Conversational speech analysis based on the formalized representation of the mental lexicon

Vadim Zahariev, Sergei Nikiforov, Elias Azarov
Belarussian State University of
Informatics and Radioelectronics
Minsk, Belarus
Email: zahariev@bsuir.by, nikiforov.sergei.al@gmail.com, azarov@bsuir.by

***Abstract*—This article is dedicated to the study of the possibilities of formalization and modeling with the help of modern approaches in the field of artificial intelligence such a structure of human consciousness as "the mental lexicon". According to experts in the field of psycholinguistics [1], [39], it is a large graph of concepts and lexemes of a language, related to each other by semantic connections, which plays an essential role in the processes of generation and perception of speech by a human [2], [40].**

The usage of semantic networks for modeling this structure allows using the advantages of this form of representation in full for solving problems of natural speech understanding by technical systems. This approach also implies the need to develop methods of analysis and processing of the speech signal, which precede the stages of semantic processing, closer integration of these stages and, in a sense, the erasing of the boundaries between them, taking into account how this occurs in the process of speech perception by the human brain [41]–[43]. This article is concerned with the possibilities to show the advantages of such a complex approach to the analysis of speech messages through its implementation based on the Open Semantic Technology for Intelligent Systems (OSTIS).

***Keywords*—mental lexicon, semantic network, voice assistant, intelligent personal assistant, spoken language understanding**

I. Introduction

Currently, the systems of intelligent personal assistants have become widespread and popular. This is proved by the large number of products related to the usage of artificial intelligence methods in the field of speech technologies, which are appearing on the market [3], as well as the attention that leading technology companies pay to the development of this direction [4].

As a rule, modern assistants have two main modes of conducting a dialogue: command and free dialogue on random topics ones [5].

The command mode is based on the search for a certain keyphrase called "intent" in the user speech and the implementation of a response action, depending on the type of this intent. This mode allows, through the speech interface, getting important for the user information, performing some useful actions related to a certain class of intent (ordering a taxi, turning on and off the music or some device in the house, requesting a location or weather for a certain date, searching for the nearest interesting for the user place, etc.) [5]. If the system cannot precisely classify the intent of the user, it transmits it in text form to an Internet search engine, which should give some relevant answer based on a combination of keywords.

A random mode, when the system tries to simulate conducting a dialogue with a human, keeping up a conversation on general topics, depending on the initial assumptions contained in the user utterance. As a rule, in this case, the recognition of input phrases is carried out with an unlimited dictionary, when it is assumed that the user can ask anything and the system, based on its internal "knowledge" and the model of conducting a dialogue, should "understand" the utterance and formulate the most probable answer from the point of view of a human. It should be noted that modern systems add various stylistic language techniques, emotions, moods or even humor to the conversation for making a dialogue realistic. To prevent errors in case of incorrect interpretation of the input phrase, it may ask the user to repeat one or more of the last phrases [5].

By default, the system starts in command mode. Switching between modes is performed automatically by some key "intents", when the system understands that the user just wants to "talk" but not give the system concrete target designations.

The modern voice assistant is a distributed software and hardware system that consists of two main parts: a client and a server ones (fig. 1).

From the point of view of the implementation of the architecture modules, modern speech assistants consist of the following main modules implemented with the help of such libraries:

- automatic speech recognition (ASR) – "DeepSpeech", "Kaldi", "Vosk" [6]–[8];
- natural language understanding (NLU) and dialogue managment (DM) – "Wit.AI", "Dialogflow", "Rasa NLP", [9], [10];
- natural language generation (NLG) – "GPT2", "BERT", "GPT3" [11], [12];
- text-to-speech synthesis (TTS) – "WaveNet", "Tacatroon", "Nvidia Nemo TTS" [13]–[15].

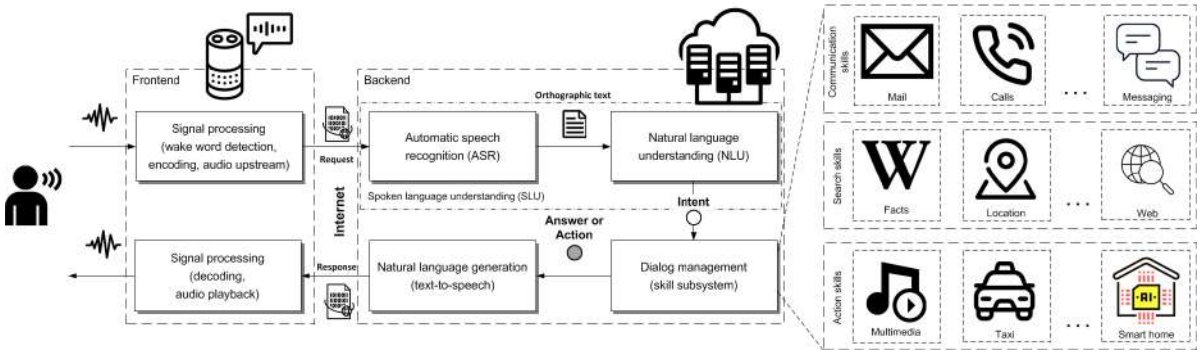

Figure 1. The intelligent speech assistant system architecture

In the current implementations of speech assistants, most of these modules are fundamentally based on neural network models and information processing methods. There are a lot of open source libraries and frameworks (including those listed above) that allow implementing the functionality necessary for the operation of these modules. Also, major technology companies have their products that are being developed on a proprietary basis [4].

It is an established matter that the currently dominant neural network approaches to solving this problem allow achieving top results relative to classical machine learning models, rule-based systems and other classical methods of building dialogue systems [16]. However, these approaches also have their peculiarities and limitations, primarily related to the strong dependence of the final characteristics of the models on the quantity and quality of training data, sample size, choice of network architecture, the complexity of formalization and understanding of the processes that occur inside in the learning process, computational constraints of using large neural network models on end devices with limited resources. Therefore, it seems quite logical and promising to use data-based methods, which include neural networks together with other artificial intelligence methods that allow effectively formalizing the processes of storing and processing information inside the dialogue assistant.

The further development of intelligent personal assistants will be quite difficult without extensive analysis of the nature of the speech message, the peculiarities of its perception at all stages, including at the semantic and even pragmatic levels.

One of the fundamental structures of consciousness, called in the psycholinguistic literature “mental lexicon” [17], [18], [20], has a great significance in the acts of speech message perception and understanding. Scientists describe this structure as an internal dictionary, which is a network of connected concepts (words), which are used by a human when enabling speech communication. In addition to the mental lexicon, there is also the concept of a “mental grammar” of some add-in over the “mental lexicon”, where the rules that we use when generating and perceiving word forms and even whole sentences are stored [53].

From our point of view, semantic networks and intelligent agents that run in connection with them can serve as a good abstraction for modeling such structures of human consciousness, the properties and purpose of which for computer systems can largely overlap with the purpose of the “mental lexicon” units for a human, as it is considered by specialists in psycholinguistics [21].

The question of displaying this structure in the framework of an intelligent system with a natural-language interface for building dialogue systems of a new class, the building of which will be discussed in this article, seems particularly intriguing.

II. Problem definition

The mental lexicon is a term that denotes, in a wide sense, how words are represented and systematized in the human consciousness [17], [18]. Most scientists agree that the mental lexicon can be described as a giant network of concepts (fig. 2), where words that are close in meaning and similar in sound (writing, in the case of written speech) are connected [19]. The affinity and organization of these concepts, primarily by semantic and phonetic markers, has been proved by a series of various experiments [22].

The mental lexicon is one of the most important components of a language ability of a human, in which notions about the world and their lexical representation are fixed. The units of the mental lexicon are connected into a single complex dynamic system that can be rebuilt, depending on the situation. The mental lexicon has a complex multi-sided and multi-level structure with intra- and interlevel connections (fig. 3) [19], [44], and in its organization, it is possible to distinguish the core and peripheral areas, for example, based on the criteria of the frequency of usage of concepts [45], [46].

Taking into account the volatility of the lexicon, it can be stated that the boundaries between the core and the periphery cannot be invariant because these structures are in relations of dynamic connection. The

core changes dynamically during human life since new units are assimilated by the individual and some, on the contrary, may fall outside the core as a result of a change of professional activity, place of residence, social status, etc. Nevertheless, the analysis of research has shown that vocabulary that reflects national, social, professional, age and other realias of the outside world of the individual is prevalent in the core [47].

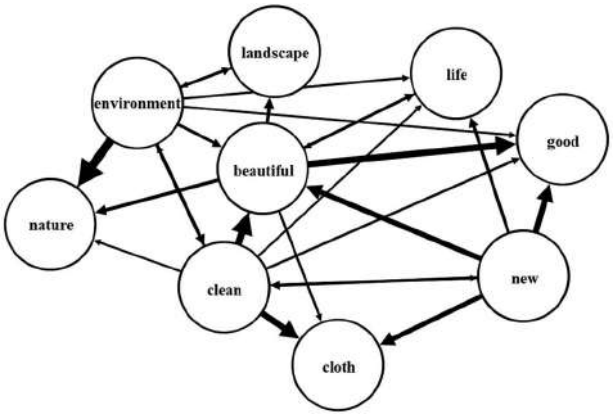

Figure 2. A fragment of the mental lexicon that includes the semantic neighborhood of the concept "environment" [19]

The agility and mobility of the units of the mental lexicon in the direction from the core to the periphery and vice versa is provided by the type of connections between them, based on the semantic similarity of the units. In the mental lexicon, those connections of units are active, which are reflected in integrations considered in traditional semantics, such as a lexical-semantic field, a lexical-semantic group, a lexical-thematic group, a thematic group, etc.

Modern models of mental language representation differ not only in their attitude to the oppositeness of lexicon and grammar but also in their directivity to the speech generation or perception as well as compatibility with languages of different structures. Although the mental and neurophysiological mechanisms for enabling speech activity are the same, the procedures vary significantly depending on the type of language [23], [24].

Due to this fact, for languages of different types, the description of the mental lexicon in the form of a single model is hardly possible. The actual models of the mental language representation are grouped into four classes [23]:

- single-system models that do not separate lexicon and grammar;
- two-system models that separate the lexicon as an inventory of units from grammatical rules;
- hybrid models that allow, along with the integrated one, elementwise storage of word forms and the usage of rules for combining morphemes in word forms;
- models without a mental lexicon.

An essential part of the research of the mental lexicon by psycholinguists correlates with hybrid models focused on speech perception [48].

Hybrid models assume both gestalt (perceived as an integral whole) storage of word forms and the decomposition of a word form into morphemes; at least productive affixes can be stored as independent units and, accordingly, act as operational units when planning an utterance. In this case, the storage of units and their operation occurs due to three levels of representation: the notion of morphemes, the notion of a complete word form, the notion of a lemma. The word form is extracted from the mental lexicon as a gestalt, although its elementwise extraction is possible both for recognition when perceiving speech and generating an utterance. The lemma is interpreted in different ways, but all interpretations are unified by its relation to semantics; rather, the lemma mediates the transition from formal representations (the sound/graphic image of a word form) to the node of the semantic network [25]. The lemma represents the general meaning of the lexeme, thus providing access to the concept as an item of the semantic network [26].

Current models of the mental lexicon, along with the generalization of the results of experiments in standardized methods, often explicitly or implicitly rely on the concepts of early psycholinguistics. The possibility of ambivalent comprehension of a derivative (and its word forms) both entirely and on the basis of the elementwise analysis is proved in the following papers [25], [49]. If a word form represents a storage unit in almost all models of the mental lexicon, then in the process of searching for access to it, it is possible to operate with other units: morphemes (quasi-morphemes) for perception and lemmas for speech generation. The storage unit may not coincide with the operational one – the unit that the speaker or listener operates with [50].

The mental lexicon is often represented as an associative-verbal network [51], as a "dynamic functional system that organizes itself due to the constant interaction between the process of processing and ordering speech experience and its products" [52].

In this case, linguistic (verbal) units can act as a storage unit in the mental lexicon. The representation of morphemes, integral word forms and lemmas in the mental lexicon at its different levels within the framework of the hybrid model suggests the ability to assess the accuracy of word form recognition based on grammatical features as well as segmental and suprasegmental phonetic characteristics. Moreover, the variability of the sound image of a word form in the conversation suggests the presence of such a mechanism of speech perception that would allow correlating a variable acoustic signal with the perceptual standard of a language item stored in consciousness and associated with the lemma level. The operation of this mechanism is based on probabilistic

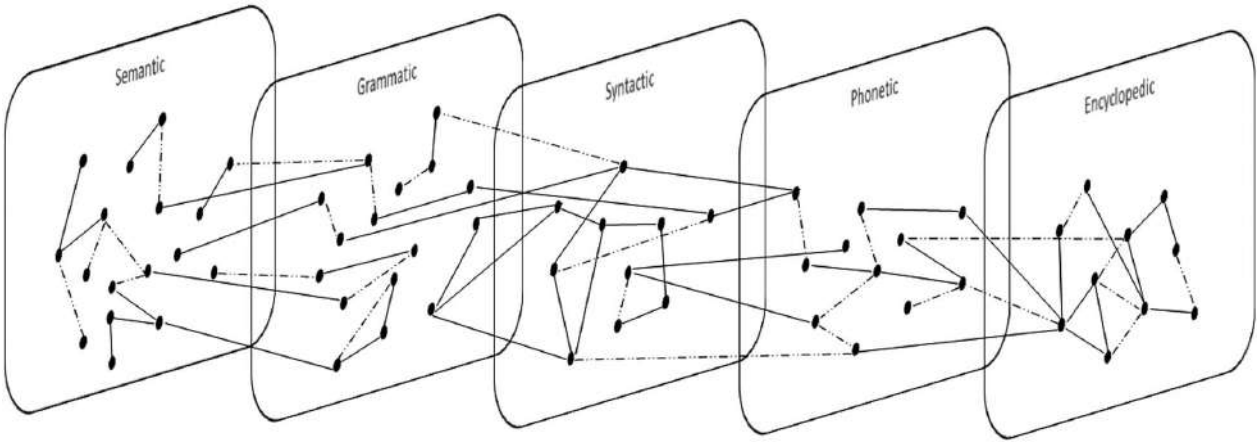

Figure 3. Possible layers of the mental lexicon [19]

forecasting of the subsequent signal of the speech chain based on various features [50]. In essence, the study of speech perception in this aspect is an analysis of the mechanisms of access to the units of the mental lexicon.

In the memory of a native speaker, the standards of the sound image of words and the main syntactic constructions (including the intonation contour) that correspond to the sentence are stored. The complex of linguistic features characterizes the expression plane of language units and, thanks to this, mediates access to a perceptual standard correlated with the content plane, that is, language semantics.

Perceptual standards combined in the perceptual language base are correlated with the units of the mental lexicon: perceptual standards of the lexical level are a unit of access to the word-form level, correlated with lemmas and thus with semantic representations. An actual linguistic problem is the modeling of the perceptual language base as a hierarchical probabilistically organized structure of multidimensional matrices of language units of different levels, permeated with multiple intersecting perceptually significant linguistic features. Characterizing the form of language units, the complex of features mediates access to the perceptual standard and intromission to the semantic level [2], [53].

The significance of the mental lexicon for the processes of speech perception and understanding, indicated above, demonstrates the relevance of the research of this psycholinguistic concept as well as the importance of the problem of implementing structures like the described one as part of intelligent systems with a natural-language interface. This can lead, including due to the synergetic effect, to an improvement in the quality of the model operation in the current implementations of the systems as well as confer them new useful properties. Thus, the purpose of this article is to study the possibilities of a modeling and implementation of the mental lexicon and the potential benefits that can enable its implementation as part of a new generation of speech assistants.

To achieve this purpose, it is necessary to solve the following main problems:

- to perform modeling and formalization of the network part of the mental lexicon as a complex system of hierarchical concepts connected by semantic, perceptual, frequency, etc. features that allow a human to carry out the processes of conceptualization, formulation, perception and articulation in the process of speech communication;
- to choose and implement an adequate model of speech perception mechanisms that is related as closely as possible to the notion of this process and its connection with the mental lexicon.

It should be noted that in the case of modeling the mental lexicon for its implementation as part of technical systems, other important and relevant from a scientific and practical points of view problems can be identified. However, the two points listed above are of the highest priority in terms of their significance and those useful properties that can be enabled by the implementation of the mental lexicon as part of an intelligent system. Approaches to their solution are the subject matter of this article.

III. Proposed approach

To solve the two abovementioned problems, we propose two main approaches for consideration:

- the usage of the semantic network appliance for modeling the structure of the mental lexicon and the connections between its elements as one of the similar models of the representation and organization of knowledge implemented in technical systems;
- taking an approach based on semantic-acoustic analysis, represented by us in a series of previous articles [30]–[32], which offers a solution to how it is possible to make a direct connection between the acoustic image of a word and its meaning in

the semantic network (in our case, in the network of the mental lexicon), as it occurs in the process of perception of speech information by a human, and how to connect the acoustic image of a word with its concept in the semantic network in the most direct way.

IV. Semantic level

A semantic network is a model for representing knowledge in a formalized form, which allows describing any information in the form of a graph structure [54], [55]. Nodes in such networks are concepts connected by semantic relations that act as edges of the graph [27]. Each node of the network has an important property from the point of view of semantics, namely, each element is associated with some of its "meaning" and "sign" (in the classification according to Ch. S. Peirce), i.e., the direct meaning of what this node means, as well as how this concept is represented in the area of perception, is its external form [57]. As an example, it is possible to introduce the concept "house" and a picture of a house as a visual image of this concept. In the case of a mental lexicon, an analogy with the acoustic or orthographic image of a word and its meaning (the place of the concept) in the mental lexicon associated with this sign can be drawn.

The usage of semantic networks for knowledge representation has the following main advantages [58]–[60]:

- they allow quickly and easily adding information, drawing conclusions based on the connections between different concepts;
- they ensure coherence and the absence of duplication of concepts;
- the information in the network can be easily interpreted by both a machine and a human;
- they can be used equally effectively to describe any subject domain.

The semantic network reflects the structure of the subject domain, the model of which it is, in an "almost isomorphic" way. So, the structure of any subject domain can be represented in the form of a semantic network quite simply, "almost isomorphically" (in the sense specified above). In other words, the correlation between the subject domain and the corresponding semantic network is quite transparent and has no unnecessary complications due not to the essence of the matter, not to the structure of the subject domain but to the peculiarities of the description language. Within the framework of this correlation, there is a fairly simple semantic interpretation of various types of elements of the semantic network (nodes, bindings, incident relations of bindings, key nodes) [61].

To an extent, semantic networks can also include the forms of information representation that humanity has been dealing with for a long time. The generally accepted ways of representing schematic electrical diagrams, logical circuits, flow diagrams are also nothing more than semantic networks represented in various alphabets. It is also obvious that semantic networks also include such ways of representing information as cognitive maps, knowledge maps and much more [28].

Such a seemingly simple and intuitive form of information representation allows formalizing knowledge from any subject domain quite fully, placing it in a model called an ontology, which has important properties for searching and extracting this information [29]. Something similar occurs in the process of human ontogenesis, when all the acquired knowledge and experience sum up in a complex system of concepts connected by nonlinear relations, where the search is carried out on the basis of semantic, associative, frequency and many other complex criteria for information search. The representation of such a structure at the linguistic level of human functioning, in fact, is the mental lexicon, the computer model of which will be nothing more than the knowledge base of the intelligent system.

At the same time, let us emphasize that the semantic network as an abstract mathematical structure should be clearly distinguished from different variants of its implementation and representation in computer memory, including graphical visualization. Various semantic networks can have different alphabets of elements (different sets of labels on the elements of semantic networks) [28], [62]. Each of these implementations has its characteristics, strengths and limitations. One of the platforms that have proven their effectiveness in the development of intelligent systems based on semantic networks is the OSTIS Technology. The features of using this technology and its advantages for modeling the mental lexicon will be considered below.

V. Acoustic level

In terms of modeling the process of speech perception, it should be noted once again that the study of speech recognition in this aspect is an analysis of the mechanisms of access to the units of the mental lexicon [2], [25], [48].

As was already mentioned above, the standards of the sound image of words and the main syntactic constructions (including the intonation contour) that correspond to the sentence are stored in the memory of a native speaker. As storage units in the mental lexicon, linguistic (verbal) units can serve. The representation of morphemes, integral word forms and lemmas in the mental lexicon at its different levels within the framework of the hybrid model implies the ability to assess the accuracy of word form recognition based on grammatical features as well as segmental and suprasegmental phonetic characteristics.

Separately, it should be emphasized that if the question is about the perception of oral communication, the nodes of the lexicon are connected directly with their corresponding acoustic images and not their analogues in

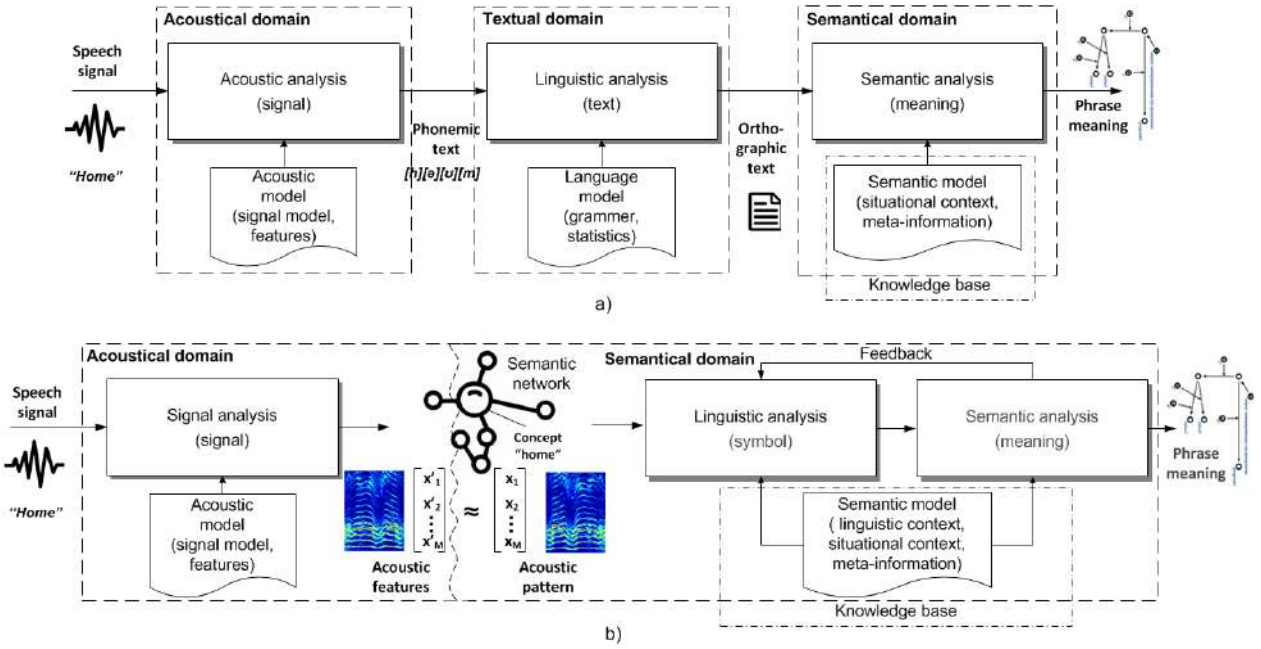

Figure 4. The conceptual architecture of the system: a) the default 3-tier approach; b) an approach based on semantic-acoustic analysis

the form of a text or other representation. Text notations associated with lexicon units are used in the process of perception and transmission of speech in written, not oral, form. There is no contradiction here because several signs of different types can correspond to the same semantic concept. For example, the concept "house" coincides with an acoustic image of the word "house", its text entry in the corresponding language, a visual image in the form of an image of a house. All these signs are equivalent in terms of the efficiency of their usage for accessing the nodes of the mental lexicon.

This fact of the organization of processes of processing auditory information differs significantly from the principles, on which modern speech recognition and understanding systems run when there is a rigid hierarchy of recognition stages. At the initial stage, the basic acoustic units of the language, phonemes and allophones, are transmitted (recognized) into their textual representation, in the future, at higher levels of language information processing (syntactic and semantic ones), only their textual representation is used.

Thus, the acoustic image of the concept is subordinated to the text image, which is not present in real life but is important for the current level of development of recognition systems, including as part of speech assistants, since it reduces the complexity of implementing current solutions. Therefore, we consider it essential to create methods of signal analysis that would allow more naturally describing the fact of the interrelation of verbal signs in acoustic form with the nodes of the mental lexicon.

In this article, we propose to use the method of semantic-acoustic analysis. The ideas of this method were proposed and considered by us in the papers [30]–[32]. It allows performing the primary parsing and parameterization of a speech message using special signal processing techniques. In the course of their application, certain "acoustic images" that correspond in our case to the concepts of the mental lexicon are separated from the conversation, which in turn will correspond to certain nodes (signs of concrete entities or concepts) in the semantic network.

In this article, at the current level of development of the proposed approach, it is recommended to extract and analyze sound images of concepts according to the principle of audio dactyloscopy systems ("audio fingerprinting") [36]. This method involves comparing the selected audio fragments of the signal, represented in parametric form, with a dictionary of standards and determining the proximity measure of the selected "acoustic image" to the standard of this image – associated with the corresponding node of the semantic network of the mental lexicon. The degree of confidence measure, in this case, is rather a proximity measure of the selected signal fragment to the reference one in the selected parametric space.

In general, the following information may be required to carry out this kind of analysis within the framework of the proposed approach:

- a set of standards for correlation with fragments of a speech signal and their specification;
- the context of the analyzed message (from whom the message was received, in what external conditions, what other sounds are present in the background,

etc.);
- a set of rules for the transition from fragments of a speech message to semantically equivalent constructions in the knowledge base;
- a semantic specification of the concepts included in such constructions.

For speech analysis, a model based on a hybrid representation of a speech signal will be used, which allows representing any fragments of a speech signal of different nature of sound generation in the most adequate form [33].

Vocalized and non-vocalized fragments of the signal are described by periodic (harmonic) and aperiodic (noise) components, which are important for the perception and differentiation of various acoustic images of concepts and are described by various components of the model. Details of the model implementation will be presented below during the description of the system architecture.

From our point of view, such a joint implementation of the abovementioned approaches focused on modeling the mental lexicon using semantic networks and semantic-acoustic analysis will allow acquiring of new useful properties by the intelligent personal assistant: additional flexibility and adaptability, the ability to overcome the limitations inherent to the systems presented on the market at the moment. All this will eventually allow such systems to come up to a whole new level.

VI. Practical application

It would be desirable to consider the possibilities of using the presented approaches on a practical example of one of the most relevant fields of application of speech assistants. At the moment, such a direction is the systems of semi-automation of the infrastructure of accommodation units, which are called "smart home" systems in common terms.

All these systems are built according to the same scheme for recognizing a selective list of commands transmitted by the user from an application with a graphical interface or through a voice control interface, which the system implements, and this scheme has a set of parameters embedded in it that are configured during the programming and deployment of the system.

For the expansion and reprogramming of such systems, special knowledge and skills are required to add certain new nodes to the system, to ensure their correct configuration depending on the operating conditions. Often, all these actions require a highly qualified user or a house-call of high-paid specialists.

In this context, speech assistants act as part of the system, providing speech input, interpretation and transmission of commands to the smart home controller as well as notification of the current state of the system. Due to the lack of flexibility and limitations of the current implementation, they do not allow the user of speech interface tools to perform "fine-tuning" of the system, changing the system parameters just-in-time and adapting the system to changing external conditions: the appearance of new users of the system who are not familiar to it, fine-tuning the system depending on changes in the weather during the day or the time of day.

As a cover of our theses, let us present several scenarios for using the smart home system that are not available in current system implementations, where improved adaptability of the system to changing conditions is necessary:

- *Example 1.* Reprogramming of the access control system just-in-time. Imagine that guests came to the owners of a detached home for dinner. A smart home equipped with the option of controlling access to rooms by identifying the user by voice or face image (this option is considered in detail in the article "Semantic analysis of the video stream based on neuro-symbolic artificial intelligence") from microphones or watching cameras of a smart home. As part of the standard implementation of the smart home system, new rules would have to be added by reconfiguring the system. Having a flexible system based on the principles described above, it would be possible to "introduce" new users to the system and order it to issue access permissions to new persons for a specified period immediately at the moment of their first appearance. The system would formulate the rules independently, according to which it can provide guests with access to the rooms of the house and adhere to them during the allotted time.
- *Example 2.* Adaptive fuzzy lighting control. During the day, significant fluctuations in the intensity of natural light can be detected within a short time. This is especially true for countries located in the midland with an oceanic or moderate-continental climate during the mid-seasons of the year when there is often a change in atmospheric phenomena. The settings of the lighting parameters of sensors and controllers are calibrated to some average values, without taking into account the complex dependencies of changes in external parameters. However, during the day, there may be situations when it is necessary to adjust the overall picture of the lighting of the rooms of the building, related to the daily round: the presence or absence of humans, work that requires lighting or not, rest schedule, etc. In modern systems, there is no functional ability to make the light in a certain room "brighter" or "more subdued" without specifying the exact intensity value. Moreover, it is often difficult for the user to give a clear command: "Make the lighting intensity 15% more". Quite often, the user does not know about the acceptable scale of intensity values and its dynamic

range. It would be much easier, in this case, to "ask" the system to "make the lighting a little brighter" or "a little more subdued", and how much is "a little", taking into account the current level of illumination of the room and the lighting parameter profile, the system should be able to determine and implement itself.

Thus, the implementation of the system based on the modeling of the mental lexicon with the help of semantic networks and speech analysis using the semantic-acoustic approach would give additional flexibility to the semantic core of the system, provide the possibility of forming new rules in the knowledge base of the system directly, by means of the language interface in the interaction process. The architecture of such a system and the analysis of its specific components are presented below.

A. Algorithm of the system operation

The general algorithm of the operation of an intelligent personal assistant that includes the implementation of a mental lexicon model (in the form of a knowledge base of an intelligent system built using the OSTIS Technology) for conducting a dialogue with the user can be described in the form of a sequence of the following actions.

- The user utters a speech message or command for the smart home system.
- Further, the acoustic module implemented within the framework of the OSTIS platform as an Agent of transition of speech into the semantic representation performs the procedure of semantic-acoustic analysis and transmits its results directly to the knowledge base of the system in the form of concept identifiers, thus displaying a natural language phrase in the semantic space of the computer "mental lexicon".
- Then the semantic analysis module, which includes the knowledge base of the system and the Message processing agent, Agent of decomposition of messages into atomic ones, Agents of allocation of entities and relations, Message classification agent and Agent of logical inference, parses, analyzes and interprets the received user message using information from the knowledge base (about the profiles of the system users, its current state, environmental factors, etc.). With the help of the Command classification agent and Agent of the control of executive devices, it executes a user command or a request.
- As a result, the response generation module, in which, based on the response structure in the knowledge base, the result of the work of the system is created in text form by means of the Message generation agent, which is then voiced by the system through the Agent of text-to-speech transition, which refers to the text-to-speech synthesis engine. As a result of the reasons for the typical implementation of the last module, consideration of its features is beyond the framework of reference of this article. The Message classification agent classifies the received message based on the rules present in the knowledge base.

Further, the features of the implementation of the two main modules, acoustic and semantic ones, will be considered in more detail to demonstrate how the designated problems of modeling the mental lexicon can be implemented.

B. Acoustic module

For speech analysis, a model based on a hybrid representation of a speech signal will be used, which allows representing any fragments of a speech signal of different nature of sound generation in the most adequate form [33]. Vocalized and non-vocalized signal fragments belong to separate parts of the model: periodic (harmonic) and aperiodic (noise) ones.

Mathematically, the main idea of the given model can be formalized in the following form:

$$s(n) = h(n) + r(n), \quad n = \overline{0,..,N-1} \tag{1}$$

where $s(n)$ is the input speech signal, $h(n)$ – a harmonic component, $r(n)$ – a noise component of the signal, n and N – the current signal sumple number and the total duration of the analysis fragment, respectively. The harmonic component can be represented by the following expression:

$$h(n) = \sum_{k=1}^{K} G_k(n) \sum_{c=1}^{C} A_k^c(n) \cos_k^c n + \phi_k^c(0)) \tag{2}$$

where G_k is an amplifier gain determined based on the spectral envelope, c – the number of sinusoidal components of the signal for each harmonic curve, A_k^c – the instantaneous amplitude of the c-th component of the k-th harmonic curve f_k^c and $\phi_k^c(0)$ – frequency and initial phase of the c-th component of the k-th harmonic curve, e_k – an actuating signal of the k-th harmonic curve. The amplitudes A_k^C are normalized to provide the sum of the harmonic energy equal to $\sum_{c=1}^{C}[A_k^c]^2 = 1$ for k = 1,...,K.

In this case, the aperiodic component is modeled over the entire frequency band, as it is demonstrated in the spectrum of a real speech signal [34]. This effect is reached by applying the technique of signal analysis through synthesis and diminution of the harmonic part from the original signal:

$$r(n) \quad = \quad \begin{cases} \max(s(n),h(n)) - h(n), & s(n) > 0 \\ \min(s(n),h(n)) - h(n), & s(n) < 0 \end{cases} \tag{3}$$

Thus, for one signal frame with the number m and the duration of N samples, a characteristic vector is formed, which includes the coefficients of the model $\mathbf{x_m} = [G_k, A_k^c, f_k^c, K, C]$. The acoustic image of a single

word is a sequence of such characteristic vectors: $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M})^T$.

It is proposed to evaluate the model parameters using the method of instantaneous harmonic analysis, which allows significantly increasing the accuracy of determining the parameters of the periodic component [63]. Its application allows obtaining a high temporal and frequency resolution of the signal as well as a clearer spectral pattern of the localization of energy at the corresponding frequencies (fig. 5) and, as a result, performing a more accurate assessment of the signal parameters (on average by 10–15%).

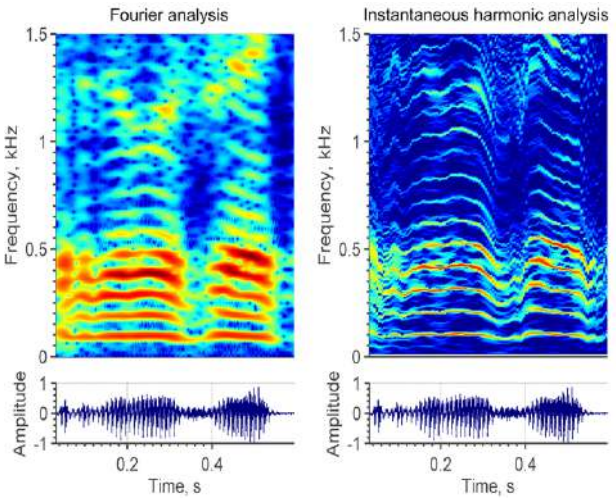

Figure 5. STFT- and IHA-based spectrograms

Unlike the classical methods of signal analysis used in modern speech recognition systems based on the determination of Mel-frequency cepstral coefficients (MFCC) or linear prediction coefficients (LPC) [64], [65], the method based on instantaneous harmonic analysis allows getting a high temporal and frequency resolution of the signal as well as a clearer spectral pattern of the localization of energy at the corresponding frequencies. Unlike classical methods, which are based on the discrete Fourier series transformation or the definition of the autocorrelation function of a signal on a short fragment, the method under consideration does not impose strict constraints related to adherence to the conditions of stationarity of the signal parameters on the analysis frame. At the same time, the parameters of the harmonic model, if necessary (for example, to describe the spectral envelope), can be relatively easily converted to other methods of representation, such as classical Mel-frequency cepstral or linear prediction coefficients.

Algorithms that implement the signal processing method described above based on instantaneous harmonic analysis have a reference implementation as part of the GUSLY audio signal analysis and synthesis framework [35].

Thus, after obtaining the parameters of the model, we can generate a spectrogram of the signal. A spectrogram is a visual representation of the frequency content of a signal as a function of time (fig. 6). The spectrogram of any audio signal can be considered unique, but this representation has too high a dimensionality to be used as a kind of fingerprint of the "acoustic image" associated with a specific node of the mental lexicon. Therefore, a more compact representation of the acoustic images of the mental lexicon is required. Similarly, this procedure is implemented in such services as, for example, the music recognition application "Shazam" [37].

The spectrogram of the signal of a phoneme, lexeme, word combination or even an entire phrase that sounds in speech can be considered its unique signature. Therefore, to determine whether two acoustic images are the same, it is possible to compare their spectrograms. However, the spectrogram is a rather large three-dimensional array (frequency, time, amplitude) and, therefore, requires a significant memory amount.

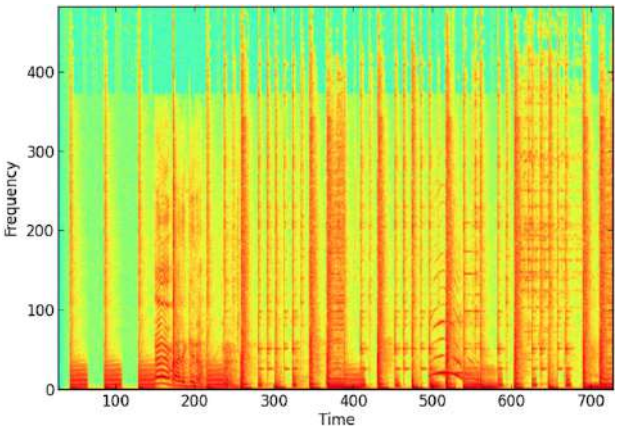

Figure 6. A spectrogram of the speech signal

Physical storage and computational comparison of unique signatures in the form of spectrograms for millions of nodes of a semantic network would be an unsolvable problem. Therefore, it is important to find methods that will allow highlighting only the most significant information from the spectrogram and finding a way to present it in tabloid form. One of the techniques of such compression involves the creation of what is called a "constellation map", i.e. an array of key points of the spectrogram, which is formed by finding local peaks in the signal spectrum (fig. 7).

The next step is to obtain a compressed representation of this array of points and create an acoustic image fingerprint that denotes an audible concept. We use a method, in which the frequency of a local peak is combined with the frequency of another local peak in its neighborhood and the time difference between the frequencies is calculated [36].

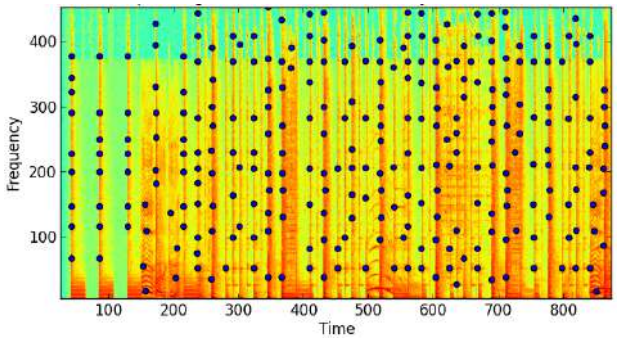

Figure 7. Highlighting of peaks on the spectrogram

For each local peak frequency (anchor), we have a set of the nearest frequencies (targets) and their time deltas. By forming sets of such triples of parameters for each peak frequency in this way, we can preserve the unique features of the analyzed signal fragment. The constellation map (fig. 8) is the information used to generate an acoustic image fingerprint by transmitting it to the hashing function. The hashing function receives data of indeterminate length and generates output data of fixed size (called a hash). In addition, hashing functions will always produce the same hash for the same input. The result of the hashing function is a sound fingerprint of the acoustic image of the spoken word, which we can compare with the standard associated with the node of the semantic network of the mental lexicon.

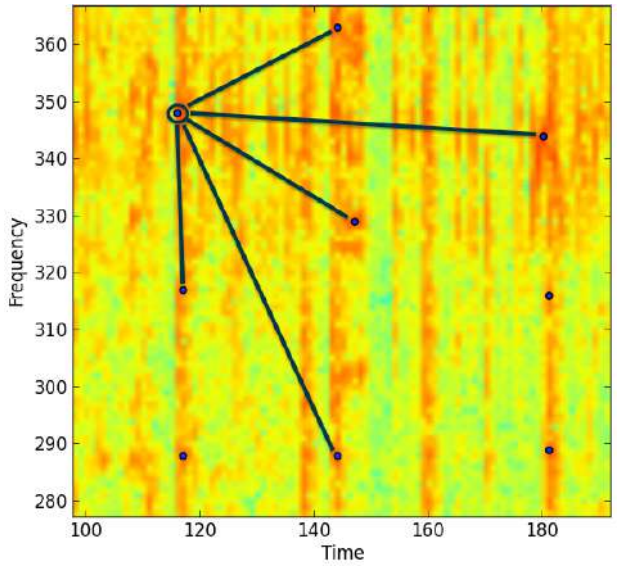

Figure 8. The constellation map

To determine the proximity between two images, we will use the Jaccard distance, which allows measuring the similarity between finite sets and is defined as the size of the intersection divided by the size of the union of these sets:

$$J = \frac{|H_sig \cap H_ref|}{|H_sig \cup H_ref|} = \frac{|H_sig \cap H_ref|}{|H_sig| + |H_ref| - |H_sig \cap H_ref|} \quad (4)$$

where J is the value of the Jaccard distance, H_sig – a hash of the acoustic image of the input signal, H_ref – a hash of the acoustic image of the standard associated with the network node.

This approach allows us to compare signals that can have different lengths, which is a very common phenomenon since words of a language can be and are often pronounced at different speeds. Another important property of the proposed algorithm is its low computational complexity, which allows it to be used in nearly real-time systems. This is especially true for creating speech assistants where the speed of the system response to user input is important.

C. Semantic module

As a technological basis for the implementation of the proposed semantic part of the system, it is proposed to use the OSTIS Technology [66]. The main advantages of the implementation of intelligent systems in the context of solving the problem of modeling the mental lexicon as a semantic network of language concepts have already been identified by us in the section “Problem definition”. Let us focus on some practical aspects of the usage of the technology and its features that provide these advantages.

The systems built on the basis of the OSTIS Technology are called ostis-systems, respectively, the module for understanding speech messages, the prototype of which is considered in this article, will be built as a reusable component that will be integrated into various ostis-systems in the future, if necessary.

As a formal basis for encoding various information in the knowledge base, the SC-code [66] is used, the texts of which (sc-texts) are written in the form of semantic networks with a basic set-theoretic interpretation. The elements of such networks are called sc-elements (sc-nodes, sc-arcs). The focus of this work on the OSTIS Technology is due to its following main advantages:

- within the framework of this technology, unified means of representing various types of knowledge, including meta-knowledge, are proposed, which makes it possible to describe all the information necessary for analysis in one knowledge base in a unified manner [38];
- the formalism used within the framework of the technology allows specifying in the knowledge base not only concepts but also any files external from the point of view of the knowledge base (for example, fragments of a speech signal), including the syntactic structure of such files;
- the approach proposed within the framework of the technology to the representation of various types of knowledge [38] and models of their processing [29]

ensures the modifiability of ostis-systems, i.e., allows easily expanding the functionality of the system by introducing new types of knowledge (new systems of concepts) and new models of knowledge processing;
- the above advantages cumulatively make it possible to perform acoustic, syntactic and semantic analysis of messages in the same memory using unified processing tools, which, in turn, allows adjusting the analysis processes at any stage using various information from the knowledge base. In turn, the developed module for speech message understanding is itself built as an ostis-system and has an appropriate architecture.

D. Knowledge base

The basis of the knowledge base of any ostis-system (more precisely, the sc-model of the knowledge base) is a hierarchical system of subject domains and their corresponding ontologies. The upper level of the hierarchy of the part of the knowledge base related directly to speech assistants is shown below.

Voice assistant knowledge base
⇐ *section decomposition**:
{• *Section. Subject domain of messages*
• *Section. Subject domain of dialogue*
⇐ *section decomposition**:
{• *Section. Subject domain of dialogue control*
• *Section. Subject domain of dialogue participants*
}
• *Section. Subject domain of a smart home*
}

The knowledge base of the ostis-based system has already been partially described in [32], but it has been expanded and has undergone some changes, which will be discussed in more detail below.

Next, let us take a closer look at some of the above subject domains.

E. Subject domain of messages

Below is the upper level of the updated and refined message classification according to various criteria that do not depend on the subject domain.

atomic message
⇒ *decomposition**:
{• *interrogative message*
• *imperative message*
• *declarative message*
}
⇒ *decomposition**:
{• *message without an emotional coloring*
• *message with an emotional coloring*
}
⇒ *decomposition**:
{• *message about the past*
• *message about the present*
• *message about the future*
}

An atomic message is a message that does not include other messages.

interrogative message
⊃ *information request message*

imperative message
⊃ *wish message*

declarative message
⇒ *decomposition**:
{• *informational message*
• *neutral message*
⊃ *greeting message*
⊃ *valedictory message*
}

A neutral message is a declarative message that does not carry new information and does not confirm or deny formerly known information. An example of a neutral message is a message that contains information known to the system, but that is not an answer to a question asked to confirm/deny this information.

informational message
⇒ *decomposition**:
{• *informing message*
• *message of denial of information*
• *message of confirmation of information*
}

message with an emotional coloring
⇒ *decomposition**:
{• *message with a negative emotional coloring*
• *message with a neutral emotional coloring*
• *message with a positive emotional coloring*
• *message with an undefined emotional coloring*
}

A message with a negative emotional coloring is an atomic message that expresses a negative emotion; negative emotions include anger, fear, hate, fright, etc. A message with a neutral emotional coloring is an atomic message that expresses a neutral emotion; neutral emotions include curiosity, surprise, indifference, etc. A message with a positive emotional coloring is an atomic message that expresses a positive emotion; positive emotions include pleasure, exultation, love, etc. A message with an undefined emotional coloring is an atomic message that has an undefined emotional coloring, on the basis of which it is difficult to determine an

emotion; this type of message can appear when a person demonstrates several, usually conflicting, emotions.

F. Subject domain of dialogue participants

To represent information about the participants of the dialogue, an appropriate model of the subject domain and ontology are created. The structure of this subject domain is shown below:

Subject domain of dialogue participants
⇒ *private subject domain**:
- *Subject domain of biography*
  ⇒ *private subject domain**:
  - *Subject domain of organizations*
  - *Subject domain of territorial entities*
  - *Subject domain of living beings*
  - *Subject domain of awards*
  - *Subject domain of education*
- *Subject domain of personal characteristics*
  ⇒ *private subject domain**:
  - *Subject domain of mental states*
    ⇒ *private subject domain**:
    - *Subject domain of emotions*
    - *Subject domain of mood*
  - *Subject domain of personality types*

The subject domain of biography contains means of describing factual data from the biography of the interlocutor, such as:

- participation in any organizations and their characteristics;
- relations that connect them with territorial entities (such as the *place of birth**);
- information about relatives and marital status;
- awards received (including honorary titles);
- educational qualifications.

The subject domain of personal characteristics contains means of describing the mental state of the interlocutor as well as their personality type. The need to store this information is due to one of the goals of the system – to maintain a good mood in the interlocutor, which results in the need to take into account their current state during the dialogue. Within the framework of this subject domain, the following classes of mental states are distinguished:

mental state
⇐ *decomposition**:
{
- *superficial mental state*
- *deep mental state*

}
⇐ *decomposition**:
{
- *conscious mental state*
- *unconscious mental state*

}
⇐ *decomposition**:
{
- *personality-related mental state*
- *mental state caused by the situation*

}
⇐ *decomposition**:
{
- *positive mental state*
- *negative mental state*
- *neutral mental state*

}
⇐ *decomposition**:
{
- *short-term mental state*
- *long-term mental state*
- *medium duration mental state*

}

Figure 9 shows a fragment of the description in the knowledge base of a specific user known to the system.

The above description contains both long-term information about the user, which will be saved after the end of the dialogue (gender, name, etc.) and short-term one, which can be updated with each new dialogue – information about the age, date of the last visit, mood, etc.

Each element of the *beginning* set is a class of temporary entities (i.e., entities with temporal characteristics: duration, initial time, final point, etc.), which have the same moment of the beginning of their existence. The concrete value of this parameter can be either an exact value or a discreet/interval one.

Each element of the *completion* set is a class of temporary entities that have the same final moment of their existence (the moment of the end of existence). The specific value of this parameter can be either an exact value or a discreet/interval one.

G. Subject domain of a smart home

For the usage of the system within the framework of the "smart home", a corresponding subject domain was introduced, which contains the means of describing both the building itself and the devices used in it as well as their operation status (the state of lighting, the presence of humans, etc.).

H. Problem solver

The problem solver of any ostis-system (more precisely, the sc-model of the ostis-system problem solver) is a hierarchical system of knowledge processing agents in semantic memory (sc-agents) that interact only by specifying the actions they perform in the specified memory.

In comparison with **??**, the structure of the problem solver has been revised to make it possible to use it within the framework of the smart home system. The top level of the updated hierarchy of agents of the problem solver of the speech assistant in the SCn-language looks like follows:

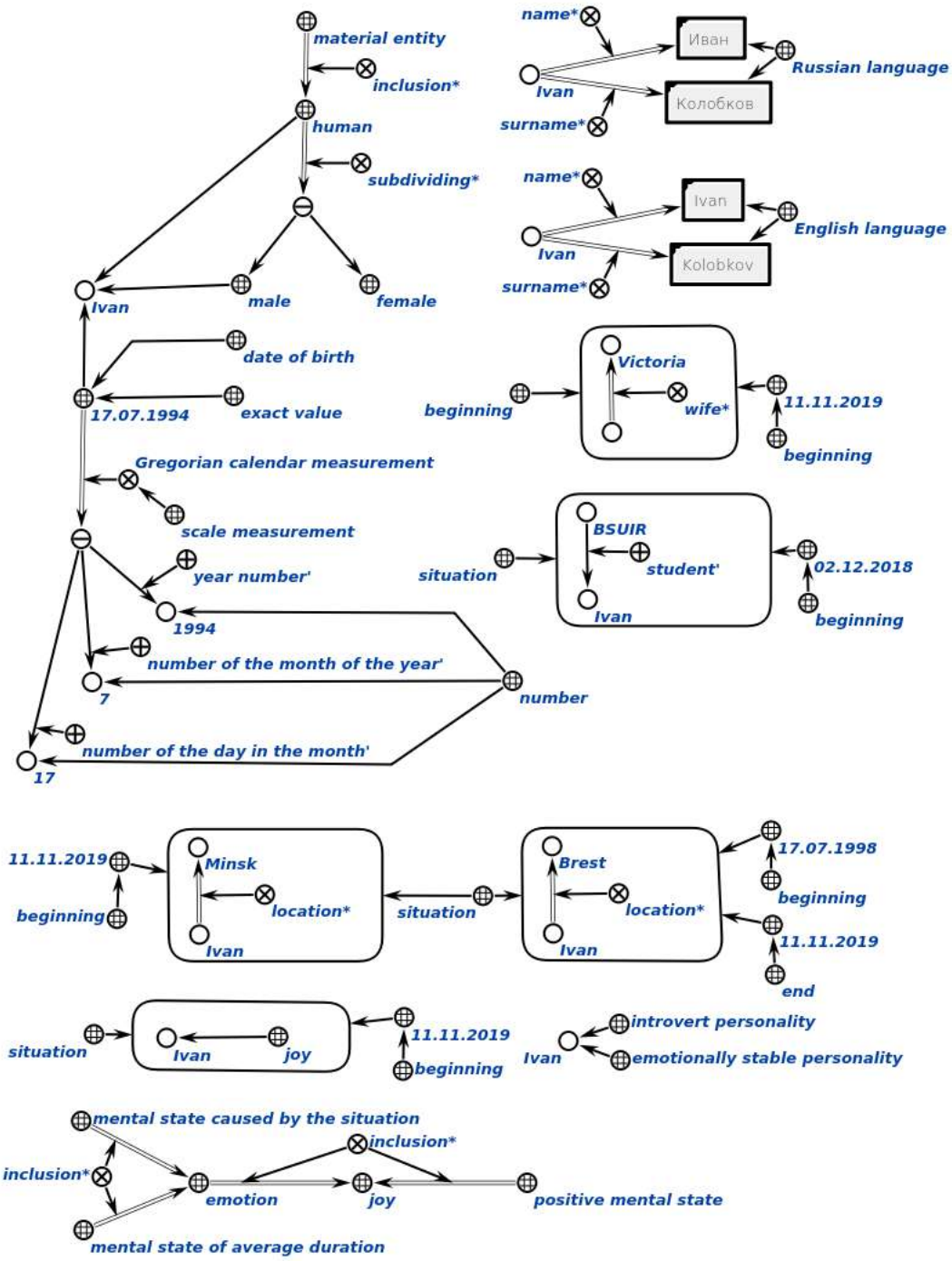

Figure 9. The first part of the model description

Speech assistant problem solver

⇐ *decomposition of an abstract sc-agent**:

{• *Agent of logical inference*

• *Message processing agent*

⇐ *decomposition of an abstract sc-agent**:

{• *Agent of decomposition of a non-atomic message into atomic ones*

• *Agent of allocation of entities*

• *Agent of allocation of relations*

• *Message classification agent*

}

• *Smart home control agent*

⇐ *decomposition of an abstract sc-agent**:

{• *Command classification agent*

• *Agent of the control of executive devices*

}

- *Message generation agent*
- *Agent of speech-to-text transition*
- *Agent of text-to-speech transition*

}

A *smart home control agent* is designed to distinguish the necessary actions depending on the events that have occurred and their performance.

The *Message processing agent* is designed to distinguish the meaning of input natural-language messages. Depending on the result of processing the received message, in particular, its class, in the future, either only the Message generation agent can be called or the Smart home control agent can be called in addition to it. In the second case, the receiving of a message that is a command (for example, to turn on lighting) is processed by the Smart home control agent as well as any other event (for example, user recognition in the frame of any of the cameras or the fall of a certain time of day). Thus, the Message generation agent also serves to generate notifications about the performance of actions.

The Agent of logical inference applies logical rules and is used by various agents, including the Message generation agent and the Smart home control agent.

The knowledge base contains the specification of these agents. As an example, the specification of the Agent of text-to-speech transition is shown in figure 10.

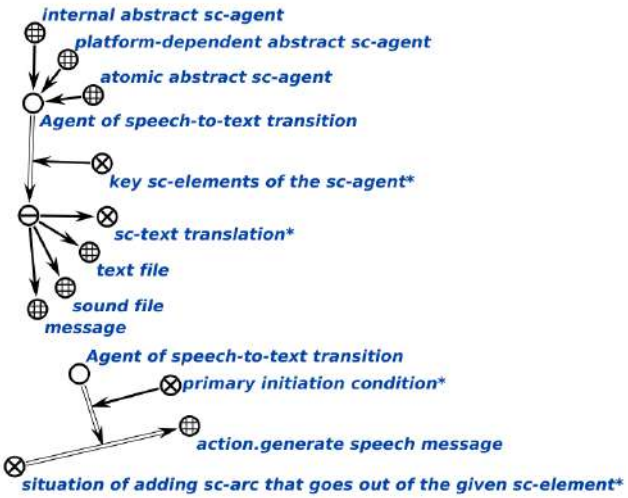

Figure 10. A fragment of the specification of the Agent of text-to-speech transition.

VII. Example of working

As an example, it is possible to give a scenario that includes the following stages:

- the user comes to the front door, the system recognizes them and unlocks the door;
- the user enters the living room, the system turns on the light in this room;
- the user instructs to increase the brightness by 30%.

At the first stage, the following rules are applied: first, according to the rule shown in figure 11, the Smart home control agent opens the door, then, according to the rule shown in figure 12, the Message generation agent generates a message notifying about the door opening. In this case, the text of the message generated by this rule is created according to the template associated with this rule by the *message text pattern** relation.

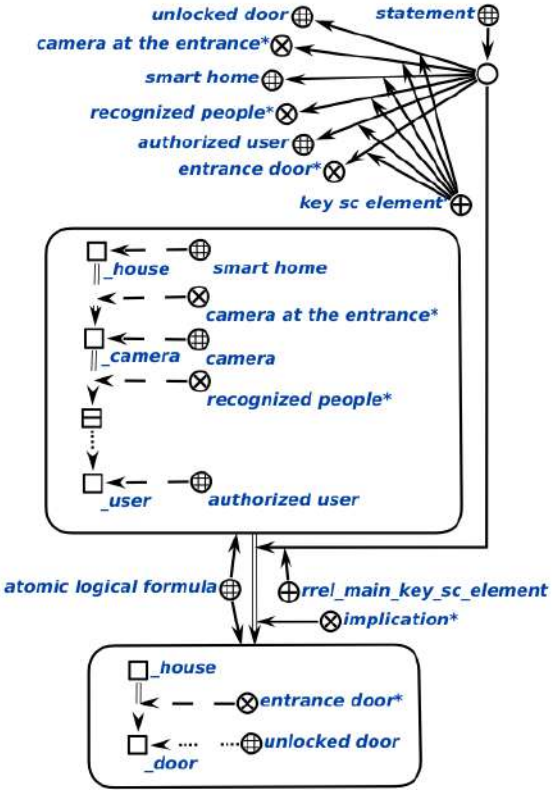

Figure 11. A rule for performing the "opening of the door" action.

To turn on the lighting in the room, if a user has entered it, the rule shown in figure 13 is applied. The message that notifies about the performed action is generated according to the rule similar to what was shown in figure 12.

An example of a rule that takes into account a user message that has been received (and processed by the message processing agent) can be the rule shown in figure 14.

VIII. Conclusion

In the article, one of the possible approaches to the implementation of the psycholinguistic concept of the mental lexicon is proposed, which plays an essential role in the process of human communication as part of intelligent personal assistants. In our opinion, the research of the possibilities of translation of this structure of human consciousness as part of intelligent systems is an important area of research in the field of AI. From a theoretical

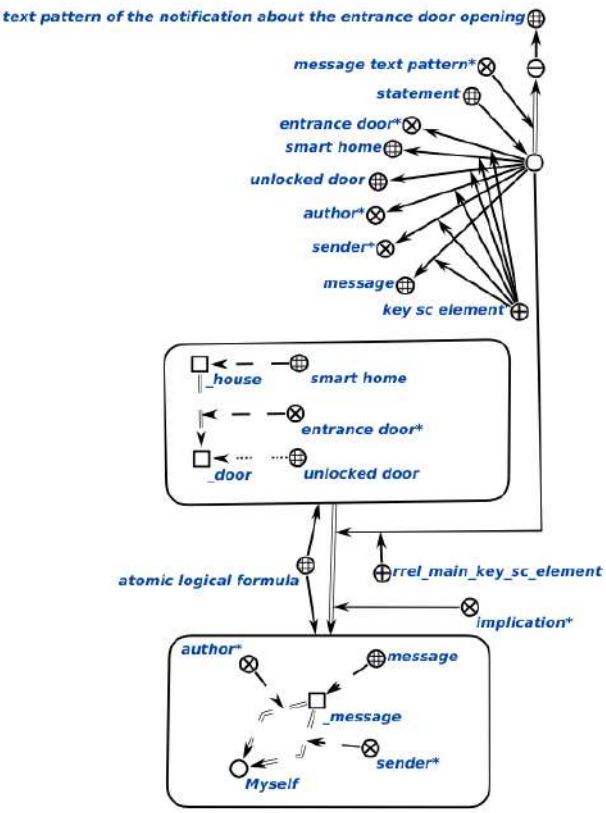

Figure 12. The rule for generating a message that notifies about the opening of the door.

point of view, this will increase the level of understanding of the processes that occur in the human consciousness when perceiving and understanding speech messages, and from a practical point of view, it will make it possible to create intelligent systems with a speech interface that have new qualitative capabilities for understanding speech messages, flexibility and adaptability, the ability to learn directly in the process of interaction with the user.

It is proposed to model the mental lexicon using the semantic network appliance, intelligent agents and knowledge bases and their practical implementation within the framework of the OSTIS Technology, which includes modern implementations of all these aspects of the semantic part of the system. To model the speech recognition process, it is proposed to use the method of semantic-acoustic analysis, which allows direct transiting from the space of acoustic images of words to the space of semantic network concepts that correspond to this image. Thus, a significant part of information processing can be carried out immediately in the semantic domain, bypassing the preliminary stages of speech-to-text transformation that are characteristic of all modern systems.

Thus, the implementation of the intelligent personal assistant system using the mental lexicon model and its implementation within the framework of the OSTIS Technology as well as an approach based on semantic-acoustic speech analysis will give it new useful properties, ensure additional flexibility and adaptability. It will make it possible to overcome some of the limitations that are characteristic of the solutions currently presented on the market and will allow intelligent systems with a speech interface to reach a qualitatively new level.

ACKNOWLEDGMENT

The authors would like to thank the Departments of Intelligent Information Technologies, Control Systems and Electronic Computing Facilities of the Belarusian State University of Informatics and Radioelectronics for the help and valuable comments.

REFERENCES

[1] Gonia J. et al. The Mental Lexicon: Core Perspectives. Gonia Jarema, Gary Libben. BRILL, Jul 1, 2007. Language Arts & Disciplines. 246 p.
[2] Eisner, F., McQueen J. M. Speech perception. Hoboken, Wiley, 2018, Vol. 3, pp. 1–46.
[3] Hoy M. B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical reference services quarterly, 2018, Vol. 37, No. 1, pp. 81–88.
[4] Global Voice Assistant Market By Technology, By Application, By End User, By Region, Competition, Forecast & Opportunities, 2024. Available at: https://www.businesswire.com/news/home/20190916005535/en/Global-Voice-Assistant-Market-Projected-Grow-1.2. (accessed 2021, May)
[5] MacTear M., Callejas Z., Griol D. The Conversational Interface: Talking to Smart Devices. Springer, 2016, 422p.
[6] Hannun A. et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.
[7] Kaldi ASR. Available at: https://kaldi-asr.org. (accessed 2021, May).
[8] VOSK Offline Speech Recognition API. Available at: https://alphacephei.com/vosk/. (accessed 2021, May).
[9] Biswas M. Wit. ai and Dialogflow. Beginning AI Bot Frameworks. Apress, Berkeley, CA, 2018, pp. 67–100.
[10] Bocklisch T. et al. Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:1712.05181, 2017.
[11] Ham D. et al. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 583–592.
[12] Floridi L., Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 2020, vol. 30, No. 4, pp. 681–694.
[13] Oord A. et al. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.
[14] Wang Y. et al. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135, 2017.
[15] Stan A., Lőrincz B. Generating the Voice of the Interactive Virtual Assistant. Virtual Assistant, IntechOpen, 2021.
[16] Deriu J. et al. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review. 2021, vol. 54, No. 1, pp. 755–810.
[17] Treisman, A. M. Contextual cues in selective listening. Quarterly Journal of Experimental Psychology 12, 1960, pp. 242–248.
[18] Oldfield, R. C. Things, words and the brain. Quarterly journal of experimental psychology 18, no. 4, 1966, pp. 340–353.
[19] Kovács L. et al. Networks in the mind–what communities reveal about the structure of the lexicon. Open Linguistics. 2021, vol. 7, No. 1, pp. 181–199.
[20] Cuyckens, H., Dirven R., Taylor J. R. Cognitive approaches to lexical semantics. Walter de Gruyter, 2009, Vol. 23.
[21] Zhang, X., Nannan, L. I. U. Exploring the Second Language Mental Lexicon With Word Association Tests. Cross-Cultural Communication. 2014, Vol. 10, No. 4, pp. 143–148.

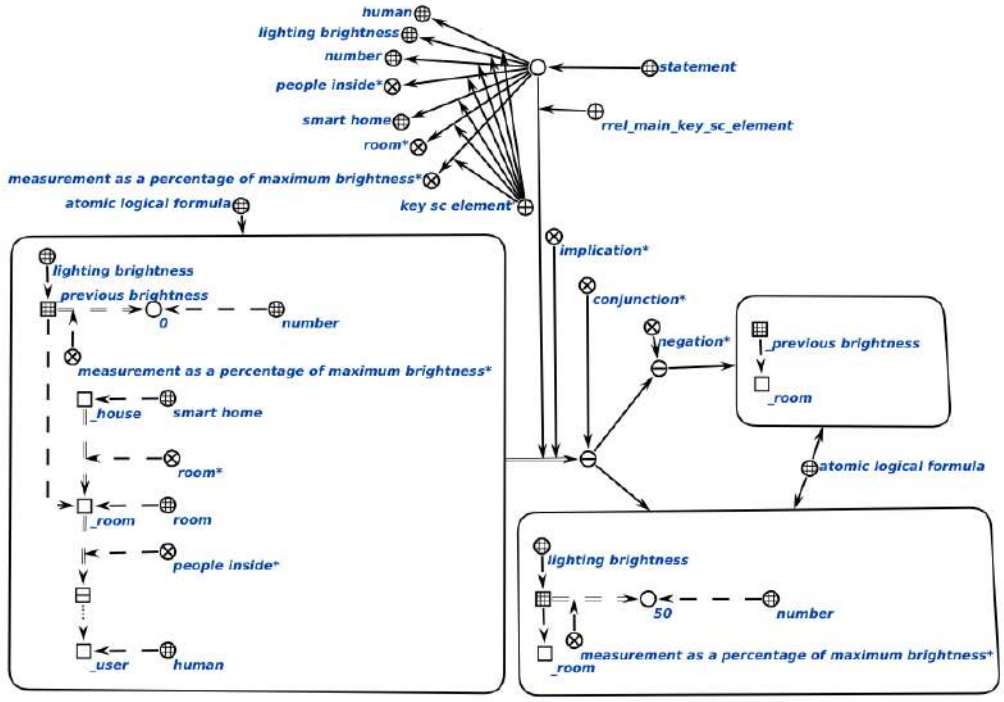

Figure 13. The rule for turning on the lighting in the room when the user enters it.

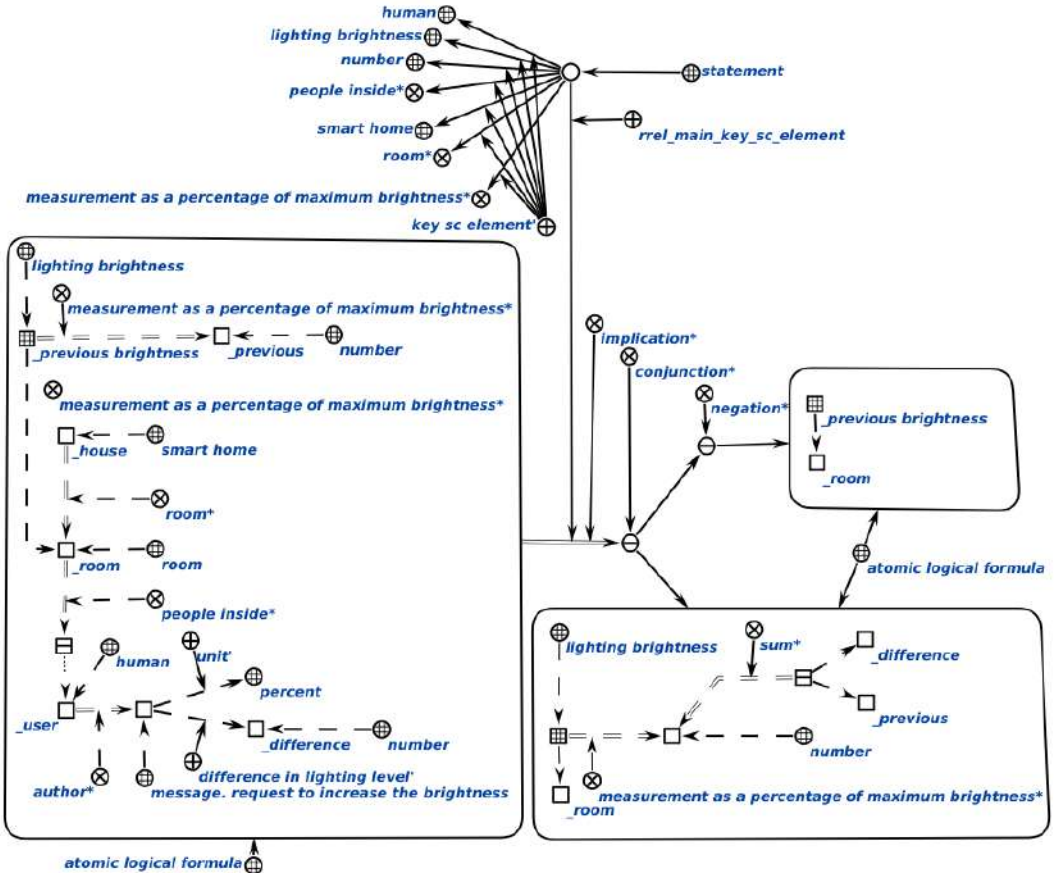

Figure 14. The rule for adjusting the brightness of lighting according to the received user message.

[22] Levelt, W. J. M. et al. Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. The time course of lexical access in speech production: A study of picture naming. Psychological review, 1991, Vol. 98, No. 1, 122 p.

[23] Baayen, R. H. Experimental and psycholinguistic approaches to studying derivation. Handbook of derivational morphology. 2014, pp. 95–117.

[24] Pirrelli, V. et al. Psycho-computational modelling of the mental lexicon. Word Knowledge and Word Usage. De Gruyter Mouton, 2020, pp. 23–82.

[25] Gaskell, G., Marslen-Wilson W. Inference processes in speech perception. Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. Routledge, 2019, pp. 341–345.

[26] Helfer, K. S., Merchant, G. R., Wasiuk, P. A. Age-related changes in objective and subjective speech perception in complex listening environments. Journal of Speech, Language, and Hearing Research. 2017, Vol. 60, No. 10, pp. 3009–3018.

[27] Sowa, J. F. Principles of semantic networks: Explorations in the representation of knowledge. Morgan Kaufmann, 2014, 594 p.

[28] Han, J. et al. Semantic networks for engineering design: A survey. Proceedings of the Design Society. 2021, Vol. 1, pp. 2621-2630.

[29] Shunkevich, D. V. Ontology-based design of knowledge processing machines. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BSUIR, 2017, pp. 73–94.

[30] Zahariev V.A., Azarov E.S., Rusetski K.V. An approach to speech ambiguities eliminating using semantically-acoustical analysis. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BSUIR, 2018, pp. 211–222.

[31] Zahariev V. A., Lyahor T., Hubarevich N., Azarov E.S. Semantic analysis of voice messages based on a formalized context. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BSUIR, 2019, pp. 103–112.

[32] Zahariev V., Shunkevich D., Nikiforov S. and Aarov E. "Intelligent voice assistant based on open semantic technology," in-OpenSemantic Technologies for Intelligent System, ser. Communicationsin Computer and Information Science, V. Golenkov, V. Krasno-proshin, V. Golovko, and E. Azarov, Eds. Springer, Cham, 2020,pp. 121–145.

[33] Stylinou, Y. Applying harmonic plus noise model in concatenative speech synthesis. IEEE Trans. on Speech and Audio Processing. 2001, Vol. 9, No. 1, pp. 21–29.

[34] Serra, X. A system for sound analysis / transformation / synthesis based on deterministic plus stochastic decomposition. PhD thesis. Stanford. 1989, 178 p.

[35] Azarov E., Vashkevich M., Petrovsky A. Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation. INTERSPEECH 2013: proceedings of 12th Annual Conference of the International Speech, Lyon, France, 2013, pp. 1697–1701.

[36] Lerch, A. An introduction to audio content analysis: Applications in signal processing and music informatics. Wiley-IEEE Press, 2012, 259 p.

[37] Froitzheim, S. A Short Introduction to Audio Fingerprinting with a Focus on Shazam. Available at: https://hpac.cs.umu.se/teaching/sem-mus-17/Reports/Froitzheim.pdf (accessed 2021, May)

[38] Davydenko I.T. Ontology-based knowledge base design. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BSUIR, 2017, pp. 57–72.

[39] Leont'ev A. A. Osnovy psikholingvistiki [Fundamentals of psycholinguistics]. Smysl, NPF Smysl, 2005, 310p. (in Russian)

[40] Vygotskii, L. C. Myshlenie i rech'. Psikhologicheskie issledovaniya. [Thinking and speaking. Psychological research] – Natsional'noe obrazovanie, 2015, 368 p. (in Russian)

[41] Chernigovskaja T. V. Chto delaet nas ljud'mi: pochemu nepremenno rekursivnye pravila? Razumnoe povedenie i jazyk [What makes us human: why recursive rules necessarily? Reasonable behavior and language]. 2008, No. 1, pp. 395–412. (in Russian)

[42] Chernigovskaja T. V. Jazyk, mozg i komp'juternaja metafora [Language, Brain and Computer Metaphor]. Chelovek [Human]. 2007, Vol. 2, pp. 63–75. (in Russian)

[43] Chernigovskaja T. V. Eshhe raz o mozge i semiozise: mozhno li najti tochku v nejrosetjah? [Once again about the brain and semiosis: is it possible to find a point in neural networks?] Voprosy filosofii [Philosophy questions]. 2021, No. 6, pp. 5–13. (in Russian)

[44] Zalevskaja A. A. Slovo v leksikone cheloveka: psiholingvisticheskie issledovanija [A word in the human lexicon: psycholinguistic research]. M. M. Kopylenko Eds. Voronezh, VGU, 1990, 204 p. (in Russian)

[45] Zalevskaja A. A. Mental'nyj leksikon s pozicij raznyh podhodov [Mental'nyj leksikon s pozicij raznyh podhodov]. Aktual'nye problemy sovremennoj lingvistiki [Actual problems of modern linguistics]. L. N. Churilina Eds. 4-e izd., 2009, pp. 311–327. (in Russian)

[46] Popova Z. D. et al. Popova i Polevye struktury v sisteme jazyka [Popova and Field Structures in the Language System]. Z. D. Popova Eds. Voronezh, 1989, 198 p. (in Russian)

[47] Zalevskaja A. A. Psiholingvisticheskie issledovanija. Slovo. Tekst: izbran. trudy [Psycholinguistic research. Word. Text: selected. labors]. Moskva, Gnozis, 2005, 542 p. (in Russian)

[48] Vencov A.V., Kasevich V.B. Problemy vosprijatija rechi [Speech perception problems]. Moskva, Izdatel'skaja gruppa URSS, 2003, 240 p. (in Russian)

[49] Saharnyj L. V. Psiholingvisticheskie aspekty teorii slovoobrazovanija: ucheb. posobie [Psycholinguistic aspects of the theory of word formation: textbook. allowance]. L., LGU, 1985, 408 p. (in Russian)

[50] Stern A. S. Perceptivnyj aspekt rechevoj dejatel'nosti: jeksperimental'noe issledovanie [Perceptual aspect of speech activity: an experimental study]. Sankt-Peterburg, Izd-vo S.-Peterb. un-ta, 1992, 236 p. (in Russian)

[51] Karaulov J. N. Aktivnaja grammatika i associativno-verbal'naja set' [Active grammar and the associative-verbal network]. M., Institut russkogo jazyka RAN, 1999, 180 p. (in Russian)

[52] Zalevskaja A. A. Vvedenie v psiholingvistiku : ucheb. dlja studentov vuzov, obuch. po filol. spec. [Introduction to psycholinguistics: textbook. for university students, training. by philol. specialist.] M., RGGU, 1999, 381 p. (in Russian)

[53] Prokopenja V. K., Sljusar' N. A., Petrova T. E., Chernova D. A. & Chernigovskaja T. V. Jeksperimental'nye issledovanija grammatiki: ustanovlenie anaforicheskih otnoshenij v processe recheponimanija [Experimental studies of grammar: the establishment of anaphoric relations in the process of comprehension]. Voprosy jazykoznanija [Linguistic issues]. 2018, No. 1, pp. 76–90. (in Russian)

[54] Golenkov V. V. Principy postroenija massovoj semanticheskoj tehnologii komponentnogo proektirovanija intellektual'nyh sistem [Principles of building mass semantic technology for component design of intelligent systems]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BGUIR, 2011, pp. 21–58. (in Russian)

[55] Skorohod'ko J. F. Semanticheskie seti i avtomaticheskaja obrabotka teksta [Semantic networks and automatic word processing]. Kiev, Nauk. dumka, 1983. (in Russian)

[56] Skrjegg G. Semanticheskie seti kak modeli pamjati [Semantic networks as memory models]. Novoe v zarubezhnoj lingvistike [New in foreign linguistics]. No. 12, M., Raduga, 1983, pp. 228–271. (in Russian)

[57] Peirce Ch. S. Logicheskie osnovanija teorii znakov [Logical foundations of the theory of signs]. SPb., Izd-vo Sankt-Peterb. un-ta: Aleteja. 2000, 352 p. (in Russian)

[58] Sapatyj P.S. Ob jeffektivnosti strukturnoj realizacii operacij nad semanticheskimi setjami [Ob jeffektivnosti strukturnoj realizacii operacij nad semanticheskimi setjami]. Tehn. kibernet. 1983, No. 5, pp. 128–134. (in Russian)

[59] Shunkevich, D. V. Vzaimodejstvie asinhronnyh parallel'nyh processov obrabotki znanij v obshhej semanticheskoj pamjati [Interaction of Asynchronous Parallel Knowledge Processing Processes in Shared Semantic Memory]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open se-

mantic technologies for intelligent systems]. Minsk, BGUIR, 2016, pp. 137–144. (in Russian)
[60] Harlamov A. A. Semanticheskie seti kak formal'naja osnova reshenija problemy integracii intellektual'nyh sistem. Formalizm avtomaticheskogo formirovanija semanticheskoj seti s pomoshh'ju preobrazovanija v mnogomernoe prostranstvo [Semanticheskie seti kak formal'naja osnova reshenija problemy integracii intellektual'nyh sistem. Formalizm avtomaticheskogo formirovanija semanticheskoj seti s pomoshh'ju preobrazovanija v mnogomernoe prostranstvo]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BGUIR, 2011, pp. 87–96 (in Russian)
[61] Golenkov V. V., Guljakina N. A. Strukturizacija smyslovogo prostranstva [Structuring the semantic space]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BGUIR, 2014, pp. 65–78. (in Russian)
[62] Golenkov V. V., Guljakina N. A. Intellektual'nye sistemy. Problemy i perspektivy [Intelligent systems. Problems and Prospects]. Informacionnye tehnologii i sistemy 2016 (ITS 2016). Minsk, BGUIR, 2016, pp. 13–20. (in Russian)
[63] Azarov I.S., Petrovskii A.A. Mgnovennyi garmonicheskii analiz: obrabotka zvukovykh i rechevykh signalov v sistemakh mul'timedia [Instant harmonic analysis: processing of sound and speech signals in multimedia systems]. LAP Lambert Academic Publishing, Saarbrucken. 2011, 163 p. (in Russian)
[64] Aificher E., Dzhervis B. Tsifrovaya obrabotka signalov: prakticheskii podkhod [Digital Signal Processing: A Practical Approach]. 2-e izd. [Digital signal processing: a practical approach. 2nd ed.]. M., Williams, 2004, 992 p. (in Russian)
[65] Rabiner L. R., Schafer R. W. Cifrovaja obrabotka rechevyh signalov [Digital processing of speech signals]. M.V. Nazarova, Ju. N. Prohorova Eds. M., Radio i svjaz', 1981, 496 p. (in Russian)
[66] Golenkov V. V., Guljakina N. A. Semanticheskaja tehnologija komponentnogo proektirovanija sistem, upravljaemyh znanijami [Semantic technology of component design of knowledge-driven systems]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]. Minsk, BGUIR, 2015, pp. 57–78. (in Russian)

Анализ диалоговой речи на основе формализованного представления ментального лексикона

Захарьев В.А., Никифоров С.А., Азаров И.С.

В работе предложен один из возможных подходов по реализации психолингвистического концепта ментального лексикона, играющего существенную роль в процессе человеческого общения, в составе интеллектуальных речевых ассистентов. Изучение возможностей воплощения данной структуры человеческого сознания в составе интеллектуальных систем, по-нашему мнению, является важным направлением исследований в области ИИ. Подобные исследовании с теоретической точки зрения позволят повысить уровень понимания процессов происходящих в человеческом сознании при восприятии и понимании речевых сообщений, а с практической точки зрения, позволят создавать интеллектуальные системы с речевым интерфейсом обладающие новыми качественными возможностями понимания речевых сообщений, гибкостью и адаптивностью, способностью обучаться непосредственно в процессе взаимодействия с пользователем.

Предлагается осуществлять моделирование ментального лексикона с применением аппарата семантических сетей, интеллектуальных агентов и баз знаний, и их практического воплощения в рамках технологии OSTIS, включающей в себя современные реализации всех данных аспектов смысловой части системы. Для моделирования процесса распознавания речи предлагается использовать метод семантико-акустического анализа, позволяющий осуществить прямой переход из пространства акустических образов слов, в пространство понятий семантической сети, соответствующих данным образом. Таким образом существенная часть обработки информации может вестись сразу в семантической области минуя предварительные этапы преобразования речи в текст.

Это позволит преодолеть некоторые ограничения, характерные для представленных в текущий момент времени на рынке решений, позволит обеспечить выход интеллектуальных систем с речевым интерфейсом на качественно новый уровень.

Received 28.05.2021