# The principles of building intelligent speech assistants based on open semantic technology

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

***Abstract*—The speech interface is one of the most natural and convenient ways of organizing user interactions with intelligent systems. Nowadays universal voice assistants are implemented by the largest world companies for various platforms and have gained popularity among users.**

**The aim of this work is to develop the principles of building voice assistants for intelligent computer systems based on OSTIS Technology [8]. The main requirements in this case are ensuring maximum independence from the subject area and providing the possibility of adaptation (primarily automatic) of the universal assistant for a specific user and the features of a specific subject area.**

***Keywords*—dialogue system, voice assistant, semantic network, spoken language understanding**

## I. Introduction

The actual field of application of voice user interfaces in modern information and communication technologies is the development of voice assistants - dialogue systems that allow users to access knowledge and interact with agents of the [11] intelligent system in a user-friendly manner.

Universal voice assistants developed by the world's largest companies are widely used in modern smartphones and operating systems. Their development trends lie on the transformation of stand-alone applications into platforms that can be deployed on devices of various manufacturers, and on the basis of which various services built on the ground of intelligent information technologies [14], [15]. This fact allows manufacturers of both electronics and software to open new niches in consumer markets using the latest advances in artificial intelligence technology.

Analytical agencies forecast that the combined annual growth rate of the global market for products using speech assistant technology will be more than 30%, increasing from $1.2 billion in 2018 to $5.4 billion by 2024 [7]. According to their estimates, this trend will contribure to two factors, such as an increase in the total number of smartphones and the expansion of the standard functions of speech interfaces: managing the dialogue context, personalization, the ability to conduct dialogue in several languages or respond not only in voice but also in text mode.

## II. Analysis of existing solutions and problem statement

Consider the architecture and principles of the voice assistant, peculiar to most modern solutions. In our description, we will focus on the voice assistant "Alexa" from the company "Amazon". This product is currently the most popular assistant (with the exception of mobile phones) in the speech technology market (about 65%) [1], based on a modern stack of speech technologies. Many other major players in the speech market such as "Google" ("Google Assistant"), "MicroSoft" ("Cortana"), "Yandex" ("Alice") are trying to adopt solutions specific to "Alexa" [10] in their products. Therefore, the architecture under consideration can be considered typical.

Modern voice assistant forms a distributed software and hardware system consisting of two main parts: front end and back end (Fig. 1).

The front end part is deployed on a specialized device or installed as an application on the user's device. The client is responsible for issues related to capturing and playing back audio information, pre-processing, wake-word system triggering activation, encoding and decoding data, and generating backend requests. Access to the server part is carried out through the corresponding program interface, most often REST [4].

The back end part includes following main components.

- spoken language understanding subsystem (SLU), which consists of automatic speech recognition module (ASR) and natural language understanding (NLU) module;
- dialogue management module (DM) includes a subsystem of "Skills" (general type like weather, music, navigation and specialized like web search, wiki, news etc.).
- natural language messages generator (NLG) module.

Let's review server components in detail. The Speech Understanding Subsystem (SLU) converts the input signal into some semantic equivalent of the spoken phrase - "intent", which reflects the meaning of the spoken phrase and the user's expectations. This subsystem includes:

The automatic speech recognition module (ASR) of speech implements the process of converting speech into text. Acoustic and linguistic models are used to describe this process. Algorithms that implement the corresponding models make it possible to determine fragments of the speech wave in the audio signal equal to the basic phonetic units of the target language and form a phonetic text from them. Then a spelling text is obtained which is based on the relevant rules of morphology and grammar. Modern ASR implementations involve the use of a combining statistical methods such as hidden Markov models (HMM) of neural network methods based on convolutional (CNN) and recurrent neural networks (LSTM-RNN) [5], [17].

The natural language understanding module (NLU) implements a natural language text processing sequence that includes the main stages: tokenization, defining the boundaries of sentences, parts of speech tagging, named entities recognition, syntactic and semantic analysis. At the output of this module, an entity is formed called "intent", which is a structured formalized user request. It conveys the main meaning of what was said and the wishes of the user. A frame model of utterances is used to form an "intention", as a result of which an object is formed which is transferred to the next module.

The dialogue management manager (DM) is a module that directly controls the dialogue, its structure and progress, fills the

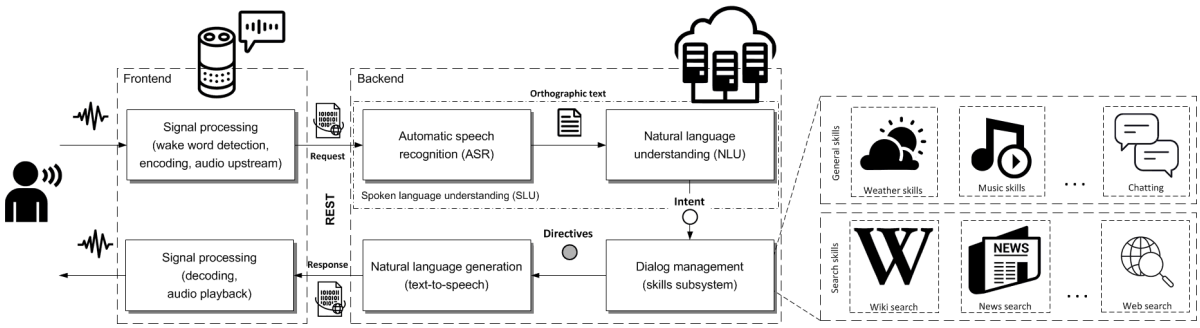


Figure 1. Dialogue system architecture

general context, contains knowledge about a specific subject area in the form of compact fragmented ontologies in the semantic network and the rules for responding to them which are known as “skills”. It receives the input data from the SLU component in the form of incoming "intentions" and must select the necessary "skill" block, then attains the global state of the dialogue process and transfers the output data to the generation module in the form of a "directive" object. The figure 2 shows examples of the description of intent and "directive".

The natural language generation module (NLG) synthesizes the voice response message in accordance with the available signal and speaker voice model and text of the response message which is located in the "directive" object. In connection with the steady tendency towards personalization of devices and programs, the possibility of adapting systems to the voice of a new speaker is one of the interesting features that we would like to consider in our work and which is not available in current solutions on the market. According to research [13] the voice of a person known to the listener is perceived 30% better than the voice of an unfamiliar person. Changing the speaker’s voice during the speech synthesis process allows you to attract the attention of system users to key information in a voice message, to emphasize especially important events or moments. It helps to improve the ergonomics of the system.

```
{
    "intent": "GetHoroscope",
    "sign": "libra",
    "date": "2019-12-31"
}
```

a) Intent object example

```
{
    "namespace": "Speech",
    "name": "speak",
    "text": "The horoscope
    for Libra."
}
```

b) Directive object example

Figure 2. Intent and directive object example

It should be noted that this architecture has proven itself in all modern solutions. However, the current situation in the market of voice assistants, from our point of view, has several unresolved problems and limitations:

- The main parts (semantic processing of information in the NLU and DM modules) of the system are proprietary closed. Developers and researchers do not have the ability to make changes to the knowledge base, supplement and modify existing ontologies. There is only the opportunity to use these modules as a service without directly participating in their development. Also, scientific and practical details related to the methods of formalization and processing of knowledge were not disclosed, which does not allow comparing the proposed solutions with alternative technologies and implementations.
- All common voice assistants have an exclusively distributed implementation, where the main part is located on the server side. There are no alternative, the so-called "on the eadge" solutions that allow you to deploy the system in an independent local environment, for example, on your servers in your own data center or directly on the client device. Such a method would make it possible to ensure stable operation of the system in cases where there is no stable Internet connection and could also be in demand if the user does not want to transmit their personal data to companies in the form of voice fragments and descriptions of their requests in the system in form of “intentions” objects ( intent) and “directives”. This thesis is of particular relevance in connection with the increasing incidence of leakage of personal information from large companies [12], [6].

In this regard, from our point of view, the urgent task is to build a voice assistant based on open semantic technologies that allow a large group of developers to participate in the design and extend the knowledge base. To solve this problem, it is necessary to formulate a number of requirements for such an assistant.

Analysis of the user needs (including various companies) allows us to present the following set of functional requirements to the developed speech assistants:

- speaker independent recognition, the ability to correctly recognize messages from various interlocutors, possibly with speech dysfunctions;
- in a situation where the message is not understood by the system or clarification of any parameters is required, the system should be able to ask the interlocutor clarifying questions;
- the ability to recognize a speaker by voice, as a result - the ability to conduct a dialogue simultaneously with a group of interlocutors;
- the ability to work in conditions of noise interference;
- the ability to accumulate and take into account the history of the dialogue with the same interlocutor for a long time (to build and store a portrait of the interlocutor);
- the ability to take into account the current state of the

user, including his emotions, as well as such individual characteristics as gender, age, etc.;
- the speech assistant can receive information about the interlocutor not only directly from the dialogue, but also have predefined information about him that is of interest in the context of the current dialogue;
- the speech assistant can conduct a dialogue of an infotainment nature (to answer user questions or conduct a conversation without a specific goal), and to pursue a specific goal that affects the dialogue process (for example, to calm or amuse a person to talk to).

The development of speech assistants that meet these requirements is hindered by a number of problems. Some problems were considered and partially solved by the authors in previous works:

- in [18], the problem of identifying and eliminating ambiguities (including those associated with speech defects, noise, etc.) in a speech signal due to a knowledge base is considered;
- in [19], an approach to the description of the context of the dialogue with the possibility of its consideration in the analysis of voice messages is considered;

In this paper, the main attention will be paid to the principles of dialogue organization (situational dialogue management), description of the user model, as well as mechanisms for adapting the dialogue process to the characteristics of a specific user and specific subject area.

An important feature that distinguishes dialogue system (in specific domain area) from universal voice assistents is the lack of the need to understand the meaning of the message completely; more often, to generate the required response, it is enough to determine the type of message and select some keywords. This feature significantly reduces the complexity of the problem being solved and will be taken into account further when detailing the proposed approach.

## III. Proposed approach

OSTIS Technology and the corresponding set of models, methods and tools for developing semantically compatible intelligent systems as the basis for building voice assistants are proposed here. The basis of OSTIS Technology is a unified version of the information encoding based on semantic networks with set-theoretic interpretation, called the SC code [8].

The architecture of each system built on OSTIS Technology (ostis-systems) includes a platform for interpreting semantic models of ostis-systems, as well as a semantic model of ostis-systems using SC-code (sc-model of ostis-systems). In turn, the sc-model of the ostis-system includes the sc-model of the knowledge base, the sc-model of the task solver and the sc-model of the interface (in particular, the user interface). The principles of engineering and designing knowledge bases and problem solvers are discussed in more detail in [9] and [16], respectively.

Models and tools application proposed by OSTIS Technology will provide, in addition to the advantages indicated in the above works, the opportunity to

- create, store and analyze the user's portrait, including both long-term information and its current state;
- save and analyze the history of the dialogue with each interlocutor;
- clearly distinguish the part of the dialogue management, depending only on the meaning of the messages and not depending on the language in which the dialogue is conducted, and the part depending on the language of dialogue;
- integrate the subject-independent part of the knowledge base with the subject-dependent part within each system, which will allow you to flexibly take into account the characteristics of the subject area when conducting dialogue.

Further in the text, we will assume that the dialogue is carried out in Russian, however, most of the models presented do not depend on the language in which the dialogue is conducted. To write formal texts within the framework of the work, we will use options for external display of SC-code constructions such as SCg (graphic version) and SCn (hypertext version).

## IV. System architecture

The general dialogue scheme can be written as follows:

- The user delivers a voice message;
- The speech analysis module, based on the dictionary of speech identifiers available in the knowledge base of the system, selects keywords within the message and correlates them with entities in the knowledge base;
- Based on the rules available in the knowledge base, the system classifies the received message;
- Based on the rules available in the knowledge base, the system generates a new message addressed to the current interlocutor (possibly non-atomic);
- The speech synthesis module generates a fragment of the speech signal corresponding to the new message and voices it;

### A. Speech analysis

To perform the processing of the speech signal inside the ASR module, it is necessary to fulfill it analysis and parametric description, i.e. represent as a sequence of characteristic vectors of the same dimension.

The proposed signal model and method for its evaluation has a number of distinctive features. They are a discrete Fourier transform or determination of the autocorrelation function of a signal in a short fragment. The method under consideration does not impose strict restrictions associated with the observance of the stationary conditions of the signal parameters on the analysis frame. This allows one to obtain a high temporal and frequency resolution of the signal, as well as a clearer spectral picture of the energy localization at the corresponding frequencies 3, and as a result, a more accurate estimate of the signal parameters (on average by 10-15 %).

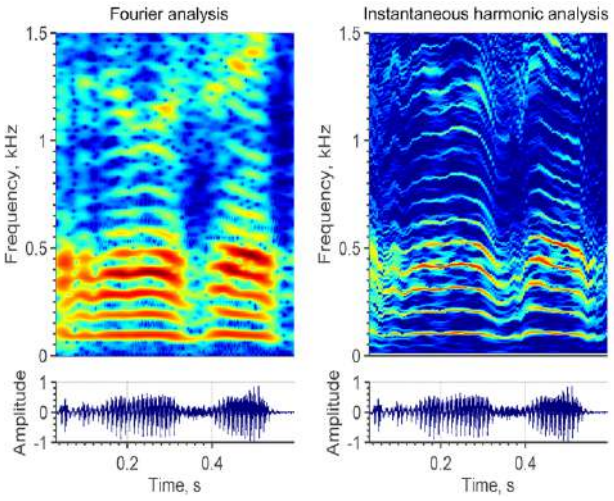


Figure 3. STFT and IHA based spectrograms

The speech signal is divided into overlapping fragments, each of which is described by a set of parameters: the spectral envelope, the instantaneous fundamental frequency (if the fragment is voiced) and the type of excitation, which can be voiced, unvoiced or mixed.

The quasiperiodic component of the speech signal is represented as the sum of sinusoids or the real part of complex exponentials with continuous amplitude, frequency and phase, and noise as a random process with a given power spectral density (PSD):

$$s(n) = \sum_{p}^{P} A_p(n) cos\phi_k(n) + r(n) = \\ = Re[\sum_{p}^{P} A_p(n) \exp j\phi_p(n)] + r(n) \quad (1)$$

where $P$ – number of sinusoids (complex exponentials), $A_p(n)$ – instantaneous amplitude of the $p$-th sinusoid, $\phi_p(n)$ – instantaneous phase of the $p$-th sine waves $r(n)$ – aperiodic component. The instantaneous frequency $F_p(n)$, located in the interval $[0, \pi]$ ($\pi$ corresponds to the Nyquist frequency), is a derivative of the instantaneous phase. It is assumed that the amplitude changes slowly, which means limiting the frequency band of each of the components. Using the obtained harmonic amplitudes of the voiced and PSD unvoiced components, a common spectral envelope is formed.

This set of parameters is extracted from a speech signal using an algorithm consisting of the following steps:

- estimation of the instantaneous fundamental frequency using the error-resistant algorithm for tracking the instantaneous fundamental frequency "IRAPT (Instantaneous Robust Algorithm for Pitch Tracking)" cite Azarov2012;
- deformation of the time axis of the signal to ensure the stationary frequency of the fundamental tone;
- estimation of instantaneous harmonic parameters of a speech signal using a DFT modulated filter bank - each harmonic of the fundamental tone of voiced speech falls into a separate channel of the filter bank, where it is converted into an analytical complex signal from which the instantaneous amplitude, phase and frequency are extracted;
- based on the analysis of the obtained instantaneous frequency values, various regions of the spectrum are classified as periodic and aperiodic;
- harmonics belonging to periodic spectral regions are synthesized and subtracted from the original signal;
- the remainder is transferred to the frequency domain using the short-term Fourier transform;
- parameters of the synthesized harmonics and the PSD of the remainder are combined into one common spectral envelope and translated into a logarithmic scale;
- adjacent spectral envelopes are analyzed to determine how to excite the entire analyzed fragment of the signal.

Each spectral envelope is represented as a vector of logarithmic energy values equally spaced on the chalk scale. For a speech signal with a sampling frequency of 44.1 kHz, a 100-dimensional vector is used. The characteristic vector consists of the fundamental tone frequency values, the spectral envelope and the sign of vocalization of the current speech fragment. The dimension of the vector determines a compromise between the quality of signal reconstruction and computational complexity. Based on practical experiments, it was found that the selected dimension is sufficient for the reconstruction of natural speech.

**Subject domain of dialogue**

⇐ section decomposition:

{

- Section. Subject domain of messages
- Section. Subject domain of dialogue control
- Section. Subject domain of dialogue participants

}

Figure 4. The hierarchy of subject areas.

### B. Knowledge base

The basis of the knowledge base of any ostis-system (more precisely, sc-models of the knowledge base) is a hierarchical system of subject areas and their corresponding ontologies. The figure 4 shows the upper hierarchy of the knowledge base part that relates directly to voice assistants.

Consider in more detail the concepts studied in each of these subject areas and examples of their use.

### C. Message subject area

The figure 5 shows the top level of message classification according to various criteria, independent of the subject area.

An atomic message refers to a message that does not include other messages, in turn, Non-atomic message is a message that includes other messages. At the same time, a non-atomic message can consist of one sentence, but have several semantic parts, for example, "Can I log in if I am not 16 years old?" (age is reported and a question is asked) or "Hello, my name is Sergey." (the name is communicated and a greeting is expressed).

In turn, the presented classes can be further divided into more private ones. The figure 6 shows the classification of interrogative sentences.

### D. Subject area of dialogue participants

To present information about the participants in the dialogue, an appropriate domain model and ontology have been developed.

Figure 7 shows a fragment of the description in the knowledge base of a specific known user system. The above description contains both long-term information about the user, which will be saved after the dialogue is completed (gender, name, etc.) and short-term, which can be updated with each new dialogue - information on age, date of the last visit, mood, etc..

### E. Dialog management area

Within the subject area of dialogue management, rules are presented according to which the analysis of user messages and the generation of response messages are carried out.

In accordance with the above general plan of work, the system distinguishes several categories of rules:

- voice message classification rules;
- rules for generating new messages within the current dialogue;
- voice message generation rules;

To simplify processing, some rules can be written not in the form of logical statements, but with the help of additional relations (for example, keywords that define the class of messages) and their corresponding knowledge processing agents. This hybridization of declarative and imperative recording options is widely used within the framework of OSTIS Technology in order to increase the efficiency of information processing while maintaining consistency of presentation at a basic level.

***message***

⇒ inclusion*:
- incentive message
- interrogative message
- declarative message

⇐ subdividing*:
- {
  - daytime message
  - evening message
  - morning message
  }
- {
  - message with respectful treatment
  - message with standard treatment
  }
- {
  - exclamatory message
  - non-exclamatory message
  }
- {
  - non-atomic message
  - atomic message
  }
- {
  - undefined language message
  - english language message
  - russian language message
  }

Figure 5. Typology of messages.

***interrogative message***

⇐ subdividing*:

{
- complete dictal question
- partial dictal question
- complete modal question
- partial modal question

}

Figure 6. Typology of interrogative messages.

Figure 8 shows an example of a simple rule for classifying messages based on keywords. The shown rule systemizes a message as a welcome class if it contains the appropriate words.

Figure 9 provides a formal definition of an atomic message.

Figure 10 shows a rule requiring you to find out the name of the interlocutor, if it is still not known to the system.

### *F. Subject-dependent fragments of the knowledge base*

If necessary, the subject-independent part of the knowledge base can be supplemented with any information clarifying the specifics of a particular subject area, if it is necessary to improve the quality of the dialogue. The figure 11 shows a fragment of the knowledge base for the speech assistant-waiter of a cafe. The given fragment of the knowledge base includes a description of some drinks composition and confectionery products available in the menu, as well as information about the order made by a specific client.

### *G. Problem solver*

The task solver of any ostis-system (more precisely, the sc-model of the ostis-system task 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.

The top level of the agents hierarchy for the speech assistant task solver in SCn is as follows:

***Voice Assistant Problem Solver***

⇐ *abstract sc-agent decomposition**:

{
- *Logical Rule Enforcement Agent*
- *Voice Message Analysis Agent*

  ⇐ *abstract sc-agent decomposition*:*

  {
  - *Keyword selection agent in a voice message*
  - *Keyword selection agent from a set of words in a speech message*
  - *Translation agent of fragments of a speech message to a knowledge base*

  }
- *Message processing agent within the framework of semantic memory*

  ⇐ *abstract sc-agent decomposition*:*

  {
  - *Message classification agent*
  - *Agent for the formation of the order of atomic messages in the framework of non-atomic*
  - *Agent decomposing non-atomic messages into atomic messages*

  }
- *Voice Concatenation Agent*
- *Voice Message Generation Agent*

}

### *H. Speech synthesis*

One of the requirements for the developed voice assistant indicated in this article is adaptation to a specific speaker. Changing the speaker's voice in the process of speech synthesis allows you to attract the attention of system users to key information in a voice message, emphasize important events or moments, according to studies. The voice of a person known to the listener is perceived 30% better than the voice of an unknown person [44, 45]. In this paper, we would like to show the applicability of the developed methods for signal synthesis with personalized speaker properties, namely, building a personal

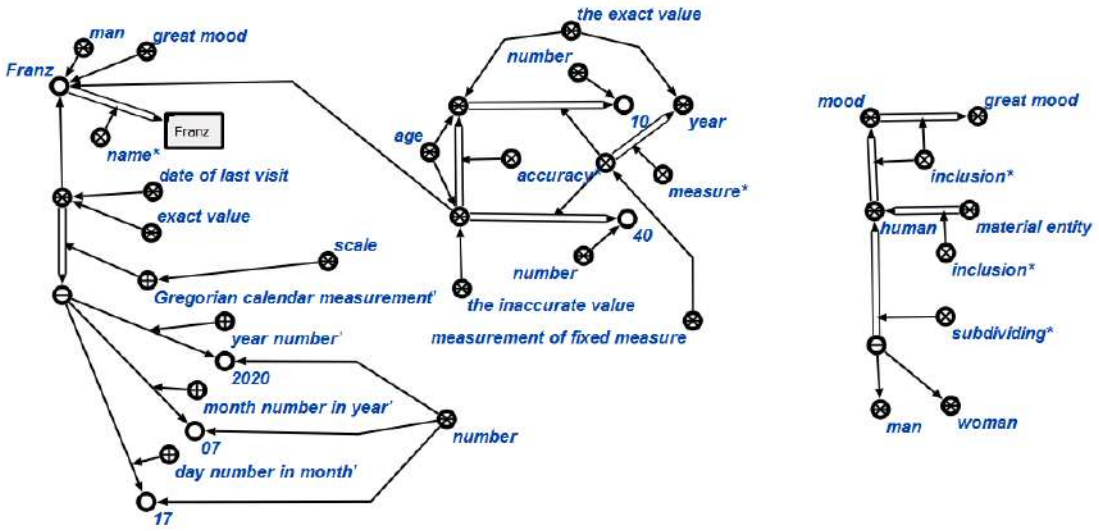


Figure 7. Portrait of a famous user system.

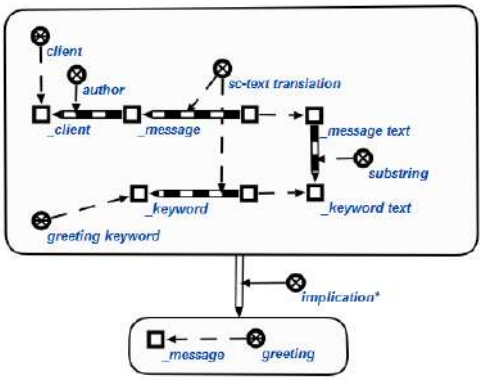


Figure 8. Example rule for classifying a message.

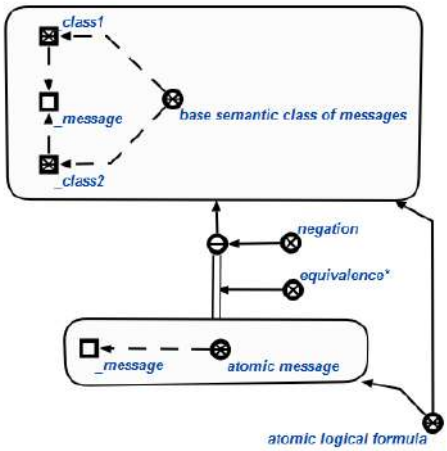


Figure 9. Definition of an atomic message.

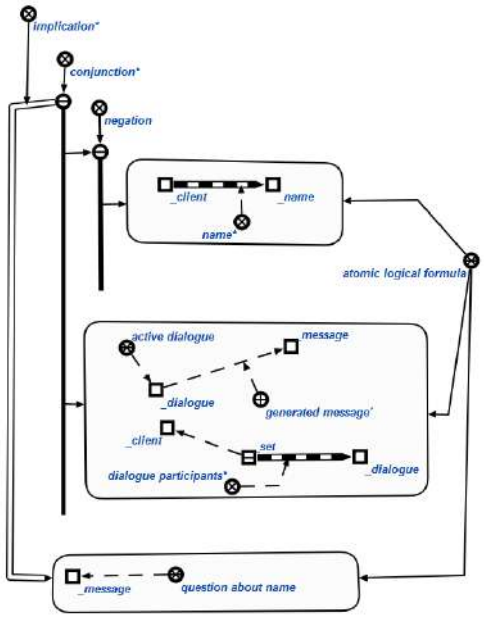


Figure 10. Rule of generating a question about a name.

speaker model that will get you around to synthesize speech with the voice of the selected target speaker.

The voice model of the speaker is based on a neural network, built on the principle of an automatic encoder. An automatic encoder is a multilayer neural network that converts multidimensional data into lower dimensional codes and then restores them in their original form. It was shown in [**?**] that data reduction systems based on neural networks have much broader capabilities, because, unlike the principal component analysis method, they permit nonlinear transformations to be performed.

The used artificial neural network configuration is shown in 12.

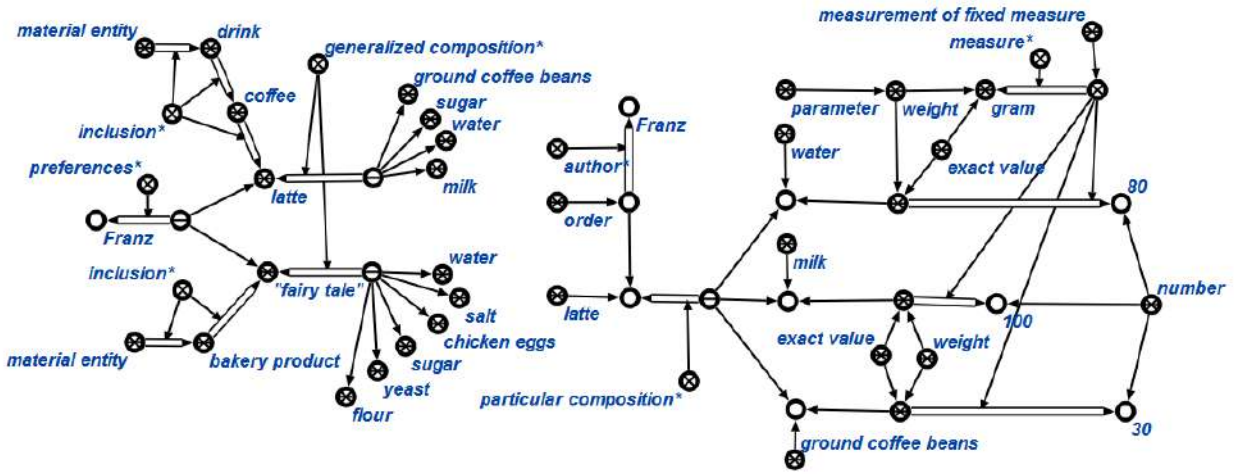


Figure 11. An example of a knowledge base fragment, depending on the subject area.

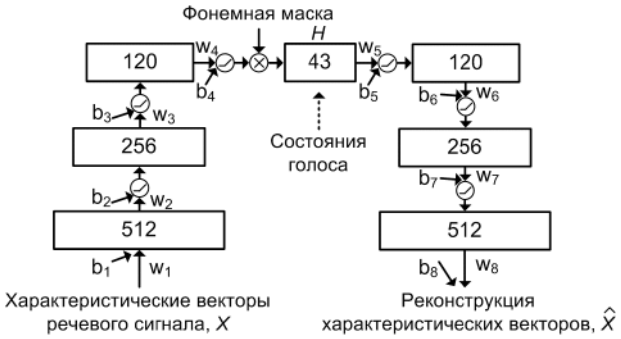


Figure 12. Speaker model based on auto-encoder

The auto-encoder performs next mapping function:

$$H = (w_4 RL(w_3 RL(w_2 RL(w_1 X + b_1) + b_2) + b_3) + b_4) \otimes M \quad (2)$$

where $X$ is characteristic vector of the speech signal, $H$ – vector of reduced dimension, $M$ – phonetic mask vector, $w_{1-4}$ and $b_{1-4}$ – weight coefficients and offsets of the corresponding network signals, $\otimes$ – stands for elementwise multiplication. The network uses a piecewise linear activation function $RL(x) = max(0, x)$, since it is shown that it provides a more effective internal representation of speech data compared to the logistic one and allows you to speed up the learning process [**?**]. At the output of the encoder, lower dimensional codes are generated, which are constrained in order to perform phonetic binding. The restriction is imposed by multiplying the signal $H$ by a phoneme mask, which is a sparse matrix, and formed on the basis of the phonetic marking of the speech corpus.

The decoder reconstructs the reduced-dimensional codes into the characteristic vectors $\hat{X}$. The corresponding display function is as follows:

$$\hat{X} = (w_8 RL(w_7 RL(w_6 RL(w_5 H + b_5) + b_6) + b_7) + b_8) \quad (3)$$

The next number of neurons in each hidden layer of the neural network was used: 512-256-120-43-120-256-512. Network training involves several steps:

- preliminary segmentation of the teaching speech corps into phonemes;
- initialization of network parameters and preliminary training;
- training of the encoder / decoder system;

As a result of training, a voice model is formed, which includes a model of each individual phoneme and the transitions between them contained in the training sample. For more details, the process of model formation is presented in [2].

## V. Example of working

The scenario of the system is proposed:

1) The user asks a question like "What is X?" (the typology of questions is still limited to one class of questions);
2) The speech signal analysis module selects a fragment corresponding to the name of entity X in the request and finds an entity with that name ( textit exactly how - see the question above) in the knowledge base;
3) The module for analyzing a speech signal in a formal language (in SC-code) forms a query to the knowledge base of the form "What is X?" for the found entity;
4) Ostis-system generates a response that is displayed to the user visually (in SCn, SCg). A subset of the answer (natural language definition or explanation for a given entity) goes to the speech synthesis module;
5) The speech synthesis module voices the received natural language text;

This dialog can be used as an example: - "Welcome. What is your name?"
–"Hello Andrey."
- "How is it going?"
- "Great!"
- "We can offer drinks and rolls. What do you prefer?"
- "I would like to order a latte."
- "Great choice, expect. Have a nice day."

Figures 13 – 19 show fragments of the knowledge base that sequentially reflect changes in it after processing each message in the dialog.

## VI. Conclusion

An approach to the development of intelligent speech assistants based on the integration of modern approaches to

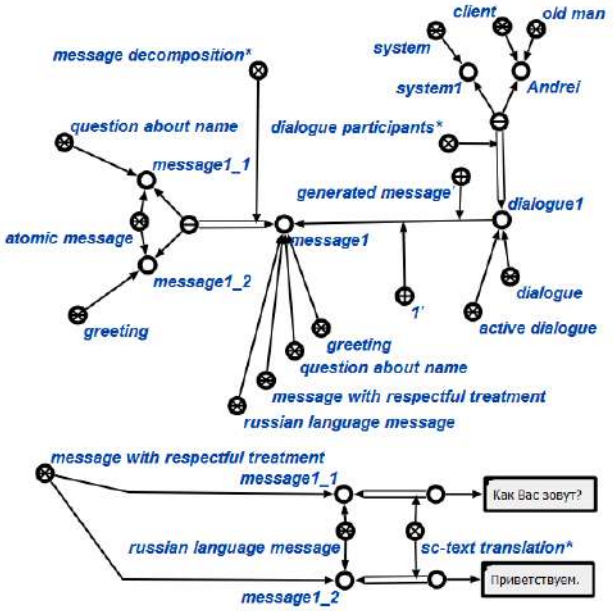


Figure 13. The portion of the knowledge base after the system generates a greeting and a question about the name.

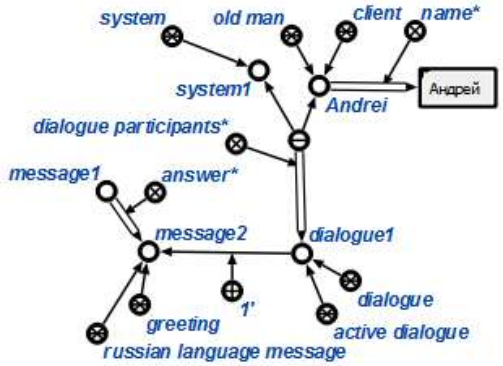


Figure 14. A part of the knowledge base after receiving and processing the user's answer to the question about the name.

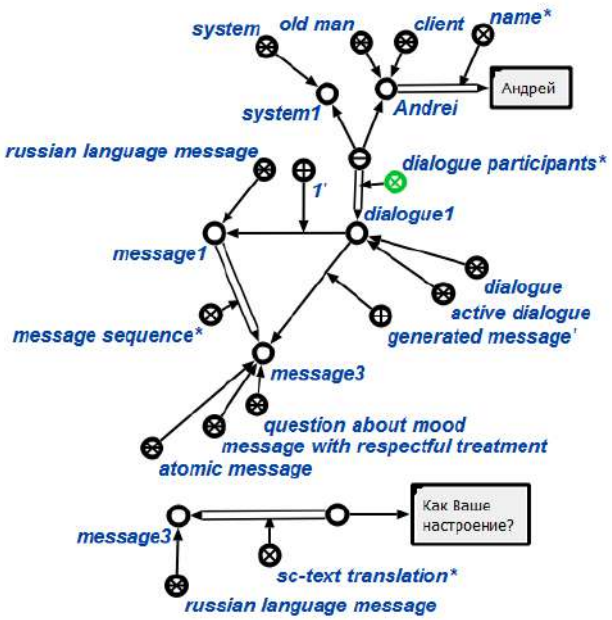


Figure 15. Section of the knowledge base after the system generates a question about the user's mood.

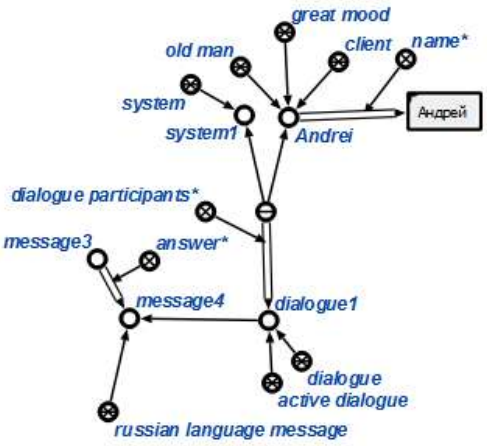


Figure 16. Section of the knowledge base after receiving and processing the user's response to a question about mood.

speech signals processing and semantic dialogue management models is proposed. Obtained results, in particular, developed set of ontologies, can be applied to the lay-out of speech assistants for various purposes wit the posibility to be adapted to the characteristics of a specific subject area.

## ACKNOWLEDGMENT

The authors would like to thank the scientific teams from the departments of intellectual information technologies, control systems and electronic computing facilities of the Belarusian State University of Informatics and Radio Electronics for their help and valuable comments.

## REFERENCES

[1] Amazon Echo & Alexa Stats [Electronic resourse]. Access mode: https://voicebot.ai/amazon-echo-alexa-stats Date of access: 08.01.2020.

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

[3] E. Azarov, A. Petrovsky, "Formirovanie personal'noj modeli golosa diktora s universal'nym foneticheskim prostranstvom priznakov na osnove iskusstvennoj nejronnoj seti [The formation of a personal model of the speaker's voice with a universal phonetic space of signs based on an artificial neural network]". Trudy SPIIRAN. 2014. Vol. 5. No. 36. P. 128-150. (in Russian)

[4] F. Bülthoff, M. Maleshkova "RESTful or RESTless – Current State of Today's Top Web APIs." arXiv:abs/1902.10514. 2019.

[5] R. Corona, J. Thomason, R. Mooney "Improving Black-box Speech Recognition using Semantic Parsing". Proceedings of

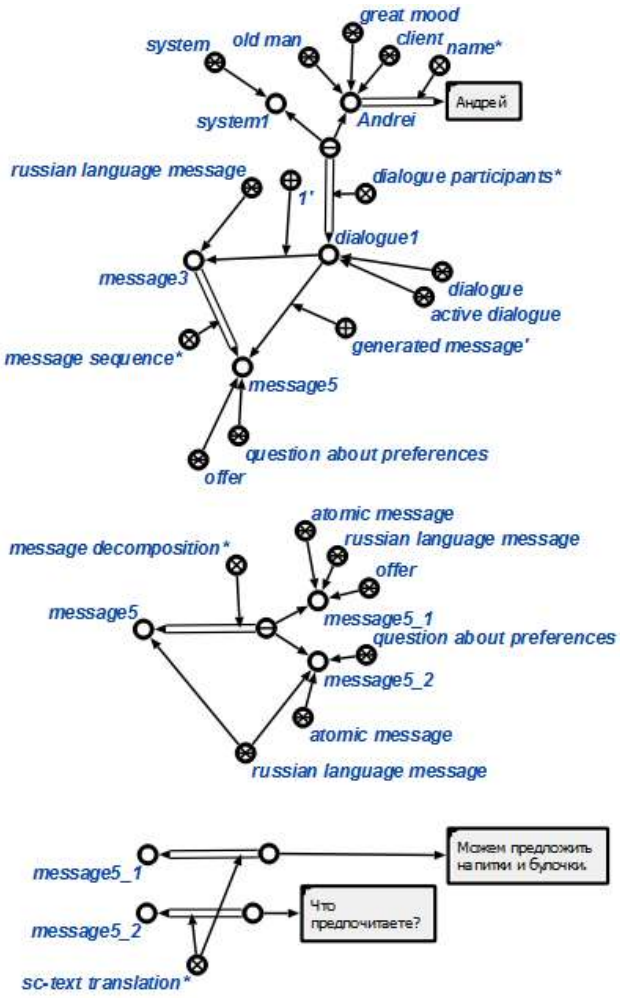


Figure 17. Knowledge base part after generating a system message containing a suggestion and a question about preferences.

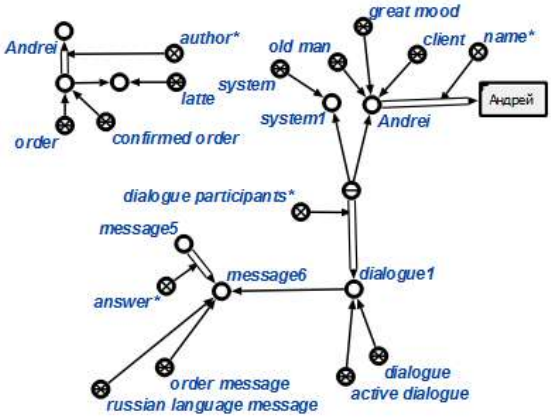


Figure 18. Knowledge base part after receiving and processing a user message containing order information.

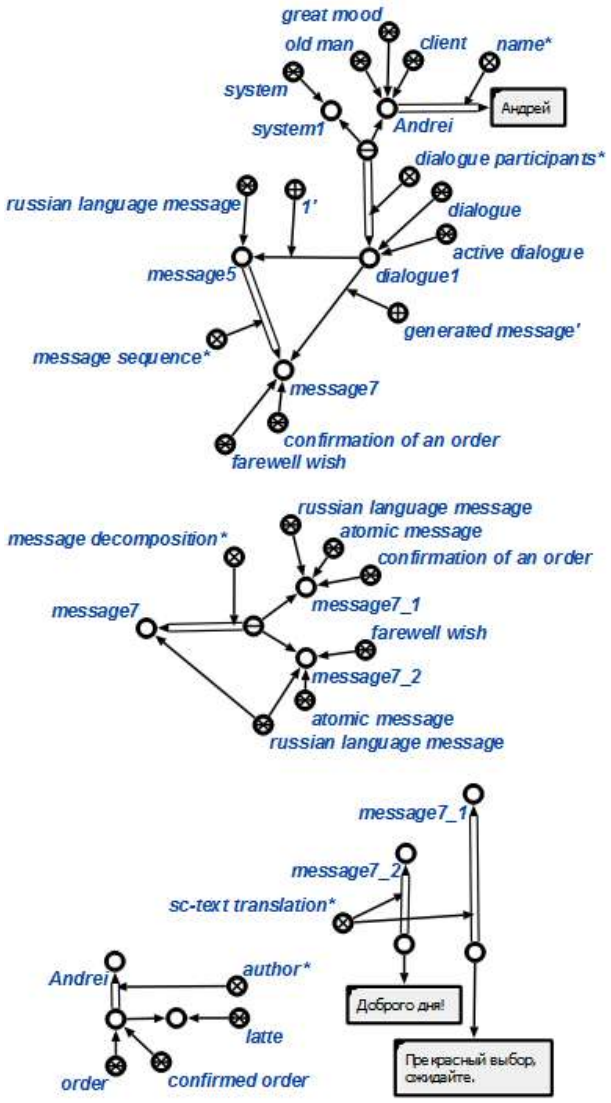


Figure 19. Knowledge base part after the system generates a message containing an order confirmation and a farewell message.

the Eighth International Joint Conference on Natural Language Processing. 2017. Vol. 2. P. 122–127.

[6] K. OЃlaherty, "Data Leak Warning Issued To Millions Of Google Chrome And Firefox Users" [Electronic resourse]. Access mode: https://www.forbes.com/sites/kateoflahertyuk/2019/07/19/data-leak-warning-issued-to-millions-of-google-chrome-and-firefox-users Date of access: 22.12.2019.

[7] Global Voice Assistant Market By Technology, By Application, By End User, By Region, Competition, Forecast & Opportunities, 2024 [Electronic resourse]. Access mode: https://www.businesswire.com/news/home/20190916005535/en/Global-Voice-Assistant-Market-Projected-Grow-1.2. Date of access: 20.12.2019.

[8] V. Golenkov, N. Guliakina, I. Davydenko, and A. Eremeev, “Methods and tools for ensuring compatibility of computer systems,” in Otkrytye semanticheskie tekhnologii proektirovaniya intellektual’nykh system [Open semantic technologies for intelligent systems], V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 25–52.

[9] I. Davydenko, "Semantic models, method and tools of knowledge bases coordinated development based on reusable components”, in Open semantic technologies for intelligent systems, V. Golenkov,

Ed. BSUIR, Minsk, 2018, pp. 99–118.
[10] M. B. Hoy "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants". Medical reference services quarterly. 2018. V. 37. No. 1. P. 81-88.
[11] M. MacTear, Z. Callejas, D. Griol, "The Conversational Interface: Talking to Smart Devices". Springer, 2016. 422p.
[12] A. Matyus, "Facebook faces another huge data leak affecting 267 million users" [Electronic resourse]. Access mode: https://www.digitaltrends.com/news/facebook-data-leak-267-million-users-affected Date of access: 22.12.2019.
[13] T. Polzehl, "On Speaker-Independent Personality Perception and Prediction from Speech", INTERSPEECH. 2012 proceedings of 13th Annual Conference of the International Speech Communication Association. Portland, 2012. P. 258–261.
[14] F. Rubin, "Amazon's Alexa and Google Assistant try making themselves the talk of CES 2020" [Electronic resourse]. Access mode: https://www.cnet.com/news/amazon-alexa-and-google-assistant-try-making-themselves-talk-of-ces-2020/. Date of access: 02.01.2018
[15] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components". IEEE Signal Processing Magazine, 2017, No. 1 (34). P. 67-81.
[16] D. Shunkevich, "Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems," in Open semantic technologies for intelligent systems, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.
[17] H. Tang, L. Lu, L. Kong, K. Gimpel, K. Livescu, C. Dyer, "End-to-End Neural Segmental Models for Speech Recognition", IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. No. 8. P. 1254–1264.
[18] V. A. Zahariev, E. S. Azarov, K. V. Rusetski, "An approach to speech ambiguities eliminating using semantically-acoustical analysis", Open Semantic Technologies for Intelligent Systems (OSTIS-2018). Minsk: BSUIR, 2018. P. 211 – 222.
[19] V. A. Zahariev, T. Lyahor, N. Hubarevich, E. S. Azarov, "Semantic analysis of voice messages based on a formalized context", Open Semantic Technologies for Intelligent Systems (OSTIS-2019). Minsk: BSUIR, 2019. P. 103 - 112.

# Принципы построения интеллектуальных речевых ассистентов на основе открытой семантической технологии

Захарьев В.А., Шункевич Д.В., Никифоров С.А., Ляхор Т.В., Азаров И.С.

Целью данной работы является разработка принципов построения речевых ассистентов для интеллектуальных систем на базе Технологии ОСТИС.

Приведено описание существующих систем, их типовой архитектуры и принципов работы. В соответсвии со сравнительной характеристикой систем данного типа, а также результатами анализа актуальных потребностей пользователей, сформулирован ряд требований, которые показывают потенциальную эффективность применения Технологии ОСТИС для разработки диалоговых систем и голосовых асистентов в часности. Ключевыми условиями в данном случае являются обеспечение, с одной стороны, максимальной независимости от предметной области, с другой стороны – обеспечение возможности адаптации (в первую очередь – автоматической) универсального ассистента под конкретного пользователя и особенности конкретной предметной области.

В работе приведен пример реализации интеллектуального речевого ассистента, основанный на интеграции современных подходов к обработке речевого сигнала и семантических моделей управления диалогом. Приведены особенности реализации этапов анализа и синтеза речевого сигнала для удовлетворения реализации вышеобозначенных требований. Полученные результаты, в частности, разработанный набор онтологий, могут быть применены для разработки речевых ассистентов различного назначения, в том числе адаптированы с учетом особенностей конкретной предметной области.

Received 12.02.2020