Audio interface of next-generation intelligent computer systems

Vadim Zahariev, Kuanysh Zhaksylyk, Denis Likhachov, Nick Petrovsky, Maxim Vashkevich, Elias Azarov
Belarusian State University of
Informatics and Radioelectronics
Minsk, Belarus
{zahariev, likhachov, nick.petrovsky, vashkevich, azarov}@bsuir.by, kuanysh.zhk@gmail.com

***Abstract*—The article is dedicated to the issues of creating audio and voice interfaces for next-generation intelligent computer systems. It is proposed to use an approach based on ontological design and formalization of a concepts system from the subject domain of audio interfaces using the OSTIS Technology. The main ideas underlying this approach, as well as their features distinguishing them from the generally accepted ones, are outlined. It is shown that in the future, the usage of this approach can provide the properties of unification, semantic compatibility, and interoperability in the development of audio and voice user interfaces, which ultimately will significantly reduce costs when creating next-generation intelligent computer systems for solving complex problems.**

***Keywords*—audio interface and voice interface of intelligent computer systems; semantically compatible components; speech processing and digital signal processing**

I. Introduction

Spoken language is one of the most natural and effective forms of information exchange between humans. This fact explains the significant interest of researchers in the development and application of voice interfaces for human-machine interaction as part of modern communication, multimedia, and intelligent systems [1], [2].

A more comprehensive form of interaction with the user and the environment through the analysis and synthesis of acoustic signals is an audio interface. This type of interface, which acts as a maternal in relation to voice ones, can be briefly defined as a hardware-software complex that analyzes and synthesizes signals in the entire available spectrum of parameters of acoustic information carriers, for example, to solve the problems of analyzing the situation and events occurring in the acoustic environment of the system, synthesizing non-speech signals (technogenic and natural sounds, warning signals, music, etc.) [3].

The following main tendencies in the development of this direction indicate the relevance of the direction of developing audio and voice interfaces:

- economic indicators and forecasts for the development of the speech technologies market, the current average annual growth rate of which, according to experts, is about 22%, and the total volume will be equal to 59.6 billion US dollars by 2030 [4];
- the appearance of a wide range of products based on the voice interface, which have gained widespread. First of all, these are personal voice assistants, such as Alexa (Amazon), Siri (Apple), Cortana (Microsoft), Alice (Yandex) [5]–[7];
- interest from the scientific community, expressed in the growth of publications in this field of research by 15% over the past 5 years [8].

It should be noted that the basic mass of scientific publications in this direction is dedicated to the development of basic technologies that are components of the voice interface, such as text-to-speech synthesis, as well as speech-to-text transformation [9]–[11]. Recent achievements in these fields are associated with the rapid development of neural network models and computing tools. They made it possible to bring the qualitative characteristics of the usage of speech technologies to a commercial level [12], [13].

II. Problem statement

Most of the existing systems, as a rule, are designed to solve a certain range of problems and are hardly compatible with each other. This fact is especially acute when designing complex systems like intelligent personal dialog assistants (Figure 1), which require using a variety of different types of processed information and different problem-solving models. Such systems, in addition to standard modules for recognition (ASR, automatic speech recognition) and synthesis (TTS, text-to-speech), at the audio interface level, should also contain models that determine the presence/absence of speech in the audio signal in a complex acoustic environment, classify environmental sounds, recognize a speaker, etc. In addition, elements of the voice interface must be compatible with higher-level modules for processing natural language information, such as modules for speech understanding (SLU, spoken language understanding) and generation (SLG, spoken language generation), dialog control (DM, dialog manager) [14].

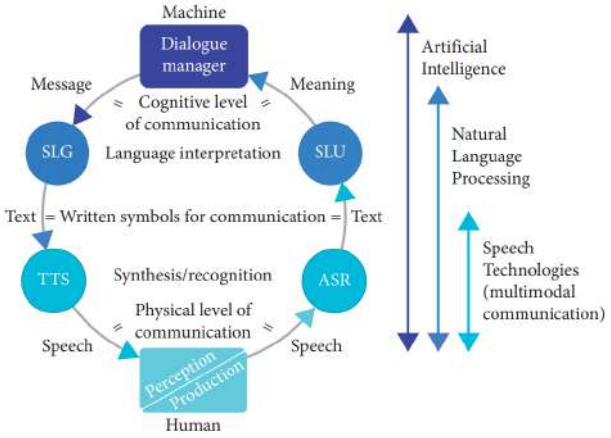

Figure 1. Components of a human-machine speech dialog system [14].

All this requires the development of approaches based not only on machine learning methods and signal processing but also on natural language processing, symbolic methods of artificial intelligence, ontological design, and formalization of the subject domain of the audio interface. This will allow the creation of systems that have a full range of knowledge in a formalized form about the types of problems, that they must solve, and the methods available for solving them.

A necessary condition for the creation of such next-generation systems with improved characteristics in terms of interoperability and flexibility is also the fact that these systems must be built on the basis of a basic technology that allows such a unity of the form of information representation at all its levels.

The combination of these factors leads to the need to create next-generation intelligent computer systems that will include audio and voice interface modules based on the principles of interoperability and semantic compatibility to solve complex problems.

III. Suggested approach

To achieve this purpose, it would be advisable to use an approach based on the principles underlying the "Standard of the Open Technology for the Ontological Design, Production, and Operation of Semantically Compatible Hybrid Intelligent Computer Systems", or briefly the "Standard of the OSTIS Technology" [15].

The essence of the approach is to consider the process of designing an audio interface as an interface subsystem within the general process of developing an intelligent computer system (ICS) and building its formal logical-semantic model.

To create such a model of next-generation intelligent computer systems, it is necessary:

- to decompose an information computer system into components. The quality of the decomposition is determined by the simplicity of the subsequent synthesis of the general formal model from the formal models of the selected components;
- to carry out the convergence of selected components in order to build compatible (easily integrated) formal models of these components;
- to perform the integration of the built formal models of the selected components and obtain a common formal model.

The general methodological principles that are the basis for the transition to next-generation ICS are:

- convergence and unification of ICS and their components;
- structural-system simplification of ICS ("Occam's Razor" principle);
- orientation to universal ICS;
- synthesis of ICS from compatible components;
- orientation towards the creation of synergetic ICS.

The following important features of the proposed approach follow from the general methodological principles, which must be taken into account in order to achieve the purpose:

- semantic knowledge representation;
- agent-oriented basic model for processing knowledge bases that have a semantic representation (insertional programming in a semantic space);
- semantic structuring of knowledge bases in the form of a hierarchical system of subject domains and corresponding ontologies that specify these subject domains;
- at the same time, a next-generation ICS interface is interpreted as a specialized intelligent information system focused on solving interface problems of the corresponding individual ICS and deeply integrated (embedded) into this ICS.

As a technological basis for the implementation of the proposed approach, the OSTIS Technology [16] will be used. Systems built on the basis of the OSTIS Technology are called ostis-systems, respectively, the audio interface subsystem will be built as a reusable component, which in the future will be built into various ostis-systems, if necessary. As a formal basis for encoding various information in the knowledge base, an SC-code [16] 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 conditioned by its following main advantages:

- within this technology, unified means of representing various types of knowledge, including meta-knowledge, are proposed, which make it possible to describe all the information necessary for analysis in one knowledge base in a unified format [17];
- the formalism used within 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 technology for representing various types of knowledge [17] and their processing models [18] ensures the modifiability of ostis-systems, i.e. allows easily expanding the functionality of the system by introducing new types of knowledge (new concepts systems) and new models of knowledge processing.

In this work, unlike the previous ones, which deal with the issues of semantic analysis of voice messages based on a formalized context [19] and the creation of dialog assistants based on a mental lexicon model [20], [21] or a multimodal system based on a neurosymbolic approach [22], OSTIS technology is used to directly build an ontology of the audio interface subsystem.

Since the next-generation ICS audio interface must have an architecture that corresponds to the general rules for building ostis-systems, the following main parts of it can be distinguished and formalized:

Audio interface of next-generation intelligent computer systems

⇒ *reduction**:
[ICS audio interface]
⇒ *generalized decomposition**:
{
- *knowledge base of the subsystem of the next-generation ICS audio interface*
- *problem solver of the subsystem of the next-generation ICS audio interface*
- *interface for interacting with other interface subsystems of the ostis-system*

}

Thus, it should be noted that the process of developing an audio interface for next-generation ICS implies, first of all, the creation of semantically structured knowledge bases in the form of a hierarchical system of subject domains and corresponding ontologies that specify these subject domains. Therefore, the first step to achieve this purpose is the phase of identifying and formalizing the entities of the audio and voice interface in order to immerse this information in the knowledge base of an intelligent computer system.

From our point of view, it is possible to decompose the subject domains and ontologies included in the knowledge base of the audio interface into the following main directions:

Subject domain and ontology of the audio interface of next-generation intelligent computer systems

⇐ *decomposition**:
{
- *Subject domain and ontology of audio interface problems*
- *Subject domain and ontology of signal parametric representation models*
- *Subject domain and ontology of signal parameter classification models*

}

As it is shown, the functional approach to the decomposition of subject domains is put at the head of the ontology, which is quite natural because it corresponds to the nature of the problems implemented by the audio interface.

The principles, represented above, together allow for the convergence and integration of components both at the level of the audio interface subsystem and at the level of the entire ICS as a whole, which, in turn, allows "transfering" an intelligent information system into a class of hybrid, interoperable, and synergistic systems.

Next, we proceed directly to the consideration of specific subject domains and the building of an ontology of the audio interface.

IV. Subject domain and ontology of audio interface problems

The first step towards building the knowledge base of the subsystem of the next-generation ICS audio interface is the formalization of the top-level ontology. This ontology is proposed to be based on a formalized representation of the main entities of the subject domain and their properties, as well as functional problems that the audio and voice interface are designed to solve.

The main entities, which require formalization and immersion into the knowledge base, include the set of concepts represented below. One of the key concepts requiring formalization is the basic definition of the signal itself, as well as the main types of signals, depending on their nature, which are of the greatest interest in the field of audio interfaces. To make the description process by means of the OSTIS Technology clearer, before proceeding directly to it, we will give examples of the main entities and concepts that require formalization and immersion into the knowledge base:

- signal;
- acoustic signal;
- audio signal;
- speech signal.

Depending on the way of mathematical description of the processed signal in the ostis-system, the following classes can be allocated:

- analog signal;
- discrete signal;
- digital signal;
- periodic signal;
- aperiodic signal;
- harmonic signal;

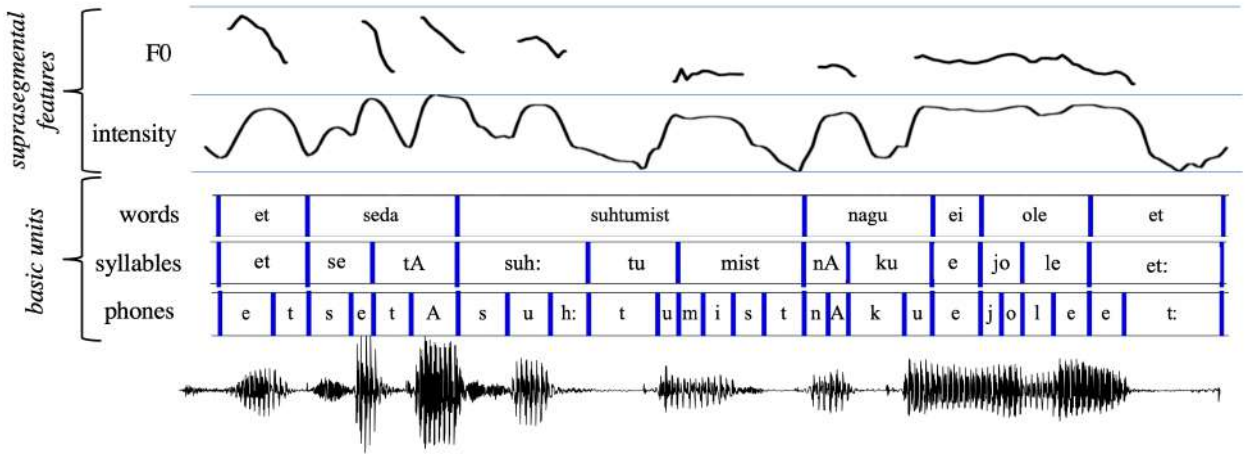

Figure 2. Segmental and suprasegmental features of the speech signal [23]

- tone signal;
- noise signal;
- pulse signal.

To successfully immerse the necessary knowledge for the operation of the audio interface, it is also necessary to formalize the basic concepts associated with the features of the signal itself, according to the following main attributes:

- signal amplitude;
- signal frequency;
- signal phase;
- signal intensity;
- signal duration;
- signal power/energy;
- signal oscillogram;
- signal spectrogram;
- signal discretization interval;
- signal quantization degree.

The key concepts of the subject domain lying in the semantic neighborhood of the subspace of the functional purpose for audio interfaces and audio signal processing are the following ones:

- audio signal analysis;
- audio signal synthesis;
- audio signal encoding;
- audio signal denoising;
- audio signal classification;
- environmental sound classification;
- acoustic scenes and events classification;
- anomalous sound detection;
- sound source localization.

The basic concepts of audio and voice interface are also closely related to its main features, which can be divided into the following main groups of concepts:

- speech signal features;
- linguistic features of the speech signal;
- paralinguistic features of the speech signal;
- extralinguistic features of the speech signal;
- segmental features of the speech signal;
- suprasegmental features of the speech signal;
- speech signal volume;
- speech signal timbre;
- speech signal rate;
- frequency of the main signal tone;
- phonemic composition of the speech signal.

The main concepts of the subject domain lying in the semantic neighborhood of the functional purpose of the speech and audio signal processing are the following ones:

- speech signal analysis;
- speech signal synthesis;
- speech recognition;
- emotional speech recognition;
- text-to-speech synthesis;
- emotional text-to-speech synthesis;
- sing synthesis;
- voice activity detection;
- key words spotting;
- wake up word detection;
- speech diarization;
- speaker recognition;
- speaker classification;
- speaker verification.

It should be noted that the above concepts are often interconnected in a complex and non-trivial way in the process of transition from information sources to direct physical parameters. Such a complex signal structure can be represented as a diagram of its information structure (Figure 3). This fact requires next-generation ICS to formalize concepts, so that the system can automatically interpret the interconnections between these features, when working with audio and speech signals, and, as a result, supply a response to the user, explaining on the

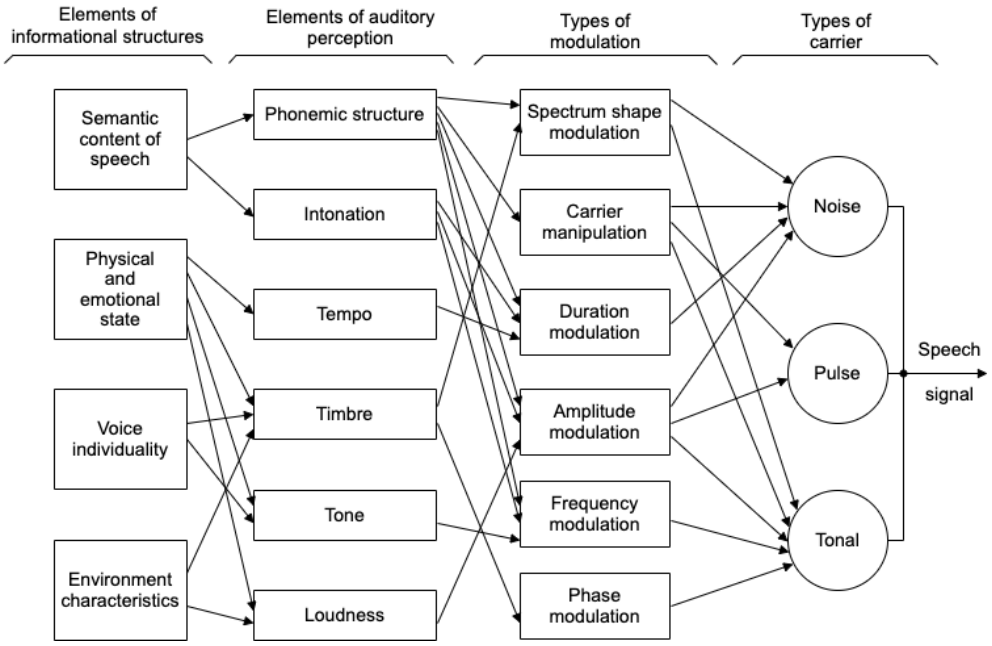

Figure 3. Speech signal features and their interconnections [24]

basis of what features the system came to a particular conclusion.

Since for the building of next-generation ICS, the focus is precisely on the problems associated with the processing of speech signals, the solution of which is necessary first of all to build a voice interface, we will try to focus on the formalization features of this subject domain.

The basis of the SC-model of the knowledge base is a hierarchical system of subject domains and their corresponding ontologies. The top level of the hierarchy for the part of the knowledge base related directly to the audio and voice interfaces is shown below.

Here is a formalized representation of some of the above concepts:

signal
⇒ *definition**:
[physical process that carries a message (information) about some event, the state of the considered object, or issues the control, alerts commands, etc.]
⊃ *acoustic signal*
⇒ *definition**:
[signal representing the propagation of elastic waves in a gaseous, liquid, or solid medium]
⊃ *audio signal*
:= [sound signal]
⇒ *definition**:
[acoustic signal whose parameters are within the range of values accessible to human senses]
⊃ *acoustic signal*
⇒ *note**:
[The frequency range of the audio signal is between 20 and 20,000 Hz.]
⊃ *speech signal*
⇒ *definition**:
[audio signal generated by the passage of air flows through the human vocal tract. As a result of various acoustic transformations, the formation of various speech sounds occurs]
∋ *verbal speech*
∋ *speech production*
⇒ *note**:
[The mechanism of human speech production is an acoustic tube with dynamically changing cross-sectional parameters, excited either by a quasi-periodic sequence of impulses generated by the vocal cords or by a turbulent flow of air pushed through constrictions in different parts of the vocal tract.]

Formalization at the level of a top-level ontology will be represented only for the basic concepts of the subject domain of signal models. The varieties of models of mathematical representation will be discussed in the corresponding subsection below.

Depending on the model for representing the signal itself in the ostis-system, the following descriptions of

the main types of signals can also be defined, the usage of which is justified by the nature of the analyzed signal, as well as the analysis of the problem to be solved:

signal model
⇐ *combination**:
{• *analog signal*
⇒ *definition**:
[signal whose parameters can be measured at any time]
⇒ *definition**:
[signal where each of the represented parameters is described by a function of time and a continuous set of possible values]
• *discrete signal*
⇒ *definition**:
[signal for which at least one of the represented parameters is described by a finite set of possible values]
⇐ *combination**:
{• *discrete in time*
• *discrete in amplitude*
}
• *digital signal*
⇒ *definition**:
[signal where each of the representing parameters is described by a discrete time function and a finite set of possible ones]
⇒ *subdividing**:
{• *signal discrete in time*
• *signal quantized (discrete) in amplitude*
}
• *periodic signal*
• *aperiodic signal*
• *tone signal*
• *harmonic signal*
• *pulse signal*
• *noise signal*
}

It should be noted that due to restrictions on the size of the material for the features of the audio signal, we will give only a hierarchy of their general interrelations, since the semantics of these concepts is quite typical for other fields of technical sciences and does not require detailed examples and explanations.

audio signal features
⇐ *combination**:
{• *signal amplitude*
• *signal frequency*
• *signal phase*
• *signal intensity*
• *signal duration*
• *signal power*
• *signal spectrum*
• *signal oscillogram*
⇒ *definition**:
[function that fixes the dependence of changes in signal features (first of all, amplitude) in time]
• *signal spectrogram*
⇒ *definition**:
[function that fixes the dependence of the power spectral density of an audio signal in time]
• *signal discretization interval*
⇒ *definition**:
[value of the frequency with which the signal was discretized over time during the analog-digital conversion]
⇐ *typical values**:
{• *8000 Hz*
• *16000 Hz*
• *22050 Hz*
• *44100 Hz*
• *48000 Hz*
}
• *signal quantization degree*
⇒ *definition**:
[permissible number of discrete signal levels expressed as a degree of two and used in the process of quantization of the signal by level in the process of analog-digital conversion]
⇐ *typical values**:
{• *8 bits*
• *10 bits*
• *12 bits*
• *16 bits*
• *24 bits*
}
}

The description of the subject domain of signal models will be considered in more detail in the next subsection of the article, so we do not consider it necessary to demonstrate it here. We formalize the ontology of the main features of a speech signal in the following form [25]:

speech signal features
⇐ *combination**:
{• *communicative features*

⇒ *note**:
[encode the meaning of the transmitted message and depend on the sender's intentions]

- *informative features*
  ⇒ *note**:
  [encode additional information of the transmitted message and do not depend on the sender's intentions]
- *informative features*
  ⇒ *note**:
  [encode additional information of the transmitted message and do not depend on the sender's intentions]
- *segmental features of the speech signal*
  ⇒ *note**:
  [carry information about the current state of the source for the duration of one or more phonetic units]
- *suprasegmental features of the speech signal)*
  ⇒ *note**:
  [carry information about the state of the source and the transitions between them throughout the entire utterance]

}

⊃ *linguistic features of the speech signal*
∋ *verbal communication means*
∋ *communicative features*
⇒ *definition**:
[features carrying information by means of usage of the human language coding system]
⇒ *note**:
[linguistic features include both phonological (segmental and supersegmental) and grammatical code (morphology and syntax). Linguistic communication informs the receiver of the sender's intentions through explicit verbal forms]

⊃ *paralinguistic features of the speech signal*
∋ *non-verbal communication means*
∋ *communicative features*
⇒ *definition**:
[features carrying information through additional means of communication, not directly related to the language]
⇒ *note**:
[transmit information about the attitude towards the subject of conversation, feelings, or emotional state of the speaker]
⇐ *combination**:
{• *speech intonation*
⇐ *combination**:
{• *frequency of the main signal tone* F_0
• *changing the frequency of the main signal tone* ΔF_0
}
• *speech volume*
⇐ *combination**:
{• *signal amplitude*
• *signal intensity*
}
• *speech tempo*
• *pause duration*
}

⊃ *extralinguistic features of the speech signal*
∋ *informative features*
⇒ *definition**:
[features that do not directly encode the meaning of the message but contain additional information about the sender and the conditions of communication]
⇒ *note**:
[transmit information about the attitude towards the subject of conversation, feelings, or emotional state of the speaker]
⇐ *combination**:
{• *narrator voice features*
⇐ *combination**:
{• *pitch*
• *timbre*
• *volume*
}
• *acoustic environment*
}

In Figure 4, the result of formalization of the subject domain and the ontology of typical problems of audio and voice interfaces for next-generation ICS by means of the SCg language is represented.

It should be noted that in this article we will give examples and focus only on the formalization of some of the concepts outlined above. The second important note is that the represented set of concepts is by no means exhaustive, since the main results of the work on the formalization of this subject domain will be discussed within more voluminous works such as a monography and the next versions of the OSTIS standard.

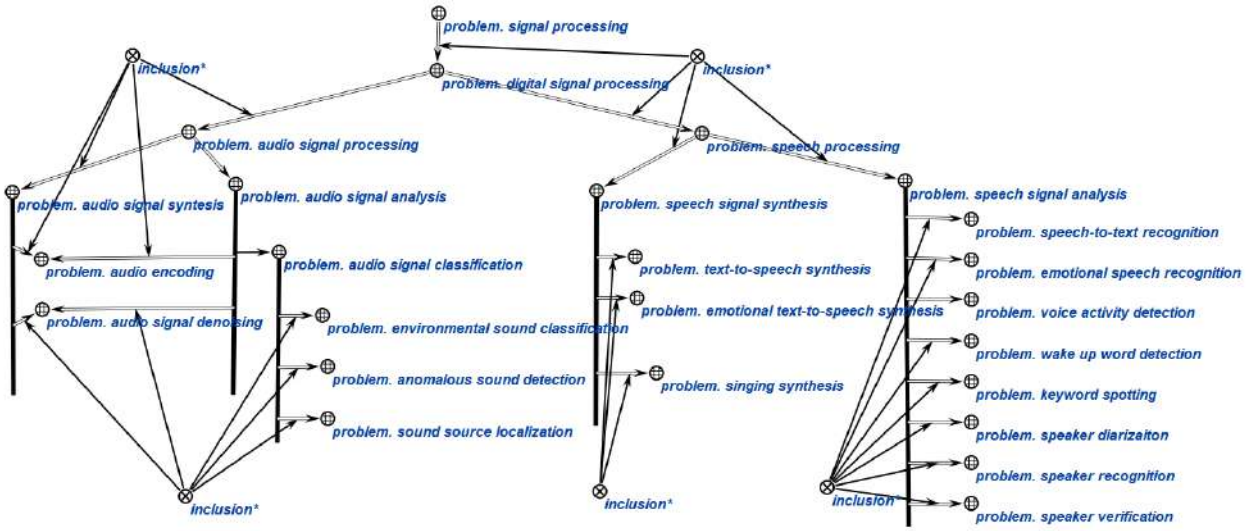

Figure 4. The top-level ontology fragment of problems of audio and voice interfaces for next-generation intelligent computer systems

V. Subject domain and ontology of signal parametric representation models

All of the above problems are interrelated, since they refer to the same object of research – the speech signal. The solution of each of them directly or indirectly depends on the effectiveness of speech modeling as a complex phenomenon in various aspects: parametric representation of the speech signal and allocation of its properties, modeling the process of phonation, perception, and interpretation of the contents of a speech message (including phonetic, semantic, emotional ones). This makes the creation of universal methods for processing speech signals a promising scientific direction. In the context of the above problems, speech modeling can be conditionally divided into three levels:

- general signal modeling using samples in the time or frequency domain;
- modeling of signal features that are specific to speech and related to the phonation process (such as frequency of the main tone, excitation sequence, and amplitude spectrum envelope);
- modeling of high-level speech features (voice, accent, expression, phonetic and semantic contents of a speech message). Each next level is based on the previous one and implies the usage of special methods of parametric description.

The first two levels include models widely known in digital processing of speech signals based on linear prediction (LP), cepstral coefficients, and sinusoidal parameters.

Among the approaches using the sinusoidal description of the signal, currently, the most promising are mixed (hybrid) models, which take into account the possibility of different modes of phonation with the participation of the vocal cords (voiced speech) and without the participation of the vocal cords (unvoiced speech), moreover, each of these two modes is described by the corresponding model (Figure 5).

Voiced speech is considered as a quasi-periodic (deterministic) signal, while unvoiced speech is considered as a non-periodic (stochastic) signal. The most famous among existing models is the harmonic + noise model, which is used to solve such complex problems as the creation of voice interfaces, speech recognition, text-to-speech synthesis, voice conversion, noise reduction, increasing the intelligibility and subjective quality of speech signals, accent correction, and so on. Its advantage is the theoretical possibility of modeling vocalized sounds in the form of continuous functions with varying parameters, which makes it possible to obtain an effective description of the phonation process and avoid the overlap of adjacent fragments, phase breaks in speech synthesis. The disadvantage of the model is the high complexity of the analysis and synthesis algorithms due to the non-stationarity of the speech signal [26]–[29].

Since voiced speech consists of quasi-periodic components with varying parameters, it is necessary to use digital filters with variable features for analysis: their bandwidth must change in accordance with the contour of the frequency of the main tone. This requires the usage of special time-frequency transformations that allow estimation of periodic components with strong frequency modulation such as Fan-Chirp and harmonic transformations. The accuracy of parameter estimation is directly related to the accuracy of estimating the contour of the frequency of the main tone, so the usage of a reliable and accurate estimation method is a necessary condition for the successful usage of this model [30]–[32].

Another difficult problem is the automatic separation of

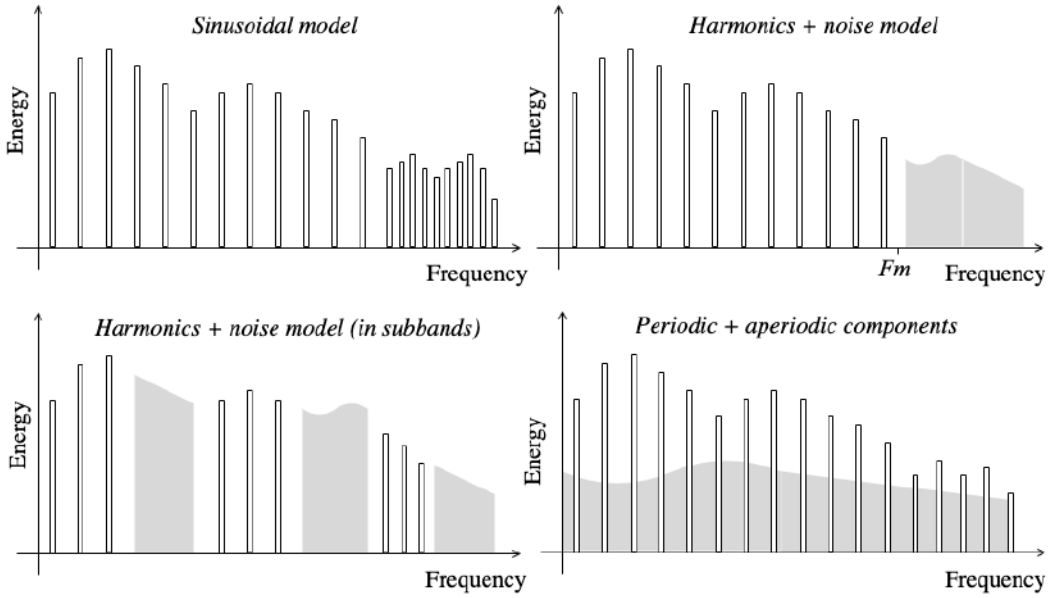

Figure 5. Classification of common speech signal models [26]–[32]

the signal into deterministic and stochastic components, for which special periodicity detectors are used.

Modeling a speech signal based on LP is a classic approach that has been used in digital speech processing for quite a long time. The main advantage of the model is the separate description of the signal in the form of the spectrum envelope and the excitation signal. The spectrum envelope determines the phonetics of the pronounced sound and characterizes the state of the vocal tract, while the excitation signal characterizes the state of the vocal cords and the pitch (intonation) of vocalized sounds. The advantage of LP is also low computational complexity.

However, despite this, recently, preference has been given to models using a sinusoidal representation of the signal, and this primarily concerns applications that involve the synthesis of a speech signal with modified parameters, such as intonation change, voice conversion, text-to-speech synthesis, and others. This fact can be explained by the point that the LP does not provide efficient methods for parametric processing of the excitation signal and continuous synthesis of the output signal. Each speech fragment (frame) of the signal is a separate independent unit, and during synthesis, there is a problem of matching adjacent frames. An inconsistent change in the envelope of the amplitude and phase spectrum during the transition from frame to frame causes the appearance of audible artifacts. In addition, the estimation of the spectrum envelope using classical LP methods is an averaging over the entire frame, as a result of which its accuracy is limited. The order of the predictor determines the complexity of the model: for low-order predictors, the spectrum envelope estimate is overly smoothed, while for high-order predictors, the accuracy becomes selective. For points of the spectrum corresponding to the harmonics of the fundamental tone, the accuracy increases, and for all other points it decreases. The optimal order of the predictor depends on the pitch of the voice, but even in the most favorable case, the accuracy of the spectrum envelope estimate has an error leading to audible distortion.

The usage of cepstral coefficients for modeling speech signals is also a classical approach. The most well developed speech modeling system using cepstral coefficients is TANDEM-STRAIGHT [33], [34]. Just as for the classical methods of analysis based on LP, when estimating the cepstral coefficients, it is assumed that the signal is stationary over the observation interval. Estimation of the envelope of the amplitude spectrum requires smoothing and is also not accurate enough compared to models based on sinusoidal parameters (Figure 6).

Due to its wide capabilities, the hybrid model based on sinusoidal parameters is the most preferred for usage in most practical cases. Nevertheless, to overcome its existing limitations associated with the complexity of estimating parameters, their interpretation in the form of specific speech features (vocal tract parameters, excitation sequence), the development of special modeling methods is required.

Depending on the application, processing of a speech signal using a particular model usually includes analysis (determining the model parameters), modification (changing the model parameters depending on the purpose of the application), and synthesis (forming a new signal from the changed model parameters). Thus, to ensure the highest practical significance, the developed modeling methods should include tools for analysis, processing of

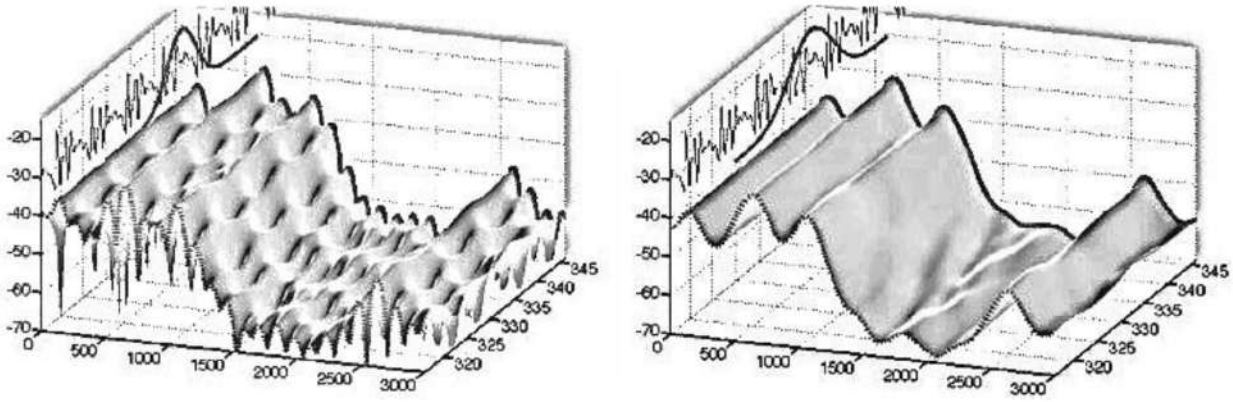

Figure 6. A conventional spectrogram (left) vs TANDEM-STRAIGHT spectrogram (right)

parameters, and synthesis.

Solving many modern applied problems requires not only the ability to describe a speech signal or the phonation process but also the usage of high-level speech features that determine the speaker's personal voice, expression, phonetics, etc. Such problems include voice conversion, text-to-speech synthesis, speaker verification, and many others. High-level speech modeling is a very complex subject domain, since it requires the usage of intelligent models and machine learning methods. At the moment, there is no single universal method used for different applications.

The vast majority of high-level speech models used in practice are problem-oriented and can only be used to solve one, highly specialized problem. The main mathematical tools used are statistical and probabilistic models.

parametric signal model

⇒ *definition**:
[mathematical expression used to represent signal samples in the time or frequency domain]

⊃ *parametric model of the speech signal*

⇒ *definition**:
[mathematical description of signal features that are specific to speech and associated with the phonation process (such as frequency of the main tone, excitation sequence, and amplitude spectrum envelope)]

⇒ *note**:
[The main speech signal models include: models based on linear prediction; based on the cepstral representation; sinusoidal and hybrid models. Among the hybrid models, the harmonic + noise model is the most well known.]

VI. Conclusion

In the article, the ideas underlying the original approach to designing audio interfaces of ICS based on ontological design and formalization of a concepts system from the relevant subject domain, using the OSTIS Technology, is represented. The main principles underlying this approach, as well as their distinctive features from the generally accepted ones, are outlined.

The following main factors can be attributed to the limitations of the proposed approach: it is obvious that in order to achieve the purpose and implement the problems of formalizing any subject domain, including audio interfaces, first of all, a large number of sources of knowledge are required to replenish them. To overcome this problem, it is necessary to involve a large number of experts with appropriate competencies and knowledge in the subject or to develop mechanisms for reliable automatic extraction of this knowledge from available sources.

Direct access to the knowledge of experts is very limited, since it requires significant efforts to select a representative sample of such experts, build effective and interoperable relationships between the parties of the process, which often depends on a large number of subjective factors, and, accordingly, requires a large amount of time and material resources.

It is known that a significant amount of information accumulated by mankind is stored in the form of natural language texts. The process of extracting this information and its representing in a formalized form – in the form of knowledge – also looks non-trivial.

Based on the nature of these problems, according to the authors, the following main directions for overcoming them are seen and, as a result, two main strategies for developing the proposed approach are:

1) Creation of specialized tools for experts working in the domain of audio and voice interfaces for formalizing and representing knowledge from a given subject domain, fixing them in the form of standards of a single form. Such tools should have qualitatively new functionality providing a high level of compatibility and interoperability in the process of accumulation and standardization of knowledge, so that the experts themselves would be interested in the application and wide distribution of this technology for knowledge representation. This item is one of the key objectives of the technology and the OSTIS standard.
2) Creation of automated and automatic means of extracting knowledge from existing sources of information, primarily natural language texts. The types of documents that contain already structured and partly formalized information are primarily standards, protocols, request for comments (RFC), instructions, etc. Therefore, the process of automating knowledge extraction should be aimed primarily at formalizing the existing industry standards for the development of audio interfaces, systems for processing and encoding audio information, speech signal processing systems, such as the standards of the International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), Institute of Electrical and Electronics Engineers (IEEE), and (Audio Engineering Society) AES series [35]–[39].

The implementation of the approach proposed in the work will ensure the properties of unification, semantic compatibility, and interoperability in the development of audio and voice interfaces (a kind of analogue of the OSI/ISO model in the field of designing ICS interfaces), which ultimately will significantly reduce costs when creating next-generation intelligent computer systems for solving complex problems.

REFERENCES

[1] C. Pearl, *Designing voice user interfaces: principles of conversational experiences*. O'Reilly Media, Inc., 2016.

[2] N. Chen, C. You, and Y. Zou, "Self-supervised dialogue learning for spoken conversational question answering," in *Proc. Interspeech 2021*, 2021, pp. 231–235.

[3] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.

[4] E. Fernandes. (2022) Speech and voice recognition market / verified market research. [Online]. Available: https://www.verifiedmarketresearch.com/download-sample/?rid=4077

[5] J. R. Bellegarda, "Spoken language understanding for natural interaction: The siri experience," *Natural interaction with robots, knowbots and smartphones*, pp. 3–14, 2014.

[6] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelligence, and computer vision." *IEEE Consumer Electronics Magazine*, vol. 6, no. 2, pp. 48–56, 2017.

[7] M. B. Hoy, "Alexa, Siri, Cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[8] S. Schoolar. Scientific papers search results by keyword "speech technology" and data range filter 2017-2022. [Online]. Available: https://www.semanticscholar.org/search?year%5B0%5D=2017&year%5B1%5D=2022&q=speech%20technology&sort=relevance

[9] V. Popov, S. Kamenev, M. Kudinov, S. Repyevsky, T. Sadekova, V. Bushaev, V. Kryzhanovskiy, and D. Parkhomenko, "Fast and lightweight on-device tts with Tacotron2 and LPCNet," in *Proc. Interspeech*, 2020, pp. 220–224.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[11] P. Deepa and R. Khilar, "A report on voice recognition system: Techniques, methodologies and challenges using deep neural network," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021, pp. 1–5.

[12] Alpha Cephei Inc. VOSK is a speech recognition toolkit. [Online]. Available: https://alphacephei.com/vosk/

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Tech. Rep., Technical report, OpenAI, Tech. Rep., 2022.

[14] V. Delić, Z. Perić, M. Sečujski, N. Jakovljević, J. Nikolić, D. Mišković, N. Simić, S. Suzić, and T. Delić, "Speech technology progress based on new machine learning paradigm." *Computational intelligence and neuroscience*, 2019.

[15] V. Golenkov, N. Guliakina, and D. Shunkevich, *Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems]*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021.

[16] 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.

[17] I. Davydenko, "Ontologicheskoe proektirovanie baz znanij [ontology-based knowledge base design]," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2017, pp. 57–72.

[18] D. Shunkevich, "Agentno-orientirovannye reshateli zadach intellektual'nyh sistem [Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems]," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2018, pp. 119–132.

[19] V. Zahariev, N. Hubarevich, and E. Azarov, "Semantic analysis of voice messages based on a formalized context," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2019, pp. 103–112.

[20] V. Zahariev, D. Shunkevich, S. Nikiforov, and E. Azarov, "Intelligent Voice Assistant Based on Open Semantic Technology," in *Open Semantic Technologies for Intelligent System*, V. Golenkov, V. Krasnoproshin, and V. Golovko, Eds. Cham: Springer International Publishing, 2020, pp. 121–145.

[21] V. Zahariev, S. Nikiforov, and E. Azarov, "Conversational speech analysis based on the formalized representation of the mental lexicon," in *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems]*, V. Golenkov, Ed. BSUIR, Minsk, 2021, pp. 141–168.

[22] A. Kroshchanka, V. Golovko, E. Mikhno, M. Kovalev, V. Zahariev, and A. Zagoskij, "A Neural-Symbolic Approach to Computer Vision," in *Open Semantic Technologies for Intelligent System*, V. Golenkov, V. Krasnoproshin, V. Golovko, and D.Shunkevich, Eds. Cham: Springer International Publishing, 2022, pp. 282–309.

[23] O. Räsänen. Linguistic structure of speech. [Online]. Available: https://wiki.aalto.fi/display/ITSP/Linguistic+structure+of+speech
[24] B. Lobanov and O. Eliseeva, *Rechevoj interfejs intellektual'nyh sistem: uchebnoe posobie [Speech User Interface for Intelligent Systems: Tutorial]*. Minsk: BSUIR, Minsk, 2006.
[25] J. Laver, *Principles of phonetics*. Cambridge university press, 1994.
[26] X. Serra, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. Stanford University, 1990.
[27] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 8, pp. 1223–1235, 1988.
[28] A. Petrovsky, E. Azarov, and A. Petrovsky, "Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding," *Signal processing*, vol. 91, no. 6, pp. 1489–1504, 2011.
[29] E. Azarov, M. Vashkevich, and A. A. Petrovsky, "Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation." in *INTERSPEECH*, 2013, pp. 1697–1701.
[30] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+ noise model," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 550–553.
[31] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.
[32] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.
[33] H. Kawahara, "Exploration of the other aspect of vocoder revisited: Az straight, tandem-straight and morphing," in *Seventh ISCA Workshop on Speech Synthesis*, 2010.
[34] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, "Development of exploratory research tools based on tandem-straight," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009, 2009, pp. 111–120.
[35] "ISO/IEC 14496-3:2005. Information technology — Coding of audio-visual objects — Part 3: Audio," International Organization for Standardization, Geneva, CH, Standard, 2005.
[36] "ISO/IEC 23003-3:2020 Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding," International Organization for Standardization, Geneva, CH, Standard, 2020.
[37] "IEEE 1857.8-2020 - IEEE Standard for Second Generation Audio Coding," Institute of Electrical and Electronics Engineers, Standard, 2020.
[38] "IEC 62087-2:2015. Audio, video, and related equipment - Determination of power consumption - Part 2: Signals and media," International Electrotechnical Commission, Standard, 2015.
[39] "AES Tech 3250-2004 Specification of the digital audio interface (AES/EBU)," Audio Engineering Society, Standard, 2004.

Аудио-интерфейс интеллектуальных компьютерных систем нового поколения

Захарьев В. А., Жаксылык К. Ж., Лихачев Д. С., Петровский Н. А., Вашкевич М. И., Азаров И. С.

Работа посвящена рассмотрению вопросов создания аудио- и речевых интерфейсов для интеллектуальных компьютерных систем нового поколения. Предлагается использование подхода на основе онтологического проектирования и формализации системы понятий из предметной области аудиоинтерфейсов посредством технологии OSTIS. Изложены основные идеи, лежащие в основе данного подхода, а также особенности, отличающие их от общепринятых.

Суть подхода заключается в рассмотрении процесса проектирования аудио интерфейса как интерфейсной подсистемы в рамках общего процесса разработки интеллектуальной компьютерной системы и построении её формальной логико-семантической модели.

Для достижения поставленной цели предлагается прибегнуть к подходу на основе принципов лежащих в основе "Стандарта открытой технологии онтологического проектирования, производства и эксплуатации семантически совместимых гибридных интеллектуальных компьютерных систем"или кратко "Стандарта технологии OSTIS" [15].

Для создания подобной модели интеллектуальной компьютерных систем нового поколения необходимо:

- произвести декомпозицию информационной компьютерной системы на компоненты. Качество декомпозии при этом определяется простотой последующего синтеза общей формальной модели из формальных моделей выделенных компонентов.
- провести конвергенцию выделенных компонентов в целях построения совместимых формальных моделей этих компонентов;
- провести интеграцию построенных формальных моделей выделенных компонентов и получить общую формальную модель.

Показано, что в перспективе использование данного подхода может обеспечить свойства унификации, семантической совместимости и интероперабельности при разработке аудио- и речевых интерфейсов, что в итоге позволит существенным образом сократить издержки при создании интеллектуальных компьютерных систем нового поколения для решения комплексных задач.

Received 30.10.2022