unlike artificial neurons, include separate information processing units and special functions. The correct action of the neuron agent is possible if the information is sequentially processed in six cognitons: recognition, estimation, goal setting, synthesis, proactive modeling and control of actions [17].

Based on the assumption that each agent at each level is focused on maintaining the balance of the dynamic system as a whole, we will transfer the described principles of building a cognitive architecture in the structure of one agent to all subsequent levels of the multi-agent system. That is, we move from elementary rational agents to the meta-level, where the process of interiorization of the events of the external environment becomes a cognitive function and a system effect arises. This is the content of the hypothesis of the invariant organizational and functional structure of the process of intellectual decision-making based on cognitive functions - the multi-agent recursive cognitive architecture [17].

Cognitive architecture proposed in [17] provides a universal set of logical techniques, methods and rules that mimic thinking activity, intellectual behavior.

Formally, the problem can be formulated as follows. Let $X = \{x_1, x_2, \dots, x_n\}$ is the set of agents, where $x_i \in \{0, 1, \dots k_i - 1\}$, $k_i \in [2, \dots, N]$, $N \in Z$ and $Y = \{y_1, y_2, \dots, y_m\}$ is the set of contracts. Each object y_i is characterized by corresponding set of features

$$x_1(y_i), \dots, x_n(y_i): y_i = f(x_1(y_i), \dots, x_n(y_i)).$$

Otherwise $X = \{x_1, x_2, \dots, x_n\}$, here $x_i \in \{0, 1, \dots, k_r - 1\}$, $k_r \in [2, \dots, N]$, $N \in Z$ is processed input data and $X_i = \{x_1(y_i), x_2(y_i), \dots x_n(y_i)\}$, $i = 1, \dots, n$, $y_i \in Y$, $Y = \{y_1, y_2, \dots, y_m\}$ is output data:

$$\begin{pmatrix} x_1(y_1) & x_2(y_1) & \dots & x_n(y_1) \\ x_1(y_2) & x_2(y_2) & \dots & x_n(y_2) \\ \dots & \dots & \dots & \dots \\ x_1(y_m) & x_2(y_m) & \dots & x_n(y_m) \end{pmatrix} \to \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}$$

Function type $Y = f(x)$ is not specified.

The dependence between contracts and agents may be represented as follows:

$$\&_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{x_j(y_i)} \to \mathrm{P(y_i), i = 1, \dots, l;\ x_j(y_i) \in \{0, 1, \dots, k-1\}},$$

where the predicate $P(y_i)$ is true, i.e. $P(y_i) = 1$, if $y = y_i$ and $P(y_i) = 0$, if $y \neq y_i$.or

$$\vee_{i=1}^{n} \bar{x}(y_j) \vee P(y_i), j \in [1, \dots, m]$$

The decisive function is the conjunction of all decisive rules:

$$\&_{j=1}^{m} x_j(y_i) \to P(y_i), i = 1, \dots, l;\ x_j(y_i) \in \{0, 1, \dots, k-1\},$$

or

$$f(X) = \&_{j=1}^{m} \left(\vee_{i=1}^{n} \overline{x_i} \vee P(y_j)\right) \qquad (1)$$

The function (1) can be interpreted as follows. If the set of contracts of k elements is described by Boolean function

$$F(x_1(y_i), \dots, x_n(y_i), P^{\sigma}(y_1), \dots, P^{\sigma}(y_n)), \text{ where}$$

$$P^{\sigma}(y_i) = \begin{cases} \overline{P(y_i)}, if\ \sigma = 0 \\ P(y_i), if\ \sigma = 1 \end{cases}$$

Then the given function is «0» on the sets $\left(x_1(y_i), \dots, x_n(y_i), P^{\sigma}(y_1), \dots, \overline{P^{\sigma}(y_i)}, \dots, P^{\sigma}(y_n)\right)$ and is «1» on all the rest sets, i.e. it allows any relations between agents and their contracts except, denying connected agents and their contracts.

The received function $f(X)$ may be written in the following form:

$$f(X) = Z_k(q_k(x), P(y_k), X);$$

$$Z_k(q_k(x), P(y_k), X_k) = Z_{k-1}\&(\vee_{i=1}^{n} \overline{x_k(y_i)} \vee P(y_k)) \vee q_{k-1}(x)\&(\vee_{i=1}^{n} \overline{x_k(y_i)} \vee P(y_k));$$

$$q_k(x) = q_{k-1}(x)\&(\vee_{i=1}^{n} \overline{x_k(y_i)};$$

$$q_1(x) = \vee_{i=1}^{n} \overline{x_1(y_i)};$$

$$Z_1 = P(y_1)$$

The goal of the presented research is to model speech recognition on the base of multiagent recursive cognitive architectures. *The task* is to create the architecture for speech recognition system that allows to focus on the fixed speaker automatically.

3. The Architecture for Speech Recognition System on the base of Multi-Agent Recursive Cognitive Models of Semantics

Speech activity is a complicate multidimensional phenomenon. It can be studied from the positions of the different fields of scientific knowledge. Moreover, studies in psychoacoustics confirm the fact that a person decoding a signal often uses auxiliary non-speech information [2, 26]. The question which of the characteristics is primary in the description of speech remains one of the most complicated problems now [2]. The spectral parameters of speech are most often used. But they are highly variable in relation to the voice of the speaker. We believe that not only the intellect, but also mastering the information redundancy inherent in the speech message, gives human a significant advantage before automatic speech systems. For successful decoding, it is necessary to analyze speech in the unity of all its aspects.

From the point of view of cognitive modeling, briefly stated above, speech recognition is a complex problem of multilevel pattern recognition. The procedure for solving this problem involves a multivariate analysis of the signal, its subsequent structuring into hierarchy of elements of a word, words, phrases etc. Figure 1 shows the cognitive architecture for speech recognition system where function of auditory analyzer is realized in the form of subsequent processing of auditory information on the following levels: previous recognition (I level), subconscious recognition (II level), conscious recognition (III level), situation (IV level). Functions and structure of each level are defined by sets of agents, actors and systems of their contact relationships.

The first layer of the architecture, *previous recognition*, is the system of agents detecting signal acoustics. We think it is necessary to distinguish four layers in the structure of the information flow of the afferent auditory tract, namely, the layers analyzing the loudness (amplitude) of the signal, the frequency of the sound wave, the duration of the sound, and finally the location of the signal source.

At the next layer, *subconscious recognition*, the signatures of the previous layer are grouped around meaningful 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 *articulatory gesture* underlying it. For each acoustic event that does not occur earlier, an agent is created that identifies events and actions that have become a source of sound, which in the long term will allow us to analyze non-verbal sound events.

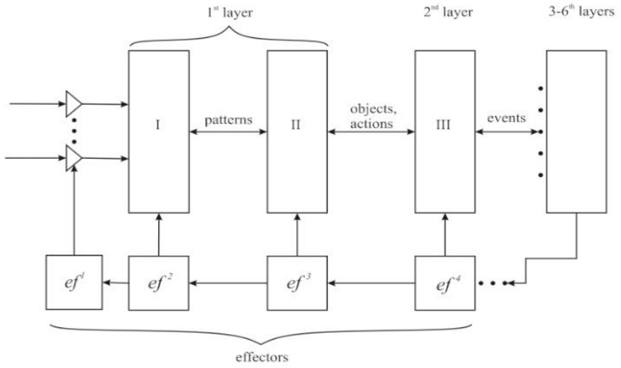

Fig. 1. The architecture for the automatic speech recognition system using multi-agent cognitive models.

At the third stage (*conscious recognition*) significant events are distinguished on the current priority, determined on the basis of the work of so-called cogniton of emotional evaluation [17]. It is a functional node of Multi-Agent Recursive Cognitive Architecture containing a priori and acquired on the basis of training information on the degree of significance of events for the realization of the objective function of an intelligent agent. The agents of the following levels of the Multi-Agent Recursive Cognitive Architecture - cognitons of goal setting and synthesis of action plans - form control commands for the effectors of fine tuning of the filtration system and amplification of the acoustic parameters of integration into the afferent tract. Control of the observing apparatus and adjustment of the afferent tracts.

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

Thus, the proposed cognitive architecture allows to make a part of the signal analysis procedure all aspects of voice communication, including the extra-linguistic component expressed in this approach in terms of the event and situation and emotional evaluation. Each subsequent level of structuring the signal into a hierarchy of word elements, words, phrases, etc. has additional time limits, for example, known word pronunciations or allowed word sequences that can compensate for errors and uncertainties at lower levels. The hierarchy of constraints is used to organize interaction between the level of decision making and on the basis of contractual relations.

A voice signal is recorded by a microphone system, and then the location of recorded signal source is estimated by the beam forming method [27]. The spectral composition of a signal is extracted using discrete Fourier transform [1]. We used, perhaps, the most popular algorithm for calculating the discrete Fourier transform, proposed by Cooley and Tukey [1].

Thus, the preliminary stage of signal processing is reduced to its transformation into a set of signatures and the creation of a matrix that fully characterizes the acoustic characteristics of a signal. The preparation of the sound for the further analysis by the multi-agent recursive cognitive architecture based on the principal of intra-modal differentiation is shown at the Fig. 2.

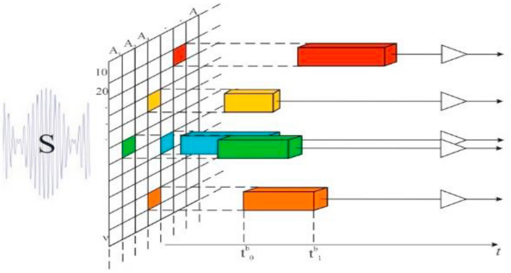

Fig. 2. Visualization of the procedure of intra-modal differentiation and creation of acoustic signal matrix

The information of the sound matrix (location of the sound source, amplitude, frequency, and duration its sound) obtained in the preliminary stage forms the first level of the architecture - previous recognition. At this stage, agents-actors are created registering the acoustics of the signal like human auditory receptors. A schematic representation of the procedure for identifying the inputs and integrating the neuron agents into the afferent pathway of the auditory analyzer is shown in Fig. 3.

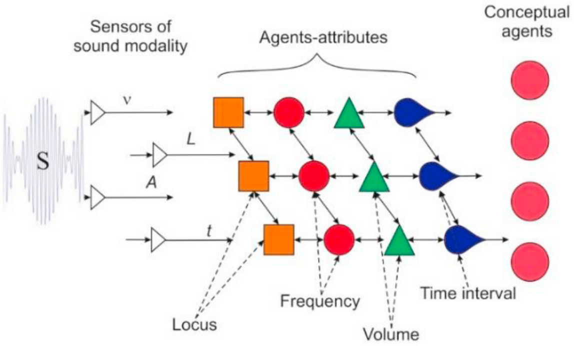

Fig. 3. Identification of inputs and creation of the afferent tract

Figure 4 shows the software implementation of the first part of the described cognitive architecture. Based on the previously described algorithms and the concept of articulatory gesture introduced by us, the parameters of phonemes are identified and identified - the simplest semantically and functionally significant components of the input speech message. The resulting integral representation of the selected phoneme parameters, visualized by the system, is based on the interaction of agents processing certain significant signs of internal modalities of sound allocated for analysis by cognitive architecture.

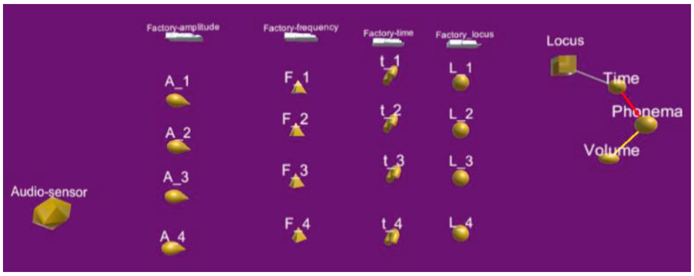

Fig. 4. Screen of the cognitive architecture of the speech recognition system based on multi-agent models of semantics.

4. Conclusion

Speech recognition system is developed on the base of multiagent recursive cognitive architecture that allows to model selectivity of perception in speech peculiarities for a speaker using multi-agent self-organization . Principles for selective signature processing inside sound modality are defined. They allow to tune on a speaker. Articulatory primitives were chosen as minimal functional pattern in the speech recognition problem. Due to multi-agent nature, use of space-time characteristics and self-learning this approach allow us to separate from each other and analyze sounds of different nature. This makes the proposed model unique and gives it advantages in tasks where noise level is extremely high.

Acknowledgements

The work was supported by Russian Foundation for Basic Research, grant №18-01-00658.

References

[1] Jurafsky, D., Martin, J. (2008) “Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.” Prentice Hall, New Jersey.

[2] Mazurenko, I. L. (1998) “Komp’uternyye sistemy raspoznavaniya rechi.” *Intellektual’nyye sistemy* **3(1-2)**: 117-134.

[3] Gupta, V. (2014) “A Survey of Natural Language Processing Techniques.” *International Journal of Computer Science & Engineering Technology* **5(1)**: 14-16.

[4] Ghai, W., Singh, N. (2012) “Literature Review on Automatic Speech Recognition.” *International Journal of Computer Applications* **41(8)**: 42-50.

[5] Reddy, R. (1976) “Speech Recognition by Machine: A Review.” *Proceedings of the IEEE* **64(4)**: 501-531.

[6] Tazhev, B. P., Gurtueva, I. A. (2016) “O nekotorych podkhodakh k resheniyu zaadachi avtomaticheskogo raspoznavaniya rechi”, in *TEL-2016. Trudy Mezhdunarodnoi Konferencii po Komp’uternoi i Kognitivnoi Lingvistike*. Kazan: 217-220.

[7] Waibel, A., Lee, K.-F. (1990) *Readings in Speech Recognition*, Burlington, Morgan Kaufman.

[8] Hemdal, J.F., Hughes, G.W. (1967) “A Feature Based Computer Recognition Program for the Modeling of Vowel Perception”, in Wathen-Dunn, W. (ed). *Models for the Perception of Speech and Visual Form,* Cambridge, MA, MIT Press.

[9] Pappas, N., Meyera, T. (2012) “A Survey on Language Modeling Using Neural Networks.” *Idiap-Research Report-32-2012*: 21.

[10] De Mulder, W., Bethard, S., Moens, M.-F. (2015) “A Survey on the Application of Recurrent Neural Networks to Statistical Language Modeling.” *Computer Speech and Language* **30(1)**: 61-98.

[11] Deng, L., Li, X. (2013) “Machine Learning Paradigms for Speech Recognition: An Overview.” *IEEE Transactions on Audio, Speech, and Language Processing* **21(5)**: 1060-1089.

[12] Mohamed A., Dahl, G., Hinton, G. (2012) “Acoustic modeling using deep belief networks.” *IEEE Audio, Speech, Lang. Process.* **20(1)**:14–22.

[13] Hinton, G., Deng, L., Yu, D., et al. (2012) “Deep neural networks for acoustic modeling in speech recognition.” *IEEE Signal Process. Mag.* **29(6)**: 82–97.

[14] Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G. (2012) “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition.” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, 4277–4280.

[15] Marti, A., Cobos, M., Lopez, J. (2012) “Automatic Speech Recognition in Cocktail-Party Situations: A specific Training for Separated Speech.” *The Journal of the Acoustical Society of America* **131(2)**: 1529-1535.

[16] Zion Golumbic, E.M, Ding, N., Bickel, S., et al. (2013) “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron* **77(5)**: 980-991.

[17] Nagoev, Z. V. (2013) *Intellektika ili myshleniye v zhyvych i iskusstvennych sistemach*. Izdatel’stvo KBNC, Nal’chik.

[18] Chomsky, N.A. (1967) “A Review of Skinner’s Verbal Behavior”, in Jakobovits, L.A., Miron, M.S. (eds.) *Readings in the Psychology of Language*, New Jersey, Prentice-Hall.

[19] Gazzaniga, M. (2009) *Conversations in the Cognitive Neuroscience*, Cambridge, MA, the MIT Press.

[20] Minsky, M. (1988) *The Society of Mind*, New York, Simon and Shuster.

[21] Haikonen, P. (2003) *The Cognitive Approach to Conscious Machines*, Exeter, UK, imprint Academic.

[22] Newell, A. (1990) *Unified Theories of Cognition*, Cambridge, MA, Harvard University Press.

[23] Schunk, D.H. (2011) *Learning Theories: An Educational Perspective*, New York, Pearson Merrill Prentice Hall.

[24] Wooldridge, M. (2009) *An Introduction to MultiAgent Systems*, Hoboken, NJ, Wiley.

[25] Kotseruba, Iu, Tsotsos, J. K. “A Review of 40 Years of Cognitive Architecture Research: Core Cognitive Abilities and Practical Applications”. *arxiv.org/abs/1610.08602*

[26] Morozov, V.P., Vartanyan, I.A., Galunov, V.I. (1988) *Vospriyatie Rechi: Voprosy Funkcionalnoi Asimmetrii Mozga,* Leningrad, Nauka.

[27] Van Veen, B.D., Buckley, K.M. (1988) “Beamforming: A Versatile Approach to Spatial Filtering.” *IEEE ASSP Magazine*. **5(2)**: 4-24.