Semantic analysis of voice messages based on a formalized context

Vadim Zahariev, Timofei Lyahor,
Nastassia Hubarevich, Elias Azarov
Belarussian State University Informatics and Radioelectronics
Minsk, Belarus
zahariev@bsuir.by, linoge@bsuir.by,
stasia@tut.by, azarov@bsuir.by

***Abstract*—The report is devoted to the problem of using formalized contextual information for the semantic analysis of voice messages in conversational systems with a speech interface. The paper proposes an approach based on several fundamental principles: the transition from the acoustic pattern to the semantic representation bypassing a separate stage of textual presentation of information, saving and processing contextual information at all levels in a single knowledge base, transferring the linguistic processing stage to the semantic analysis block (which allows to take into account not only statistical but also semantic links at this level), of applying feedback from the semantic level to the lower level to adjust the result of links work. To implement the approach, the original signal processing technique based on instantaneous harmonic analysis, convolutional neural networks for solving the classification problem, as well as the capabilities of the OSTIS methodology and technology were used.**

***Keywords*—natural language understanding, context formalization, automatic speech recognition, neural networks**

I. Introduction

The ability to conduct a dialogue with the user is one of the key and distinctive features of intelligent systems. This process can be realized effectively only when the dialogue flows in the most natural way – using the verbal channel of communication, i.e. through the speech interface.

The latest achievements in the field of machine learning and artificial intelligence, connected primarily with the development of neural network approaches and methods of formalisation of semantics, made it possible to bring qualitative characteristics of dialogue systems with speech interface to the level of commercial solutions [1]. This fact in it's turn contributed to the rapid spread of this technology on the mass market, primarily in the form of personal voice assistants such as "Alexa" (Amazon), "Siri" (Apple), "MicroSoft", "Alice" (Yandex) [2].

An important component of the dialogue system with a speech interface is a module of recognition and comprehension of speech signal [3], [5]. It allows to distinguish the basic semantic entities in the statement, to define relations between them, and to take into account peculiarities of context. The latter possibility is of particular importance, due to the variety of conditions in which dialogue systems are currently used (indoors and outdoors, in the car, in the office, etc.), which leads to an increase in ambiguity in recognition and, as a result, the errors are caused [4]. The use of contextual information (contained both in the message itself and in external sources – meta-information) allows to increase the accuracy of recognition of [6], [7], [8].

In previous works, the authors addressed to the issues related to the understanding of the speech signal based on the proposed method of semantic-acoustic analysis [9]. The main motivation of the authors is an attempt to confirm or decline the following hypothesis. Since the textual and speech forms of presenting the message are equivalent in terms of the message load, there is a shorter way to go from the speech signal to the semantic presentation ("speech" – "meaning"), than a three-level scheme for translating a speech signal into a representation in the form of spelling text and the further implementation of its semantic processing ("speech" – "text" – "meaning"). Such a transition is carried out with the direct perception of speech by human. According to the author's opinion this approach should give a number of advantages. For example, reduction of errors due to imperfection of linguistic statistical models, the ability to take into account additional information that is present in the speech signal (intonation, pauses, acoustic environment), but not in the spelling text in the process of semantic analysis, and vice versa, at the level of linguistic models to take into account information about the formalized representation of the context placed in the form of ontology in the knowledge base of intelligent system.

In this paper, obtained results and formulated ideas are used to solve a broader problem of understanding speech fragments of a larger volume while the interactive system is used in the mode of transcribing information (interview, lecture, etc.). It is also potentially one of the promising options for such systems use [2]. To solve this problem, we may take into account the predefined context (the topic of the lecture or interview), which will allow us to narrow the number of possible options while understanding specific terms.

II. Problem of context consideration

A comprehensive consideration of the context in the process of dialogue with a user is the key for his statements understanding and interpreting. The main problems of the current work in the subject domain include: identifying the topic of conversation based on the analysis of semantic information [10], using semantic information to reduce recognition errors [11], tracking the state of dialogue by means of a formalized context [12], building contextual models based on speech [13].

However, there are some problems associated with the fact that contextual information in modern systems, in our opinion, is not used in its entirety. The first part of the context (so-called linguistic context) is modeled not at the level of semantic data presentation, but at the level of statistical language models that do not allow fully capturing many relationships, the complexity and diversity of contextual links, unlike semantic models, but only reflect certain distribution of following some words (or parts of words) after others. And only the second part of the context (situational context and meta-information) is described at the level of semantic models.

There are also certain limitations connected with the storage and processing of context information to be implemented [7], [8], [15]. In modern systems, dictionaries (or even ontologies) corresponding to specific topics, are stored separately from each other, meta information from one ontology is not available for the usage in another, their number in each system is limited [15]. Thus, the contextual information is also stored in various isolated parts of the system, databases and knowledge bases, containing ontologies from certain subject areas. In the existing voice assistants, they are called «abilities» or «skills» (picture 1) [2].

The dialogue system tries to determine which one of the all «skills» the user is currently accessing. Then it connects the corresponding ontology and context, for example: «search», «news», «weather», «navigation». Thus, the system becomes task-oriented, designed for a specific subject area, and does not always allow for a «seamless» transition between different application areas.

This causes the following main problems:

- Topics can strongly intersect by concepts. If we exclude this possibility, then it is very difficult to decide where to place a specific concept. On the other hand, if we duplicate concepts in each topic, the percentage of duplications can be very large;
- Selection of a topic is rather arbitrary, which, in turn, makes it necessary to have the following possibilities:
  - it is easy to change the boundaries of specific topics, both with the addition of new concepts, and with the use of existing concepts;
  - it is easy to add new topics, also without adding new concepts, and using the concepts that already exist in other topics;
- Since the selection of topics is rather arbitrary, it is likely that the restriction of the context to only one topic may turn out to be too strong restriction; a person can use terms from different areas even in thematic speech;
- Modern approaches take into account a rather limited context. As a rule, they don't consider meta-information, for working with which a common knowledge base, accessible throughout the system, is necessary.

III. Proposed approach

The previous work of the author [9] reported, that in modern speech interfaces the task of meaning comprehension is most often solved by the «bottom-up» method. Firstly, the recognition of the speech segments of the signal takes place, converting them into text in the linguistic processing module. Then a recognized fragment is transferred to the semantic module, which is implemented separately from the linguistic. It has a knowledge base which is independent of the linguistic module. The information at the input of the linguistic module is represented as a matrix that is made of the recognition probability vectors of each speech flow segment. With successful segmentation it corresponds to a phoneme, allophone, diallophone, etc. Subsequent processing involves building a list of meaningful sentences based on some grammar from these probability vectors. Spontaneous speech, especially in flow and natural surroundings, is often agrammatic. For example, case endings in flexive languages are most often “swallowed”, i.e. do not pronounce clearly [16]. Additionally, in the Russian language there is almost free order of words in sentences. It results in not effective use of only statistical n-gram models for this language [17]. Therefore, it is not enough to use only one grammar, without taking into account the context and semantic means, especially at the linguistic level, as it complicates the process of understanding and introduces additional distortions into it. But in traditional architecture it is not possible to solve this problem due to the different ways of information storing in the linguistic and semantic processing modules (Figure 2 a).

Therefore, the approach, that is proposed in this work, implies primarily consideration of the background, context identification and case-role relationships, as well as using various available meta-information (audio description in tags: genre, author, speakers, recording transcripts), with all this information stored in a single knowledge base. Additionally, the authors offer to use the feedback of the semantic module with the recognition module: the search list for probable words in recognition

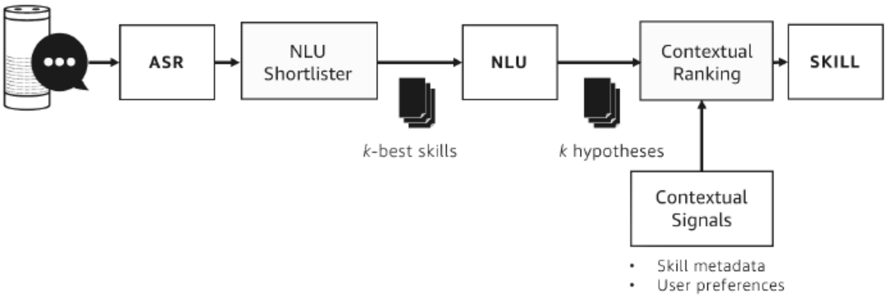


Figure 1. Natural language understanding with «skills» based context formalization [14].

is updated with associative vocabulary with subsequent recalculation of probability vectors. By repeating the cycle it is possible to achive higher percentage of correct understanding of the meaning.

Since the approach proposed by the authors is a relatively new approach to solving the problems considered, especially with regard to the systems of recognition and understanding of the Russian language, it is not possible to translate a comprehensive review of the literature on this subject. Among the existing publications, ideologically close to this work, are works devoted to the construction of direct models for transforming sounds into words [18], end-to-end models of segmentation of speech parameter sequences [19], increasing the accuracy of speech recognition through semantic analysis [20].

As already was mentioned in the previous work, the limitations of the proposed approach are homonymy and so-called information "garbage" (words that are not in the recognition vocabulary), as well as various kinds of interferences, both speech and non-speech. To eliminate various kinds of noise and artifacts while working with a signal in the most efficient way is possible, including the use of more advanced models and methods of signal processing. In particular, in this work it is proposed to use a hybrid model of signal representation and a method for estimating its parameters based on instantaneous harmonic analysis.

The approach for solving the problem of resolving paronyms and homonyms in the context of the voice messages understanding is described in work [9]. However, when analyzing a specific message, it is proposed to use not the entire knowledge base, but some part of the knowledge base corresponding to a specific topic or set of topics. In accordance with the approach to the development of knowledge bases, used in the framework of the OSTIS Technology, the knowledge base is defined by a hierarchical system of subject domains and their relevant ontologies [23]. Thus, the topic corresponds to the subject domain model and the family of ontologies corresponding to this subject domain.

Thanks to the listed components of the technology, it becomes possible not only to solve the problems discussed above, but also to get some additional benefits, namely:

- completely eliminate duplication of information (one of the fundamental principles of the SC-code);
- remove the restriction on the number of possible topics, even for a given set of concepts;
- to be able to specify the degree (believability) of the concept correlation to a particular topic and take this into account when analyzing messages;
- to be able to specify various meta-links between topics, for example, to indicate related topics with an indication of the closeness degree (both qualitative and quantitative). This will allow to intellectualize the process of choosing the most appropriate concepts, i.e. if the contradiction cannot be resolved within the framework of one topic, then system can try to expand the search context by related topics;
- to be able to analyze the correctness of the knowledge base fragments of arbitrary configuration, set complex rules and relationships between objects.

In addition, the SC code allows storing and specifying any external files in the knowledge base. Thus an analyzed file can be specified (for example, the author of the lecture is indicated), as a result, the system can independently choose more or less suitable topics based on the analysis of this specification.

IV. Implementation of the proposed approach

A. General system architecture

The architecture of the system implementing the proposed approaches is presented in the figure 2.

It is easy to see that the standard architecture of the understanding subsystem (2) a), in which the recognition process precedes the understanding stage, includes three types of transformations: signal analysis with selecting basic units of speech flow (phonemes, morphemes), linguistic processing and translation into spelling text, and only then translation into semantic models. Approaches based on statistical models, that allow to take

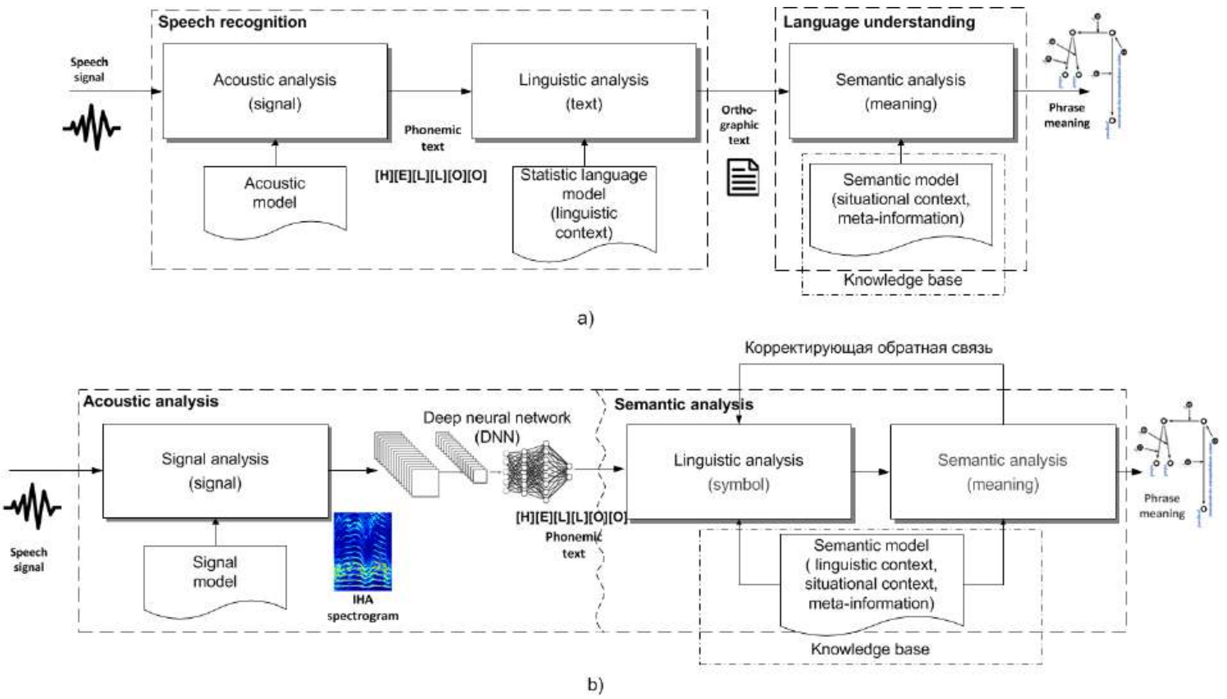

Speech recognition
Speech signal
Acoustic analysis (signal)
Linguistic analysis (text)
Phonemic text
[H][E][L][L][O][O]
Acoustic model
Statistic language model (linguistic context)
Ortho-graphic text
Language understanding
Semantic analysis (meaning)
Semantic model (situational context, meta-information)
Knowledge base
Phrase meaning
a)
Корректирующая обратная связь
Acoustic analysis
Semantic analysis
Deep neural network (DNN)
Signal analysis (signal)
Signal model
IHA spectrogram
[H][E][L][L][O][O]
Phonemic text
Linguistic analysis (symbol)
Semantic analysis (meaning)
Semantic model (linguistic context, situational context, meta-information)
Knowledge base
Speech signal
Phrase meaning
b)

Figure 2. Speech understanding system architecture: a) implementing standart approach; b) implementing proposed approach.

into account only a fraction of the possible links, are used at the stage of the linguistic processor. In the proposed architecture, the use of semantic models instead of statistical ones permits at an early stage (when moving from acoustic models to immediately semantic ones) to carry out a detailed analysis of the context, not based solely on statistical relationships in word sequences.

The system consists of two main parts: modules for acoustic and semantic analysis.

B. Acoustic processing part

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 $\vec{x_m}$ is formed for the current frame, which is placed in a sequence of similar vectors $\vec{X}$,

For speech analysis it is proposed to use a model based on a hybrid representation of speech signal with multiband excitation, which allows the most adequate representation of any fragments of the speech signal of a different nature of sound formation [24]. Voiced and unvoiced fragments of the signal refer to separate parts of the model: periodic (harmonic) and aperiodic (noise).

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

where $s(n)$ – input speech signal, $h(n)$ – harmonic component, $r(n)$ – noise component of the signal, n and N – current sample number and the total duration of the analysis frame in samples, respectively.

The harmonic component can be represented by the following expression:

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

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 $k = 1, ..., K$. It is easy to see that one of the features of the models is the fact that each harmonica is described not by one but with c sinusoidal somponents (multiband excitation).

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 [26]. The apperiodical component is determined, according to the expression (1), as the signal remainder $r(n) = s(n) - \hat{h}(n)$. Model implies the use of signal analysis through synthesis technique and subtraction of the harmonic part $\hat{h}$ from the original signal.

The aperiodic component $r(n)$ in the frequency domain $R(w)$ can be approximated using its spectral envelope and parametrized using the linear spectral frequencies $R_p^{LSF} = LSF(r(n))$, where p is the number of spectral envelope coefficients [27].

The estimation of parameters of the model is proposed to be carried out using the original method of instantaneous harmonic analysis (IHA), which allows to significantly increase the accuracy of the definition of parameters of the periodic component [28]. 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. This allows to obtain a high temporal and frequency resolution of the signal, as well as a clearer spectral picture of the localization of energy at the appropriate frequencies 3, and as a result, to perform a more accurate estimation of the signal parameters (on average above standard methods for 10-15 %) [29].

Consequently, for one frame of signal with the number m and duration N of counts the characteristic vector which includes coefficients of model $\mathbf{x_m} = [G_k, A_k^C, f_k^C, K, C, r_p^{LSF}]$ is formed. And the acoustic image of a signal duration M is a sequence of such characteristic vectors: $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_M})^T$.

In terms of signal processing, this sequence can be represented as a analogue of the spectrogram with an extended number of parameters, where the values of normalized instantaneous amplitude harmonic A_k^C and linear spectral frequencies R_p^{LSF}, characterizing the distribution of energy in the periodical and the aperiodic part of the signal respectively (which equals the area of low and high frequencies in the Fourier spectrogram), supplemented with information about their instantaneous frequency f_k^C, and energy G_k in the band (excitation signal parameters).

In contrast to the previous work, where the fragment of the signal was converted into a phonetic word, and the method of comparison of the acoustic image with the benchmark was used, which was quite applicable to the problem of recognition with a limited dictionary, in this work, to realize the possibilities of working with unlimited dictionary, and fragments of high-length audio recordings, a classic approach is used based on obtaining a sequence of spelling words in sequence characteristic signal parameters vectors. This sequence arrives at the input of a deep neural network to solve the problem of classifying the obtained sequence of parameters and converting it into a sequence of phonemes of units, on the basis of which the trigram model is built a sequence of spelling words. Linguistic modelling was partially carried out using the statistical models of the HMM, and partly in the semantic processing unit. And each of the words will be associated with some node in the semantic network, which will later perform the procedure of linguistic modeling of the statement already taking into account the contextual information available on both the linguistic and semantic levels.

The neural network architecture was chosen on the basis of the structure of the network proposed in the work [30], which is a combination of RCNN and BLSTM networks. This architecture has proved to be effective for solving the problem of recognition of Russian spontaneous speech [31], compared to the approaches based on one type of networks [32]. The architecture is presented in the figure 4.

Features were transformed into tensors of a dimension 40×11 and were sent to RCNN with $T = 3$. Then, there was a unit that was consist of two convolutional layers with a batch-normalization and ReLU, 3×3 kernel with padding and $1 : 1$ stride. Then, convolutional layer with 2×2 and $1 : 1$ stride. Finally, BLSTM's stack (three layers with 512 units in each layer) was applied. Initialization and training of the network is carried out according to the schemes presented in the work [33]. For the input feature vector, the procedure of lowering the dimension based on the principal component analysis to the dimension of the input layer is applied. To initialize the training, limited Boltzmann machines were used. The network was trained using the criterion for minimizing mutual entropy. The implementation of the network was carried out using Kaldi and CNTK software packages in accordance with the methods described in the paper [30]. Two corpuses of speech phonograms were used for model training:

- The first corpus contains about 100 hours of audio recordings received from the video lectures on YouTube by automatic extraction of audio tracks. For the training of models and context accounting also both user-provided and auto-generated subtitles (automatic captions) for the extracted audio were used. The enclosure is characterized by variability and includes recordings containing voices of 80 speakers obtained in different acoustic environments. The thematic area of the video was taken in accordance with the main subject domain considered in the article: lecture materials, reports of conferences from various sections of mathematics including algebra, geometry, graph theory.
- The corpus of training data and test data was made up of available «Voxforge», «SPIIRAS» and «STC» speech corpus fragments [35]. Total duration of audio training set was about 30 hours. The lexicon corresponds to the common form of speech. Since a comparatively small amount of data were available for experiments with this corpus, the system dictionary at the moment was about 1000 words.

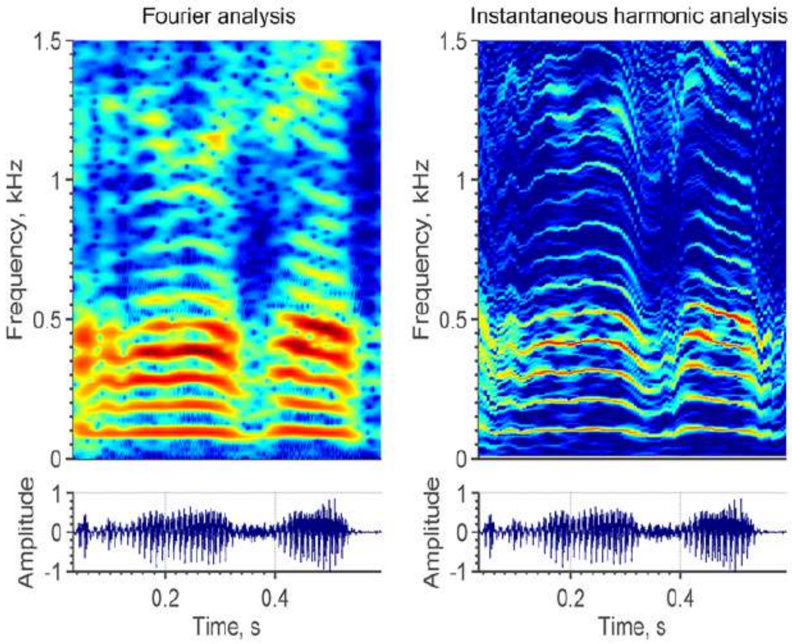


Figure 3. STFT and IHA based spectrograms

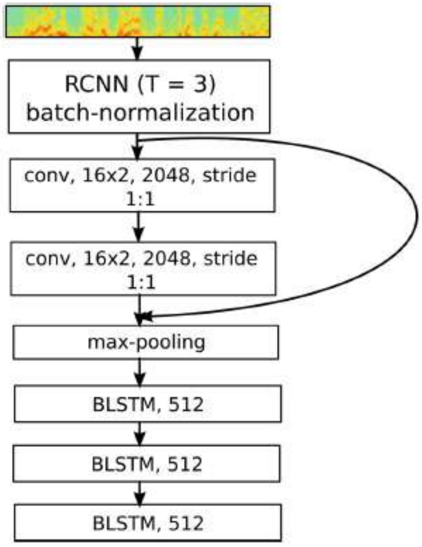


Figure 4. Neural network architecture based on a combination of RCNN and BLSTM networks [30]

The recordings were characterized by a large inter-dictatorial variability, as well as by the diversity of the acoustic environment.

All phonograms in both corpora were recorded at a sampling frequency of 16,000 Hz, 16 bits per sample. To test the system, 500 phonograms were selected from each enclosure containing phrases ranging from 10 seconds to 1 minute. The remaining phonograms were used to train the neural network. The main feature of the training sample preparation process was the fact that their lexical composition was selected in such a way as to maximally reuse the existing knowledge bases made on the base of OSTIS Technology, for example, geometry and graph theory [23], [37]. The lexical composition of the training set corresponded to the concepts and relations available in the knowledge base. For example, the most frequent words in the training sample corresponded to the main nodes of the ontology, containing such concepts as: geometric shape, point, segment, ray, line, plane, polygon, triangle, quadrilateral, etc. As a result, there was no need to carry out the procedure of forming a knowledge base from scratch, but it was possible to supplement it with new concepts.

As a phoneme alphabet was chosen a set of 54 phonemes: 16 vowel phonemes, 36 consonant phonemes, one phoneme for pauses and one for speech noise. This set of phonemes has been successfully used in the development of an automatic subtitle generation system for real-time television shows [34]. For the simulation of vowel sounds, 6 nuclear, 4 postnuclear, 5 prenuclear and 1 preprenuclear phoneme were used. Consonant sounds were modeled using 21 hard and 15 soft phonemes. This separation of vowels and consonants improves the quality of speech signal modeling, since both vowel sounds (stressed and unstressed) and consonant sounds (hard and soft) have noticeable differences in spectral and temporal characteristics.

The integration of the neural and semantic parts is carried out using the approaches presented in [36]. Further processing is carried out in the semantic processing module.

C. Semantic processing part

The semantic processing module, in accordance with the architecture of systems built on the OSTIS technology, includes a knowledge base, which is interpreted as a hierarchical system of subject areas and corresponding ontologies, and also a problem solver, which is interpreted as a hierarchical system of agents, managed by situations and events in the knowledge base.

The main classes of agents that make up the task solver of the semantic processing module were considered in [9].

Let us consider several fragments of the knowledge base of the semantic processing module, illustrating the possibility of solving the problems formulated earlier.

Figure 5 shows an example of correlating several concepts to several subject domains (SD) with an indication of the membership believability to a particular subject domain. As it can be seen from the example, each concept can be part of an arbitrary number of subject domains. At the same time, topics separation can significantly narrow the search area when resolving paronyms and homonyms, for example, the word «graf» in the meaning of a noble title will be considered in the last order for a lecture on discrete mathematics, and the word «graph» in the sense of the mathematical structure will be the last to be considered in the framework of the lecture on history.

Figure 6 shows an example of specifying meta-links between subject domain. The main relations in this case are the relations *particular subject domain* and *related subject domain*. It can also indicate the degree of closeness, which can be taken into account, for example, when expanding the analysis context (a particular subject domain is automatically considered to be related with the maximum degree of closeness).

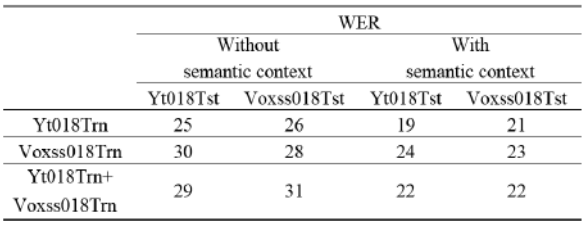
Figure 7 illustrates an example of an audio file specification (lecture recording) with attribution. In turn, the author of the record with an equal degree of confidence corresponds to a set of subject domains to which the lecture will most likely be devoted.

It is important to note that such characteristics as the degree of correspondence of a concept and a subject domain and the degree of relationship between subject domains can be established both by an expert and automatically calculated on the basis of analysis, for example, subject texts corpus. Thus, the task of building knowledge base fragments corresponding to different topics can be significantly simplified.

V. Experimental results

Experiments were conducted to identify the efficiency of proposed approach, i.e. the application of an additional formalized context at the semantic and linguistic levels. The Word Error Rate (WER) metric is used to evaluate perfomance. Since at the moment the performance was evaluated only for the signal and linguistic levels processing modules, this type of metric (widespread for testing ASR systems) was used. It represents a normalized levenshtein distance between two word sequences that is averaged for all samples. The WER is defined as follows $WER = (I + D + S)/N$, where I – is the number of insertions, D –is the number of deletions, S –is the number of substitutions, and N is the number of words in the reference. The speech corpus collected on the basis of the lecture material on YouToube had the name «YtO18Trn» and «YtO18Tst» for training and testing respectively. The speech corpus collected from fragments of cases «Voxforge» was called, «SPIIRAS» and «STC» for training «VoxssO18Trn» and testing «VoxssO18Tst». Two modes of work were considered, taking into account the semantic context in the linguistic module and without accounting. The results of the experiments are presented in Table 1.

Table 1. Experimental results.



	WER			
	Without semantic context		With semantic context	
	Yt018Tst	Voxss018Tst	Yt018Tst	Voxss018Tst
Yt018Trn	25	26	19	21
Voxss018Trn	30	28	24	23
Yt018Trn+ Voxss018Trn	29	31	22	22

The obtained results allow us to assert that the use of formalized context, based on the approach suggested in the article, allows to reduce the word error rate by 5-7% on average, depending on the size and composition of the training sample. To obtain more representative experimental results, it is necessary to expand the training sample of audio recordings and conduct additional experiments. For learning of deep neural networks, shells are used that usually include from 500 to 2000 hours of audio recordings[20], [31]. This fact also explains the relatively low percentage of recognition accuracy and the WER metric value.

VI. Conclusion

An approach to the problem of semantic analysis of voice messages with the use of formalized context is proposed. This approach involves saving and processing of contextual information at all levels in a single knowledge base, transferring a stage linguistic processing into a block of semantic analysis, using joints specific acoustical, statistical and semantic models and methods. This approach allows acheiving deduplication of context information, taking a degree (believability) of the concept correlation to a particular topic and take this into account when analyzing messages, analyzing the correctness of

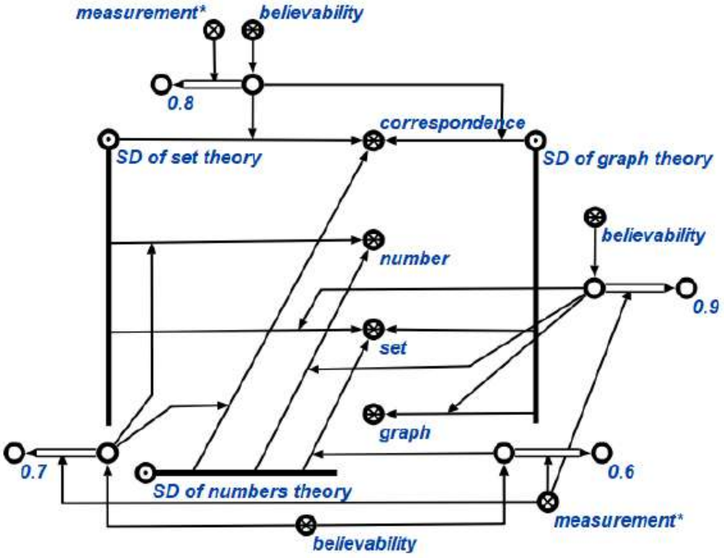


Figure 5. Correspondence of concepts with subject domains

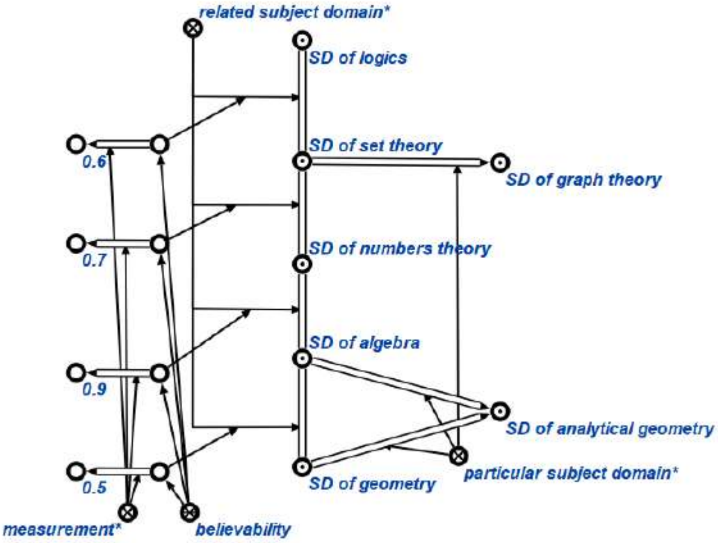


Figure 6. Metalinks between subject domains

the knowledge base fragments of arbitrary configuration, setting complex rules and relationships between objects. The original system architesture developed with the help of signal processing technique based on hybrid speech model and IHA, deep neural network for solving the classification problem, as well as the capabilities of the OSTIS methodology and technology are used. This allows to reduce the word error rate by 5-7% on average. Further work will be aimed to improving the quality characteristics of the proposed approach and testing it on large corpus of speech data.

Acknowledgment

The research presented in this paper was conducted in close collaboration with the Department of Intelligent Information Technologies of Belarusian State University of Informatics and Radioelectronics. Authors would like to thank the research group of the Department of Intelligent Information Technologies for productive cooperation.

References

[1] Lemley, J. Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision / J. Lemley, S. Bazrafkan, P. Corcoran // IEEE Consumer Electronics Magazine. — 2017. ‘ -– Vol. 6. -– No. 2. -– pp. 48-56.

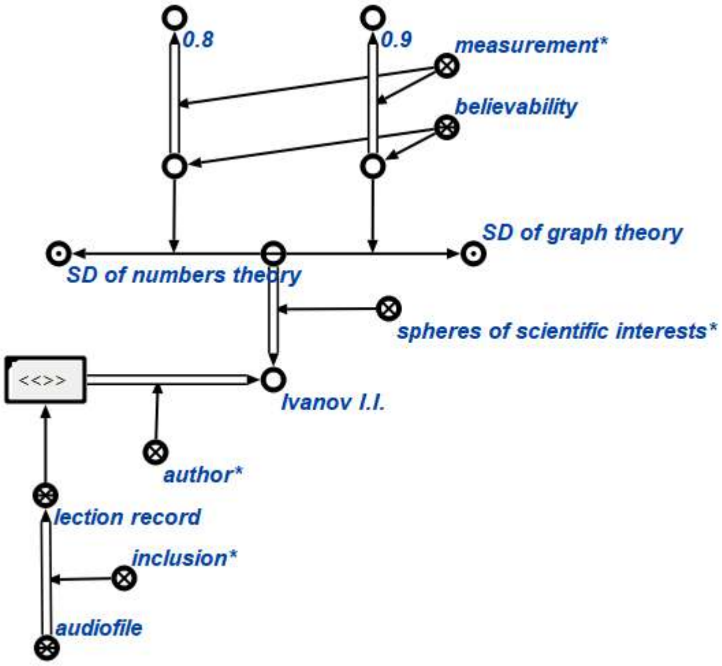


Figure 7. An audio file specification

[2] Hoy M. B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants / M. B. Hoy // Medical reference services quarterly. -– 2018. -– Vol. 37. -– No. 1. -– pp. 81-88.

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

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

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

[6] 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

[7] Using context information to facilitate processing of commands in a virtual assistant. / T. R. Gruber et al. – U.S. Patent No. 9,858,925. – 2018.

[8] Bringing contextual information to google speech recognition / P. Aleksic et al. // Sixteenth Annual Conference of the International Speech Communication Association. -– 2015.

[9] Zahariev, V. A. An approach to speech ambiguities eliminating using semantically-acoustical analysis / V. A. Zahariev, E. S. Azarov, K. V. Rusetski // Open Semantic Technologies for Intelligent Systems (OSTIS-2018). – Minsk: BSUIR, 2018. -– pp. 211 – 222.

[10] Guinaudeau. C. Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations / C.Guinaudeau , G. Gravier , P. Sebillot // Proc. Annual Intl. Speech Communication Association Conference (Interspeech). -– 2010.

[11] ASR error management for improving spoken language understanding / Simonnet E. et al. // arXiv preprint arXiv:1705.09515. -– 2017.

[12] Vodolan M. Hybrid dialog state tracker with asr features / M. Vodolan, R. Kadlec, J. Kleindienst // arXiv preprint arXiv:1702.06336. -– 2017.

[13] Weigelt S. Context Model Acquisition from Spoken Utterance / S. Weigelt, T. Hey, W.F. Tichy // International Journal of Software Engineering and Knowledge Engineering. -– 2017. -– Vol. 27. -– pp. 1439-1453.

[14] Sarikaya, R. The Role of Context in Redefining Human-Computer Interaction [Electronic resourse]. Access mode: https://developer.amazon.com/blogs/alexa/post/3ac41587-f262-4fec-be60-2df2f64b9af9/the-role-of-context-in-redefining-human-computer-interaction. Date of access: 10.12.2018

[15] Sarikaya, R. The technology behind personal digital assistants: An overview of the system architecture and key components. / R. Sarikaya // IEEE Signal Processing Magazine. — (2017) -– 34. 1. – pp. 67-81.

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

[17] Ronzhin A. L. Russian voice interface / A. L. Ronzhin, A. A. Karpov //Pattern Recognition and Image Analysis. -– 2007. -– Vol. 17. -– No. 2. -– pp. 321-336.

[18] Direct acoustics-to-word models for english conversational speech recognition / Audhkhasi K. et al. // arXiv preprint arXiv:1703.07754. -– 2017.

[19] End-to-End Neural Segmental Models for Speech Recognition / Tang H. et al. // arXiv preprint arXiv:1708.00531. – 2017.

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

[21] From training intelligent systems to training their development tools / V. V. Golenkov et. al. // Open Semantic Technologies for Intelligent Systems (OSTIS-2018) – Minsk: BSUIR, 2018. – pp. 81 - 98.

[22] Shunkevich, D. V. Agent-oriented models, method and tools of compatible problem solvers development for intelligent systems / D. V. Shunkevich // Open Semantic Technologies for Intelligent Systems (OSTIS-2018) – Minsk: BSUIR, 2018. – pp. 119 – 132.

[23] Davydenko, I. Semantic models, method and tools of knowledge bases coordinated development based on reusable components / I. Davydenko // Open Semantic Technologies for Intelligent Systems (OSTIS-2018) – Minsk: BSUIR, 2018. – pp. 99 - 118.

[24] Sercov, V.V. An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate

MBE coders / V. V. Sercov, A. A. Petrovsky // Proceedings of the EUROSPEECH 1999. -– 1999. -– pp. 1479–1482.

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

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

[27] 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)

[28] Azarov, I.S., Petrovskii, A.A. Mgnovennyi garmonicheskii analiz: obrabotka zvukovykh i rechevykh signalov v sistemakh mul'timedia [Instant harmonic analysis: processing of sound and speech signals in multimedia systems]. LAP Lambert Academic Publishing, Saarbrucken. – 2011. -– 163 p. (in Russian)

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

[30] Deep neural networks in Russian speech recognition / Markovnikov N. et al. // Conference on Artificial Intelligence and Natural Language. -– 2017. -– pp. 54-67.

[31] Kipyatkova I. DNN-based acoustic modeling for Russian speech recognition using Kaldi / I. Kipyatkova , A. Karpov //International Conference on Speech and Computer. -– Springer, Cham, 2016. -– pp. 246-253.

[32] Kipyatkova, I. S. Raznovidnosti glubokih iskusstvennyh nejronnyh setej dlja sistem raspoznavanija rechi [Deep artificial neural networks for speech recognition systems] / I. S. Kipyatkova, A. A. Karpov // Trudy SPIIRAN. -– 2016. -– Vol. 6. -– No. 49. -– pp. 80-103. (in Russian)

[33] Medennikov I., Prudnikov A. Advances in STC Russian Spontaneous Speech Recognition System / I. Medennikov , A. Prudnikov // Proceedings of SPECOM-2016. – 2016. – pp. 116–123.

[34] Automated closed captioning for Russian live broadcasting / K. Levin, et al. // Proc. Annual Conference of International Speech Communication Association (INTERSPEECH). -— 2014. -— pp. 1438–1442.

[35] Tatarinova A. Building Test Speech Dataset on Russian Language for Spoken Document Retrieval Task / A. Tatarinova, D. Prozorov //2018 IEEE East-West Design & Test Symposium (EWDTS). -– 2018. -– pp. 1-4.

[36] Integration of artificial neural networks and knowledge bases / V. A. Golovko and et al. // Open Semantic Technologies for Intelligent Systems (OSTIS-2018). – Minsk: BSUIR, 2018. -– pp. 133 - 146.

[37] (2018, Dec.) IMS metasystem. [Online]. Available: http://ims.ostis.net/

СЕМАНТИЧЕСКИЙ АНАЛИЗ РЕЧЕВЫХ СООБЩЕНИЙ НА ОСНОВЕ ФОРМАЛИЗОВАННОГО КОНТЕКСТА

Захарьев В.А., Ляхор Т.В., Губаревич А.В., Азаров И.С.

Доклад посвящен проблеме применения формализованной контекстной информации для семантического анализа речевых сообщений в диалоговых системах с речевым интерфейсом. В работе предлагается подход на основе нескольких основополагающих принципов: перехода от акустического образа к семантическому представлению минуя отдельный этап текстового представления информации, сохранения и обработки контекстной информации всех уровней в единой базе знаний, переноса этапа лингвистической обработки в блок семантического анализа (что позволяет учесть не только статистические но и семантические связи уже на данном уровне), применения обратной связи от семантического уровня к нижестоящем для корректировки результатов их работы. Для реализации подхода используются оригинальная техника обработки сигналов на основе мгновенного гармонического анализа, свёрточные нейронные сети для решения задачи распознавания, а также модели, средства и методы технологии OSTIS.

Received 29.12.18