RECOGNITION AND SPEECH SYNTHESIS IN THE DIALOGUE SYSTEMS STRUCTURES

Roman V. Mescheriakov, Vladimir P. Bondarenko, Vladislav P Kotsubinsky

Tomsk, Russia. Tomsk Station University of the Control Systems and Radioelectronics,
Lenin-avenue, 40, phone +7(3822)413426, fax +7(3822)414638
e-mail: mrv@keva.tusur.ru, bvp@online.tomsk.net,
kvp@sapr.fet.tusur.ru

Abstract.

The systems of recognition and speech synthesis, essentially, concern to intellectual systems. They should, at first, understand speech, despite of possible interference's, inaccuracy diction, deviation from normative syntax etc., and, secondly, to provide qualitative speech having high articulation and naturalness of a voice signal. The man integrally uses as language knowledge (phonetics, lexicon, syntax, semantics, prosody etc.), and not language, i.e. knowledge of data domain of the dialogue. Main psychology-acoustic by the characteristic of speech its articulation, i.e. degree of correct perception of sounds, words and sense of speech is. The maximum articulation is characteristic of continued speech perception — phrase articulation. If the man perceives the isolated words, percent of articulation appears less. It is even more reduced at perception of the isolated phonetic units of speech such as syllables. However, the majority of the approaches is based on sequential execution of recognition of words and syntactical analysis. But the experience of the linguists shows, that the man in the beginning seizes sense of a phrase and only then begins purely the task of comprehension of speech, which frequently is considered as the procedure independent of a stage of recognition. This point of view naturally reduces in weak outcomes. From psychology speech perception becomes obvious, that the recognition and comprehension — is two tightly coupled procedures.

2. Basic positions of conceptual model of the speech dialogue systems.

External entry and output data of systems of the speech dialogue are: semantic space of words both phrases of the preset language and data domain; a voice call; and also for the channel of synthesis — parameters speech formation system. Generally accepted to represent the speech system as hierarchy of levels: pragmatic; semantic; syntactical; phonetic; physical. Each level has the data set and rules providing

information processing. Accordingly for reaching the purposes of information processing on top levels at perception of speech the problem solving of lower layers and on the contrary is necessary. Similar arises at speech synthesis — for creation and pronouncing of the expression the problem solving of top levels is necessary. These two processes are so strongly interconnected, that at solution of the direct task the solution return is required. Thus, for a basis of the dialogue system can be multilevel hierarchical model, that assumes:

- the complete description of the language as hierarchy of the tightly integrated events with definition of appropriate transformation rules of the information at appropriate levels;
- the registration of metalanguage knowledge — data domain of the dialogue, i.e. prognosis of development of events, as a feature of speech perception and other external items of information.

In outcome the main positions at construction of model of the dialogue system can be reduced to the following:

- the models of the world of the subjects who are carrying on the dialogue, are intersected, i.e. for them are partially common knowledge of the world defined by concrete data domain of the dialogue [1];
- the subjects always, on a previous history of the dialogue, with the defined probability, predict speech response of the opponent.

Therefore, the effective procedure of perception and speech synthesis should include:

- complete knowledge of the language;
- parameters of the external environment;
- the prognosis of response of the opponent.

These three positions will allow to limit area of valid solutions in each current moment of the dialogue practically at all levels of hierarchy and, as it is necessary to expect to boost reliability of perception of speech.

3. Conceptual model of the dialogue.

Outgoing from the requirements, which are entered in the previous section the simplified model of the dialogue system can be represented by the scheme in a fig. 1, in which by arrows are indicated directions of driving of the information. The driving of the information from below upwards characterizes the channel of perception (recognition), and from above downwards — the synthesis channel. Each selected object is defined by the set of the items of information about the language, rules of conversions and links with other levels. In the scheme the immediate correlation's only of two levels are reflected: higher and low, in a reality exists more correlation's. So, at higher levels the large value has the language knowledge, in particular, items of information on current data domain. Let's mark, that on low levels this knowledge of the language loses the value.

Let's consider a cross-coupling of top levels on lower and on the contrary. It is obvious, that the sense of sentence superimposes limitations on structure of words contained in sentence, and also on their syntactical links. Undoubtedly, as the

syntactical links of words in sentence determine sense of the expression and its goal function. It is known [2], that the characters in the text have various probability of appearance, and also the probability of appearance of the following character is determined with the defined degree prior. However, already at three and more letter combinations the probability is reduced, too most takes place and for words in sentences. I.e. the current prognosis at a level of characters or words allows to lower areas of decision making at recognition.

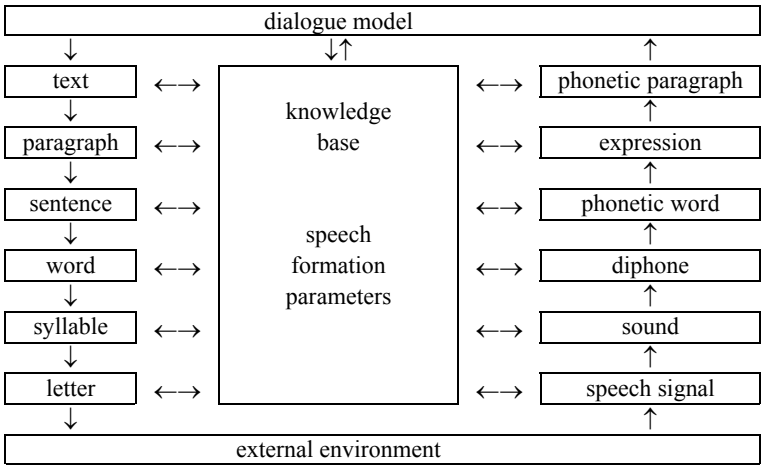

Fig. 1.

Represents the defined interest influence of outcomes of the prognosis of development of the dialogue at a level of model of the world, for example, on lexical base of area of valid solutions at a level of phonetic words. It is known, that it is enough for home dialogue 100-200 most frequently of used words. It is natural, that any other special data domain will require inclusion of the terms and concepts, i.e. new words, that will reduce in the extension of lexical base. It appears, that the common lexical base is enlarged insignificantly. With this purpose on the basis of algorithms morphological and parse considered in operation [3], the rating of lexical base for enough wide data domains was carried out: computer equipment; a radioelectronics. Has appeared, that most frequently met words in these areas enlarge lexical base on 30-50 of %. It is natural, that if the current prognosis of the dialogue is carried on, the area of valid solutions even on this lexical base will considerably is limited, therefore, it should to reduce in rise reliability of recognition at a level of words.

The reduced scheme of the dialogue system allows to lead preliminary classification of speech systems:

- systems of perception of speeches realizing completely the channel of processing from a voice call up to model of the world;
- systems of complete speech synthesis realizing the channel from model of the world up to a voice call;
- systems of recognition of words and phrases;

- systems of synthesis in the printed text etc.

Recognition systems of words and phrases now are to some extent realized, and also synthesis in the printed text, which comparative analysis is reduced in operation [4].

It is possible to mark, that the systems operating only a limited amount of levels, for example, only phoneme have low reliability of recognition. At usage of the information on syllabic recognition the reliability is boosted. Similarly happens at speech synthesis: at immediate conversion of the character in a sound quality and the articulation of a voice signal will be low. At the registration of influence more high levels the quality of the synthesized speech is boosted.

There are correlation's not only levels of one channel, but also channels among themselves. For example, at speech synthesis it is required permanently to be set up under the changed characteristics of articulation of a signal. For this purpose the feedback's with usage of the channel of recognition of a voice call are entered, on the basis of which the synthesis is adjusted [5].

Thus, summarizing all earlier we shall define, that the existing feedback's are determined:

1. by a level of hierarchy in a data reduction system;
2. by a degree of necessity of language knowledge.

The necessity of feedback's can be shown on an example of the speech formation system of the man. The man, making speech adjusts her through some feedback's: at first, through the acoustical system: by air, on bones; secondly, takes into account location of speech formation organs; in third, fixes response of the interlocutor on visual channels. According to the obtained information the man brings in corrective amendments to made speech.

4. Frame of levels of the dialogue.

The dialogue assumes interaction as a minimum of two subsystems, which frames are generally similar also models of the world are intersected. Each of subsystems can be represented as:

$$A_i = \left(A_{ic}, A_{ip}\right); \tag{1}$$

Where Aic, Aip - frame and behavior of a subsystem.
The dialogue system in this case will be represented as association of subsystems (1).

$$A = \bigcup_i A_i = \bigcup_i \left(A_{ic}, A_{in}\right); \tag{2}$$

Each of subsystems during the dialogue will aspire to realize the purpose Aig. It allows to speak about frame and behavior of systems Ai, which generally can not coincide. The implementation of the private purpose Aig assumes appropriate frame and behavior of a subsystem Ai, which can, somewhat be presented as optimal exchange [6]. It is clear, that it requires definition, on the one hand, of functional of quality, with another — of area of valid solutions. It is most difficult to determine a

functional of quality for the dialogue as a whole, however it is possible decomposition on levels of hierarchy, having presented as for hierarchy of qualities, and, therefore, there should be a hierarchy of areas of valid solutions.

The main purpose is determined by a pragmatically component of the dialogue, which on more high level of hierarchy is determined by global sense of the dialogue. At lower levels the global sense represented by the parts (concrete sense).

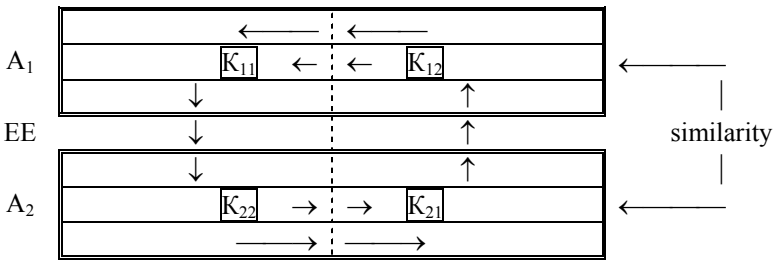

Fig. 2.

In a figure 2 two subsystems represented: A1 and A2, participating in the dialogue through the external environment EE. The levels K11 and K21 are similar, since realize similar functions, such as: creation and message passing. The levels also are subsystems. The levels K12 and K22, realizing functions of the receiver, also are similar. Thus, the functions of the subjects of the dialogue are similar with some degree. It is possible to assume, that at creation of the messages the part of the information on the channel of perception is used. Thus, the message source takes into account possibilities of the receiver of the message. In this case it is possible to tell, that at creation and message transfer the receiver not only tends to reach(achieve) the purpose of an optimum coding, but also reaching of a sufficient level of reliability of recognition, using the items of information, available in his(its) system, on possibilities of the receiver, which some parameters it can use, allocating frame and behavior of the perception.

Let's consider, that the system resources, for example, complexity, are limited. In particular, at generation of the expression for the man the amount of air in mild is limited, articulator organs speed etc. It reduces that it is necessary to signal a message with a sufficient degree of reliability of perception with limitation on operational life's, that allows to speak about optimal exchange in sense Phleishman [6].

In turn, it is necessary to consider subsystems of generation and recognition, as hierarchically constructed according to a fig. 1. That each level of hierarchy of the dialogue system is selected as independent, it is necessary to determine the goal function of its origin and to determine how its optimization will be carried out at the dialogue. It can be considered as reaching of some quality of the system at preset limitations.

Usage of the given approach assumes the registration of hierarchy of qualities of the complex system. The considered type of systems concerns to the class deciding, which should have the following hierarchy [6]:

- O-quality (presence of feedback, automatic system), possessing properties of stability, accuracy etc.;

- R-quality (reliability) possessing property of a structural stability;
- I-quality (noise stability) switching on property of coding and decoding;
- C-quality — controllability (behavior).

Thus the role of qualities varies with body height of a level of hierarchy of the dialogue system. If on a signal level, sound, diphone the main role is played O, R, with I-qualities, at levels of the expression, phonoparagraph the essential value, except for these qualities, gains C-quality.

The dialogue of two subsystems is carried on through the environment by elementary signals. The environment brings in distortions and interference's to these signals. Limitations on complexity of the system as a whole reduces in limitation of complexity of levels. Therefore, in the limited period the level can treat a finite information content. For this purpose it is required to use layering, at which the information content required for the description entry and output signals in unit of time, is reduced at driving on the channel of recognition and is enlarged at driving on a backward channel of synthesis. The considered property is coupled to necessity of representation of the information for the complex system with a different degree abstraction, and this property have all complex hierarchical systems, that is a corollary of limitations on structural complexity of any systems (1). The given property can be considered as special M-quality of complex hierarchical systems.

5. Frame of representation of the information at different levels of hierarchy.

The quality assumes, that the condition of a reversibility should be kept at perception and synthesis of a voice signal, i.e.

$$\delta(x, x^*) \le \delta_0; \quad (3)$$

Where d — an error of the synthesized signal x* under the description of a signal x at absence of its transformations; d0 — a preset valid error.

Any level of hierarchy is characterized by a couple of conversions, which describe direct and return channels of interaction in the given system

$$s: X \to Y;\ s^*: Y^* \times Z \to X^*; \quad (4)$$

Where X and Y — an input and output of the system from below, i.e. for the channel of perception;

$Y^* \times Z$ and X* — an input and output of the system from above, i.e. for the channel of synthesis;

Z — set of transformations assigned to higher levels.

The sizes of the description $\mu(X)$ and $\mu(Y)$ should be in the ratio:

$$\mu(X) > \mu(Y); \quad (5)$$

It is possible, if on set X the equivalence relation, i.e. splitting into classes $X_Y \in X$ is installed. Therefore, the description Y, and is possible and X_Y, should contain as a minimum two components: the class name; listing or order of generation of units *x*,

inhering to the given class X_Y. But if the units x can be somehow generated, they can be stored at an appropriate level of hierarchy as some knowledge base permitting to place(install) correspondence between the class name and his(its) contents. Then Z will characterize valid conversions of a subset X_Y, not quitting from the given class. If set Z homeomorphic to a single segment, the mapping s* becomes a homotopy, that is equivalent to splitting of entry set into homotopic equivalence classes. Most brightly it is tracked on such concepts, as a phoneme and allophone. In view of it, in the description Y it is necessary to include an additional component describing its) states, and also component describing a position X_Y in set X. It reduces in the form of the description of a unit Y (called forming), offered in operation [7].

$$\mathbf{O(Y)=\{ N_Y , P_Y , S_Y\};} \quad (6)$$

Where N_Y — a name;

P_Y — tags;

S_Y — of link X_Y with other equivalence classes of set X.

Representation of entry set as mapping (4), fitting to conditions (3) and (5), at a level of signals, i.e. when the units of set XY represented in a scale of intervals, does not call doubts. In particular rating (3) can be produced by calculation of distance. The task sharply becomes complicated on more high levels, when the procedure of a rating (3) becomes not obvious.

In the considered setting each level is characterized by some variable amount of units (6), depending from a concrete voice intelligence. The unit, as it was marked earlier, can have links with other units of the selected level of hierarchy. On occasion frames of the configuration of units find property of regularity, that allows qualitatively to transfer to more level of hierarchy, accepting for a unit regular frame low of a level. Similarly happens and on a backward channel: the unit is divided on some units of a low level, each of which has the name and tags, and also which the links have among themselves. However, the key information can have the large value at all levels of the dialogue system. An example of such information is the frequency of main hue, which is required at different levels and is saved in tags of units - forming.

Conclusion

The carried out analysis of conceptual model of the dialogue system allows to make outputs, that the construction of the qualitative system speech perception is impossible without implementation of the system of synthesis and on the contrary. Thus it is necessary to take into account such appearances, as the prognosis of development of events. Not only at each level of hierarchy (chain of words, phonemes, characters, words etc.), but also between levels. In particular, the knowledge of data domain of the dialogue allows to limit lexical base at decisionmaking at a level of words. The considered conceptual model allows purposefully to formulate the requirements on recognition and speech synthesis at different levels of hierarchy of the dialogue system.

The analysis of conceptual model shows, that between the system speech perception and speech formation there is a rather deep link: a common knowledge base; tight interaction them at the dialogue. The introduction of concept of M-quality of hierarchical systems allows to use mathematical methods of search of an optimum, since the criteria of a rating of quality have also hierarchical character. Thus not only the range of values is reduced, but also there is a possibility to prognosticate an index point of search of a best value.

Reference

1. Artificial intelligence. - In 3 Book. B2. Models and methods: the quick reference - M.: a wireless and link, 1990. - 304 l: an ooze.
2. Shannon To. Operations on information theory and cybernetics. - M.: In the foreign literature, 1963. - 489c.
3. Belonogov G.G., Novoselov A.P. Automation of processes of accumulation, search and generalization of the information. - M.: Science, 1979. - 257c.
4. Bondarenko V.P., Kotsubinsky V.P., Mescheriakov R.V. Hierarchical structures of recognition and synthesis roar. // the Collection: the intellectual automated systems of designing, handle and training: - Tomsk: NII AEM, 2000. With. 115-125.
5. Bondarenko V.P., Kotsubinsky V.P., Mescheriakov R.V. Synthesis of a voice call in the printed text. //Book: automatic and automated management by complex systems: - Tomsk: TSU, 1998. with 204-217.
6. Phleishman B.S. Units of the theory of potential efficiency of complex systems. M.: the Soviet wireless, 1971. 223c.
7. Grenander Г. The lectures under the theory of images. Synthesis of images. / M.: the world, 1979. 383c
8. Recleitis G., Reidvindran A., Regel To. Optimization in engineering- M.: the world, 1986г. - 350c, ooze.