

Р.В. МЕЩЕРЯКОВ
**КРИТЕРИЙ СТРУКТУРНОЙ СЛОЖНОСТИ
ИНФОРМАЦИОННЫХ СИСТЕМ**

Мещеряков Р.В. Критерий структурной сложности информационных систем.

Аннотация. В статье предлагается критерий оптимизации для информационных систем. Показывается возможность использования показателя энтропии обрабатываемой информации при оценке сложностных показателей. Дается пример оценки информационной системы по обработке речевого сигнала. Показывается его эффективность.

Ключевые слова: сложность, модель, информационная система.

Mescheriakov R.V. **Criterion for the structural complexity of information systems.**

Abstract. The article suggests a criterion for the optimization of information systems. Demonstrates the possibility of using the entropy index of the processed information in assessing the complexity indicators. An example of evaluation of an information system for processing the speech signal. Demonstrates its effectiveness.

Keywords: complexity, model, information system.

Введение. Обработка информации в различных информационных системах требует оценки требуемых вычислительных ресурсов. При этом используются классические показатели: вычислительная сложность как число элементарных операций (с вариациями «минимальная», «средняя» и др.), сложность по памяти как требуемые объемы для работы системы и вариации этих оценок. Однако при обработке сложноструктурированной информации, например при решении задач создания искусственного интеллекта, применения нейронных сетей и т.п., эти показатели не позволяют адекватно оценивать сложностные параметры систем. Последние необходимы для определения (перераспределения) требуемых ресурсов информационной системы.

Цель настоящей работы — определение критерия оценки информационной системы. Отметим, что в настоящей работе оценка ориентирована именно на информационные системы, обеспечивающие взаимодействие различных сторон. Явным примером систем являются интеллектуальные телекоммуникационные вопросно-ответные системы. Основная особенность подобных систем — непосредственное участие в речевом диалоге человек—машина.

Другие распространенные представители этого класса — системы безопасности. Одними из основных элементов практически любой системы являются подсистемы аутентификации и идентификации. Их основа — идентификатор объекта, используемый субъектом

для доступа в систему. Система при идентификации проверяет соответствие субъекта (объекта) и предъявляемого идентификатора. При этом вступают во взаимодействие две стороны:

- 1) субъект (объект),
- 2) вычислительная система.

Очевидно, что отсутствие учета факторов взаимодействия элементов всей системы безопасности может привести к рассогласованию этих элементов.

Очевидно, что для полноценного проектирования систем необходимо учитывать природу информационных систем, основанных на обработке естественного языка [1]. В зависимости от назначения и способа обработки данных информацию можно представить различными уровнями иерархии. При этом конфигурации нижних уровней формируются из элементов верхних уровней конфигураций нижних уровней и наоборот. Таким образом, при наличии прямой и обратной цепочек преобразований проводится сравнение их результатов.

Примем, что преобразование является достоверным с заданной точностью, если оно обеспечивает покрытие характеристиками элементов поля сочетаний не менее заданного [2, 3]. Важно отметить, что при увеличении числа последовательных преобразований между уровнями возникают рассогласования, обусловленные не только ошибками преобразований, но и особенностями интерпретации знаний.

При формировании сообщений используется часть информации по каналу восприятия. Таким образом, источник сообщения учитывает возможности приемника сообщения. В данном случае можно сказать, что при формировании и передаче сообщения приемник не только стремится достичь цели оптимального кодирования [4], но и достаточного уровня надежности распознавания. При этом используются имеющиеся в системе приемника сведения о его возможностях, некоторые параметры которых он может использовать, располагая структурой и поведением своего восприятия.

Чем ближе между собой значения параметров (или их покрытие) тем более эффективна процедура преобразований. Процедуры сравнения обработанных параметров могут измеряться с помощью различных мер. Данное качество может быть интерпретировано как ограничения, накладываемые на систему обработки знаний. Необходимо учитывать особенности преобразований различных шкал, в которых выражены параметры элементов системы [5].

Кроме того, можно определить достижение конечной цели при работе информационной системы. Однако не всегда ясна и формулируется настоящая цель и не всегда может быть реализован функционал ее достижения.

Отметим, что данный подход не всегда эффективен при использовании в информационных системах. Так, в случае выброса показателя одного из параметров объекта значение итоговой функции значительно влияет на саму меру близости. Кроме того, значения могут выражаться в интервальных шкалах или шкалах наименований (например, на уровнях семантики или прагматики).

При определении требуемых вычислительных ресурсов для реализации систем необходимо учитывать шкалы, в которых выражены сигналы на соответствующих уровнях иерархии информационной системы. Сложность представления и распределения ресурсов требует сравнения результатов определения различной информации в системе.

Предлагается использование показателя энтропии для оценки трудоемкости выполняемых операций в информационной системе. Очевидно, что этот показатель можно применять и к любым элементам систем обработки информации. Рассмотрим дискретный случай. Примем, что блок обработки информации, на вход которого поступает s потоков, в каждом из которых может быть q состояний, производит некоторое преобразование B , в результате которого на выходе получаем s' потоков, в каждом из которых может быть q' состояний:

$$B: s \times q \rightarrow s' \times q'.$$

Оценим количество входящей обрабатываемой информации в битах, при этом сделаем допущение, что вероятность p нахождения каждого потока s в состоянии q одинакова. Тогда:

$$H = - \sum_{i=1}^s p_i \log_2 p_i = -s \times p \times \log_2(p). \quad (1)$$

Таким образом, постановка задачи разбиения блока обработки информации на несколько блоков может быть представлена как задача нахождения оптимальной иерархии [6]. При этом частная задача определения оптимального числа блоков обработки и уровней иерархии представляет собой задание функционала:

$$\arg \min_{G \in \Omega} R(G),$$

где Ω — множество представлений процесса обработки информации (с представлением в иерархическом виде) с заданным функционалом

$$R: \Omega \rightarrow [0, +\infty).$$

При этом определение функционала представляет собой постановку задачи определения ограничений на оптимизацию процесса обработки информации, и по существу он представляет собой функционал стоимости.

Предположим, что блок обработки информации разбили на m блоков, сокращая информационную нагрузку на каждый блок. Вместе с тем необходимо выделить особый блок — блок управления B_c , который будет согласовывать работу различных блоков, а также собирать результаты (рис. 1). Необходимо отметить, что распараллеливание процесса обработки информации подобным образом проводится достаточно успешно на различных уровнях обработки: как на уровне элементов процессора (например, на уровне ядра), так и при реализациях в виде кластерных систем.

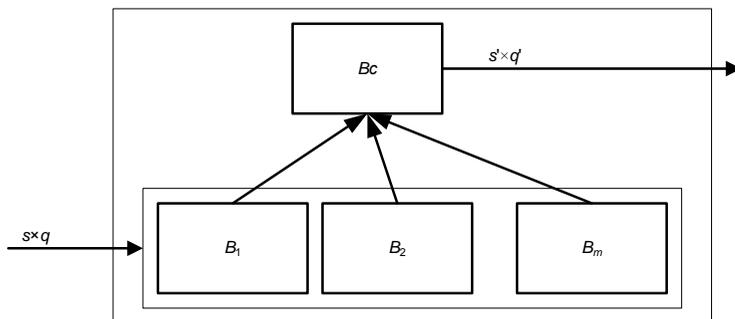


Рис. 1. Блоки обработки информации.

Предлагается использовать критерий оценки количества обрабатываемой информации для оптимального распределения вычислительных ресурсов системы. Будем считать критерием оптимальности (или функционалом для задачи оптимизации иерархии) равенство количества обрабатываемой информации, определенной по формуле (1):

$$H_1 = H_2 = \dots = H_m. \quad (2)$$

Очевидно, что при увеличении числа блоков m , на которые разбивается исходный блок обработки информации, количество входящей информации в блок управления растет. Оптимальным в общем случае будет равенство:

$$H_1 = H_2 = \dots = H_m = H_{B_c}. \quad (3)$$

Тем не менее количество обрабатываемой информации может превысить вычислительные возможности блока управления. Для их

учета необходимо таким образом изменить функционал нахождения оптимальной иерархии, чтобы функционал стоимости учитывал вычислительные ресурсы каждого блока иерархии системы обработки информации. Для практического применения целесообразно приводить следующий критерий, дополняющий критерий (3):

$$V_{Bc}H_{Bc} = V_1H_1 = V_2H_2 = \dots = V_mH_m, \quad (4)$$

где V_i — вычислительная сложность, реализованная в блоке i .

Таким образом, учитывается не только количество входящей информации, но и вычислительная сложность реализуемой функции обработки данных в блоке. В ряде случаев система обработки информации представляет собой реализацию последовательного или параллельного процесса, который включает в себя преобразование данных, представленных на рис. 2.

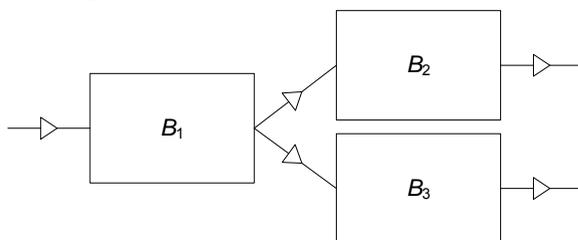


Рис. 2. Разбиение блока обработки информации.

Очевидно, что обработка данных в классической постановке вопроса включает в себя расчет информационной энтропии для дискретного и непрерывного случаев [4]:

$$H = -\sum_{i=1}^n p_i(x) \log_2 p_i(x),$$

$$H = -\int_{-\infty}^{+\infty} p(x, y) \log_2 p(x, y) dy$$
(5)

Учитывая, что ценность обрабатываемых данных может быть различна от блока к блоку, предлагается согласовывать их функции по показателю информационного наполнения, т. е. формировать обработчики таким образом, чтобы количество обрабатываемых данных было одинаковым:

$$H(B_1) = H(B_2),$$

или стремилось к данному соотношению. При такой постановке вопроса возникает вопрос согласования уровней обработки данных в выделенных блоках.

Примем по известному соотношению 7 ± 2 из [7] предположение о числе одновременно обрабатываемых объектов. Однако обрабатываемые данные могут иметь различную природу и выражаться в различных шкалах. При этом получаем, что для вероятного дискретного случая (n принимаем равным числу объектов) информационная энтропия изменяется от 11,6 до 28,5. Соответственно интеграл также должен изменяться в заданном диапазоне. Кроме того, могут быть сложные сочетания получаемых и обрабатываемых данных.

Очевидно, что информационная ценность получаемых данных после обработки повышается и при последующей обработке может быть сформирована таким образом, что количество информации будет увеличиваться (так как неопределенность результата в общем случае увеличивается).

Предлагается разнесение блоков обработки по информационно-разделенным функциям с соблюдением условий качества системы обработки информации, например в смысле [2]. Тогда получаем линейный рост функции информационной энтропии по всему процессу обработки информации.

В частности, можно рассмотреть подкласс иерархических систем, в которых на каждом уровне иерархии обрабатываются данные, выраженные в одних шкалах и обладающие одинаковой (или близкой) ценностью. Тогда получаем систему обработки, представленную на рис. 3.

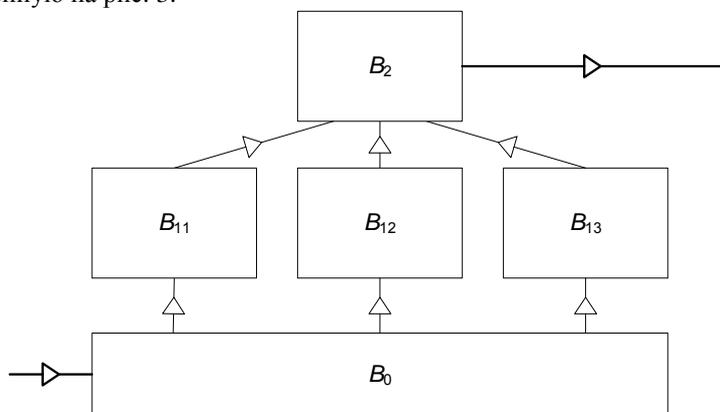


Рис. 3. Система обработки информации.

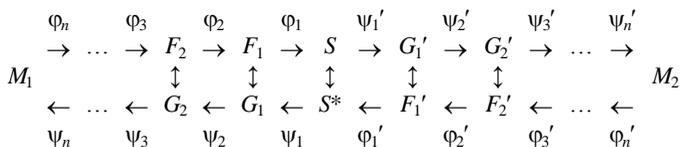
Предположим, что первоначально поступают числовые данные. На первом этапе в блоке B_0 производится разбиение поступающей информации на взаимно независимые участки по условию равенства ее количества. Далее информация в виде данных поступает на блоки B_{11} , B_{12} , B_{13} уровня 1, информация преобразуется в порядковых шкалах и ее количество соответственно уменьшается, а так как число возможных исходов сокращается, информационная ценность одного блока увеличивается.

Получаемая после обработки информация с трех блоков поступает на уровень 2 и далее обрабатывается в шкале наименований. Эта информация также повышает информационную энтропию, но диапазон представлений ее сокращается. Очевидно, что использование прямого расчета по числу возможных исходов не может быть оценено в виду различных видов оценки количества информации, а также ее неопределенности (оценки). Однако в общем случае информационная энтропия может быть сопоставлена и соотносена таким образом, что управляемость процессом сохраняется. Рассмотренная реализация для информационных систем может быть использована и для других видов систем обработки информации с возможностью оценки и проверки работоспособности и нагруженности блоков обработки.

Критерием данного подхода может быть качество структурной сложности системы [1], которое позволяет оценить возможность реализации системы обработки информации с точки зрения информационной энтропии. Из этого можно сделать вывод, что система будет работать не только эффективно и прогнозируемо, но и с соблюдением условий надежности при обработке данных.

Однако в информационной системе необходима абстракция данных и знаний. Это может быть реализовано при помощи категорий, морфизмов и функторов.

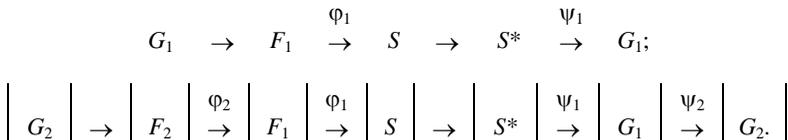
Для иллюстрации предлагаемого подхода рассмотрим систему синтеза и распознавания речи. Речевая система в общем случае рассматривается как иерархическая многоуровневая система и представляется цепочкой отображений:



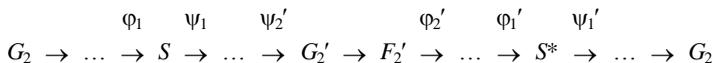
где F_i , F_i' , G_i , G_i' — описание высказываний на разных уровнях иерархии в системах речеобразования и речевосприятия; Φ_i' , Φ_i , Ψ_i' ,

ψ_i — отображение описаний; S, S^* — речевые сигналы; M_1, M_2 — модели мира участников диалога; стрелки между описаниями F_i и G_i характеризуют взаимосвязи между ними, реально за этими стрелками должны быть соответствующие модели.

Возможны различные типы связей в речевой системе, например:



Кроме того, возможны биологические обратные связи, включающие в себя участников диалога:



Приведенные структуры определяются структурой естественного языка. Более того, они замыкаются на очень низком уровне, т.е. на уровне образования звонких звуков речи, поэтому для нижних уровней иерархии диалоговой системы необходимо обеспечить устойчивость и точность управления, т.е. соблюдение условий (1).

Движение потоков информации от звукового сигнала к цели разговора — это канал восприятия (распознавания), а от цели высказывания к звуковому сигналу — канал синтеза. Каждый выделенный объект определяется своим набором сведений о языке, правилах преобразований и связей с другими уровнями.

Рассмотрим взаимное влияние верхних уровней на нижние и наоборот. Очевидно, что смысл предложения накладывает ограничения на состав слов, содержащихся в предложении, а также на их синтаксические связи. Бесспорно, что и синтаксические связи слов в предложении определяют смысл высказывания и его целевую функцию. Буквы в тексте имеют различную вероятность появления, а также вероятность появления следующей буквы зависит от предшествующих букв [6]. Однако уже при трех и более буквенных сочетаниях вероятность появления следующей буквы снижается, то же самое имеет место и для слов в предложениях, т.е. текущий прогноз на уровне букв или слов позволяет снизить области принятия решений при распознавании.

Представляет определенный интерес влияние результатов прогноза развития диалога на уровне модели мира, например, на

лексическую базу области допустимых решений на уровне фонетических слов. Для бытового общения достаточно 100–200 наиболее часто употребляемых слов. Естественно, что любая другая специальная предметная область потребует включения своих терминов и понятий, т.е. новых слов, что приведет к расширению лексической базы. Оказывается, что общая лексическая база увеличивается незначительно. С этой целью на основе алгоритмов морфологического и синтаксического анализа, рассмотренных в работе [8], проведена оценка лексической базы для предметных областей: компьютерная техника; радиоэлектроника. Оказалось, что наиболее часто встречаемые слова в этих областях увеличивают лексическую базу на 30–50 %. Естественно, что если ведется текущий прогноз диалога, то область допустимых решений даже на этой лексической базе будет значительно ограничена, следовательно, это должно привести к повышению надежности распознавания на уровне слов.

Можно предположить, что при использовании структурной сложности число объектов, одновременно обрабатываемое каждым уровнем, приблизительно одинаково [9]. Однако объекты имеют различную природу и различные свойства. Очевидно, что представление информации на физическом уровне отличается от ее описания на уровне слов. В частности, информация на уровне сигналов представлена в шкалах интервалов, а на уровне слов — в шкалах наименования и порядка.

Для выделения элементарных объектов физического уровня (уровня звуков) чаще всего используют предварительную сегментацию на квазиоднородные участки звукового сигнала с последующей их классификацией. Имеет место множество разнородных сегментов

$$S = \{s_i\}, \quad i = 1, 2, \dots, I,$$

где I конечно.

Для каждого выделенного объекта можно определить свое тело кластера K_i . Используя какую-либо меру близости [3], в данном пространстве можно и нужно определить образ, в который попадает искомый сегмент речевого сигнала. Очевидно, что на данном уровне сравнение может происходить в шкалах интервалов и в качестве меры близости может использоваться расстояния [3].

Однако количество информации, обрабатываемой на данном уровне, ограничивается числом состояний введенных сегментов, включая основные блоки обработки информации [10, 11]. При этом

предлагаемое разбиение блоков обработки информации последовательно классифицирует участки, подлежащие обработке: на первом блоке определяется характер сигнала — голос или не голос (*Voice Activity Detector*), затем тип участка, его однородность и пр.

Следующий уровень описания речевого сигнала — параметрический, входом которого является множество элементов

$$M_N = \{m_i\}.$$

В зависимости от длительности цепочки необходимо учитывать связи не только двух последовательно стоящих элементов, но и все находящиеся в цепочке элементы. В настоящем случае целесообразно применять меру близости [3], формирующую отнесение элемента к группе [12]. Аналогично действуем при повышении уровня к самому верхнему уровню прагматики высказывания.

Для реализации иллюстрации разбиения блоков рассмотрим часть системы, отвечающую за синтез речевого сигнала. Известно, что длительность ударного гласного, а также его произнесение зависят от местоположения в слове относительно ударного гласного (предударный, ударный, заударный, неударный), в высказывании [10, 11].

Коммутативные диаграммы после определения правил преобразования и баз данных являются идеализированной моделью системы синтеза речи. Она может служить в качестве базы для синтеза речи в общем случае (без настройки на диктора). В реальности всегда происходит подстройка параметров речеобразующего тракта во время произнесения.

Таким образом, выберем блок обработки информации: транскрибирование текста. При произнесении текста человек генерирует определенное сочетание частот для создания какого-либо звука, которое в фонематическом тексте выражается фонемой. Отметим, что не существует взаимно однозначного отображения множества звуков во множество фонем. Сложность для русского языка нахождения фонетической транскрипции состоит в том, что 90 % слов написаны фонематически, а 10 % — нефонематически (традиционно) [11].

В нашем случае рассмотрим систему, на вход которой поступает орфографический текст с размеченными ударениями и границами синтагм, а на выходе — фонетическая транскрипция текста. На рис. 4, *а*, представлены блоки системы до применения предложенного метода, на рис. 4, *б*, — после.

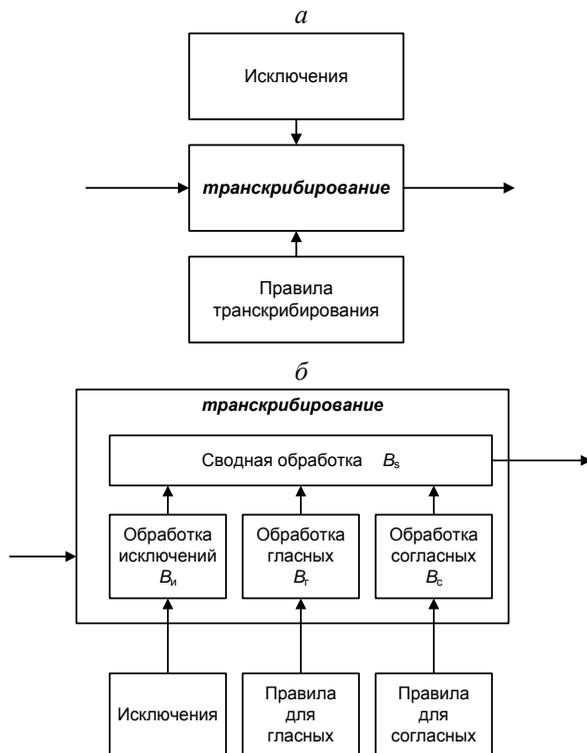


Рис. 4. Транскрибирование текста

Система транскрибирования проводит различные операции и включает блок «обработка слов-исключений», которые не могут быть транскрибированы верно по существующим правилам языка. Правила фонетического транскрибирования могут быть разделены на две группы: 1) для гласных и 2) согласных, так как для согласных звуков при фонетическом транскрибировании в большей степени влияют последующие звуки, а для гласных — предыдущие и их степень ударности. Блок сводной обработки собирает полученные данные с других блоков и принимает окончательное решение о выдаче последовательности транскрипционных знаков. Оценки по критерию (1) составляют для B_n — 780, B_r — 1060, B_c — 1382, B_s — 864.

Таким образом, на выходе системы транскрибирования получаем множество фонетических символов, объединенных в фонети-

ческие слова и далее в высказывания. Кроме того, процесс обработки информации с применением критериев (1)–(3) оптимален.

Закключение. Сформулированный критерий структурной сложности может быть использован в информационной системе. Он позволяет перераспределять ее ресурсы в соответствии с количеством информации, обрабатываемой в каждом блоке.

Перспективой реализации предложенного подхода может послужить система речевого диалога. Кроме того, перспективным представляется формирование дополнительных требований на человеко-машинный интерфейс, учитывающий возможности восприятия информации человеком.

Литература

1. *Мещеряков Р.В., Бондаренко В.П.* Диалог как основа построения речевых систем // Кибернетика и системный анализ, 2008. № 2. С. 30–41.
2. *Флейшман Б.С.* Элементы теории потенциальной эффективности сложных систем. М.: Советское радио, 1971. 223 с.
3. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.
4. *Шеннон К.* Работы по теории информации и кибернетике. М.: Изд. иностр. лит., 1963. 830 с.
5. *Пфанцгаэль И., Бауман В., Хубер Г., Кузьмин В.Б. и др.* Теория измерений. М.: Мир, 1976. 250 с.
6. *Воронин А.А., Мишин С.П.* Оптимальные иерархические структуры. М.: ИПУ РАН, 2003. 214с.
7. *Миллер Д.Ж.* Магическое число семь плюс или минус два: О некоторых пределах нашей способности перерабатывать информацию // Инженерная психология / Под ред. А.Н. Леонтьева. М.: Прогресс, 1964.
8. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. М.: Наука, 1979. 257 с.
9. *Клацки Р., Память человека: Структуры и процессы / Под ред. Е. Соколова.* М.: Мир, 1978. 319 с.
10. *Taylor Paul, Text-to-Speech Synthesis.* Cambridge: University Press, 2009. 597 p.
11. *Лобанов Б.М., Цирюльник Л.И.* Компьютерный синтез и клонирование речи. Минск: Издательский дом «Белорусская наука». 2008. 343 с.
12. *Бондаренко В.П., Конев А.А., Мещеряков Р.В.* Сегментация и параметрическое описание речевого сигнала // Изв. вузов. Приборостроение. 2007. Т. 50, № 10. С. 3–7.

Мещеряков Роман Валерьевич — канд. техн. наук, доцент кафедры комплексной информационной безопасности электронно-вычислительных систем Томского государственного университета систем управления и радиоэлектроники (ТУСУР). Область научных интересов: анализ, синтез речевого сигнала, медицинские технологии, информационная безопасность. Число научных публикаций — 234. mrv@security.tomsk.ru; КИБЭВС ТУСУР, пр. Ленина, д. 40, Томск, 634050, РФ; р.т. +7 (3822)413-426, факс +7 (3822)900-111

Mescheriakov Roman Valerievich — PhD, assistant professor, Dept. of Complex Security of Electronic-computing Systems of Tomsk State University of Control Systems and Ra-

dioelectronics (TSUCSR). Research interests: speech analysis, speech recognition, medical technology, information security. The number of publications — 234, IEEE Senior Member. mr.v@security.tomsk.ru; KIBEVS Dept. TSUCSR, 40, Lenin-avenue, Tomsk, 634050, Russia; office phone +7(3822)413-426, fax +7(3822)900-111.

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией д-р техн. наук, доцент А.Л. Ронжин.
Статья поступила в редакцию 15.11.2010.

РЕФЕРАТ

Меццержков Р.В. **Критерий структурной сложности информационных систем.**

Современные методы оценки недостаточно учитывают особенности естественного взаимодействия человека и вычислительной системы. В статье рассматриваются возможности нового критерия.

Информационные системы работают с данными, которые можно оценить с помощью энтропии. Таким образом показывается, что трудоемкость обработки данных связана с количеством информации. Кроме того, информация представляется в различных шкалах. Для сопоставления количества информации предлагается использовать категории и функторы, являющимися абстракциями данных. Таким образом формализуется критерий структурной сложности информационной системы.

Приводится пример оценки структурной сложности на задачах анализа и синтеза речевого сигнала. Показывается разбиение блоков обработки информации по предложенному критерию. Он позволяет перераспределять ресурсы в соответствии с количеством информации, обрабатываемой в каждом блоке.

Перспективой реализации предложенного подхода может послужить системы речевого диалога. Кроме того, перспективным представляется формирование дополнительных требований на человеко-машинный интерфейс, учитывающий возможности восприятия информации человеком.

SUMMARY

Mescheriakov R.V. **Structural complexity criterion of information systems.**

Modern methods of assessment not take into account the peculiarities of the natural interaction between man and computer systems. The paper deals with the possibility of new criterion.

Information systems work with data that can be evaluated using the entropy. Thus it is shown that the complexity of the data related to the amount of information. In addition, information is presented in different scales. To compare the amount of information offered to use categories and functors, which are data abstractions. Thus, the criterion is formalized structural complexity of the information system.

Is an example of assessing the structural complexity on the analysis and synthesis of speech signal. Shows the partition blocks of information processing on the proposed criteria. It allows you to reallocate resources in accordance with the amount of information processed in each block.

The prospect of implementing the proposed approach could serve as a system of verbal dialogue. In addition, promising to formation of additional requirements for human-machine interface that takes into account the possibility of human information perception.