УДК 681.391

В. Е. Гай, В. А. Утробин, А. В. Лукьянчикова, И. В. Поляков

Распознавание изолированных речевых команд с позиций теории активного восприятия

Работа посвящена описанию метода распознавания изолированных речевых команд в условиях априорной неопределенности. В отличие от существующих методов распознавания, работающих на уровне отсчётов, предлагаемый метод реализует концепцию грубо-точного анализа сигнала, описанную в теории активного восприятия.

Ключевые слова: распознавание голосовых команд, теория активного восприятия

V. E. Gai, V. A. Utrobin, A. V. Lukjanchikova, I. V. Polyakov

Recognition of isolated voice commands from the standpoint of the theory of active perception

This abstract related to a description of the method for recognition of isolated voice commands in conditions of a priori uncertainty. In contrast to a signal samples methods this method represents coarse-to-fine conception of signal analyze described in active perception theory. Provides results of computing experiment for confirming the efficiency of this method.

Key words: speech command recognition, theory of active perception

Введение

Задача распознавания образов является одной из актуальных задач теоретической информатики. В рамках данной общей задачи выделяют задачи распознавания зрительных образов (изображений) и слуховых образов (речи) и т. д. Процесс распознавания образов, с позиций системного анализа, можно разделить на три этапа: формирование исходного описания, нахождение системы признаков и построение решающего правила. Существуют две формулировки задачи распознавания: в узком и широком смыслах [1]. В узком смысле задача распознавания сводится к построению классификатора, в широком смысле – к распознаванию в условиях априорной неопределённости (в данном случае не известны множество признаков и множество классов).

Известны проблемы, связанные с применением существующих методов распознавания образов [2]:

1) проблема формирования исходного описания, связана с тем, что существующие модели и методы распознавания адаптированы к конкретному классу прикладных задач и требуют априорного знания свойств анализируемых сигналов;

2) проблема формирования системы признаков, связанная с выбором конечного множества признаков, обеспечивающих однозначность решения задачи классификации на этапе распознавания и отвечающая требованиям необходимости и достаточности. Этап выбора системы признаков необходим для сокращения размерности входного описания. Учитывая, что задача сокращения размерности – оптимизационная задача, то для её решения необходимо использование критерия информативности. Отсутствие модели априорной неопределённости и модели её раскрытия породило большое количество методов в выборе критерия информативности, что, в свою очередь, породило большое число возможных вариантов признаков [3,4];

3) проблема принятия решений в условиях априорной неопределённости. Этап принятия решения заключается в сравнении с имеющимся эталоном признакового описания анализируемого сигнала. Предполагается, что эталону соответствует компактное множество точек в системе признаков. Однако помехи, структурные изменения одного и того же представителя класса приводят к перекрытию классов. Поэтому проблема принятия решения замыкается на проблемы формирования системы признаков, позволяющей сформировать эталон, имеющий компактное представление.

Теория активного восприятия предлагает решение описанных проблем [1]. Настоящая работа посвящена применению данной теории к анализу речевых сигналов.

1. Обзор методов распознавания речевых сигналов

Рассмотрим методы, применяемые на разных этапах решения задачи распознавания [5]:

1) этап предварительной обработки звукового сигнала, обычно, заключается в фильтрации сигнала и выделении границ речевой активности [6, 7]. Учитывая, что задача распознавания решается в условиях априорной неопределённости (информация о помехе отсутствует), выбрать подходящий фильтр сложно;

2) для создания описания входного сигнала вычисляются признаки: коэффициенты спектра Фурье, кепстральные коэффициенты, мел-частотные кепстральные коэффициенты, коэффициенты линейного предсказания (linear predictive coding), коэффициенты вейвлет-спектра

и т. д. Необходимо отметить, что существующие методы обработки речевых сигналов основаны на стратегии точно-грубого анализа, который заключается в том, что признаки вычисляются по участку сигнала, длительность которого составляет около 25 мс [4, 5];

3) на этапе классификации в системах распознавания речи взаимодействуют несколько модулей [8]:

а) модуль акустической модели позволяет по входному речевому сегменту определить наиболее соответствующие ему шаблоны отдельных звуков. При акустическом моделировании используется скрытая марковская модель, модель гауссовой смеси, нейронная сеть, метод опорных векторов. Использование данных моделей предполагает их предварительное обучение и выбор параметров, что, в условиях априорной неопределённости является не тривиальной работой;

б) модуль модели языка – используется для определения наиболее вероятной последовательности слов. Необходимость использования языковой модели объясняется ростом словаря распознаваемых слов, в результате чего увеличивается число слов похожих по звучанию. Выделяют дискретные (модель с конечным числом состояний, на основе теории формальных языков, на основе лингвистических знаний) и статистические модели (*n*-граммная модель, модель на основе деревьев решений, статистическое обобщение формальных языков);

в) декодер – объединяет данные, поступающие от акустической и языковой моделей, и формирует результат распознавания.

2. Метод распознавания речевых сигналов на основе теории активного восприятия

Выше указано, что современные методы распознавания речи обрабатывают анализируемый сигнал на уровне отсчётов. С другой стороны, известны факты, свидетельствующие о целостности механизма зрительного восприятии человека, выполняемом зрительной системой грубо-точном анализе сенсорных данных [1]. В теории активного восприятия (ТАВ) описан метод грубо-точного анализа, который используется для распознавания изображений. Предполагается, что похожие механизмы работают в слуховой системе, исходя из чего, данный метод может быть применён и к анализу речевых сигналов. Рассмотрим предлагаемую реализацию этапов системы распознавания с точки зрения ТАВ.

2.1. Предварительная обработка

В условиях априорной неопределённости процесс раскрытия неопределённости звукового сигнала заключается в дихотомии его

области определения G на равные части. Поскольку все отсчёты сигнала находятся в отношении эквивалентности, множество отсчётов можно разбить на любое число подобластей $G_{ij} \subseteq G$ без пересечения этих областей между собой. Последовательное применение операции дихотомии позволяет сгенерировать пирамидальную структуру (см. рис. 1, i – уровень разложения, j – номер области на i-ом уровне, T – длительность сигнала). Таким образом, этап предварительной обработки заключается в выполнении операции дихотомии и формировании подобластей G_{ij}.

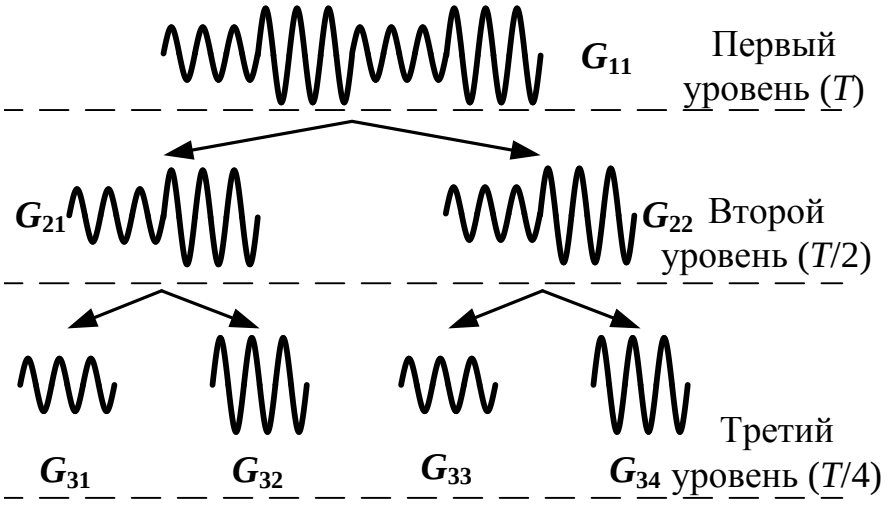

Рис. 1. Пирамида описания сигнала

2.2. Вычисление признаков

Рассмотрим предлагаемый метод вычисления признакового описания подобласти $G_{ij} \subseteq G$:

1) отсчёты сигнала, относящиеся к подобласти G_{ij}, разбиваются на множество сегментов $\mathbf{g} = \{g_k\}$ длиной L * 16 отсчётов, со смещением в S отсчётов, $k = \overline{1, N}$, где N – число сегментов в подобласти G_{ij};

2) к каждому сегменту g_k применяется U-преобразование (U-преобразование является базовым в теории активного восприятия), в результате формируется спектральное представление каждого сегмента: $u_k = U[\, g_k \,]$, $\boldsymbol{u} = \{\, u_k \,\}$, где U – оператор вычисления U-преобразования;

3) по вычисленному спектральному представлению u_k сегмента g_k определяются замкнутые группы: $p_k = P[\, u_k \,]$, $\boldsymbol{p} = \{\, p_k \,\}$, где P – оператор вычисления замкнутых групп;

4) вычисляется гистограмма замкнутых групп: $d_{ij} = H[\, \boldsymbol{p} \,]$, где H – оператор формирования гистограммы замкнутых групп, которая и является признаковым описанием области G_{ij};

5) признаковые описания областей G_{ij} объединяются в вектор x.

Отметим, что при создании признакового описания используется принцип рекурсии, т. е. к сигналу последовательно применяется одна и та

же операция – дихотомия. Таким образом, для выявления структуры сложного сигнала применяется одна и та же операция.

2.3. Принятие решения (классификация)

Этап классификации основан на использовании метода классификации *k*-ближайших соседей (см. рис. 2). Решающее правило метода *k*-ближайших соседей записывается следующим образом:

$$a(u; X^l, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} \left[y_u^{(i)} = y \right],$$

где u – классифицируемый объект, k – параметр алгоритма (количество соседей), $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$ – обучающая выборка, заданная в формате "объект-ответ", $Y = \{y_i\}$, $y \in \overline{1,C}$ – множество классов, C – количество классов. Определение близости между объектами x и x' выполняется с помощью расстояния Евклида:

$$\rho(x, x') = \sum_{i=1}^{M} \sqrt{(x_i - x_i')^2}.$$

Оптимальное значение параметра k определим по критерию скользящего контроля с исключением объектов по одному (leave-one-out, LOO):

$$LOO(k, X^l) = \sum_{i=1}^{l} \left[a(x_i; X^l \setminus \{x_i\}, k) \neq y_i \right] \to \min_k.$$

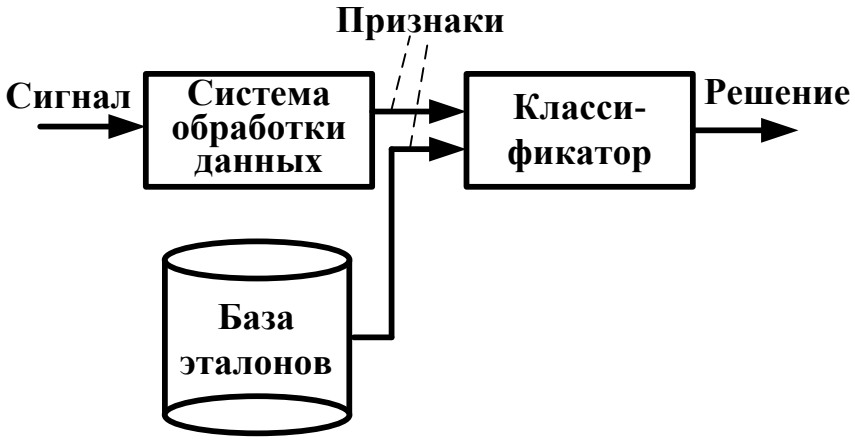

Рис. 2. Классификация принятого сигнала

3. Вычислительный эксперимент

3.1. Описание тестовых данных

В вычислительном эксперименте использовались звуковые записи следующих слов: Антенна, Вперёд, Декабрь, Жизнь, Карапуз, Кокс, Кокос, Корабль, Кресло, Лево, Мама, Машина, Москва, Назад, Насос, Наташа, Ночь, Пакет, Парниша, Право, Свет, Сон, Стол, Стул, Стоп, Тесно, Тосно,

Экономия, Энтропия, Ягода. Выполнено 50 записей для каждого слова.

Вычисления и запись базы данных выполнялись на ноутбуке Asus K70AD, процессор – AMD Turion(tm) II Dual-Core Mobile M500, 2200 МГц (два ядра), объём оперативной памяти – 4 Гб. Вычислительный эксперимент заключается в проверке точности работы описанного и известных методов распознавания.

3.2. Тестирование известных методов

В табл. 1 приведены результаты вычислительного эксперимента по распознаванию слов, выполненного с помощью известного алгоритма распознавания. Каждый эталон описывается мел-частотным кепстральными коэффициентами, моделирование вероятностного распределения признаков происходит с помощью модели гауссовой смеси (Gauss Mixture Model, GMM), на этапе классификации используется метод динамического искажения времени (Dynamic Time Warping, DTW). Вычислительный эксперимент выполнялся на основе метода перекрёстной проверки (тестовая выборка разбивалась на 10 частей).

Табл. 1. Результаты распознавания на основе мел-частотных кепстральных коэффициентов

№	Слово	Точность	№	Слово	Точность
1	Антенна	96	16	Наташа	98
2	Вперёд	50	17	Ночь	98
3	Декабрь	98	18	Пакет	98
4	Жизнь	98	19	Парниша	96
5	Карапуз	98	20	Право	94
6	Кокс	2	21	Свет	98
7	Кокос	16	22	Сон	86
8	Корабль	46	23	Стол	74
9	Кресло	94	24	Стул	90
10	Лево	98	25	Стоп	2
11	Мама	96	26	Тесно	88
12	Машина	98	27	Тосно	86
13	Москва	98	28	Экономия	98
14	Назад	98	29	Энтропия	98
15	Насос	98	30	Ягода	98

Средняя точность распознавания составляет 82 %. Низкую точность распознавания слов "Кокс", "Кокос", "Корабль" и "Стоп" объясняется тем, что они похожи произношением на другие слова.

3.3. Тестирование предлагаемого метода

Рассмотрим результаты тестирования предлагаемого метода. Данные для определения оптимального значения параметра k приведены в табл. 2. В табл. 3 показана точность классификации для одномерной гистограммы замкнутых групп (первая строка – нормализованные признаки, вторая строка – ненормализованные признаки, $L = 4$, $S = 2$, анализируемый сигнал делится на 4 части).

Табл. 2. Выбор значения параметра k

k	1	2	3	4	5	6	7
Норм.	73	77	89	79	64	63	55
Не норм.	71	70	94	86	61	54	54
k	8	9	10	11	12	13	14
Норм.	52	54	51	50	52	50	53
Не норм.	53	57	49	52	50	58	51

Табл. 3. Точность классификации изолированных команд (в %)

L / S	1/1	1/2	1/4	1/8	2/1	2/2
Норм.	88	88	85	83	93	91
Не норм.	88	89	87	79	93	93
L / S	2/4	2/8	4/1	4/2	4/4	4/8
Норм.	92	92	89	89	88	88
Не норм.	91	90	94	94	93	92

В табл. 4 приведена зависимость точности классификации от числа дихотомий входного сигнала ($L = 4$, $S = 2$). В последней строке таблицы приведена суммарная ошибка для указанного числа дихотомий.

Табл. 4. Точность классификации в зависимости от числа дихотомий (в %)

№	Слово	Число дихотомий					
		0	2	4	8	16	32
1	Антенна	76	90	98	98	98	98
2	Вперёд	72	90	94	92	92	92
3	Декабрь	56	90	98	98	98	98
4	Жизнь	94	98	96	98	98	98
5	Карапуз	70	90	98	96	96	96
6	Кокс	78	82	94	94	94	94
7	Кокос	60	70	86	86	86	86
8	Корабль	76	86	92	96	96	96

9	Кресло	46	54	76	86	86	86
10	Лево	70	76	84	92	92	92
11	Мама	72	82	96	90	90	90
12	Машина	66	74	92	92	92	92
13	Москва	86	86	96	98	98	98
14	Назад	90	96	96	98	98	98
15	Насос	86	90	94	96	96	96
16	Наташа	64	80	96	98	98	98
17	Ночь	86	96	98	98	98	98
18	Пакет	94	98	98	98	98	98
19	Парниша	76	94	90	92	92	92
20	Право	90	76	96	98	98	98
21	Свет	90	96	96	96	96	96
22	Сон	54	54	84	90	90	90
23	Стол	66	86	92	88	88	88
24	Стул	96	90	98	96	96	96
25	Стоп	74	90	88	94	94	94
26	Тесно	96	96	98	98	98	98
29	Тосно	82	86	88	90	90	90
27	Экономия	90	96	98	98	98	98
28	Энтропия	90	92	92	94	94	94
30	Ягода	80	82	84	90	90	90
	Среднее	77	85	92	94	94	94

Не очень высокую точность распознавания некоторых групп слов (например, «Стол», «Стул», «Стоп» и «Кресло», «Тосно») можно объяснить их подобием в произношении (на русском языке).

В табл. 4 приведены результаты, из которых следует, что с увеличением числа дихотомий точность распознавания в определённый момент перестаёт увеличиваться. Отметим, что с увеличением числа дихотомий увеличивается число вычислительных операций, необходимых для формирования признакового описания сигнала. Определим точку оптимальности, т. е. число дихотомий, которое позволит минимизировать отношение «точность распознавания / вычислительная сложность».

В табл. 5 приведена зависимость точности распознавания от размера базы эталонов. Из таблицы следует, что с увеличением базы эталонов, точность распознавания также увеличивается.

В табл. 5. Зависимость точности классификации от размера базы эталонов

Число реализаций одного слова	10	20	30	40	50

Размер базы эталонов	300	600	900	1200	1500
Количество верно распознанных слов	233	491	776	1093	1414
Точность распознавания	77	81	86	91	94

Рассмотрим устойчивость предложенного метода к локальным искажениям анализируемого сигнала. Вычислим точность распознавания для случаев, когда одна или несколько областей G_{ij} пропущены (см. табл. 6).

Табл. 6. Точность классификации в зависимости от числа используемых областей G_{ij}

G_{ij}	1	2	3	4	1, 2	1, 3	1, 4
Точн. классиф.	66	72	74	50	84	88	74
G_{ij}	2, 3	2, 4	3, 4	1, 2, 3	1, 2, 4	1, 3, 4	2, 3, 4
Точн. классиф.	86	82	80	92	86	89	90

Описанный эксперимент показывает, что предложенный метод распознавания устойчив к локальным искажениям сигнала. Известно, что слуховая система также устойчива к выпадению отдельных сегментов речевого потока за счёт включения разных компенсаторных процессов.

Заключение

В статье дано описание модели информационных преобразований для решения задачи распознавания изолированных команд. Предложенная модель основана на использовании операции дихотомии, позволяющей выделить однородные участки анализируемого сигнала для их дальнейшего описания. При создании признакового описания используется *U*-преобразование, которое является базовым в теории активного восприятия. Разработанный метод является дикторозависмым, поэтому дальнейшая работа будет направлена на его модификацию для снятия указанного ограничения. Использование предложенного метода позволяет, с одной стороны, получить инвариантность описания реализации речевого сигнала в рамках одного участка дихотомии, а с другой стороны сохранить грубую временную структуру сигнала. Алгоритмическая реализация предложенного метода обладает высоким потенциалом к распараллеливанию.

Статья рекомендована оргкомитетом XXI Международной научно-технической конференции «Информационные системы и технологии. ИСТ-2015». Публикуется при поддержке гранта РФФИ № 15-07-20095.

Список литературы

1. Утробин В. А. Элементы теории активного восприятия изображений // Труды Нижегородского государственного технического университета им. Р.Е. Алексеева. – 2010. – Т. 81. – № 2. – С. 61-69.

2. Верхаген К. [и др.] Распознавание образов: состояние и перспективы // М.: Радио и связь, 1985. – 104 с.

3. Загоруйко Н. Г. Методы распознавания и их применение / Н.Г. Загоруйко. – М.: Советское радио, 1972. – 208 с.

4. O'Shaughnessy D. Acoustic Analysis for Automatic Speech Recognition // Proceedings of the IEEE. – 2013. – Vol. 101. – N. 5. – P. 1038-1053.

5. Saon G., Chien J.-T. Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances, // IEEE Signal Processing Magazine. – 2012. – Vol. 29. – N. 6. – P. 18-33.

6. Котомин А. В. Распознавание речевых команд с использованием сверточных нейронных сетей // Труды Молодежной конференции «Наукоёмкие информационные технологии» SIT-2012. – Переславль-Залесский: Университет города Переславля, 2012. – С. 17-28.

7. Котомин А. В. Предобработка звукового сигнала в системе распознавания речевых команд // Труды XV Молодежной конференции «Наукоёмкие информационные технологии» SIT-2011. – Переславль-Залесский: Университет города Переславля, 2011. – С. 25-38.

8. Мазуренко И. Л., Холоденко А. Б., Бабин Д. Н. О перспективах создания системы автоматического распознавания слитной устной русской речи // Интеллектуальные системы. – 2004. – Т. 8. – № 1-4. – С. 45-70.