Integral Robot Technologies and Speech Behavior

Integral Robot Technologies and Speech Behavior

Edited by

Alexander A. Kharlamov
and Maria Pilgun

Cambridge
Scholars
Publishing

Integral Robot Technologies and Speech Behavior

Edited by Alexander A. Kharlamov and Maria Pilgun

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-5436-8
ISBN (13): 978-1-5275-5436-8

TABLE OF CONTENTS

ABSTRACT

The monograph presents papers on the subject domain “Integral robot. Speech behavior”. These cover issues of a theoretical nature, including representation and processing of speech information in the human mind in the process of both text analysis and text generation, and specifically the need to use jointly working linguistic and extralinguistic models of the world, the use of temporal summation of signals along with spatial summation for information processing in natural neural networks. The use of fuzzy representations for decision making is also considered, as well as some issues of the current state of speech analysis and synthesis and related areas: image analysis and speech dialogue. These problems are addressed in terms of using both rules and artificial neural networks of various paradigms. A few relevant (based on neural networks with temporal summation of signals) implementations of these theoretical concepts are presented: software technology for automatic semantic analysis of texts TextAnalyst, which uses the mechanism of temporal summation of signals in text analysis, as well as a microelectronic implementation of associative filters, where temporal summation of signals takes place. Material is also presented on a promising microelectronic implementation of synaptic connections based on memristive structures, which will ensure the creation of an effective neurochip based on neurons with temporary summation of signals in the future. The materials considered provide an extensive review of the integral problem of the implementation of speech behavior not only in terms of existing technical solutions, but also from the perspective of information processing in the human mind, which makes it possible to deepen the understanding of the mechanisms of information processing in the human brain.

PREFACE

Integral robot. Speech behavior

"Integral Robot" and "Integral Robot 2" were the titles of collected translated papers published by Mir Publishing House in Moscow in 1978 and 1982. They presented materials describing the state of the subject domain at that time. They did not generalize the idea of the situation, but rather drew a "sketch" of the state: what tools were available and what could be available in the future.

Now the situation in the subject domain has changed very much. A lot has been done. The subject matter has been highly detailed. But, unlike the 70-80s of the past century, there is no general idea of the direction of such research. There are many small "branches" in the general trend, but there is no idea about the subject domain as a whole.

Speech behavior is purposeful behavior, and therefore is studied within the most important direction of modeling human intellectual activity. The two most important human intellectual functions – understanding meaningful texts and generating meaningful language texts – are the two main functions that distinguish humans from higher primates (with no language education). Therefore, using speech behavior, it is possible to demonstrate both achievements and trends in the subject domain, which (both the former and the latter) characterize its state at the current time, on the one hand, and should be something integral and complete, on the other hand.

The issues of intentional behavior remained outside the scope of the monograph: it does not pose or resolve conflicts of opposition between natural and artificial intellect. Needs and associated behavior, and therefore emotions in speech, are not the subject of consideration. Models of intellectual behavior are considered only from the perspective of representing the mechanisms of information processing. They are not opposed to humans, but are considered in terms of strengthening human intellectual endowments.

Two main intellectual functions are inherent in humans: the generation of meaningful coherent texts (spoken or written) and the understanding of texts generated by others (humans or artificial entities). All other intellectual endowments of the human are connected in one way or another

with these two main functions, or stem from them.

It is quite clear that the speech behavior of humans is not limited to the analysis and synthesis of texts, which are based only on the linguistic model of the world. It is clear that speech behavior implemented by the mechanisms of analysis and synthesis of texts is subject to a mechanism that forms an extralinguistic model of the human world and manipulates it.

The extralinguistic model of the human world is the basis, on which the linguistic model of the world is built, generalizing and simplifying it. The actual language model of the world is only a part of the description of the world in which a person exists, and it omits purely physical details, which in the language are called core knowledge. That is why the currently existing systems that implement speech functions – that is, systems for analyzing and synthesizing texts – are far from perfect: there are no details in the linguistic form presented in the extralinguistic model of the world (and there will be none in the near future).

Therefore, it seems appropriate to talk about the mechanisms that should be implemented in integral speech robots – that is, speech systems that have all the necessary capabilities for the implementation (albeit potentially) of high-quality speech behavior, including, along with the language model and the functions of analysis and synthesis of texts using it, an extralinguistic model of the world, specifically, its analyzing and synthesizing parts.

This is all the more necessary, since it is of vital importance for understanding the structure of the actual language model. Modern ideas of Russian linguists (Russian, as an inflectional language, is less amenable to formalization, for example, compared to English as a non-inflectional language) show that the formalisms of the syntactic level do not exhaust the linguistic details of clustering a set of identical morphological phenomena into separate subsets that differ greatly in their semantic content. In the Moscow school, a separate independent level of deep syntax was introduced for this, which many are now trying to exclude and replace with the combination of the syntactic and semantic levels, without suspecting (or not willing to realize and recognize and trying to solve everything within the language framework) that the level of deep syntax lies not in the linguistic, but in the extralinguistic plane: it represents the physical limitations of the real world, not described (or described with great difficulty) in the language model. Therefore, it must be recognized that semantics is not constructions that describe the physics of the real world, but only restrictions on the pairwise compatibility of phenomena (objects and events) of this world, and in the language it is reflected in restrictions on the pairwise compatibility of word-concepts that describe

these restrictions on the compatibility of the world phenomena.

However, the purpose of this publication is not an attempt to clarify the formalism of the linguistic description of the real world at the syntactic and semantic levels of the language model of the world, but to demonstrate the structure and mechanisms of linguistic behavior, which include extralinguistic representations along with purely linguistic ones: to show that only the totality of these representations (linguistic, based on multimodal ones) provides a full-fledged purposeful behavior, which indeed speech behavior is. An example of such behavior is dialog behavior.

One of the directions covered in the publication is demonstration of not only individual mechanisms for representing extralinguistic facts about the world a person lives in, using the example of video information analysis as one of the extralinguistic modalities that form the first human signal system (another such modality, the somatosensory one, can be modeled using, for example, 3D models), but an attempt to represent them in interaction, including interaction with the language model, in the formation of the second signal system.

Along with purely linguistic mechanisms (analysis and synthesis of texts, both spoken and written), the book also presents mechanisms for analyzing information of other modalities, including visual and somatosensory ones (the latter is described in a somewhat unusual interpretation in the form of 3D models), as well as mechanisms that enable combinations of the above-described representations in dialogue behavior and such a three-modal system for analyzing the visually perceived world.

In turn, the extralinguistic model of the world is formed by the neural mechanisms of the human brain, which operate using the same principles as in the formation of the language model of the world (contra credere: the language model is based on mechanisms similar to the mechanisms of the multimodal model formation): in both the language and extralinguistic models, through structural analysis, hierarchies of dictionaries of images of objects and events are formed, where the words of the dictionaries of the next level are “grammars” of the words of the previous level, since both text and other quasi-text sensory sequences (for example, video sequences) are sequences of repeating words (quasi-words) of various levels of the language (quasi-language). In the linguistic modality, these are speech sounds (or language symbols in written texts), morphemes, words, syntactic groups, pairwise word compatibility. In the visual modality, these are the simplest images (point, line segments, arc segments, various kinds of intersections), fragments of images of objects,

objects themseleves, scenes.

The extralinguistic model, in addition to the main visual information, includes information about the physics of the world – that is, somatosensory information: information, the content of which reflects the reaction of the human body to physical phenomena: manipulating objects, a person's movements in space through walking and changing the position of the body and its individual limbs.

Therefore, the mechanisms for controlling the speech articulation are part of the mechanisms of the language model of the human world, on the one hand, and on the other hand, they are part of the extralinguistic model (that is, to be precise, the total linguistic-extralinguistic model of the world), the formation of which is part of a more general process: the formation of the world model, which includes, along with the linguistic part, also the extralinguistic part. Therefore, it is possible to consider the mentioned mechanisms only in the context of the processes of formation of the entire model of the world.

The model of the human world is not formed in parts (first its extralinguistic part, and then the linguistic part), but as a whole, where the linguistic part relies on its extralinguistic part that was formed a little earlier and continues its own formation.

The formation of the world model is based on two main principles (one is physiological, the other is psychological): (1) the neuroelectrophysiology of the human brain implements the structural associative principle of the formation of the world model (both its extralinguistic and linguistic parts); (2) human psychology implements the mechanism of imitative behavior as the basis of motivation (goal setting) during its formation.

In some sense, the process of generating texts (and quasi-texts) is symmetrical to the process of text analysis. However, it is broader in the sense that text generation is an active process (as opposed to analysis), which is therefore purposeful behavior.

Here, it must be borne in mind that articulation itself, as the control of articulatory organs, is the lowest part of the synthesizing part of the language model. It is this lower part that is more associated with imitative human behavior. Actually, the synthesis of meaningful spoken texts results from the work of the entire model of the human world, which goes through a number of stages; the consideration of these makes it possible to understand the criticality of certain stages of the formation of the world model for the formation of speech behavior.

The stages of synthesizing a spoken meaningful text differ from the similar process for a written one only in the lower level: in the case of generating a spoken text, the process is controlled by articulatory organs,

and in the case of generating a written text, the process is controlled by the hand of the dominant arm.

The stages of text generation are as follows: (1) formation in the anterior cortex of a plan of speech purposeful behavior as a chain of sentence names describing a chain of situations from the original to the target one, represented by their names, which (purposeful behavior) is realized in the anterior cortex by a fuzzy search through the available fragments of the chains formed previously; (2) implementation of the formed plan by unfolding a chain of names into a chain of syntagmas corresponding to these names of situations stored in the dominant hippocampus; (3) fuzzy linkage of linguistic representations of various levels (from semantic to morphological) with mutual concord of elements of various levels, satisfying (concord) the norms of the spoken (or standard) language; (4) formation of sequences that control articulation proper in the motor cortex; (5) projection of the identified lexical elements onto the extralinguistic part of the world model (taking into account the fuzzy processes of the language concords of various language levels: from morphological to semantic one); (6) the reverse process (perhaps iterative) of pronouncing inwardly the formed motor representation of the text with its projection onto the analyzing part of the language model, in order to assess the accuracy of the description of the represented situation by the generated text; (7) assessment of discrepancies in the text obtained in the process of synthesis from the generated plan, with the content of a fragment of the extralinguistic model of the world that gave rise to this plan; (8) correction of discrepancies and formation of a refined plan with the passage of all the above (1-5) stages; and finally, (9) involvement (possibly at earlier stages) of the recipient's model to refine the formed plan from the perspective of its possible perception by the recipient.

The formation of a plan for speech purposeful behavior, in turn, includes the following processes: (1) identification of a fragment of the extralinguistic model of the world (semantic network), which is determined by the objective of speech behavior (the objective is introduced into the system from the outside: the psychology of filling needs does not affect the speech mechanism); (2) breaking it down (via situation templates stored in the hippocampus) into separate situations; (3) fuzzy linkage of these situations into a chain (or several parallel chains), which allows one to move from the initial situation on the fragment of the extralinguistic model of the world, intended for its description by text, to the target situation (various ways of description are possible: along the time axis, against the time axis, mixed, parallel); (4) naming these situations by their names; (5) finding situations suitable for these names in

the dominant hippocampus (during text synthesis). Or vice versa: (4) finding situations suitable for names in the dominant hippocampus (during text analysis); (5) naming these situations by their names.

At the moment, only some mechanisms of purposeful speech behavior are implemented technically. Therefore, the monograph attempts to describe some technologies that could be involved in speech behavior. First of all, these are technologies for automatic speech recognition (using the inflectional Russian language as an example, the complexity of the speech recognition mechanism is shown), text-to-speech synthesis, automatic text analysis, automatic image and video recognition. The theory and practice of dialogue implementation are presented. And, finally, an attempt is made to describe the implementation of a three-modal system, where, along with the language model, a visual model functions, supplemented by a 2.5D model, which represents an equivalent (though a primitive one) of the somatosensory system.

In this regard, the following sections are presented in the monograph. Chapter 1 "Speech as Purposeful Behavior" considers issues of organizing the processing of specific information of various levels and of various modalities, including issues of the signal level of processing, that is, information processing in the periphery of the auditory and visual analyzers. Also issues of the symbolic level of processing are considered: structural processing in the columns of the cerebral cortex with the formation of representations up to the semantic level inclusive in the sensory cortex (both linguistic and multimodal non-linguistic), with the formation of models (templates) of situations in the hippocampus. The representations of information in the anterior cortex are considered, including the formation of a model of the world in terms of purposeful behavior as a set of elements of chains of situations of various levels of complexity. Some issues of the formation of the first and second signal systems are considered: that is, the extralinguistic model of the world and the language model of the world. On the basis of the described mechanisms of information representation in the sensory cortex and hippocampus, issues of the formation of speech purposeful behavior are considered using the example of inner speech.

Chapter 2 "Intelligent Analysis and Scenario Modeling of Problem Situations Based on a Combination of Neural Network Technology for Text Processing, Dynamic Clustering Methods and Fuzzy Cognitive Analysis" discusses issues of fuzzy choice of optimal behavior strategies using the analysis of the Covid situation based on texts from social networks provided by A. N. Raskhodchikov, head of the *Moscow Center of Urban Studies "City"*, for three epochs of observations in the initial

period of the COVID-19 pandemic, characterized by high dynamics of change in the subject domain. The corpora of texts processed using the TextAnalyst technology for automatic semantic analysis of texts, with the main concepts identified using neural network analysis that characterize the mentioned Covid situation in a particular context, were the basis for the subsequent fuzzy cluster analysis of pairs of mutually interacting entities described by these concepts that characterize the arrangement of the process participants, and their dynamics makes it possible to form chains of situations (and whole scenarios), which are plans for purposeful behavior; the most effective of them are to be implemented.

Chapter 3 “Text and Quasi-Text Analysis. Upper Levels of Speech Information Processing” presents the mechanisms of structural information processing in the columns of the cerebral cortex on the example of text information processing. This means that the chapter does not contain information about signal-level processing: only mechanisms for detecting the structure of the language (which are also suitable for processing quasi-linguistic quasi-texts). A large volume of language units implies a multilevel structure of the language in order to eliminate (or reduce) the influence of noise and interference. That is, the multilevel structure of the language allows, by introducing several levels of the code space structure, to reduce the dimension of the code space at each level by using the units of the next level as a context for the units of the previous level. In addition to describing the structural processing algorithm itself, the chapter describes a technology for automatic semantic analysis of texts, including examples of its application in various subject domains from linguistics to genetics (represented by genetic quasi-texts, that is, signaling networks).

Since both analysis and synthesis of texts rely on an extralinguistic representation of the world we live in, Chapter 4 “Computer Vision” presents the available technical solutions that implement the mechanisms of extralinguistic information processing of various levels in terms of only one visual modality. The representation is rather simple, taking into account the development of this subject domain at the present time, but sufficient to determine its boundaries.

At present, solutions in the field of automatic speech recognition have reached a high point of development. In this subject domain, as in many others, rule-based solutions and solutions based on artificial neural networks compete. Both of them are very difficult, since the focus in speech recognition is still mainly on the signal level of information processing. Chapter 5 “On the Issue of Continuous Speech Recognition” presents a comprehensive solution to the problem of recognizing sounding

Russian-language speech, based both on the rules and (where this is the optimal solution) on the apparatus of artificial neural networks.

Speech recognition algorithms using artificial neural networks (see Chapter 6 “Neural Network Language Models for Speech Recognition Systems”) are presented with the transition from a separate description of pronunciation and language models to a description of an integral artificial neural network that implements both mechanisms in one. Rule-based speech recognition algorithms turned out to be not free from the use of artificial neural networks (see Chapter 5), but within reasonable limits: at each level of language representation (and sometimes several at a level), a specific type of artificial neural networks was formed, working effectively precisely within the system of rules of a given level. Perhaps such a variety of approaches turned out to be effective only for the analysis of texts in Russian as an inflectional language.

Improving the performance of speech analysis and synthesis systems is associated not only with the inclusion of an extralinguistic model of the world, but also with the involvement of other (besides visual) modalities, primarily the somatosensory modality. Direct modeling of information processing of the somatosensory modality is a very difficult task. But there is a workaround: the somatosensory modality includes physical representations of the world into the extralinguistic model of the world. At the moment, the so-called 3D models have appeared, which form ideas about the world in terms of just physics: a track of the agent movement, impassability of obstacles (for example, walls), smoothness of trajectories (assembled from various parts, even with gaps). Chapter 7 (Three-Modal System for The Formation of a Textual Description of the Visual World) presents the architecture for involving a 2.5D model for a video analysis task as a filter of important events at the analysis stage before describing this video representation with the texts of unfolding scenarios; it also presents the very procedure for this very description and subsequent analysis of the texts thus represented. Re-describing the video with text leads to a sharp decrease in the variability of the analyzed information, and therefore, to a jump-start improvement in the quality of the analysis results.

The second main intellectual function of human is the generation of a coherent meaningful text. Text generation is a task that is divided into two non-overlapping areas: text-to-speech synthesis and synthesis of a coherent meaningful text. Unlike text analysis, which (solely for the purpose of removing noise- and interference-induced uncertainties at the input of spoken text) is necessary in analysis systems, text-to-speech synthesis is a relatively simple procedure, although the technical details of

this process are quite numerous. However, quite a good picture of the subject domain (linguists have been studying the subject for one and a half hundred years) made it possible to solve the problem of text-to-speech synthesis with a fairly good quality. This subject is covered in Chapter 8 ("Rule-Based Multi-Wave Speech Synthesis"). As for the generation of a coherent meaningful text, which is the result of purposeful behavior, this task will not be solved as long as intentional behavior (that is, needs) is taken off the table. We do not consider this issue, since the introduction of needs into an artificial system gives rise to contradictions between artificial and natural intelligence, which should never be done. However, if the objective is introduced into the dialogue from the outside by a person, then a coherent meaningful text may well be generated without any detriment to the latter.

To close the issue of text-to-speech synthesis, Chapter 9 ("Neural Networks for Speech Synthesis of Voice Assistants and Singing Machines") presents survey on an alternative to rule-based synthesis approach: text-to-speech synthesis based on artificial neural networks. In contrast to the rule-based approach, where everything is obvious, neural networks require huge amount of data to train and depends on such data. Architectures of neural networks for speech synthesis quite broad, and complex, so Chapter 9 shows description of building blocks for such networks. Also, ideas to synthesis a singing voice are presented.

Finally, chapters 10 and 11 are the last chapters dealing with theoretical issues of speech analysis-synthesis. Chapter 10 (Processing in a Dialogue System) represents formally and accurately the two branches of information processing both in analysis and in synthesis and describes the attention mechanism, which allows either to stay focused on a certain topic or to move on to another. However, since both the intentional part of the behavior and the extralinguistic part of the world model are absent in this (purely linguistic) mechanism, it is not possible to implement dialogue behavior as purposeful in good quality. We have tabooed the introduction of the former, and the latter within the system under consideration will not be implemented soon, according to the corresponding assessments regarding the implementation of speech purposeful behavior.

These assessments do not apply to private dialogue systems, the goal of which is explicitly embedded by the structure of the dialogue. Such systems, for example, include systems of communication between clients of a bank and its administration in order to meet the needs of the former or the latter. An approach to implementing such systems is presented in Chapter 11 ("Transformation of Dialogue Communications: from a Person

to a Virtual Assistant"). This chapter is a transition from theory to implementations, and in addition to theoretical considerations for the implementation of dialogue systems, it also describes a specific approach, where a dialogue is represented as a tree of transitions, each of which is determined by the specific needs of its participants. In the event that the client goes beyond the prescribed dialog behavior, the services for him/her are further provided by a human expert.

Finally, the last two chapters (12 and 13) deal with the implementation of the theoretical provisions presented in the collective monograph, in the form of specific microelectronic devices. Unlike the TextAnalyst software technology for automatic semantic analysis of texts, the description of which is an integral part of the theoretical section of the monograph, since the capabilities of the software technology were comprehended as experience in analyzing texts with its help was gained, microelectronic implementations were developed by independent developers. Chapter 12 (Design of Matched Filters for Equipment of Navigation, Communication and Radars. Task Decomposition of Correlation and Convolution Computation in ASICs) was written on the basis of ideas similar to those presented in Chapters 1 and 3, which arose within the same team in the 80s of the past century. The direction of development of these ideas, presented in Chapter 12, was driven by the conditions of their micro-integral implementation in relation to a certain class of information, for the implementation of which the devices were intended. The title of this chapter is "Design of Matched Filters for Equipment of Navigation, Communication and Radars. Task Decomposition of Correlation and Convolution Computation in ASICs". Briefly, I would title Chapter 12 as follows: "Micro-Integrated Associative Filters". It describes one of the possible directions for the implementation of micro-integrated artificial neural networks based on neuron-like elements with temporal summation of signals. Other approaches to their implementation are also possible, depending on the task under consideration.

And, finally, Chapter 13 ("Neuromorphic Systems and Neurointerfaces Based on Memristive Devices") presents another approach to the implementation of such artificial neural networks (the idea of which is based on dynamic associative memory, DAMs): an approach based on the use of so-called memristors as a way to implement analog memory, in contrast to digital, which requires much more hardware resources for its implementation.

Thus, speech behavior based on the mechanisms of structural analysis of the cortex and associative combination of images of objects and events in space and time of the hippocampus is a convenient material for studying

purposeful behavior. At the same time, the analysis of texts (spoken and written), as well as the synthesis of coherent meaningful texts, are the two main intellectual functions of humans, the study of which provides valuable insights into the essence of other information processes in the human mind and, therefore, consider them also as intellectual.

Thus, the flows of sensory information of modalities other than textual (code sequences obtained as a result of processing sensory information of various modalities by sensory organs) become easy to interpret if they are considered as quasi-texts: that is, sequences of level-forming elements of levels of quasi-languages, for example, sequences of video frames of visual modalities.

The use of a linguistic representation of the world simultaneously with a multimodal one, among other things, simplifies the identification of structural patterns of a highly variable multimodal representation by projecting it onto an isomorphic linguistic representation.

These sensory flows presented at the upper levels of the processes of structural analysis in the form of semantic networks provide an opportunity for another level: the representation of the analyzed world in the form of restrictions on the compatibility of their elements, which, by limiting the combinatorics of the represented objects and events, makes an infinite increase in processor power unnecessary, but requires associativity of information addressing: the transition from von Neumann calculators with the AND-OR-NOT basis to neural network structures with the association-correlation-extremation basis, which successfully solve the problem of finding an effective solution through purposeful behavior.

The search for an effective solution in the human mind is performed through fuzzy clustering of a set of pairs of mutually interacting objects: fragments of chains of situations that form plans of purposeful behavior, which indeed human behavior is – that is, speech behavior in particular. An approach to intellectual analysis, predictive analytics and scenario modeling of the development of problem situations is proposed, which is based on a combination of neural network text processing technology, dynamic clustering methods and fuzzy cognitive analysis.

In this case, the neural network technology is used for a preliminary analysis of the subject domain, which consists in processing heterogeneous textual information, analyzing, identifying and evaluating the significance of the subject domain concepts that characterize the analyzed problem situations.

Problem situations are structurally determined by stable clusters of interrelated and interdependent concepts of the subject domain, the representation of the relationships and interdependencies between which,

in the form of consistent fuzzy binary relations of mutual interaction, provides the possibility of using fuzzy causal algebra methods to substantiate the corresponding clustering indicators of these concepts.

From a set of indicators based on consistent fuzzy mutual interaction relations between the concepts of the problem domain, the following indicators were reasonably chosen for clustering the concepts of the problem domain: influence of the concept on the subject domain, influence of the subject domain on the concept, consistent influence of the concept on the subject domain.

A method for clustering subject domain concepts using indicators based on fuzzy binary relations of mutual interaction between concepts is presented.

In the process of "evolution" of the subject domain from one epoch to another, the observations are changed by the relations and interdependencies between the identified concepts. To accomplish this task, a method is proposed for monitoring the dynamics of changes in the cluster structure of the subject domain and its stability, based on the use of dynamic clustering methods for various epochs of observation in order to analyze and identify the following events: transition of concepts from one cluster to another; drift of cluster centers; disappearance and emergence of new clusters; merging and separation of clusters; change in the stability of clusters of concepts and the cluster structure of the subject domain as a whole.

For each identified cluster of concepts of the subject domain (identified problem situation) in each epoch, its own "personalized" model (for example, a fuzzy cognitive model) can be built that properly reflects the relations and interdependencies between the corresponding concepts of the subject domain to ensure predictive analytics and scenario modeling of problem situations.

An example of the implementation of the proposed approach is considered for the analysis of the subject domain with the conditional name "Covid situation" in the initial period of the COVID-19 pandemic, characterized by high dynamics of change in the problem domain.

The problems of automatic speech recognition in their modern statement are mainly related to relevant signal processing, while text analysis almost completely solves the problems of representing information of upper language levels, from morphology to semantics. Here, both the form of representation of language knowledge (morphology and syntax) and their content (vocabulary and semantics) play an important role. The formal side of the language is very important for its standard templates, which in their consistent use at various levels

correspond to the standards of the language. And the content side, in the form of lexical and grammatical semantics, fills these templates with specific meaning, addressing to the non-linguistic part of the world model, which represents this world in the human mind.

The language model of the world is filled with content by projecting onto the extralinguistic part of the world model: words into images of objects and events reflected in the human mind, pairwise compatibility of words of the semantic level into pairwise compatibility of images of objects and events, that is, scenes also presented in the human mind.

The world model of a particular person, represented by linguistic and extralinguistic components, is a tool for manipulating the world (its description and impact algorithms), which ultimately makes it possible to form purposeful behavior as a prediction of future events that are useful to this person.

The extralinguistic part of the world model (due to the identity of the tools for processing information of various modalities) is, like the linguistic part, a hierarchy of dictionaries of event images of various levels and complexity of various modalities, combined at the upper (semantic) level into a semantic network (semantic networks of individual modalities, combined into a single representation), which is combined with the semantic network of the language model into a single whole.

Unlike semantics represented by dictionaries of words and pairwise compatibility of words (events and objects, and their pairwise compatibility), pragmatics, not being a language structure, is built on top of this combined semantic network as chains of situations that include the combinatorics of objects and events (and the words that describe them) in space and time.

The semantic network of the text (representation of the language part of the model) characterizes the compatibility of words in this text, forming a semantic portrait of the text. Semantic portraits of texts can be used to compare texts by meaning, to classify texts (assign a text to a subject domain by comparing its semantic network with the semantic networks of subject domains described by text corpora), to cluster texts into subject domains.

The use of intelligent speech systems is unvaryingly associated with other channels of interaction between intelligent systems and a human user; it requires a low-level development of multimodal interfaces and ensuring a high-quality level of sensory information processing, including that received from various vision systems.

The availability of a full-featured vision system will allow the robotic system or complex to evaluate the characteristics of the external

environment, determine obstacles, own coordinates and coordinates of the robot elements, and use these data as a payload for monitoring and control.

The vision system is a complex software and hardware system that includes a matrix for receiving the light flux, ensuring its correction at various wavelengths and spatial correction of distortion and other deformations, as well as additional active systems. When processing information, it is necessary to take into account both the work of "analytical" models and work on training data, including artificial intelligence technologies as well.

The presence of such a vision system in the robot control loop allows not only to monitor and control the state of the agent and the environment, but also to actively select objects of interest and form commands for the agent and its operating mechanisms.

Analysis of the main problems of the speech control channel (speech recognition system) led to the following conclusions.

The "bottleneck" of existing recognition systems for Russian continuous speech is ensuring stability in relation to acoustic variability and speaker change. In addition, public systems for automatic transcription generation do not take into account all the features of the Russian language.

At the moment, there is no sufficient number of annotated Russian-language speech and normalized text corpora that are in the public domain (as compared to similar corpora in English and Chinese).

The development of new methods and models for systems of speaker-independent recognition of continuous Russian speech, that account for the features of the Russian language and adapt to any subject domain while ensuring the invariance of acoustic models to the speaker change and acoustic environment, will improve the accuracy of speech recognition.

One of the most promising trends in the formation of an acoustic model is neural network parameterization based on multilingual corpora belonging to the same language group. Another promising trend is data-driven fine-tuning of an acoustic model based on neural network parameterization using the reinforcement learning approach, i.e. monitored unsupervised learning.

The creation and further development of neural network models in speech recognition have entailed a qualitative improvement in the characteristics of language models: the perplexity value for the main test data corpora decreased several times compared to n-gram statistical representation, up to values that recently were deemed unlikely and far from potentially achievable.

An important advantage of neural network models over statistical ones is the use of continuous vector representations of words with similarity measures that correlate with their semantic and syntactic features. Another important advantage is the use of longer (in time) contexts. The quality of modern neural network language models significantly depends on the size and properties of training text corpora.

For language models based on simple neural networks with relatively small training data corpora, interpolated statistical n-gram and neural network language models had minimal perplexity, while modern results on very large text data corpora for models with LSTM/GRU cells or convolutional attention-based networks can suggest that there is no further need to use statistical models.

However, it cannot be argued that neural network models have become the basis of modern language modeling in recognition systems. In practical, production speech recognition systems, statistical n-gram models are still widely used. One of the main reasons for this is the increased requirements for computing resources and the size of data corpora that neural network models impose.

Most of developers' efforts are aimed at creating models with significantly lower requirements for memory and computing power. The main lines of research for solving this problem and the results obtained in the form of suboptimal solutions (for example, hierarchical Softmax and its modifications, importance sampling, Noise Contrastive Estimation, the use of subword-based dictionaries, etc.) still work worse than solutions based on Softmax calculation for all dictionary words.

Nevertheless, neural network language models based on reference text corpora have quality indicators (perplexity and entropy) several times lower than statistical models. It can be expected that this will be accompanied by a symmetrical increase in the accuracy of speech recognition. Therefore, one of the most promising areas for improving speech recognition systems in the near future will be the development and use of neural network language models.

An integrated system designed for video analysis of the behavior of agents on the scene and its re-description in the form of a textual scenario for subsequent analysis, when analyzing real scenarios, forms rather complex representations that turn out to be noisy due to a large amount of information not related to the analyzed process; this makes it difficult to work with this representation, and, most importantly, makes it impossible to use automatic procedures for analyzing the generated representations. The introduction of somatosensory modality as a filter into the system makes it possible to remove unnecessary information, which is eliminated

from subsequent analysis by this filter. The scenario of the agent behavior obtained in the form of a text in this case is quite easy to read by the user and, most importantly, can be subjected to the now automatic analysis procedure for subsequent operations with it.

The system presented is implemented in order to combine the processing of information of various modalities (including linguistic and extralinguistic), which makes it possible to increase the reliability of identifying some patterns in the behavior of complex objects (such as people, for example) in complex conditions of real production processes, which makes it possible to automate the procedure for taking into account the behavior of agents in terms of relatively simple parameters (execution time, for example), and thereby intellectualize the process of interaction between the system and the operator.

Speech synthesis is the second unique intellectual function inherent in human. The solution of the lower levels of the text-to-speech synthesis task is a task fairly well-developed theoretically and implemented practically. The task of synthesizing a meaningful connected text has not been solved as yet, no matter how much the developers of artificial neural networks (like GPT) would like to think the opposite, since there is still no solution to the problem of forming a purposeful sequence of situations that solves the problem of describing a specific type of behavior: description or algorithm.

The development of the program model of rule-based multi-wave speech synthesis described in the monograph dates back to the beginning of the 1st decade of the 21st century. Employees of the Laboratory of Speech Recognition and Synthesis of the Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus took part in its creation. The developed software tools made it possible to customize the sound of synthesized speech. The statement of the problem of speech cloning was previously presented in the work.

The achieved level of intelligibility and quality of the synthesized speech were quite satisfactory in comparison with the speech of other well-known synthesizers of that time.

As in the task of speech analysis, the text-to-speech synthesis task involves, along with rule-based approaches, approaches based on artificial neural networks. Researchers working on applications of neural networks for voice synthesis are considering various approaches often combined together. These are architectures based on recurrent networks and the attention mechanism, and generative networks and approaches based on the diffusion process as well, and many others Representations both as a sound wave and a mel-spectrogram are used. The quality of the

synthesized voice has almost reached the level of human perception. Synthesizers of singing voices appeared, also with a fairly high sound quality. However, it should be noted that such high results were obtained for a small set of languages (English and Chinese), for other languages the success is much more modest. This is mainly due to the insufficient volume of the training data sets and the computational resources required for training the models.

Despite the fact that the problem of synthesizing a coherent meaningful text is still not solved, the implementation of dialogue behavior can be completely solved if the dialogue is initiated by human, or the objective of the dialogue is determined by the operator. Separate parts of the dialogue system (linguistic processor) may well be implemented using modern speech technologies.

However, the construction of a high-quality speech perception system is impossible without the implementation of a synthesis system and vice versa. In addition, it is necessary to take into account such phenomena as the prediction of events not only at each level of the hierarchy of the linguistic processor that models the language (chains of words, phonemes, letters, words, etc.), but also between levels. There is a deep connection between the systems of speech perception and speech production: a common knowledge base and their close interaction in dialogue. The common knowledge base assumes that when speech signal is recognized at the sound level, a parametric description should be formed that corresponds to the models of movements of the articulatory organs of the speech-producing system. In other words, the description of sounds in a speech flow should be close in many respects to the description of the state of the articulators; otherwise the feedback during recognition and speech perception may turn out to be ineffective.

Dialogue communications, which are fundamental to human interaction, are increasingly dependent on technological innovations, in particular on generative artificial intelligence. Generative artificial intelligence, in turn, is turning into a widespread UX paradigm for interacting with most technical products, although it certainly has nothing to do with Artificial general intelligence, the development of which is obviously a task of the very distant future. There is no doubt that the generation of modern artificial intelligence models in the near future will become a human interface for the perception of the information flow, which cannot but have a huge impact on changing communication processes.

The specificity of communicative roles today is largely determined by the nature of the actors, since human is increasingly forced to contact artificial entities (chat bots, voice assistants, robots, artificial intelligence

algorithms, etc.). The polyphony of human communication is blurred, thus changing the structure of discourse and ways of representing meanings.

It is obvious that, in the near future, especially after the appearance of new versions of GPT, the focus will be on the transformation of social communications in the digital environment and at various levels of interaction: human-human, human-robot, natural-virtual actors, robot-robot, etc. New forms and types of communications, in turn, will have an impact on legal, socio-cultural, political, international processes, conditions of socialization and life strategies of the younger generations, and many others.

So far, we have been talking about theoretically determined capabilities in the subject domain of speech behavior. Now, it becomes possible to state that the intellectual principles of information processing considered in the first part of the monograph, are beginning to penetrate into the practical areas of information processing. Representation of text semantics in the form of a semantic network (a set of pairs of root stems that characterize the permissible combination of words that describe objects and events of the world) is used in the TextAnalyst software technology for automatic semantic analysis of texts, which makes it possible to generate a text abstract, compare texts by meaning and classify texts. Correlation-extreme signal processing can be used to form a convenient representation of texts in their various encodings (for example, in the form of so-called noise-like signals) in text comparison problems. And finally, we are talking about the implementation of micro-integrated devices that implement artificial neural networks, where analog type synapses (memristors) are used instead of digital device synapses.

Correlation-extreme signal processing is a powerful tool widely used in radio navigation, radar detection and location, communications, image recognition, information security, that is, wherever there is a need to analyze meaningful character sequences in a certain spatial context.

A technique is proposed for the synthesis of homogeneous processor structures based on the calculation of correlation and convolution using a graphical representation of the calculation process in the space of indices for the analysis of texts represented by noise-like signals.

A number of specialized processor structures for calculating the correlation and convolution functions have been synthesized and studied; these have been implemented in a number of specialized very large-scale integrated circuits. The results can be used in the development of FIR interpolation and decimation filters. A special class of calculators is identified: ring processor structures widely used in the implementation of matched filters.

These results were used in the development of specialized very large-scale integrated circuits for a number of radar devices, broadband communications, satellite navigation, local navigation systems.

Innovative advances in the development of memristive devices and systems based on them have shown that memristive devices can be used to create neuromorphic systems that have a significant potential for solving hardly formalized and non-formalizable problems in the field of artificial intelligence. The main advantage of such systems is the shifting away from the von Neumann architecture and the organization of a new computing principle, that is, in-memory computing, where the memory and the processor are physically the same device.

As can be seen, the theoretical and practical results presented in the monograph only partially cover the subject domain of "speech purposeful behavior". These included information relating only to the mechanism of information processing by a linguistic processor, which is based on the idea of information processing in the human mind (the subject domain of Neuroinformatics). The intentional part of the problem is left outside the scope of the monograph. This was done for a good reason: ethical considerations of opposing artificial and natural intelligence should not affect the practical feasibility and use of the information processing mechanism. That is, we must be able to put into practice everything we know, bearing in mind that this will not lead to the opposition of artificial and natural intelligence.

As is traditional, having collected the dummy copy of this collective monograph, the editors realized that the idea to present such an extensive subject domain in one volume is futile, even despite the deliberate qualification of readers. Therefore, both editors and readers cannot but accept everything as it is and hope that at least part of the material will be useful to someone.

The editors thank the authors of the collective monograph for their participation, and the publishing house for their foresight that helped publish such a complicated monograph in such a complicated subject domain as speech purposeful behavior.

Alexander A. Kharlamov,
Editor

CHAPTER ONE

SPEECH AS PURPOSEFUL BEHAVIOR

ALEXANDER A. KHARLAMOV

Introduction

Purposeful human behavior is formed as a reaction to an external or internal situation that involves a transition from the current situation to the planned final one. The capability to implement purposeful behavior appeared as a result of the ability of a person to form an idea about individual situations and manipulate these ideas both in mind and in the real world.

Representations of situations in the human mind are formed by the information processing mechanisms of the hippocampus, which, in turn, manipulates the representations of individual objects and events that are formed in the sensory cortex of the cerebral hemispheres in the process of structural processing of sensory information. These images of objects and events in the hippocampus are combined in combinations in space and time; such representations of situations are named by the names of one or another modality, and chains of names of situations in the anterior cortex further form plans for purposeful behavior.

The mechanisms for representing and processing information in the cortex and in the hippocampus are very different. If the structural processing of input sensory information is performed in the cortex, when hierarchies of dictionaries of images of events of various modalities are formed in the columns of the posterior cortex, then combinations of image indices of these dictionaries are formed in the hippocampal lamellae, as they are combined in space and time in specific situations. These representations in the hippocampal lamellae, being multimodal, include individual events that are characteristic of a particular situation and can be considered its name. The chains of these names, corresponding to the sequences of situations represented in the anterior cortex, form the same (as in the sensory cortex) hierarchies of representations of varying degrees

of detail of representation, which are the material for searching for the chain that most effectively connects the source and target situations, that is, one that characterizes purposeful behavior. The search for such a chain is implemented by fuzzy choice mechanisms.

To show all these processes, this section presents the mechanisms of structural processing of sensory information in the posterior cortex, the mechanisms of formation of the representation of situations in the hippocampus, the unification of representations that form the model of the world within the framework of the first signaling system, which also form the model of the world within the framework of the second signaling system, which also includes, in addition to non-linguistic modalities, a linguistic representation of the world, which differs from the extralinguistic one in terms of lesser variability, that is, it makes it easier to identify the structure of the world. The use of structural processing in the anterior cortex is also shown, which differs from the same representation in the posterior cortex in terms of the information being processed, which takes processing to another level: this not only simplifies the logical representation of the processes analyzed, but also makes it possible to implement a fuzzy search for effective solutions.

Since speech is purposeful behavior and language is a very well-studied subject, speech behavior turns out to be a convenient model for studying purposeful behavior.

1.1. Neuroinformatics

Speech is a very complex phenomenon. But if one looks at it from a convenient point of view, then a lot can become clear. First, speech processing can be divided into signal and character parts. Secondly, the character part of the speech process strongly resembles other similar events, for example, text processing (which is the most obvious example). It is not so obvious that the character part of text processing is similar, for example, to the processing of video sequences, which are also quasi-texts.

But let’s consider everything in order.

1.1.1. Signal processing

1.1.1.1 Information processing in the periphery of the auditory analyzer

The processing of a sound signal is extremely simple: in the cochlea of the inner ear, hair cells form a spectral representation of a sound wave in terms of its frequency components, which becomes sharper as switching

occurs in the subcortical nuclei (Labutin, Molchanov, 1973; Bekhtereva, 1978). The sequence of sounds encoded in this way (including speech sounds) enters the auditory cortex.

Of course, the mechanism of auditory information processing is not so simple (it includes details of the so-called sound masking (Giraud & Poeppel, 2012)), but in this case we can restrict our consideration to such a representation.

1.1.1.2 Information processing in the periphery of the visual analyzer

Signal processing in the visual analyzer is much more complicated than in the auditory one. It includes, as in the auditory analyzer, an increase in the dynamic range of input signals by averaging the image brightness on the retina, the formation of an optimal light flux in terms of brightness by changing the pupil diameter, and the convolution of color channels into one (Shkolnik-Yarros, 1986). But the main thing in the signal part of processing is the formation of the image scanning direction, which is ensured partly by automatisms of the periphery of vision (saccades and tremors) and partly by the oculomotor system, taking into account the semantics of the image itself and the problem statement (Podvigin et al., 1986; Yarbus, 1975).

There are two channels of visual information processing: coarse and fine; each in their own way participate in this processing. The coarse channel calculates the coordinates of the points with the greatest informativity of the image (for example, by defocusing (Zavalishin, 1974)), while the fine channel determines the direction of transition from point to point, depending on the problem statement and training background.

1.1.2. Character processing

The code sequences of various modalities obtained by different methods of primary processing (depending on the type of information being processed), reduced to a uniform representation, are then subjected to structural processing in the columns of the cortex, the mechanisms of which are the same in terms of the representations of various modalities (Kharlamov, 2017; Kharlamov & Pilgun, 2020).

1.1.2.1. Structural processing

The column of the cortex (more precisely, the set of pyramidal neurons of the third layer of the cortex, which perform the main processing of

specific information in the column) models, with the addresses of its neurons, the points of the multidimensional signal space, onto which the code sequences of sensory modalities formed at the signal level are mapped (which are different in columns of different sensory modalities) (Kharlamov, 2017; Kharlamov & Pilgun, 2020). Let us consider the formalism of information processing in columns in the simplest – binary – model (as convenient for understanding and interpretation). In this case, the mapping onto an n-dimensional unit hypercube takes place.

Let we have an n-dimensional signal space R^n and a unit hypercube $G_e^n \in R^n$ in it. For further presentation, we introduce some notations and definitions.

Denote by $\{A\}$ the set of code sequences formed by the signal periphery of the analyzers, the elements of which are the characters that make up the codes of the input sequences $A = (\dots\ a_{-1},\ a_0, a_1, \dots, a_i, \dots)$. In the binary case, $a_i \in \{0, 1\}$. In what follows, for simplicity, we consider everything using the example of processing binary sequences. As such a sequence, for example, the text of the novel "War and Peace" by L.N. Tolstoy can be taken, encoded in binary code.

Denote by $\{\hat{A}\}$ the set of trajectories of sequences corresponding to the set of input sequences $\{A\}$, whose elements $\hat{a}_i$ are points in the space R^n, i.e. vertices of the unit hypercube $\hat{a}_i \in G_e^n$, where $\hat{a}_i = (a_{i-n+1}, a_{i-n+2}, \dots, a_i)$ are successive fragments of the sequence A of n characters in length, shifted relative to each other by one character (by one period of time) – the coordinates of points of the multidimensional space R^n (due to the binary representation, these are vertices of the unit hypercube G_e^n).

Definition 1.1. A **trajectory** is a sequence of points $\hat{a}_i$ of the multidimensional space R^n, corresponding to the input code sequence A.

We introduce an associative transformation F_n:

$$F_n: A \to \hat{A}, F_n(A) = \hat{A}, \tag{1.1}$$

where $A = (\dots, a_i, \dots: a_i \in \{0, 1\})$, a $\hat{A} = (\dots, \hat{a}_{-2}, \hat{a}_{-1}, \dots \hat{a}_i, \dots) =$

$$= \begin{pmatrix} \dots, (a_{-n-1}, a_{-n}, \dots, a_{-2}), (a_{-n}, a_{-n+1}, \dots, a_{-1}), \dots \\ \dots, (a_{i-n+1}, a_{i-n+2}, \dots, a_i), \dots \end{pmatrix}.$$

The introduced transformation F_n, which forms a trajectory in the n-dimensional signal space, with the coordinates of the points given by n-term fragments of the original binary sequence A, is the basis for structural information processing. It has the associativity property of addressing to points of the trajectory $\hat{A}$ along an n-term fragment of the sequence A: any