An approach to speech ambiguities eliminating using semantically-acoustical analysis

Zahariev V.A.
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
zahariev@bsuir.by

Azarov E.S.
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
azarov@bsuir.by

Rusetski K.V.
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
rusetski@bsuir.by

***Abstract*—An approach to the problem of elimination of ambiguities in speech messages by application of semantically-acoustical analysis is presented in this paper. Authors propose the architecture of the intelligent system that implements this principle. According to this principle the direct transition from speaking to meaning of given phrase is possible with the help of digital signal processing techniques, as well as knowledge formalization methods using semantics networks (semantically-acoustical analysis). A prototype of intelligent system to resolve speech ambiguities of a certain type (homonyms and paronyms) based on the tools provided by the OSTIS technology and GUSLY signal processing framework has been implemented. The main advantages of the proposed solution in comparison to the standard automatic speech recognition systems and possible ways of further development for natural language understanding problem are also reported in this paper.**

***Keywords*—speech processing, semantic technologies, acoustic analysis, semantic analysis, semantically-acoustical analysis, instantaneous harmonic analysis, natural language understanding, ostis-system, SC-code**

I. Introduction

Speaking is the one of the most natural and effective forms of communication between people. This fact explains the significant interest of researchers to the development and practical use of speech interfaces to provide human-machine interaction in the modern communicational, multimedia and intelligent systems [10], [9]. The majority of scientific publications in this direction is devoted to the issues of basic technologies development. This technologies are the components of the speech interface, such as text-to-speech synthesis (TTS), as well as speech-to-text (STT) recognition [4], [8].

Especial interest cause the tasks that are associated with understanding the sense of the natural linguistic message in the context of the practical use of the speech interface as a part of the intelligent systems.(NLU) [33], [18], [30]. In terms of intelligent systems, it is interesting to investigate the integration of the message's information content into the current state of the intellectual system of knowledge base for the purpose of further processing by the knowledge machine and intelligent agents.

Most modern systems of sense understanding are built on the basis of three-tier architecture, when the voice message consistently passes the stages of acoustic analysis of speech signal, then linguistic analysis. The linguistic analysis results in the textual form of the original message are presented, and only then semantic analysis of the message is performed. However, from works on psycholinguistics and cognitive psychology It is known, that processes of perception and comprehension in human consciousness proceed continuously [21], [27], and in general it is not necessary to bring speech message preliminary to a text form to perform semantic analysis of its contents. Oral and written forms of speech with equal success can be processed by sensory and cognitive systems of the person [22]. Therefore, the question of creation of approaches and methods for systems in which the direct transition from processing of the message in a speech form to the analysis of its semantic content is actual.

II. Problem statement

The classical three-tier approach (acoustic, linguistic, semantic analysis), in case of solving the problem of comprehension of speech messages, has a number of significant disadvantages:

- introduction of intermediate stage: conversion of speech signal to text, entails extra costs associated with the need of linguistic processing, thereby increasing the overall computational complexity of the algorithm.
- the presence of a text processing stage causes additional errors and distortions due to the limitations and incomplete of correspondence linguistic models to the process used to navigate to the textual representation of information on different stages of transformation (phoneme-to-morpheme, morpheme-to-word, word-to-phrase, etc.) [13]
- an approach when we translate a speech signal into text, we could lose some of the information that may be important for understanding the meaning of the message, such as volume, duration, intonation, pauses between words that may not always appear in the text clearly expressed with punctuation marks, etc. This problem is especially relevant when analyzing messages that are not complete sentences, but can be interpreted by the listener. For example, in everyday speech a sentence consisting of only a sound [ah] depending on the volume, intonation and duration of sounding can express pain, wonder, question, act as a conjunction or a particle («ah,

leave him...» - «a... оставь его», «ah, who is it?» - «a... кто это?», «ah, and if we did a different...» - «а если бы сделали по-другому...») [26].

- the translation of the sound signal into the text makes it impossible to analyze audio, which are not only speech messages, but also carry potentially important information for the system, for example:
  - information about conditional signals, issued by objects of external environment, in particular, equipment on manufacture, cars on road, etc;
  - sounds that can correspond to emergency situations or alarm signal (rumbling, clang, hiss, explosions, etc.);
  - other sounds that potentially carry information about the state of the environment of the automated system.

  The lack of this type of signal analysis greatly limits the ability of automated systems that are oriented on work in a constantly changing environment, including environment that is difficult to predict.

A. The problems in the analysis of voice messages

Until now, many problems related to the understanding of natural language, and, in particular, speech messages, remain unresolved. These problems can be divided into two groups: (1) problems caused by the properties of natural language and inherent to the understanding of natural language messages presented in any form, both textual and verbal; (2) problems inherent to the directly understanding of speech messages.

Problems of analysis and understanding of natural language texts are widely considered in the literature [3] and their full review goes beyond the scope of this paper. In more detail, consider those that will be partially resolved within the framework of the approach proposed in the work. Such problems include:

- the problem of solution for homonyms identifying [23];
- pronoun identification problem [34];
- the problem of proper names identifiaction (for example, the word «Slava» at the beginning of a sentence can mean both a diminutive form of the name «Vyacheslav», and a common noun);
- the problem of homographs (zAmok-zamOk);
- the problem of understanding the different forms of the same word (which in some cases may coincide with the forms of other words);
- the problem of understanding terms that consist of two or more words («acceleration of free fall»).

In addition, there are a number of additional problems related directly to the characteristics of the speech signal:

- the problem of paronyms (similar in sounding words) (koza-kazak-kazan-Kazan', dictant-dictat, postel'-pastel', etc.);
- the problem of proper names resolution is complicated by the lack of capitalization;
- the speech signal is more complex in terms of presentation and processing than textual information, because of the greater variability and expressiveness of oral speech compared to the written form. This fact, in particular, is connected with the rich intonational (prosodic) possibilities of oral speech. The intonation is formed by melody (change of frequency) of speech, intensity (loudness) of speech, duration, increase or slowing down the tempo [28]. A large role is played by the place of logical stress, the degree of clarity of pronunciation, the presence or absence of pauses. The speech has such an intonational variety that it can convey the whole wealth of human experiences, moods and emotions.

Obviously, the listed problems cannot be solved without an analysis of the context of the use of a particular word, while the context in general can be quite broad and go beyond one sentence.

Analysis of the current state of publications about the understanding of oral speech shows that to solve the problem of contextual recognition in the absence of a linguistic stage of processing, approaches based on popular machine learning techniques based on the neural network classification (Connectionist Temporal Classification - CTC) based on recurrent neural networks with Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) and Deep Neural Network (DNN) methods [1], [5], [16]. And in this case, we mean the recognition of individual phonemes, morphemes or words in the speech stream. However, in these works, questions of the semantic processing of such information and integration into the intellectual system are not posed and are not considered, which leaves this issue open and indicates that there are unsolved problems.

III. Semantically-acoustical analysis

In this paper, problems that were mentioned above are proposed to be solved by performing a semantically-acoustical analysis. This process involves the primary analysis of a voice message using special signal processing techniques. In the course of their application, words are isolated from the stream of individual «acoustical pattern». This acoustical pattern will correspond to certain nodes (signs of concrete entities or concepts) in the semantic network. It is assumed that the results of the acoustic analysis phase will be iteratively corrected considering the information stored in the knowledge base of the system, including the semantic analysis of context-sensitive information.

By «acoustical pattern» we mean a fragment of the speech signal, usually represented in some parametric form, which corresponds in duration to the phonetic word in the speech stream. One of the hypotheses put forward in the work is the assumption that from the point of view of formalizing the semantics of a message one can work with a signal at the level of the whole word (and not its separate parts, such as a phoneme or morpheme used in classical speech recognition systems on a linguistic basis stage of processing). And a word, in this case, is the minimal sense distinctive unit of speech, i.e. denotes the concept or specific entity represented in the knowledge base of the system.

In this work, at the current level of development of the proposed approach, the selection and analysis of words is proposed to be carried out based on recognition systems in comparison with the vocabulary of standards and further determining the degree of certainty (correspondence) of the allocated "acoustic pattern" to the reference of this pattern – the content of the node of the semantic network in the knowledge base. The use of such a measure as the degree of certainty instead of the immediate value of the likelihood of the pattern matching to the reference is due to the fact that the concept of probability in this case can not be used in the full sense correctly, because the standards presented in the dictionary do not represent a complete group of events, the sum of the probabilities of which gives unity. The measure of the degree of confidence in this case is rather a measure of the proximity of the selected signal fragment to the reference signal in the selected parametric space.

In the general case, in order to carry out this kind of analysis, the following information may be required in the proposed approach:

- a set of standards for comparison with the fragments of the speech signal and their specification;
- the context of the message being analyzed (from whom the message was received, in what external conditions, what other sounds are present on the background, etc.);
- a set of rules for the transition from fragments of a voice message to semantically equivalent constructs in the knowledge base;
- the semantic specification of the concepts that make up such constructions.

For speech analysis will be used a model based on a hybrid representation of a speech signal, which allows the most adequate representation of any fragments of the speech signal of a different nature of sound formation [15]. Vocalized and unvoiced fragments of the signal refer to separate parts of the model: periodic (harmonic) and aperiodic (noise).

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

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

where $s(n)$ – input speech signal, $h(n)$ – harmonic component, $r(n)$ – noise component of the signal, n and N – current signal reference 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 – gain coefficient on the basis of the spectral envelope, c is the number of sinusoidal signal components for each harmonic, A_k^c – instantaneous amplitude of the c-th component and k-th harmonic, f_k^c and $\phi_k^c(0)$ – frequency and initial phase of the c-th component of the k-th harmonic, e_k is the excitation signal of the k harmonic. The amplitudes A_k^C are normalized in order to provide the sum of the energy of the harmonics equal to $\sum_{c=1}^{C}[A_k^c]^2 = 1$ for k = 1,...,K.

In this case, the aperiodic component is modeled in the whole frequency band, as it is observed in the spectrum of the real speech signal [12]. This effect is achieved by applying the technique of signal analysis through synthesis and subtraction 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 a single frame of the signal with the number m and the length of N samples, a characteristic vector is formed, including the coefficients of the model $\mathbf{x_m} = [G_k, A_k^c, f_k^c, K, C]$. And the acoustic pattern of one word is a sequence of such characteristic vectors: $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M})^T$.

The model parameters are proposed to be estimated using the original method of instantaneous harmonic analysis, which makes it possible to significantly improve the accuracy of determining the parameters of the periodic component [19].

In contrast to the classical methods of signal analysis used in modern speech recognition systems based on the definition of the mel-cepstral coefficients (MFCC) [32], [20] or linear speech prediction (LPC) [31], the method based on instantaneous harmonic analysis (IHA) makes it possible to obtain a high temporal and frequency resolution of the signal, as well as a more precise spectral picture of the localization of energy at the appropriate frequencies. In contrast to classical methods based on a short-time Fourier transform (STFT) or the definition of the autocorrelation function of a signal on a short fragment, the method in question does not impose strict limitations connected with observance of the stationary conditions of the signal parameters on the analysis frame 1. In this case, the parameters of the harmonic model, if necessary (for example, for describing the spectral envelope) can be relatively easily converted to other presentation methods, such as classical mel-cepstral or linear prediction coefficients.

Algorithms implementing the above-described method of signal processing based on instantaneous harmonic analysis have a standard implantation in the framework of analysis and synthesis of audio signals GUSLY [2], the individual components of which will be used for the demonstration example presented below.

As a technological basis for implementing the proposed approach, OSTIS [24] Technology will be used. Systems based on OSTIS technology are called ostis-systems, respectively, the module for understanding voice messages, the prototype of which is considered in this work, will be built as a reusable component, which in future will be integrated into various ostis-systems if necessary.

As a formal basis for encoding various information in the knowledge base, the SC-code [24] is used, the texts of which (sc-texts) are written in the form of semantic networks with a basic set-theoretical interpretation. Elements of such networks are called sc-elements (sc-nodes, sc-arcs).

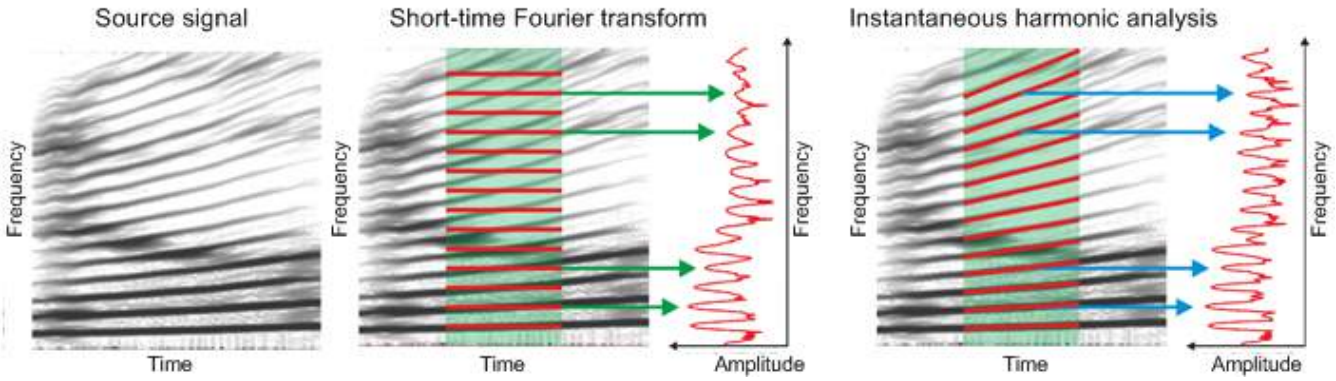

Figure 1. Comparison of signal analysis methods based on short-time Fourier transform vs instantaneous harmonic analysis.

The orientation of this work to 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 allows to describe all the information necessary for analysis in one knowledge base in a unified way [6];
- used in the framework of technology formalism allows you to specify in the knowledge base not only concepts, but also any external files from the point of view of the knowledge base (for example, fragments of the speech signal), including the syntactic structure of such files;
- the approach to representation of various kinds of knowledge [6] and models of their processing [14] offered in the framework of technology provides modifiability of ostis-systems, i.e. allows to easily expand the functionality of the system, introducing new types of knowledge (new concepts systems) and new models of knowledge processing;
- the above-mentioned advantages allow to provide acoustic, syntactic and semantic analysis of messages in the same memory with the help of unified processing facilities, which in turn allows you to correct the analysis processes at any stage using different information from the knowledge base.

In its turn, the developed module for understanding voice messages is itself built as an ostis-system and has an appropriate architecture.

IV. Limitations of the proposed approach

In general, the task of understanding the meaning of a message in a language external to the ostis-system involves the following main steps:

1) Syntactic (semantic-syntactic) analysis of the message. The result of this stage is the sc-text describing the syntactic structure of the external text.
2) Semantic analysis (translation) of the message. The result of this stage is the sc-text semantically equivalent to the original message in the external language.
3) Immersion (integration) of the received sc-text into the knowledge base of the ostis-system. At this stage, there is a "glueing together" of synonymous sc-elements contained in the initial knowledge base of the system and the corresponding sc-elements included in the sc text, which is the result of the translation of the original message [25].
4) Alignment of concept systems. At this stage, the system of concepts used in the analyzed message is brought to the system of concepts accepted as the main one in the knowledge base of the ostis-system.
5) The analysis of the value of the information received, the logical conclusion. At this stage, new information is generated based on the logical inference mechanisms and existing in the system.

Some or all of the listed steps can be performed iteratively, taking into consideration the context, which is generally formed both using information from the knowledge base and information obtained from the external environment, including, explicitly requested from the user.

The focus of this work is on the first two of these stages of understanding, while the task of understanding speech messages is refined with several requirements that allow focusing attention on solving the problems listed above. Let us enumerate the specified requirements:

- voice messages do not contain noises, the words in the phrase are pronounced by the speaker, so that the pauses are clearly distinguishable in the speech signal, since the task does not set the construction of a robust algorithm for extracting words from an arbitrary speech signal;
- it is assumed that the analyzed speech message and the standards of the fragments of the speech signal contained in the knowledge base are recorded by the same speaker, e.g., those, in the speaker-dependent mode;
- only those speech messages that contain phrases of the form "subject-predicate-object" are analyzed. A predicate may be an action performed by a subject with respect to an object, or some relation connecting a subject and object. It is assumed that the components of the phrase follow exactly in this order, i.e. the first in order always

follows the subject, the last is the object, although in the natural language, in the general case, an inverse sequence is possible. It is believed that each of the components of the phrase is called with one word;
- it is assumed that all concepts used within the message are known to the system (specified in the knowledge base) and are the basic concepts, thus there is no need to align systems of concepts;
- the system of concepts used within the message does not change during the analysis, i.e. a priori, the information stored in the knowledge base is considered to be truth, and the meaning of the message is specified considering this information;
- it is assumed that there is a dictionary of lexemes corresponding to the words used in the analyzed phrase in the knowledge base, so all words are known to the system. In this case, each token corresponds to a set of fragments of the speech signal (standards) describing all the word forms of the corresponding word, thus morphological analysis is not carried out;
- it is assumed that the analyzed message is correct from the point of view of the current state of the knowledge base.

A. Test cases

In view of the problems and limitations discussed above, a number of test phrases has been selected and these phrases will be used to demonstrate the proposed approach to understanding voice messages (for phonetic transcription, the X-SAMPA notation will be used [17]):

1) «The goat ate cabbage» - «Коза съела капусту» - [kaza] [sj'Ela] [kapustu]
2) «Grozny captured Kazan» - «Грозный захватил Казань» - [grozn1j'] [zaxvat'il] [kazan']
3) «Kazan is located on the Kaban (Kaban - in this case the lake)» - «Казань расположена на Кабане (Кабан - в данном случае озеро)» - [kazan'] [raspaloz'Ena] [na] [kaban'E]
4) «Boar is the ancestor of a pig» - «Кабан – предок свиньи»- [kaban] [pr'Edak] [sv'in'i]
5) «Grozny is the capital of Chechnya» - «Грозный – столица Чечни» - [grozn1j'] [stal'itsa] [ts /Etsn/h'i]
6) «Shepherd bought a goat» - «Чабан купил козу» - [ts/aban] [kup'il] [kazu]
7) «Shepherd bought kazan» - «Чабан купил казан» - [ts/aban] [kup'il] [kazan]

Examples of correct constructions semantically equivalent to some of these examples are shown in the figures 2–4:

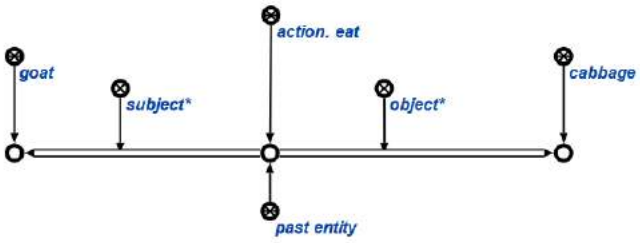

Figure 2. A construction equivalent to a phrase «The goat ate cabbage» - «Коза съела капусту».

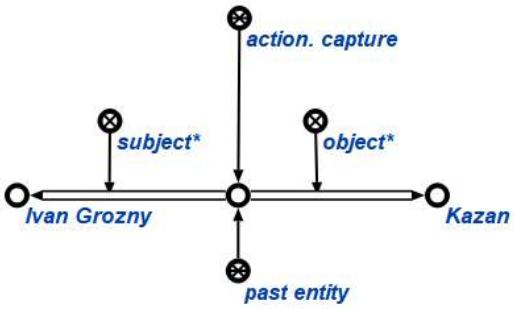

Figure 3. A construction equivalent to a phrase «Grozny captured Kazan» - «Грозный захватил Казань».

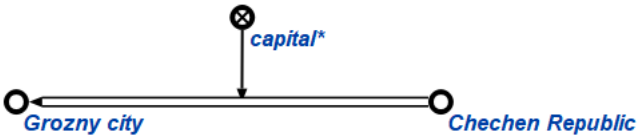

Figure 4. A construction equivalent to a phrase «Grozny - the capital of Chechnya» - «Грозный – столица Чечни».

V. Algorithm for analyzing the voice message

In general, the proposed algorithm for understanding speech messages with elimination of ambiguity includes the following steps:

1) Selection of the individual words (fragments of the speech signal between pauses) in the original speech message. The result of this step is the specification in the knowledge base of the syntactic structure of the message being analyzed, which, given the previously mentioned limitations in this work, is given by a set of words and the order of their following.
2) The received words are compared with the standards stored in the knowledge base and corresponding to some lexemes. For each of the standards, the degree of confidence (from 0 to 1) is calculated in that the analyzed fragment of the speech signal and the reference coincide. Coincidence, the degree of confidence for which below a certain threshold is discarded, the rest are fixed in the knowledge base.
3) Considering the received confidence levels, those pairs «signal fragment» - «reference» are chosen, for which

the confidence levels are maximal and their translation into semantically equivalent sc-text is carried out. Broadcasting in the general case can be carried out both with the help of a universal mechanism based on implicative rules, and with the use of specialized translation agents, each of which is oriented only to the construction of a certain type.

4) The received sc-text is verified by the existing means of the knowledge base verification in the system. In case of contradictions, return to step 3 is performed, and the following degree of confidence of the pair «signal fragment» - «reference» is chosen. In addition, it is considered in which fragments of the received sc-text a contradiction has arisen, and first of all those fragments of the speech signal that correspond to the specified fragments of the sc-text are taken into account.
   If contradictions are not revealed, then the conclusion is made that the received sc-text does not contradict the current system of concepts and is integrated into the knowledge base.

Below is the application of this algorithm to the example of a specific phrase.

VI. The architecture of the speech message comprehension software module

The architecture of the software module for understanding voice messages is shown on the figure 5.

The system consists of two main components: the acoustic component and the semantic analysis component. In its turn, the first component includes a signal analyzer, a module that implements the patter matching algorithm with a reference, a knowledge base, a configuration file, and initial parameters of the signal representation model (length and type of the analysis window, sampling and oversampling frequencies, etc. constants included in the configuration of the signal representation model).

The speech signal is fed to the input of the analysis module, where the procedures for dividing the signal into frames with a duration of 50 msec with 25% overlap are performed, the signal frames are weighted by multiplying the current signal fragment by the Hamming window, and the pitch frequency is searched. Next, the parameters of the signal model are estimated and a characteristic vector $\mathbf{x_m}$ is formed for the current frame, which is placed in a sequence of similar vectors $\mathbf{X}$, characterizing the entire word. Since the number of analysis frames, and accordingly the size of the sequence of characteristic vectors, will fluctuate depending on the duration of the word, it is necessary to perform the procedure for normalizing the number of vectors in the sequence. For all words, the size of the sequence is reduced to 10 vectors by applying the vector quantization procedure over the sequence [7].

Further in the comparison module, the obtained normalized sequence is compared with a number of standards stored in the database, which are a collection of the same sequences, but prepared and written to the database in advance. The comparison is performed using the dynamic time warping (DTW) procedure, based on the dynamic programming algorithm for multidimensional data [11]. For the three most appropriate standards, the degree of assurance of the pattern's compliance with the reference is calculated, the values obtained are recorded in the contents of nodes of the semantic network, which are appropriately specified in the knowledge base. Further, the semantic processing is carried out on the basis of information from the knowledge base, as mentioned above.

A. The knowledge base of the speech interface

To implement the semantically-syntactical analysis in the knowledge base, lexemes (sets of word forms) are specified that correspond to words that will be used in the framework of the speech message. Examples of the specification of lexemes (Figures 6-9):

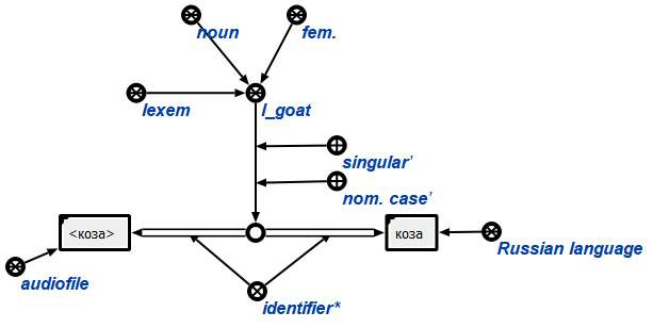

Figure 6. Lexeme «goat» - «коза».

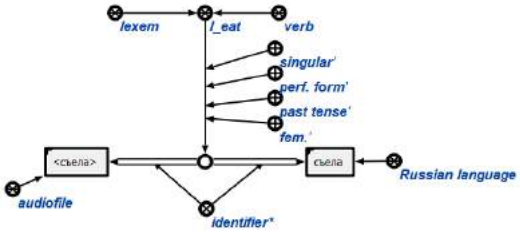

Figure 7. Lexeme «eat» - «есть».

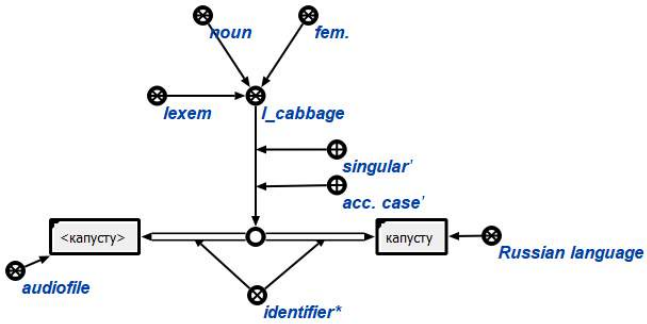

Figure 8. Lexeme «cabbage» - «капуста».

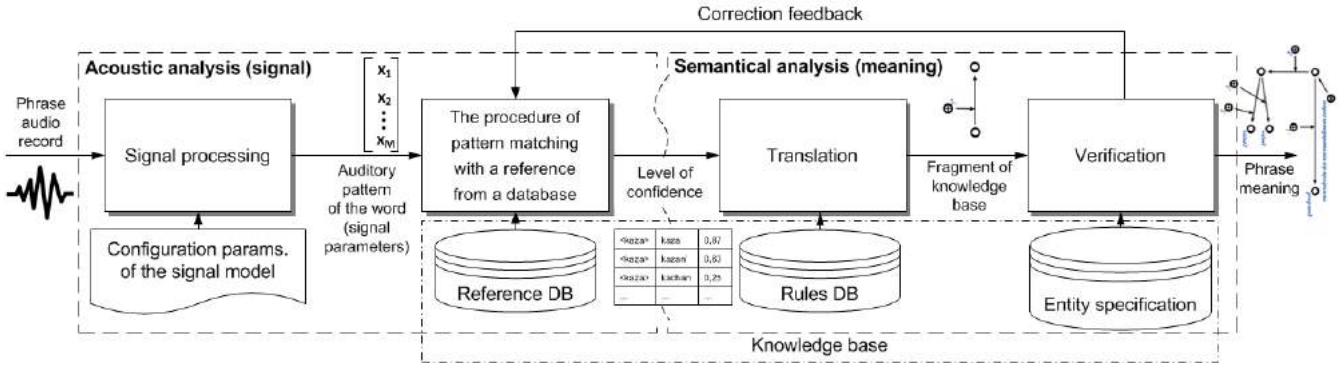

Figure 5. System architecture.

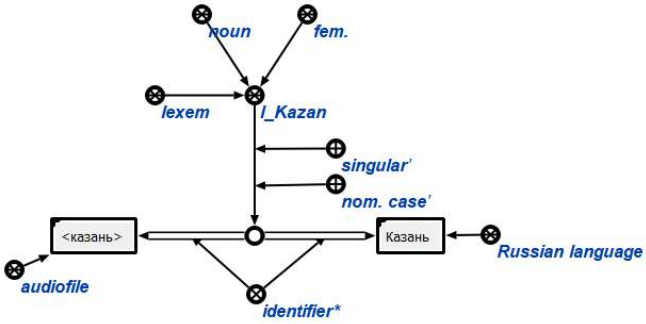

Figure 9. Lexeme «Kazan» - «Казань».

In the examples considered, the angular brackets (< goat> - <коза>) conditionally indicate fragments of the speech signal stored in the knowledge base as standards with which the comparison will be carried out at the analysis stage.

In addition, the knowledge base contains a specification of concepts corresponding to lexemes, which is used in the analysis process to verify the result of translation. The specified specification can include, for example, specifying the relationship domains (figure 10) or specifying the class of possible subjects and objects for the action class (figure 11).

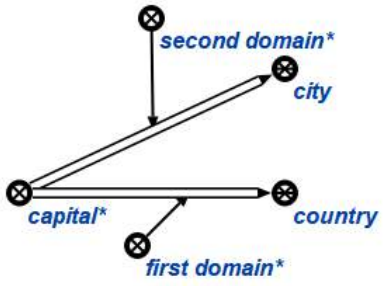

Figure 10. Relationship specification «capital*» - «столица*».

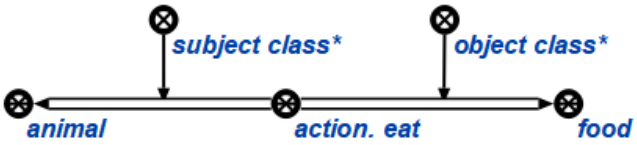

Figure 11. Activity class specification «eat» - «есть».

B. The knowledge processing machine of the speech interface

The knowledge processing machine for the speech interface according to the considered analysis algorithm currently has the following structure:

The speech interface knowledge processing machine *<=abstract sc-agent decomposition**:

{

- *Abstract sc-agent for audio preprocessing*
- *Abstract sc-agent for recognizing audio fragments*
- *Abstract sc-agent for generating the translation task*
- *Abstract sc-agent for translating external files to the knowledge base*
- *Abstract sc-agent of knowledge base verification*
  *<=abstract sc-agent decomposition**:
  {
  - *Abstract sc-agent for checking the matching of bindings to its domains*
  - *Abstract sc-agent for checks the compliance of the activity specification with its class*

  }

}

VII. Example of the algorithm for understanding the voice message

As an example, consider the process of analyzing the phrase "the goat ate cabbage". In this example, double angular brackets (<<goat>> - <<коза>>) conditionally identify fragments of the analyzed speech signal. The above illustrations are recorded using one of the SC-code visualization variant, the SCg [29] language.

Step 1. In the first step, the original speech message is decomposed into separate words, the decomposition result is fixed in the knowledge base, as shown in the figure 12.

Step 2. The received fragments of the speech signal are compared with the standards, the result of the comparison is fixed in the knowledge base, as shown in the figure 13. In this example, we will assume that the analyzed word <<goat>> - <<коза>> coincides with the reference <goat> −<коза> with a confidence level of 0.55 and a reference <kazan> - <казань> with a confidence level of 0.65.

Step 3. Considering the confidence levels obtained, those pairs « signal fragment » - « reference » are chosen for which the confidence levels are maximal and their translation into semantically equivalent sc-text is carried out. In the example under consideration, the reference <kazan> - <казан> was first chosen, since the confidence level for it turned out to be larger (figure 14). Thus, the meaning of the phrase being analyzed is interpreted as «Kazan ate cabbage».

Step 4. For the obtained sc-text, the sc-agent for the verification of knowledge bases is initiated (figure 15), which is based on information from the knowledge base (figure 16) that Kazan is a city, not animal, and only animals can eat, so the agent concludes that the resulting structure is incorrect (figure 17).

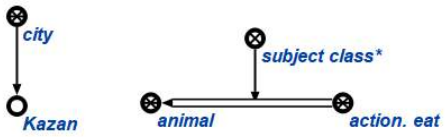

Figure 16. Verification context

Step 5. Since the received sc-text was incorrect, the translation result is deleted and a return to step 3 is made, where the next set of pairs «signal fragment»-«reference» is translated (figure 18) and verified (figure 19). In the example under consideration, the resulting sc-text turned out to be correct within the current state of the knowledge base, otherwise the translation and verification steps would be repeated.

VIII. Conclusion

An approach to the problem of eliminating ambiguities such as homonymy and paronymy in a speech signal with the use of semantically-acoustical analysis for speech understanding systems is proposed. This approach involves using the knowledge base to verify the results of message translation into the internal representation of an intelligent system, with subsequent correction of the results of recognizing fragments of the speech signal. Its main feature is the fact that it suggests the transition to semantic processing by passing the stage of processing textual information, the presence of which is a characteristic feature of all modern solutions to understanding speech. This implementation let to avoid the loss of part of the information and reduces the number of errors introduced at the text processing stage due to the imperfection of linguistic models. For the analysis and parametrization of the signal, a hybrid model is used. This model is based on the hybrid representation of the speech signal, which allows the most appropriate representation of any fragments of the speech signal of a different nature of sound formation, both vocalized and unvoiced. The method of instantaneous harmonic analysis is used to estimate the parameters of the model, which makes it possible to significantly improve the accuracy of determining the parameters of the periodic component.

For semantic analysis, OSTIS technology is used, which provides unified means for representing various types of knowledge, including meta-knowledge, which makes it possible to store and process all necessary information in one knowledge base in a unified way. In addition, technology allows you to specify in the knowledge base not only concepts, but also any external forms of knowledge representation, for example acoustic patterns of words, also allows you to easily expand the functionality of the system by introducing new types of knowledge and new models of knowledge processing. This technology provides for the modifiability of ostis-systems, i.e. allows you to easily expand the functionality of the system, introducing new types of knowledge and new models of knowledge processing.

The main differences of this work from the existing works in the field of understanding of voice messages and elimination of ambiguities in such messages include the following:

- the paper proposes an original approach that assumes consistent use of acoustic and semantic analysis, which allows to take into account the context at different stages of the speech message understanding and to adjust the results of each stage with use of the context;
- the proposed approach, unlike modern methods of speech recognition and understanding, excludes the need for an intermediate stage of the message presentation in text form, which allows to expand the context of the message analysis taking into account various parameters of the voice message (loudness, intonation, emotional coloring, etc.), and also allows in the future to analyze messages that do not have a unique text equivalent (sounds of the environment);
- the means of knowledge representation and processing used in the approach provide an ability to enhance the functionality of the system and the quality of understanding easily, including by specifying within the knowledge base of different types of context and their subsequent use in the analysis process, as well as expanding of intermediate information verification means at different stages analysis;
- the proposed approach to eliminating ambiguities in voice messages is part of the solution of a more general problem related to the learning and self-learning of intelligent systems by understanding of information obtained from various external sources, including speech and sound.

Further work will be focused on development, improving and expanding the proposed approach for a more general case, detailed comparative studies with existing systems of

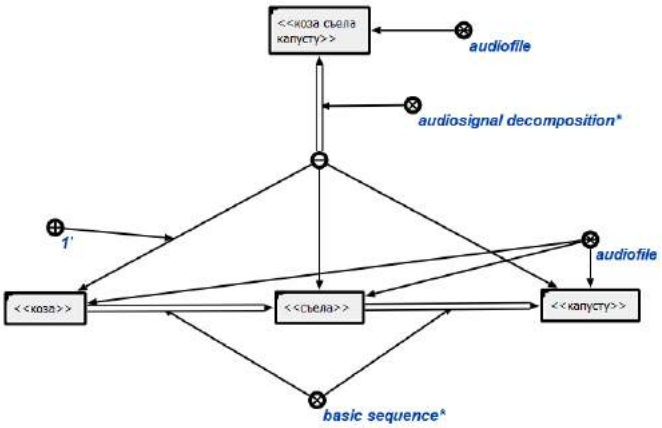

Figure 12. Analysis and parameterization of the speech signal

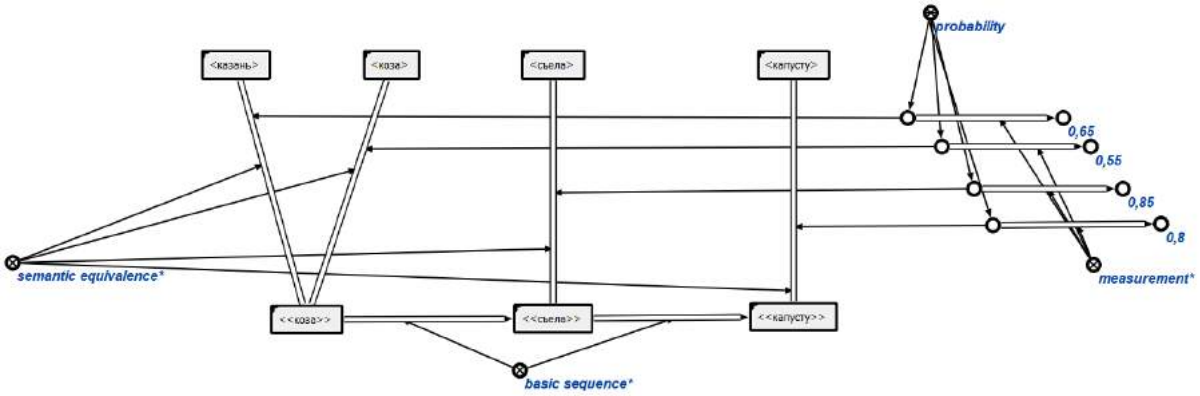

Figure 13. The result of matching the words of the original message with the standards from the knowledge base

speech recognition and understanding. Future improvements of the proposed approach should be in closer integration with approaches in the field of signal processing, psychoacoustics, psychosemantics and artificial intelligence to solve problems associated with pattern recognition and the formalization of meaning from information sources of any kind.

REFERENCES

[1] Audhkhasi, K. et al., Direct Acoustics-to-Word Models for English Conversational Speech Recognition. Proceedings Interspeech 2017, pp. 959–963.

[2] 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. P. 1697–1701.

[3] Barr A. Natural language understanding. AIMagazine. 1980. vol. 1. N. 1. P. 5.

[4] Bellegarda, J. R. Spoken language understanding for natural interaction: The siri experience. Natural Interaction with Robots, Knowbots and Smartphones. Springer, New York, 2014. P. 3–14.

[5] Corona, R., Thomason, J., Mooney, R. Improving Black-box Speech Recognition using Semantic Parsing. Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017. vol. 2. P. 122–127.

[6] Davydenko, I. T. Ontology-based knowledge base design. Open semantic technologies for intelligent systems. Minsk: BSUIR, 2017. P. 57–72.

[7] Gersho, A., Gray, R.M. Vector quantization and signal compression. Springer Science & Business Media, 2012. vol. 159. 732 p.

[8] Lin, K. An Examination of Natural Language Processing as an Alternative to Traditional UI's for User Interaction with Applications. – 2017.

[9] MacTear, M., Callejas, Z., Griol, D. The Conversational Interface: Talking to Smart Devices. Springer, 2016. 422p.

[10] Pearl, C. Designing Voice User Interfaces: Principles of Conversational Experiences. O'Reilly Media, 2016. 287 p.

[11] Rabiner, L. R., Rosenberg, A. E., Levinson, S. E. Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition. IEEE Trans. Acoust. Speech Signal Processing, 1978. Vol. ASSP-26, N. 6. P. 575–582.

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

[13] Schmitt, A., Wolfgang, M. Towards Adaptive Spoken Dialog Systems. Springer, 2012. 251 p.

[14] Shunkevich, D. V. Ontology-based design of knowledge processing machines. Open semantic technologies for intelligent systems. Minsk: BSUIR, 2017. P. 73–94.

[15] Stylinou, Y. Applying harmonic plus noise model in concatenative speech synthesis. IEEE Trans. on Speech and Audio Processing, 2001. Vol. 9, N. 1. P. 21–29.

[16] Tang, H. et al. End-to-End Neural Segmental Models for Speech Recognition. IEEE Journal of Selected Topics in Signal Processing, 2017. vol. 11. N. 8. P. 1254–1264.

[17] Wells, J.C. Computer-coding the IPA: a proposed extension of SAMPA. Revised draft, 1995. vol. 4. N. 28. P.19-25.

[18] Winograd, T. Understanding natural language. Cognitive psychology. 1972. vol. 3, N. 1. pp. 1–191.

[19] Azarov, I.S., Petrovskii, A.A. Mgnovennyi garmonicheskii analiz: obrabotka zvukovykh i rechevykh signalov v sistemakh mul'timedia

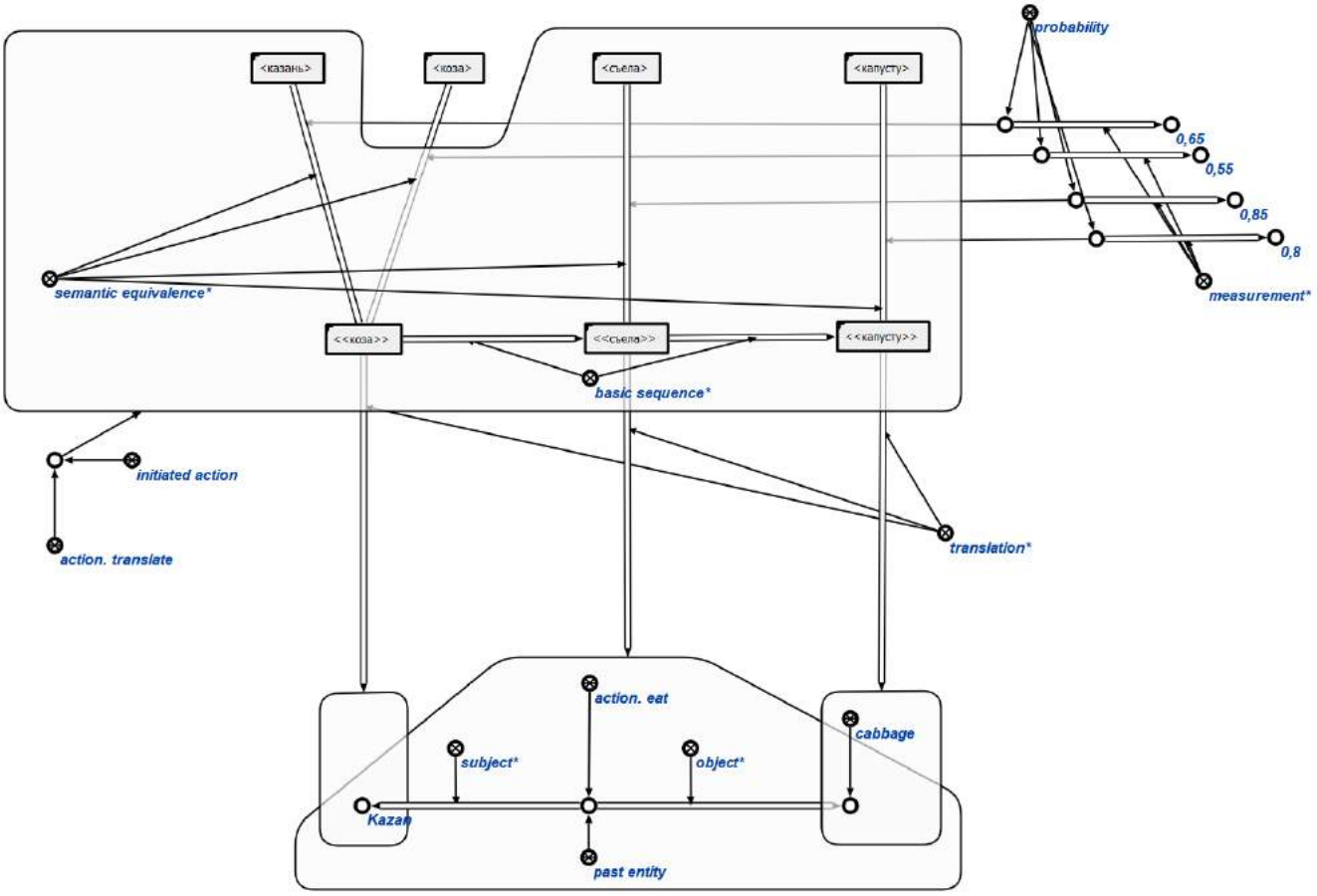

Figure 14. Translation result

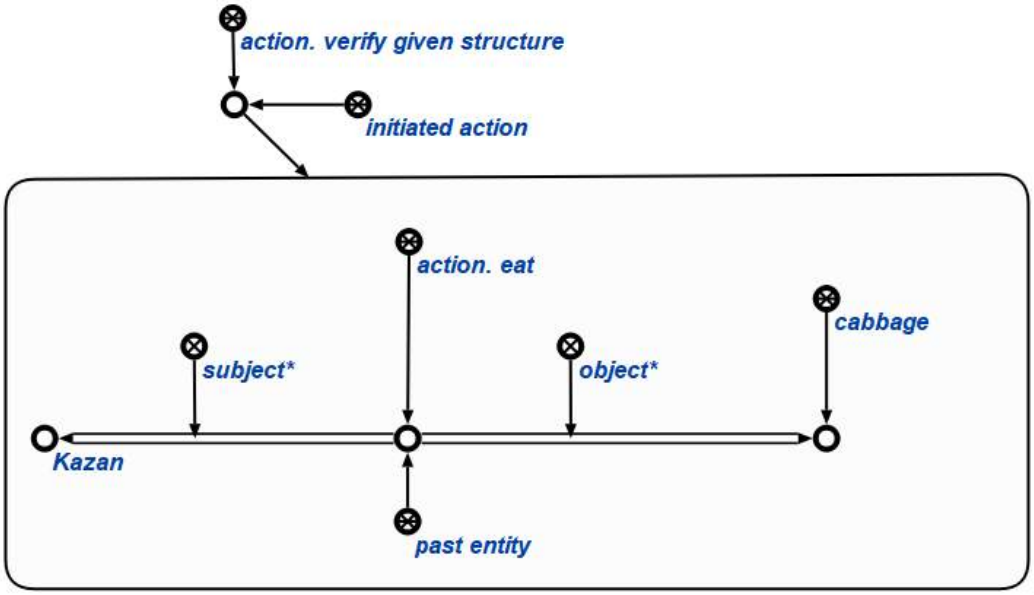

Figure 15. Assignment for verification

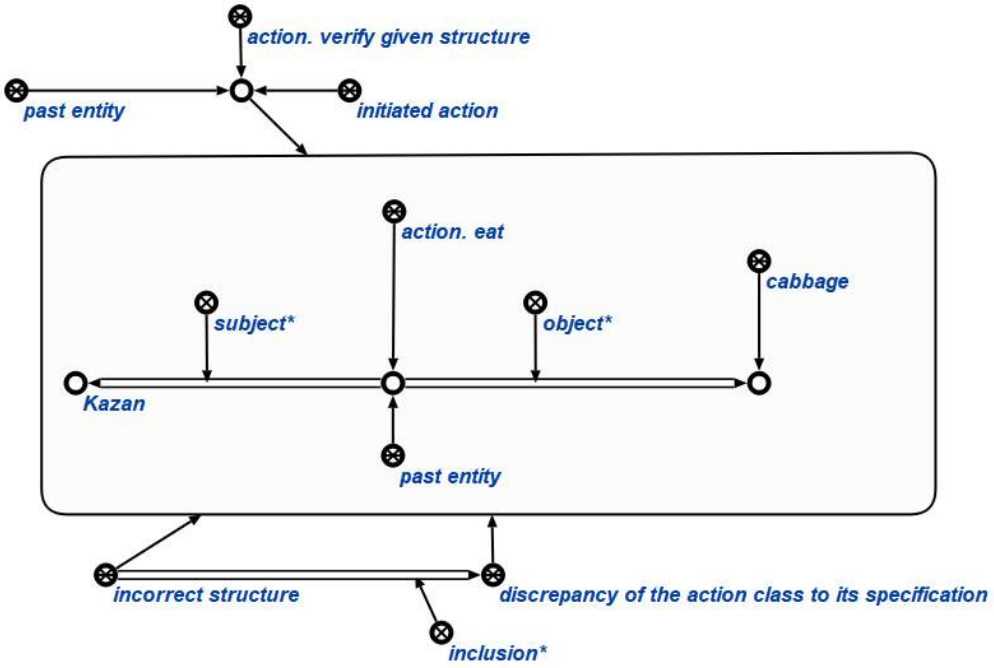

Figure 17. Verification result

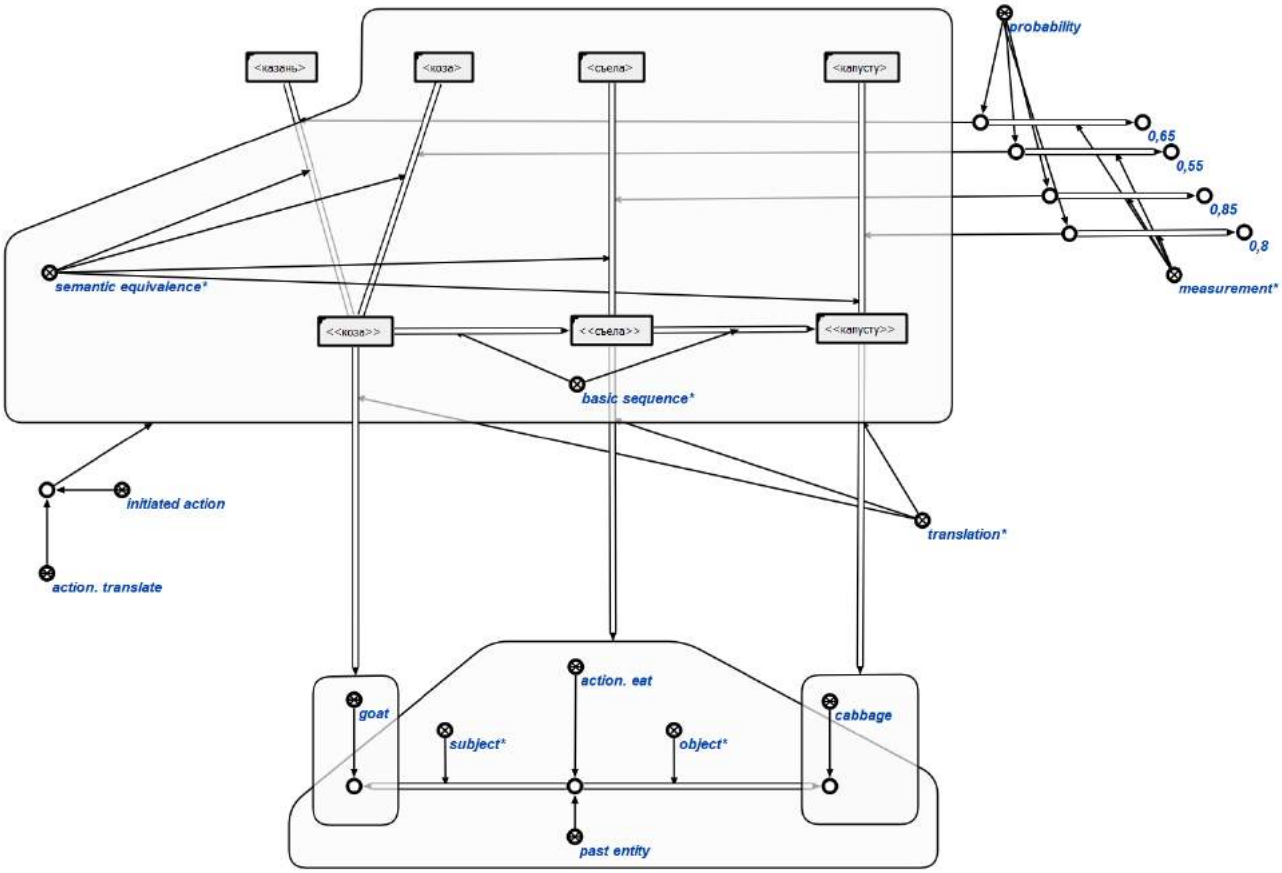

Figure 18. Repeated translation result

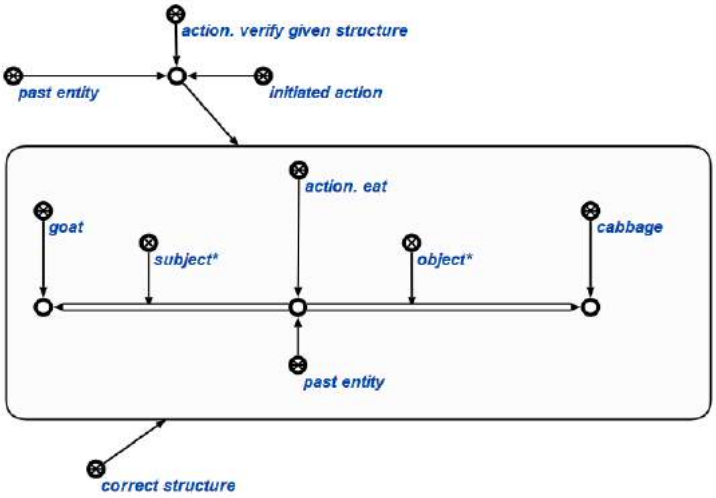

Figure 19. Repeated verification result

[Instant harmonic analysis: processing of sound and speech signals in multimedia systems]. LAP Lambert Academic Publishing, Saarbrucken. 2011. 163 p. (in Russian)

[20] Aificher, E., Dzhervis, B. Tsifrovaya obrabotka signalov: prakticheskii podkhod. 2-e izd. [Digital signal processing: a practical approach. 2nd ed.]. M.: Williams, 2004. 992 p. (in Russian)

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

[22] Goikhman, O. Ya., Nadeina, T. M. Rechevaya kommunikatsiya [Speech communication]. M.: Infra -m., 2007. 207 p. (in Russian)

[23] Zin'kina, Yu. V., Pyatkin N. V., Nevzorova O. A. Razreshenie funktsional'noi omonimii v russkom yazyke na osnove kontekstnykh pravil [Permission for functional homonymy in the Russian language on the basis of contextual rules]. Trudy mezhd. konf. Dialog. 2005. pp. 198–202. (in Russian)

[24] Golenkov, V.V., Gulyakina N.A. Semanticheskaya tekhnologiya komponentnogo proektirovaniya sistem, upravlyaemykh znaniyami [Semantic technology of component design of knowledge-driven systems]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem [Open semantic technologies for intelligent systems]. Minsk: BSUIR, 2015. pp. 57–78. (in Russian)

[25] Ivashenko, V.P. Modeli i algoritmy integratsii znanii na osnove odnorodnykh semanticheskikh setei [Models and algorithms of knowledge integration based on homogeneous semantic networks]: PhD thesis: 05.13.17. Minsk, 2015. 151p. (in Russian)

[26] Kodzasov, S. V. Razmeshchenie tonal'nykh aktsentov v russkom slove. In G.E. Kedrovoi and V.V. Potapov (Eds.). Yazyk i rech': problemy i resheniya: sb. nauch. trudov k yubileyu professora L.V. Zlatousovoi. [Placement of tonal accents in the Russian word. Language and Speech: Problems and Solutions: Sat. sci. works for the jubilee of Professor L.V. Zlatausova]. M.:Maks-Press, 2004. pp.62–76. (in Russian)

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

[28] Lobanov, B. M., Tsirul'nik L. I. Komp'yuternyi sintez i raspoznavanie rechi [Computer synthesis and recognition of speech]. Minsk: Belorusskaya nauka, 2008. 344 p. (in Russian)

[29] (2017, Jun.) IMS metasystem. [Online]. Available: http://ims.ostis.net/

[30] Popov, V. E. Obshhenie s EVM na estestvennom jazyke [Communication with a computer in natural language]. M.:Nauka, 1982. 360 p. (in Russian)

[31] Rabiner, L., Gould, B. Teorija i primenenie cifrovoj obrabotki signalov [Theory and application of digital signal processing]. Perevod s angl. Zajceva A. L.; pod red. Aleksandrova J. N. M.: Mir, 1978. 848 p. (in Russian)

[32] Rabiner, L. R., Schafer R. W. Cifrovaja obrabotka rechevyh signalov [Digital processing of speech signals]. Perevod s angl.; pod red. M. V. Nazarova; Ju. N. Prohorova. M.: Radio i svjaz', 1981. 496 p. (in Russian)

[33] Tarasov, V. B. Problema ponimanija: nastojashhee i budushhee iskusstvennogo intellekta [The Problem of Understanding: The present and future of artificial intelligence]. Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem [Open semantic technologies for intelligent systems]. Minsk: BSUIR, 2015. pp. 25–42. (in Russian)

[34] Tolpegin, P. V., Vetrov D. P. and Kropotov D. A. Algoritm avtomatizirovannogo razreshenija anafory mestoimenij tret'ego lica na osnove metodov mashinnogo obuchenija [Algorithm for the automated resolution of the anaphora of third person pronouns based on machine learning methods]. Trudy mezhd. konf. Dialog. 2006. pp. 504–507. (in Russian)

ПОДХОД К УСТРАНЕНИЮ РЕЧЕВЫХ НЕОДНОЗНАЧНОСТЕЙ НА ОСНОВЕ СЕМАНТИКО-АКУСТИЧЕСКОГО АНАЛИЗА

Захарьев В.А. (БГУИР), Азаров И.С. (БГУИР), Русецкий К.В. (БГУИР).

В работе рассмотрен подход к проблеме устранения неоднозначностей в речевых сообщениях путем применения семантико-акустического анализа. Предлагается архитектура интеллектуальной системы, в которой, с использованием методов цифровой обработки речевого сигнала, а также формализации и обработки знаний с помощью семантических сетей, осуществляется непосредственный переход от обработки сообщения в речевой форме к анализу смыслового его содержимого (семантико-акустический анализ). На основе инструментов, предоставляемых в рамках технологии OSTIS и фреймворка обработки сигналов GUSLY, реализован прототип интеллектуальной системы для разрешения речевых неоднозначностей определенного типа: омонимов и паронимов. Показаны основные достоинства предлагаемого подхода по сравнению со стандартными системами автоматического распознавания речи, а также возможные пути дальнейшего развития предлагаемого подхода для решения задачи понимания речи.