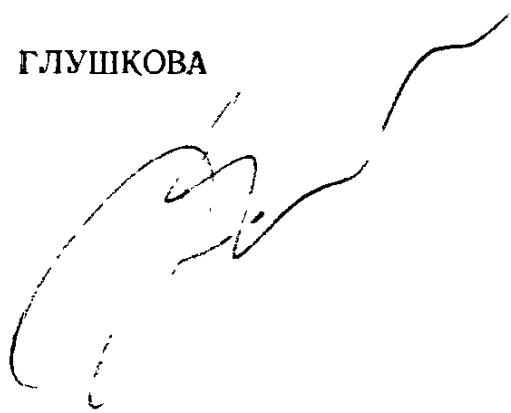


АКАДЕМИЯ НАУК УКРАИНСКОЙ ССР
ИНСТИТУТ КИБЕРНЕТИКИ имени В. М. ГЛУШКОВА

Т. К. ВИНЦЮК



АНАЛИЗ,
РАСПОЗНАВАНИЕ
И ИНТЕРПРЕТАЦИЯ
РЕЧЕВЫХ
СИГНАЛОВ

КИЕВ НАУКОВА ДУМКА 1987

УДК 534.4 : 621.391

Анализ, распознавание и интерпретация речевых сигналов / Винценок Т. К.— Киев : Наук. думка, 1987.— 264 с.

В монографии рассматриваются вопросы автоматического анализа, распознавания, смысловой интерпретации, синтеза и компрессированной передачи речевых сигналов применительно к устному диалогу человека и ЭВМ на формализованных и естественных языках предметных областей для использования в человеко-машинных системах сбора, обработки информации и управления. В рамках единого подхода, основанного на задании множеств сигналов с помощью автоматных порождающих грамматик и на применении динамического программирования, излагаются методы и средства поэлементного и пофонемного распознавания слов и слитной речи, сегментации и самосегментации сигналов, обучения и самообучения распознаванию речи, настройки на голос оператора, многодикторного и кооперативного распознавания, распознавания дикторов по речевым сигналам, выбора мер сходства и способов анализа сигналов.

Для аспирантов, научных работников и специалистов, занимающихся вопросами обработки сигналов, распознавания образов и искусственного интеллекта, создания интеллектуальных ЭВМ.

Ил. 1—40. Табл. 10. Библиогр.: с. 255—262 (172 назв.).

Ответственный редактор *В. И. Скурихин*

Рецензенты *В. И. Васильев, Н. Ф. Кириченко*

Редакция физико-математической литературы

в 1503000000-070
M221(04)-87 168-87

© Издательство
«Наукова думка», 1987

ОГЛАВЛЕНИЕ

Предисловие	6
Введение	9
Глава 1. Методологические основы композиционно-оптимального подхода (КДП-подхода) к распознаванию и смысловой интерпретации речи	16
§ 1.1. Сущность и особенности проблемы	16
§ 1.2. Основы КДП-подхода	18
§ 1.3. Математические модели речевых сигналов и их свойства	20
§ 1.4. Непрерывность сигналов и дискретность решений	23
§ 1.5. Необходимость моделирования	24
Глава 2. Поэлементный метод распознавания слов	26
§ 2.1. Описание речевых сигналов	26
§ 2.2. Кусочно-постоянная математическая модель речевых сигналов	32
§ 2.3. Поэлементный метод распознавания	34
§ 2.4. Анализ различных модификаций метода	40
§ 2.5. Возможные усовершенствования метода	44
§ 2.6. Эксперименты	47
Глава 3. Обучение поэлементному распознаванию речи	51
§ 3.1. Постановка и анализ задачи обучения	51
§ 3.2. Оптимальная сегментация реализаций. Обучение по одной реализации	53
§ 3.3. Точное решение задачи обучения	57
§ 3.4. Основной алгоритм обучения	61
§ 3.5. Выбор длины исходного эталона слова	64
§ 3.6. Формирование темпоральной транскрипции	65
§ 3.7. Примеры решений задачи обучения	65
§ 3.8. Другие варианты организации процедуры обучения	68
§ 3.9. Общие замечания	69
Глава 4. Пофонемное распознавание слов речи	71
§ 4.1. Переход от поэлементного к пофонемному распознаванию	71
§ 4.2. Роль транскрипций слова	72
§ 4.3. Преимущества пофонемного метода	73
§ 4.4. Постановка и анализ задачи обучения и самообучения пофонемному распознаванию	74
§ 4.5. Итерационный алгоритм обучения	75
§ 4.6. Случай незаданных длин транскрипций	82
§ 4.7. Задача дообучения распознаванию речи	84
§ 4.8. Экспериментальные исследования	86

Глава 5. Распознавание слитной речи, составляемой из слов выбранного словаря	97
§ 5.1. Постановка задачи поэлементного распознавания слитной речи. Порождающие грамматики и графы слитной речи	97
§ 5.2. Алгоритм распознавания слитной речи. Основные свойства	101
§ 5.3. Пофонемное распознавание слитной речи	106
§ 5.4. Обучение распознаванию слитной речи	106
§ 5.5. Особенности использования метода	108
§ 5.6. Результаты экспериментальных исследований	110
Глава 6. Глубокое пофонемное распознавание речи	115
§ 6.1. Проявление фонемности	115
§ 6.2. Освобождение от лексических ограничений. Общий фонемный граф	116
§ 6.3. Распознавание произвольных последовательностей фонем	120
§ 6.4. Использование фонетической транскрипции слова. Переход к глубокому пофонемному распознаванию слов и слитной речи	124
§ 6.5. Обучение глубокому пофонемному распознаванию	125
§ 6.6. Взаимосвязь задач распознавания и синтеза речи	127
Глава 7. Смысловая интерпретация слитной речи	129
§ 7.1. Взаимосвязь задач распознавания и смысловой интерпретации речи. Генеративная модель смысловой интерпретации	129
§ 7.2. Задание информации о языках устного диалога	132
§ 7.3. Альтернативные пути решения задачи смысловой интерпретации	135
§ 7.4. Алгоритм смысловой интерпретации, основанный на речевой ориентированной семантической сети	137
§ 7.5. Двухэтапный алгоритм смысловой интерпретации	142
§ 7.6. Обобщенная задача распознавания слитной речи	142
§ 7.7. Алгоритм многозначной смысловой интерпретации	146
§ 7.8. Результаты моделирования	148
Глава 8. Проблема диктора в распознавании речи	153
§ 8.1. Способы проявления и учета индивидуальных особенностей голоса	153
§ 8.2. Взаимосвязь задач обучения распознаванию и настройки на голос оператора	154
§ 8.3. Настройка на голос оператора при поэлементном и пофонемном распознавании	155
§ 8.4. Постановка задачи создания многодикторных систем распознавания	162
§ 8.5. Кооперативное распознавание	163
§ 8.6. Проблема распознавания диктора по речевому сигналу	165
Глава 9. Исследования по первичному анализу речевых сигналов и выбору меры сходства	169
§ 9.1. Выбор интервала анализа и моделирование различных анализаторов речевых сигналов	169
§ 9.2. Сравнительный анализ различных описаний речевого сигнала и элементарных мер сходства	176
§ 9.3. Синтез табличной элементарной меры сходства	180
§ 9.4. Модели анализа речевого сигнала. Вычисление признаков тональности (признака тон—шум и периода основного тона)	182
§ 9.5. Общая математическая модель речевого сигнала. Необходимость в кусочно-линейных моделях	189
Глава 10. Низкоскоростные системы компрессированной передачи речи	194
§ 10.1. Постановка задачи	194
§ 10.2. Нуль-полюсные вокодеры на 2400 и 1200 бит/с	196
§ 10.3. Квазифонемный вокодер на 600 бит/с	200
§ 10.4. Результаты моделирования. Перспективы создания фонемного вокодера	204

Глава 11. Экспериментальные системы распознавания, смысловой интерпретации и компрессированной передачи речи	207
§ 11.1. Универсальный моделирующий стенд	207
§ 11.2. Экспериментальные системы 1966—1983 гг.	209
§ 11.3. Использование моделирующего стендса и экспериментальных систем	213
Глава 12. Проектирование систем распознавания. Системы речевого диалога	216
§ 12.1. Архитектура устройств и систем распознавания речи. Требования, предъявляемые к устройствам и системам	216
§ 12.2. Параллельная машина для распознавания слов и слитной речи	221
§ 12.3. Структура квазифонемного вокодера на 600 бит/с	223
§ 12.4. Необходимость систем речевого диалога, объединяющих функции распознавания и синтеза речи	225
§ 12.5. Разработка и применение систем речевого диалога серии «Речь»	226
Заключение	239
Приложения	245
Список литературы	255

ПРЕДИСЛОВИЕ

Проблема создания средств устного диалога человека с машинами является одной из наиболее актуальных проблем кибернетики, информатики и вычислительной техники. Оснащение ЭВМ средствами распознавания и синтеза речи имеет и в еще большей степени будет иметь огромное экономическое и социальное значение. Это обеспечит доступность ЭВМ всему населению, возможность программирования и решения задач на естественном языке, безбумажную технологию управления, сокращение сроков обучения пользователей ЭВМ, повышение производительности труда в сферах производства, распределения и в быту, повышение эффективности использования техники, создание благоприятных условий труда.

Речь идет, таким образом, о создании и применении интеллектуальных ЭВМ и человеко-машинного интерфейса на естественном языке прежде всего в системах САПР, АСУ, АСУТП, ИСС, ГАП, робототехнических комплексах, различного рода АРМах, персональных ЭВМ, системах контроля и испытаний техники и т. д.

Благодаря созданию таких машин высвободится значительная часть общественных сил и увеличится эффективность поиска оптимальных решений во всех областях народного хозяйства, повысится производительность труда в сферах распределения благ, обслуживания, управления.

История науки и техники насчитывает немало попыток создания «слушающих» и «говорящих» машин начиная еще с XVIII в. Этому в значительной мере способствовали становление и развитие электроники и электросвязи. Однако наибольший интерес к проблеме и ее развитие начинаются одновременно с появлением ЭВМ и их широким распространением, с автоматизацией различных областей деятельности человека.

Устный диалог человека с ЭВМ в наиболее удобной и привычной для человека форме — голосом — стал технико-экономической и социальной необходимостью. Чисто в научном плане конечной целью исследований является создание средств устного диалога человека и ЭВМ на естественных языках, например¹ автоматической машинки, печатающей и редактирующей тексты под диктовку, или машин-переводчиков с голоса.

Настоящая монография содержит результаты исследований, которые были выполнены автором и возглавляемой им группой исследователей в Институте кибернетики им. В. М. Глушкова АН УССР в последние 20 лет.

В книге представляется и развивается разработанный автором КДП-подход к распознаванию и смысловой интерпретации речи. В рамках этого подхода предлагаются

ся методы решения следующих основных задач: распознавания отдельно произносимых слов, распознавания слитной речи, составляемой из слов выбранного словаря, распознавания и смысловой интерпретации слитной речи применительно к устному диалогу человека и ЭВМ на формализованных или усеченных естественных языках. Расматриваются возникающие задачи обучения и самообучения распознаванию речи и обосновываются алгоритмы их решения. Вводятся новые понятия эталонного элемента, эталонного сигнала фонемы, эталонных сигналов слова и слитной речи, акустической, темпоральной, громкостной и тональной транскрипций слова и указана связь этих транскрипций с фонетической транскрипцией. В процессе сравнения распознаваемых сигналов с эталонными используются понятия об элементарном сходстве элементов речи и интегральном сходстве сигналов, потенциально-оптимальном слове и потенциально-оптимальном индексе. В рамках КДП-подхода предложены и исследованы конструктивные методы пофонемного распознавания речи. Показано, что при КДП-подходе реализация принципов пофонемного распознавания приводит к увеличению надежности распознавания и существенному снижению требований к объемам памяти и вычислений. При смысловой интерпретации речи применены понятия о типах смысла и типах предложения, о потенциально-оптимальных подсловарях. Уделено внимание предварительной обработке сигналов, выбору интервала анализа и способов описания речевых сигналов, обоснованию мер сходства. Решаются возникающие в рамках КДП-подхода вопросы учета индивидуальных особенностей голоса, настройки на голос оператора, распознавания диктора по голосу, создания многодикторных систем распознавания. Отдельная глава посвящена разработке низкоскоростных систем компрессированной передачи речи, основанных на распознавании элементов речи. Эти системы приобретают важное значение в связи с внедрением в практику низкоскоростных цифровых каналов связи. Приводятся результаты исследований экспериментальных систем распознавания и смысловой интерпретации речи, разработанных в рамках КДП-подхода. Изучаются вопросы проектирования систем распознавания и смысловой интерпретации речи и систем устного диалога, выбора архитектур этих систем, элементной базы. Обсуждаются проекты мульти микропроцессорной параллельной машины для распознавания слов и слитной речи и квазифонемного вокодера на 600 бит/с. Описываются разработка, освоение в производстве и применение систем речевого диалога (СРД) серии «Речь».

Наибольший интерес представляют разделы монографии, посвященные распознаванию и смысловой интерпретации слитной речи, пофонемному распознаванию речи, обучению (самообучению) распознаванию речи. Развитие этих работ позволило создать в Институте кибернетики им. В. М. Глушкова АН УССР одну из первых и лучших систем распознавания и смысловой интерпретации слов и слитной речи, оперирующую со сменным словарем предметной области объемом в 1000 слов. Эта часть работы получила в свое время высокую оценку академика В. М. Глушкова.

Монография характерна тем, что она рассматривает автоматическое распознавание и синтез речи как естественно-научную проблему. Само исследование проводится по следующей схеме. Сначала строится математическая модель сигналов классов, для чего используются сведения из теории речеобразования и восприятия речи и из теории преобразования речевых сигналов, используются знания о свойствах речевых сигналов. Затем формулируется кибернетическая модель обработки речевых сигналов с целью их анализа, распознавания, осмысления и синтеза. Следующие этапы связаны с моделированием разработанных моделей, созданием экспериментальных систем и проведением натурных экспериментов. Затем модели уточняются и весь цикл исследований повторяется.

Подход автора к распознаванию речевых сигналов получил всеобщее признание и распространение как в СССР, так и за рубежом. Фактически в лучших разработках в области распознавания речи в той или иной форме используется КДП-метод. Аналогичные методы в Советском Союзе, а затем в Японии, Франции, США, Англии, ФРГ, ГДР стали разрабатываться несколькими годами позже.

Думается, что представленный в монографии материал будет интересным как для специалистов в области автоматического распознавания и синтеза речевых сигналов, так и для тех, кто занимается проблемами цифровой обработки сигналов, распознавания образов, искусственного интеллекта, интеллектуальных ЭВМ, технической диагностики машин по излучаемым ими сигналам.

В. И. Скурихин

ВВЕДЕНИЕ

В создании интеллектуальных ЭВМ, интеллектуальных САПР, коллектического разума одна из главенствующей ролей отводится речевому общению (устному диалогу) человека и ЭВМ на естественном языке.

Хотя решение проблемы устного диалога человека и ЭВМ вне связи с развитием ЭВМ и имеет самостоятельное значение, все же объективно эта проблема будет решаться синхронно, одновременно с созданием машин логического вывода и принятия решений, работающих с базами знаний.

Типичными примерами применения средств устного диалога могут быть следующие:

а) управление работой графического дисплея и граffопостроителя при автоматизированном проектировании или автоматизированной обработке изображений;

б) составление паспортов изделий при контрольно-измерительных испытаниях (человек делает устный комментарий показаний приборов в ходе эксперимента);

в) сбор данных с рабочих мест и управление в АСУ и АСУТП;

г) устный запрос в ИПС, в том числе посредством телефона;

д) управление машинами и механизмами на производстве и в быту (включение-выключение станка, управление работой телевизора, сортировка посылок, управление в конвейерном производстве и т. п.).

Отметим, что устный диалог человека с ЭВМ возможен как на естественном, так и на искусственном языках со строгими правилами следования слов.

Сегодня мы далеки от решения проблемы устного диалога человека и ЭВМ на естественном языке в полном объеме.

Анализ результатов, полученных в СССР и зарубежных странах, позволяет утверждать, что в плане автоматического распознавания и смысловой интерпретации речи актуальными, как и 15 лет тому назад, однако разрешимыми в ближайшие пять лет, являются следующие задачи:

1) распознавание отдельно произносимых слов (объем словаря до 1000 слов);

2) распознавание слитной речи, составляемой из слов выбранного словаря (объем словаря до 1000 слов);

3) распознавание и смысловая интерпретация слитной речи для устного диалога человека и ЭВМ на формализованных и усеченных естественных языках предметных областей (объем словаря до 1000 слов).

В последующие пять лет сохранятся эти же задачи, однако существенно увеличится объем словаря — до 10 000 слов и более.

Автору удалось разработать метод распознавания речевых сигналов, развитие которого привело к решению основных задач обработки речевых сигналов: распознавания отдельно произносимых слов речи, распознавания слитной речи, составляемой из слов выбранного словаря, смысловой интерпретации слитной речи применительно к устному диалогу человека и ЭВМ на формализованных и (или) естественных языках предметных областей, пофонемного распознавания речи, обучения и самообучения распознаванию речи, оптимальной сегментации речевых сигналов, настройки системы распознавания речи на голос диктора, многомерного квантования речевых сигналов.

Представляемый метод, названный КДП-методом, основан на экономном задании (составлении, композиции (К)) разнообразных и изменяющихся так называемых эталонных речевых сигналов с помощью автоматных порождающих грамматик и на применении динамического программирования (ДП) для направленного поиска и разбора эталонного сигнала речи, наиболее правдоподобного по отношению к распознаваемому речевому сигналу. Являясь по своему существу конструктивным воплощением идеи анализа сигналов посредством их синтеза в цепи обратной связи, этот метод позволил рассмотреть различные задачи анализа, распознавания, смысловой интерпретации речевых сигналов с некоторых единых позиций и указать конструктивные пути решения этих задач.

Разработка КДП-метода была начата еще в 1966 г. [1—4]. Этот метод, получивший распространение под названием ДП-метод, стал позже использоваться в различных модификациях в других организациях СССР [5—6], а затем в Японии [7], Франции [8], США [9, 10], ГДР [11]. Однако наиболее полная разработка ДП-метода была выполнена в рамках КДП-подхода. Это — распространение метода на распознавание слитной речи и смысловую интерпретацию слитной речи, реализация принципов пофонемного распознавания, обучения и самообучения распознаванию, создание действующих экспериментальных систем, разработка и внедрение систем распознавания речи. Так, метод распознавания слитной речи, составляемой из слов выбранного словаря [12—13], предложенный еще в 1970—1971 гг., и на сегодняшний день превосходит по быстродействию и надежности распознавания другие, последовавшие затем, ДП-методы [14—15]. Действующие системы пофонемного распознавания и смысловой интерпретации слитной речи, разработанные на основе КДП-подхода [16—18], по-прежнему остаются единственными в СССР и одними из немногих в мире [19—21].

Распознавание речи с помощью динамического программирования наиболее успешно разрабатывалось в СССР, Японии, США, Франции, ГДР. Именно в этих странах и созданы лучшие экспериментальные,

исследовательские и промышленные системы распознавания и смысловой интерпретации речи [20—25].

Другое направление, объединяющее значительную часть исследований как в СССР, так и особенно за рубежом, основано на признании иерархического (И) принципа переработки информации и на введении многозначных (МЗ) решений на всех уровнях этой переработки [19, 20, 26]. Для краткости этот подход назовем ИМЗ-подходом. На первом уровне вводится многозначная сегментация речевого сигнала на части (сегменты), соответствующие фонемам или их фазам. Затем, на втором уровне, осуществляется многозначное распознавание сегментов как фонем или фаз фонем. Далее, на третьем уровне, осуществляется переход от многозначных фонемных решений к многозначным словесным решениям. На четвертом уровне вырабатываются многозначные решения о последовательностях слов, передаваемых речевым сигналом. Наконец, на последнем, пятом, уровне с использованием синтаксиса, семантики и прагматики языка диалога принимается окончательное решение о последовательности слов и смысле, передаваемых речевым сигналом.

Типичными работами, выполненными в рамках ИМЗ-подхода, являются система KEAL (1978, Франция) [20], основанная на использовании сильных синтаксических ограничений в языке диалога из 120 слов, и особенно система HARPY (1976, США, Университет Carnegie — Mellon) [19]. Последняя считается наиболее удачной среди систем смысловой интерпретации, разработанных в США в рамках проекта ARPA (системы HWIM, HEARSAV, HARPY и др.). Надежность смысловой интерпретации в системе HARPY в условиях сильных синтаксических ограничений (коэффициент ветвления равен 33) — около 95 %.

Интересно, что, как в системах KEAL и HARPY, так и в других, разработанных в рамках ИМЗ-подхода, на отдельных этапах принятия решений используется динамическое программирование. В дальнейшем будет показано, что многозначные решения могут использоваться и в рамках КДП-подхода. Вообще же, КДП- и ИМЗ-подходы представляются теоретически эквивалентными по своим возможностям в распознавании речевых сигналов.

Решение задач в рамках ИМЗ-подхода характеризуется значительной трудоемкостью. Так, в системе HARPY, реализованной на ЭВМ типа PDP, распознавание выполняется в замедленном (в 80 раз) масштабе времени.

В целом же оказалось, что разработки КДП-подхода являются более подготовленными для реализации в системах распознавания слов и слитной речи и системах смысловой интерпретации. Объясняется это относительной простотой перехода от распознавания элементов речи к распознаванию слов и, далее, к распознаванию слитной речи. Что же касается ИМЗ-подхода, то здесь по-прежнему остаются недостаточно отработанными нижние, акустические, уровни, связанные с сегментацией речевого сигнала и фонемным распознаванием, имеет место увеличение верхними, лексико-семантическими, уровнями, которые не позволяют устраниТЬ те фатальные ошибки распознавания, которые возникают на нижних уровнях из-за их несовершенства.

Из других подходов, предлагаемых для решения основных задач распознавания и смысловой интерпретации, заслуживает внимания бионический подход, поскольку он характерен для ряда исследований. Основан он на моделях восприятия речи человеком, на попытках моделировать некоторые функции человека. Например, предполагается, что осмыслившее восприятие речевого сигнала — это активное выдвижение гипотез интеллектом распознающей системы по законам грамматики и семантики предметной области и последующие их проверка, отбрасывание и верификация [27--28]. Нужно сказать, что бионический подход основан на формулировке некоторых весьма общих принципов обработки информации, которые, предположительно, реализуются человеком. В целом этот подход характеризуется как малоконструктивный. Конкретно выполненные работы не пошли дальше распознавания отдельно произносимых слов, да и то в этом последнем случае часто применяются приемы, аналогичные используемым в КДП- или ИМЗ-подходах.

Можно еще упомянуть аппаратурно-программный подход в распознавании речевых сигналов [29]. Однако представляется, что такое название подхода не отражает существа дела, оно лишь указывает способ реализации предлагаемых методов, а именно: с помощью вычислительной техники, оснащенной соответствующими техническими и программными средствами. Но по такому способу создаются все системы обработки информации, в том числе речевой, независимо от того, в рамках какого подхода они разработаны.

Таким образом, конструктивных подходов, приемлемых для решения всех задач основного перечня, не так уж много. Зато существует огромное количество работ, которые не идут дальше распознавания отдельно произносимых слов, в основном по той причине, что не ясно, как используемые методы распространить дальше на распознавание и смысловую интерпретацию слитной речи.

В СССР наиболее весомый вклад в разработку методов распознавания отдельно произносимых устных команд внесли исследовательские группы под руководством В. М. Величко и Н. Г. Загоруйко [30], Т. К. Винценко [31], В. И. Галунова [32], С. В. Голубцова [33], М. Ф. Деркача [34], А. Г. Какауридзе [35], Г. М. Петрова [36], Г. С. Рамишвили [37], Г. С. Слуцкера [38], В. Н. Трунина-Донского [39], Г. Д. Фролова [40], Г. И. Цемеля [41]. Из зарубежных наиболее известны работы, возглавляемые Мартином (T. B. Martin) [42] в США, Сакоэ и Чиба [43] в Японии, Мерсье (G. Mercier) [44] и Лиенаром (J. S. Lienard) [45] во Франции и др. Все же наиболее результативными разработками по распознаванию слов оказались те, которые были выполнены в рамках КДП-подхода, в несколько меньшей степени — в рамках ИМЗ-подхода. Другие конкурентоспособные методы оперируют с описанием слова в целом, используют так называемую временную нормализацию описаний слова либо выделяют характерные признаки в речевом сигнале, в пространстве которых строятся логические решающие правила.

При всех успехах в распознавании отдельно произносимых слов по-прежнему остаются актуальными проблемы повышения надежнос-

ти распознавания, увеличения объема словаря, обучения распознаванию слов, учета индивидуальных особенностей голоса, упрощения алгоритмов.

Гораздо хуже обстоят дела с распознаванием слитной речи. Между тем ввод данных в ЭВМ слитной речью в два — пять раз производительнее пословного. Именно слитная речь наиболее свойственна человеку, удобна для общения с ЭВМ. Наибольшие успехи в распознавании слитной речи получены в рамках КДП-подхода.

В практическом плане проблема распознавания слитной речи подчинена третьей проблеме, имеющей наибольшую значимость. Это — распознавание и смысловая интерпретация слитной речи для устного диалога человека и ЭВМ на формализованных и (или) естественных языках предметных областей. Проблема же распознавания слитной речи выносится отдельно в силу ее фундаментальной значимости для решения указанной главной проблемы, с решением которой раскрывается полная эффективность речевого ввода информации.

Между тем нужно отметить, что с решением проблемы распознавания и смысловой интерпретации слитной речи не все обстоит благополучно. Разрабатывается слишком мало проектов и экспериментальных систем. Конечно, проблема эта необычайно сложна. По сравнению с отдельно произносимыми словами дополнительно необходимо учитывать вариативность слов в слитной речи под влиянием соседних слов, синтагматического и фразового ударений, учитывать изменяемость темповых и мелодических характеристик речи. Важную роль приобретают проблемы создания языковых моделей предметных областей, быстрой перестройки языковой модели на новую предметную область, построения семантико-сintаксических сетей. Центральное место в этих проблемах отводится задаче экономного задания множеств предложений естественного языка, выражаяющих один и тот же смысл.

Проблема смысловой интерпретации слитной речи существенно отличается от смысловой интерпретации текста. В речи нет заглавных букв, точек, запятых, пробелов (пауз) между словами. Зато содержится много мешающей информации о голосе человека, интонации, функциональном и эмоциональном состояниях человека. Нельзя представлять дело так, что задачу смысловой интерпретации слитной речи можно решить в два этапа. На первом этапе распознать слитную речь, т. е. указать последовательность слов, которая передается речевым сигналом, не прибегая при этом к синтаксису, семантике и прагматике, а ограничиваясь только лексикой языка. Затем, на втором этапе, используя синтаксис, семантику и прагматику, скорректировать возможные ошибки в распознавании отдельных слов слитной речи и представить передаваемый смысл в определенной канонической форме, удобной для последующего использования в ЭВМ. Ясно, что в таком подходе будем иметь дело с сильно искаженной последовательностью слов и для ее восстановления уже не хватит априорной информации о языке, используемой на втором этапе. Легко видеть и то, что семантико-сintаксическая сеть в случае анализа речи должна быть более мощной и действенной, чем в случае анализа неискаженных текстов.

Методологические задачи распознавания и смысловой интерпретации слитной речи должны решаться в едином взаимосвязанном процессе, что в принципе достигается в КДП- и ИМЗ-подходах.

В СССР в области распознавания и смысловой интерпретации слитной речи предпринимаются определенные усилия. Здесь можно выделить исследовательские группы под руководством Т. К. Винценко [46], В. И. Галунова [47], М. Ф. Деркача [48], Н. Г. Загоруйко и В. М. Величко [49], Г. М. Петрова [6], А. Н. Петрова [50], В. Н. Трунина-Донского [51]. За рубежом наиболее результативно работают группы Лиенара (J. S. Lienard) [52], Мерсье (G. Mercier) [44], Редди (D. R. Reddy) [53], Сакоэ и Чиба (H. Sakoe and S. Chiba) [54]. Наибольшие успехи достигнуты в работах, выполненных в рамках КДП-подхода, более скромные результаты получены в ИМЗ-подходе.

В настоящее время в СССР пока нет массового выпуска и применения устройств и систем распознавания, смысловой интерпретации и синтеза речи. Однако подготовительный период уже пройден, и, по крайней мере, несколько систем распознавания слов и слитной речи и синтеза речи либо производятся мелкими сериями, либо осваиваются в производстве. За рубежом получил распространение целый ряд устройств распознавания устных команд, слитной речи, синтеза речевых сигналов.

Выполнение технико-экономического и социального заказа на системы распознавания и смысловой интерпретации речи сдерживается не только трудностями принципиального характера, обусловленными сложностью речевой коммуникации, но и чисто инженерными проблемами. Для реализации методов распознавания и смысловой интерпретации требуется значительное быстродействие вычислительной техники. Возникает проблема распараллеливания вычислений, выбора определенной архитектуры вычислительных средств. Создание мульти микропроцессорных систем распознавания и смысловой интерпретации речи с возможностями обучения (настройки) на желаемый словарь, голос человека, на ту или иную предметную область должно стать основным направлением разработок.

Важным свойством систем устного диалога является объединение функций распознавания и синтеза речи в одной системе. Благодаря этому достигается новое качество — нечто большее, чем простая сумма распознавания и синтеза речи. Необходимость разработки и применения именно систем устного диалога, как обеспечивающих двустороннее взаимодействие человек — ЭВМ, отстаивается в монографии.

Примыкающей проблемой является распознавание и верификация дикторов по речевому сигналу. Эта проблема рассматривается как в связи с ограничениями на доступ лиц к системам устного диалога и системам обработки информации и управления, так и в связи с обучением и настройкой систем распознавания на словарь и голоса дикторов.

Общим методом представляемого в монографии исследования явилось: 1) построение математических моделей речевых сигналов, использование для этих целей теории порождающих грамматик и теории графов, знаний о свойствах речевых сигналов, теории их преобразования;

2) применение статистических процедур принятия решений, основанных на математическом программировании; 3) натурное моделирование с помощью ЭВМ процессов обработки речевой информации; 4) разработка и исследование систем распознавания, смысловой интерпретации, компрессированной передачи, синтеза речевых сигналов и систем речевого диалога для выработки рекомендаций по их проектированию и применению.

ГЛАВА 1

МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ КОМПОЗИЦИОННО-ОПТИМАЛЬНОГО ПОДХОДА (КДП-ПОДХОДА) К РАСПОЗНАВАНИЮ И СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ РЕЧИ

В данной главе излагается методология КДП-подхода к распознаванию и смысловой интерпретации речевых сигналов, основанного на составлении (композиции (К)) эталонных сигналов из элементарных частей и на сравнении их с распознаваемым сигналом с помощью динамического программирования (ДП). Работы в рамках этого подхода, начатые в 1966 г. [1—4], привели к решению основных задач анализа, распознавания и смысловой интерпретации речевых сигналов.

§ 1.1. СУЩНОСТЬ И ОСОБЕННОСТИ ПРОБЛЕМЫ

Чтобы обеспечить устный диалог человека и ЭВМ, необходимо разработать методы и средства распознавания и смысловой интерпретации речевых сигналов.

Распознавание речи — это процесс автоматической обработки речевого сигнала с целью указания последовательности слов, кою оная передается этим сигналом.

Аналогично смысловая интерпретация речи — это процесс автоматической обработки речевого сигнала с целью указания смысла, передаваемого речевым сигналом, и представления этого смысла в определенной канонической форме, удобной для последующего использования. Очевидно, что смысловая интерпретация речи является более высокой ступенью обобщения информации, чем распознавание, поскольку одну и ту же мысль можно выразить различными последовательностями слов. Поэтому распознавание речи будем рассматривать как проблему, подчиненную смысловой интерпретации.

Распознавание речи использует знания об акустике, фонологии, фонетике и лексике языка устного диалога, в то время как смысловая интерпретация речи дополнительно предполагает использование синтаксиса, семантики и прагматики предметной области. Конечно, надежное распознавание речи не может быть достигнуто без привлечения семантико-синтаксической информации, однако уж если эта информация используется для распознавания, то почему бы одновременно с ним не осуществить и смысловую интерпретацию.

Распознавание и смысловая интерпретация речи должны выполняться в едином взаимосвязанном процессе, при котором одновременно достигаются наилучшие результаты как распознавания, так и смысловой интерпретации речи. Однако в дальнейшем, чтобы не отождествлять эти два понятия, условимся, что распознавание речи не идет дальше лексических ограничений и что, таким образом, распознавание и смысловая интерпретация — это разные процессы обработки информации.

Приведенные определения распознавания и смысловой интерпретации речи не используют понятий фонемы и слога, пофонемного и послогового распознавания. Это обусловлено тем, что конечной целью распознавания и интерпретации речи является смысл сообщения, который скорее передается не последовательностью фонем или слогов, а последовательностью слов. Это дает основание рассматривать задачи пофонемного или послогового распознавания речи как подчиненные (вспомогательные, промежуточные) по отношению к распознаванию последовательностей слов и смысловой интерпретации речи.

В дальнейшем будут изучаться следующие проблемы, имеющие самостоятельное научное и практическое значение:

- 1) распознавание отдельно произносимых слов устной речи;
- 2) распознавание слитной речи, составляемой из слов выбранного словаря;
- 3) распознавание и смысловая интерпретация слитной речи для устного диалога человека и ЭВМ на формализованных или усеченных естественных языках.

Таким образом, классами распознаваемых сигналов (искомыми параметрами при распознавании и смысловой интерпретации речи) выступают слова, последовательности слов и передаваемый смысл.

Чтобы решать задачи распознавания, необходимо сначала тем или иным способом задать (описать) множества сигналов, соответствующих классам распознаваемых сигналов, а затем распознаваемый сигнал отнести к тому классу, множеству сигналов которого этот сигнал принадлежит.

Этот универсальный прием решения задач распознавания сигналов является, однако, весьма неконструктивным. Требуется указать конкретные способы задания множеств сигналов классов и конкретные способы проверки предъявляемого сигнала на принадлежность этим множествам.

Может показаться приемлемым следующий универсальный способ решения задач распознавания. Множества сигналов классов задавать простым перечислением (запоминанием) всех возможных сигналов, а проверку на принадлежность множеству осуществлять путем проверки на совпадение распознаваемого сигнала с ранее запомненными.

Однако такой, казалось бы, действенный, прием решения задач распознавания является не реализуемым. Во-первых, не найдется такой памяти ЭВМ ни теперь, ни в будущем, которая позволила бы запомнить все возможные речевые сигналы отдельных классов. Во-вторых, даже если такая память станет возможной, то все равно не найдется такой быстродействующей ЭВМ ни теперь, ни в будущем,

ни последовательной, ни параллельной, которая была бы в состоянии сравнить в реальном времени такое огромное количество сигналов.

Принципиально ничего не изменится, если запоминать не все, а наиболее характерные речевые сигналы классов, и при распознавании производить не сравнение на точное совпадение, а находить наиболее похожие на распознаваемый характерные сигналы классов. Объясняется это все тем же огромным разнообразием даже характерных сигналов классов.

Значит, должны быть найдены приемы, которые позволяют преодолеть возникающие проблемы памяти и вычислений.

Речевые сигналы классов, например одного и того же слова или предложения, характеризуются чрезвычайным разнообразием и изменчивостью. Они зависят не только от языка, диалектных особенностей, что очевидно, а и от индивидуальности голоса, функционального и эмоционального состояний говорящего, способа и манеры, темпа и громкости произнесения, причем темп и интенсивность произнесения изменяются нелинейно во времени. Речевые сигналы классов изменяются под влиянием соседних сигналов в последовательностях. Так, речевые сигналы фонем подвергаются изменениям вследствие явлений коартикуляции и редукции, а сигналы слов варьируются под влиянием внутрисинтагматических и фразовых ударений. Наблюдается также изменяемость под влиянием интонации: перечисления, обращения, завершенности, незавершенности, вопроса, восклицания и т. п. В довершение отметим, что даже двум подряд произнесениям одного и того же слова одним и тем же диктором соответствуют всегда разные сигналы.

В речевом сигнале содержится информация не только о том, что сказано, а и о том, кто говорит, каково его состояние, каков темп речи и т. п. Вся эта информация выступает как избыточная и мешающая по отношению к информации о том, что говорится.

Большое разнообразие и изменяемость сигналов определяют основные трудности в распознавании речи.

Пусть нереализуемой, но все же привлекательной и единственной кажется схема распознавания и смысловой интерпретации, основанная на запоминании всех возможных сигналов классов и на сравнении этих сигналов с распознаваемым. Можно попытаться обойти возникающие трудности по объемам памяти и вычислений, например, множества сигналов классов описывать некоторыми экономными по памяти средствами, а сравнение сигналов производить не полным перебором, а средствами направленного поиска.

Конкретизация этого замысла приводит к методу распознавания и смысловой интерпретации речевых сигналов, названному КДП-подходом.

§ 1.2. ОСНОВЫ КДП-ПОДХОДА

Возможность экономного задания сигналов классов основана на предположении о том, что эти сигналы носят далеко не случайный характер, что они связаны сильными детерминированными зависимо-

стями, определяемыми не очень большим числом варьируемых параметров, что изменчивость и разнообразие сигналов поясняются определенными закономерностями их преобразования.

Из этого следует, что принципиально возможно создание экономных структур, позволяющих генерировать разнообразные сигналы, аппроксимирующие с той или иной степенью точности реальные множества сигналов классов. Чтобы сгенерированные с помощью упомянутых структур сигналы отличать от наблюдаемых, подлежащих распознаванию, в дальнейшем будем называть их эталонными (модельными, прототипными).

Пусть нам удалось тем или иным способом экономно задать множества сигналов (точнее, множества эталонных сигналов) классов. Тогда возникает проблема эффективного направленного поиска эталонного сигнала, являющегося наиболее правдоподобным по отношению к распознаваемому, с одновременным указанием класса, которому этот наиболее правдоподобный эталонный сигнал принадлежит.

Выбор процедуры направленного поиска и ее эффективность в значительной мере определяются способом описания множеств эталонных сигналов, к которым, в свою очередь, кроме экономности задания, предъявляются требования адекватного соответствия реальным множествам сигналов. Оказывается, что не для всяких способов задания множеств могут быть указаны процедуры направленного поиска. Поэтому выбирать экономный способ задания множеств эталонных сигналов приходится с оглядкой на то, найдется ли для него приемлемая процедура направленного поиска.

Компромиссом, удовлетворяющим требованиям адекватности, экономности задания и поиска, является КДП-подход. Он основан на экономическом задании множеств эталонных сигналов классов с помощью автоматных порождающих грамматик, составляющих (синтезирующих) эталонные сигналы из элементарных частей, представляющих фонемы или их фазы и в дальнейшем называемых эталонными элементами. Сравнение же наблюдаемого речевого сигнала с эталонными и поиск (вместе с разбором-анализом) наиболее правдоподобного эталонного сигнала осуществляются с помощью процедур динамического программирования.

При КДП-подходе реализуется распознавание речевых сигналов путем направленного синтеза эталонных сигналов речи.

В дальнейшем будет показано, как в рамках КДП-подхода учитываются основные детерминированные закономерности изменяемости речевых сигналов, обусловленные явлениями коартикуляции и редукции, нелинейного изменения темпа и интенсивности произнесения, индивидуальными особенностями голоса, и как учет этих закономерностей приводит к решению задач распознавания и смысловой интерпретации речи.

В рамках КДП-подхода возникают новые понятия эталонного элемента, эталонного сигнала фонемы, слова и слитной речи, акустической, темпоральной, громкостной и тональной транскрипций слова, согласованных с понятием фонетической транскрипции слова, новые понятия элементарного и интегрального сходств распознаваемого и эталонного

сигналов, понятия типов предложений и типов смысла. Все эти понятия возникают естественным образом в процессе постановок и решения задач распознавания и смысловой интерпретации.

§ 1.3. МАТЕМАТИЧЕСКИЕ МОДЕЛИ РЕЧЕВЫХ СИГНАЛОВ И ИХ СВОЙСТВА

В формулировках методов распознавания и смысловой интерпретации речи в рамках КДП-подхода главенствующая роль отводится средствам экономного задания множеств сигналов классов и установлению связей между эталонными и наблюдаемыми сигналами. Упомянутые средства и связи устанавливаются математической моделью речевых сигналов.

В математической модели речевого сигнала прежде всего должны быть отражены детерминированные закономерности, которыми связаны наблюдаемые сигналы.

Обозначим через X наблюдаемый сигнал и будем исходить из следующей модели речевого сигнала:

$$X = f(k, a, d, \tau, h, \omega, \dots) + R, \quad (1.3.1)$$

где k — искомый параметр (слово, последовательность слов или смысл), который должен быть оценен при распознавании; a — параметры, которые являются неизменными, постоянными в конкретной задаче распознавания или смысловой интерпретации, например, набор эталонных элементов, эталонных сигналов фонем, транскрипции слов и т. п.; $d, \tau, h, \omega, \dots$ — параметры, которые при фиксированных k и a могут менять свое значение от произнесения к произнесению, от одной реализации X к другой (Пусть, для определенности, d, τ, h и ω представляют индивидуальность голоса, темп, интенсивность и тональность произнесения соответственно. Эти параметры выступают как мешающие по отношению к искомым параметрам.); $f(k, a, d, \tau, h, \omega, \dots)$ — функция, устанавливающая детерминированную зависимость эталонных сигналов $E = f(k, a, d, \tau, h, \omega, \dots)$ от искомых, постоянных и меняющихся от произнесения к произнесению параметров. (Именно эта функция выражает закономерности коартикуляции и редукции, темпа и интенсивности произнесения, индивидуальных особенностей речи.); R — шум с постулируемым распределением $p(R)$, посредством которого объясняются возможные несовпадения наблюдаемых сигналов X с эталонными сигналами E , на шум R обычно списываются все неточности модели.

В математической модели сигнала не очень много мешающих параметров и тем не менее их изменением можно пояснить все разнообразие и изменчивость речевых сигналов.

В модели главенствующая роль принадлежит детерминированной части. Именно благодаря ей все возможные сигналы классов представляются как такие, которые образуют маломерные, зависящие от малого числа параметров, многообразия в пространстве сигналов X . Детерминированная часть задает как бы гиперповерхность регрессии для этого многообразия. Статистический же характер модели проявляется

лишь в том, что гиперповерхность регрессии как бы окружается облаком определенной «толщины». Структура же этого облака по-прежнему определена детерминированной частью. Сказанное иллюстрируется рис. 1.1.

Допустим, что нам удалось тем или иным способом построить математическую модель в виде (1.3.1). Следующий важный момент: как пользоваться моделью в процессе принятия решения об исскомом параметре k на основании предъявляемого сигнала \mathbf{X} .

Теория статистических решений [55—60] в условиях мешающих параметров предполагает принятие решения по максимуму апостериорной вероятности $p(k|\mathbf{X})$, вычисляемой по формуле Бейеса с учетом интегрирования по мешающим параметрам. К сожалению, мы вынуждены констатировать, что ни априорные вероятности классов, ни распределения мешающих параметров не являются известными либо не существуют вовсе. В этих условиях, считая, что основной вклад в решение должна давать детерминированная часть модели, будем придерживаться метода наибольшего правдоподобия в некоторой его наиболее свободной трактовке, т. е. решение будем принимать по формуле

$$k(\mathbf{X}) = \operatorname{argmax}_k \max_{\mathbf{d}, \tau, h, \omega, \dots} p(\mathbf{X}/f(k, \mathbf{a}, \mathbf{d}, \tau, h, \omega, \dots)), \quad (1.3.2)$$

где $p(\mathbf{X}/\mathbf{E})$ — вероятность наблюдения \mathbf{X} при условии эталонного сигнала \mathbf{E} .

В обычном методе правдоподобия нам бы пришлось применить операцию интегрирования по мешающим параметрам $\mathbf{d}, \tau, h, \omega, \dots$. Заменив операцию интегрирования операцией максимизации, мы избежим использования неизвестных или несуществующих распределений мешающих параметров. Однако в этом случае получается, что искомыми (оцениваемыми в процессе распознавания) параметрами становятся не только класс k , которому сигнал \mathbf{X} принадлежит, а и мешающие параметры $\mathbf{d}, \tau, h, \omega$. Как будет показано в дальнейшем, возникающая таким образом комплексная задача распознавания имеет вполне оправданный содержательный смысл, когда одновременно с указанием класса $k(\mathbf{X})$ указывается диктор \mathbf{d} , произнесший сигнал \mathbf{X} , темп τ , громкость h , тональность ω , состояние человека и другие параметры, в условиях которых анализируемый сигнал \mathbf{X} был произнесен.

Таким образом, согласно критерию (1.3.2) распознавание сигнала \mathbf{X} определяется наиболее правдоподобным для него эталонным сигналом $f(k, \mathbf{a}, \mathbf{d}, \tau, h, \omega, \dots)$.

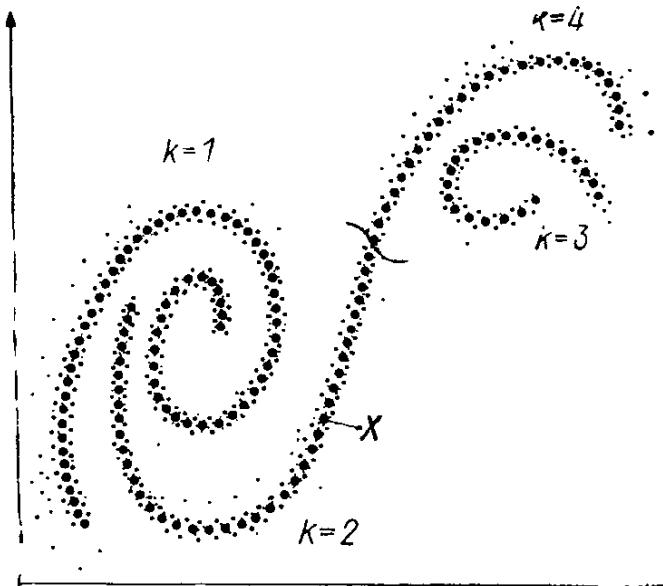


Рис. 1.1. Предполагаемая структура множеств речевых сигналов классов.

Чтобы воспользоваться математической моделью речевых сигналов 1.3.1 и методом распознавания (1.3.2), необходимо указать конкретный способ задания эталонных сигналов $f(k, a, d, \tau, h, \omega, \dots)$ и конкретный способ вычисления максимального сходства

$$G_k(X) = \max_{d, \tau, h, \omega, \dots} G(X, f(k, a, d, \tau, h, \omega, \dots)), \quad (1.3.3)$$

где

$$G(X, f(k, a, d, \tau, h, \omega, \dots)) = \ln p(X/f(k, a, d, \tau, h, \omega, \dots)) - \quad (1.3.4)$$

величина, характеризующая сходство распознаваемого X и эталонного $f(k, a, d, \tau, h, \omega, \dots)$ сигналов. При постулировании сферически-симметричного распределения $p(R)$ сходство $G(X, f(k, a, d, \tau, h, \omega, \dots))$ выражается через евклидово расстояние сигналов X и $f(k, a, d, \tau, h, \omega, \dots)$, взятое со знаком минус.

Даже располагая математической моделью речевого сигнала, не так уж просто указать конструктивные способы задания эталонных сигналов и вычисления максимального сходства.

В КДП-подходе, исходя из математических моделей речевого сигнала, множества эталонных сигналов задаются с помощью автоматных порождающих грамматик, составляющих эталонные сигналы из эталонных элементов, а распределения $p(R)$ постулируются такими, чтобы сходство сигналов выражалось аддитивно через элементарные сходства наблюдаемых и эталонных элементов, что делает возможным применение динамического программирования для нахождения максимальных сходств.

Итак, чтобы решать задачи распознавания и смысловой интерпретации, необходимо располагать математической моделью речевого сигнала.

В общем виде математическая модель должна описывать параметрический процесс порождения речевых сигналов на выходе микрофона, отражающий разнообразие и основные детерминированные закономерности преобразования речевых сигналов, обусловленные особенностями языка и его диалектов, фонемного и словарного составов, индивидуальными и эмоциональными особенностями, нелинейными изменениями темпа и интенсивности произнесения, явлениями коартикуляции и редукции, просодическими явлениями и т. п. Такая модель должна отражать процессы речеобразования и не противоречить известным данным по восприятию речи человеком. Она должна также учитывать акустику помещения и преобразовательные свойства микрофона. Необходимо, чтобы модель была наглядной, стохастической, в которой главенствующая часть выражается детерминированными зависимостями.

В настоящее время общие математические модели речевого сигнала еще не найдены. Формулируются лишь частные модели, ориентированные на постановку и решение частных задач распознавания речи. Примерами таких задач являются распознавание отдельно произносимых слов для одного диктора, распознавание диктора по парольной фразе, определение функционального состояния одного данного человека по его голосу.

В случае распознавания того, что говорится, как правило, имеют дело с иерархией моделей, в основном двух моделей: 1) на уровне предварительной обработки; 2) на уровне собственно распознавания.

Модели речевого сигнала на уровне предварительной обработки используются для получения так называемого описания речевого сигнала — представления речевого сигнала в виде набора значений признаков, сохраняющих информацию о передаваемом речевым сигналом смысле. Эти признаки чаще всего представляют передаточную характеристику речевого тракта и характеристики источников его возбуждения.

Модели речевых сигналов на уровне собственно распознавания задают множества описаний речевых сигналов классов, тех описаний, которые были получены на уровне предварительной обработки.

Иерархия моделей объясняется сложной природой речевого сигнала.

При построении математических моделей на уровне собственно распознавания в рамках КДП-подхода мы будем идти по пути постепенного их усложнения, отправляясь от простейших базовых моделей. Сначала для фиксированных микрофона, помещения, диктора изучим разнообразие описаний отдельных слов. Затем перейдем к слитной речи для одного диктора, затем к речи многих дикторов, к смысловой интерпретации и т. д.

Путь же к построению математических моделей на любом уровне подробности подобен обычному естественно-научному исследованию: изучение свойств и закономерностей преобразований речевых сигналов, выдвижение гипотез, обращение к эксперименту, уточнение гипотез.

При всем разнообразии факторов, определяющих изменчивость речевых сигналов, есть один наиболее специфичный фактор, в дальнейшем положенный в основу построения всех математических моделей речевого сигнала — нелинейное изменение темпа произнесения.

§ 1.4. НЕПРЕРЫВНОСТЬ СИГНАЛОВ И ДИСКРЕТНОСТЬ РЕШЕНИЙ

Исходный речевой сигнал и его описание по существу представляются непрерывными функциями времени, тогда как принятие решений характеризуется явно выраженной дискретностью — на основании предъявленного сигнала необходимо ответить на вопрос, какая последовательность фонем, слогов или слов или какой смысл передаются сигналом. В связи с этим возникает вопрос о границах фонем, слогов, слов, предложений в непрерывном речевом сигнале, о сегментации речевого сигнала на участки, соответствующие предложениям, синтагмам, словам, слогам, фонемам.

В рамках КДП-подхода исходя из того, что границы между распознаваемыми фонемами, слогами, синтагмами и предложениями выражены нечетко и носят условный характер, вообще исключается такой прием, как предварительная сегментация речевого сигнала на участки, соответствующие предложениям, словам, фонемам, и последующее

дискретное распознавание этих участков. В КДП-подходе сегментация непрерывного речевого сигнала и дискретное распознавание осуществляются в едином взаимосвязанном процессе, в котором из эталонных сигналов фонем и слов составляется наиболее правдоподобный эталонный сигнал слитной речи и решение о последовательности дискретных элементов, переданных непрерывным речевым сигналом, осуществляется на основе анализа этого наиболее правдоподобного эталонного сигнала, с одновременным указанием, если необходимо, и границ дискретных элементов в непрерывном речевом сигнале. Таким образом, сегментация речевого сигнала в КДП-подходе не предшествует собственно процессу распознавания, а осуществляется одновременно с ним и является одним из результатов, сопутствующих распознаванию. Получаемая таким образом сегментация является наилучшей в том смысле, что она соответствует наиболее правдоподобному эталонному сигналу речи. Все это — следствие анализа (распознавания) предъявленного речевого сигнала через синтез эталонных сигналов.

Отказ от сегментации речевого сигнала, предшествующей распознаванию, и включение ее в единый с распознаванием процесс — одно из достоинств КДП-подхода.

§ 1.5. НЕОБХОДИМОСТЬ МОДЕЛИРОВАНИЯ

Как бы тщательно ни были обоснованы математические модели речевого сигнала, окончательный ответ об адекватности модели, ее пригодности и вытекающего из нее метода распознавания и смысловой интерпретации ставится в зависимость от эксперимента и результатов по надежности распознавания и смысловой интерпретации речи.

Распознавание и смысловая интерпретация речи — это естественно-научная проблема, которая решается в условиях определенных предпосылок и предположений. Выводы относительно пригодности тех или иных методов распознавания, вытекающих из этих предположений, должны следовать из результатов натурного моделирования этих методов.

Натурное моделирование с помощью ЭВМ является необходимым условием разработки методов и средств распознавания и смысловой интерпретации речи. Это действительно натурное моделирование, поскольку, оснастив ЭВМ преобразователем аналог-код, на ней можно полностью воссоздать информационный процесс обработки речевого сигнала, определяемый методом распознавания. Однако моделирование с помощью ЭВМ будет использоваться не только для отработки методов, а и для опробования вариантов архитектур устройств и систем распознавания и смысловой интерпретации речи.

ВЫВОДЫ

1. Для решения задач распознавания и смысловой интерпретации речи предлагается КДП-подход, заключающийся в экономном задании множеств сигналов классов с помощью автоматных порождающих грамматик, позволяющих получать так называемые эталонные сигна-

лы речи путем их составления (композиции) из эталонных элементов, и в использовании динамического программирования в процессе направленного поиска для распознаваемого сигнала наиболее правдоподобного эталонного сигнала. Процесс распознавания и смысловой интерпретации завершается анализом (разбором) этого наиболее правдоподобного эталонного сигнала.

КДП-подход эквивалентен другому, нереализуемому, однако теоретически мощному подходу, в котором сначала все возможные сигналы классов запоминаются, а затем, при распознавании, предъявленный сигнал сравнивается со всеми запомненными сигналами.

В КДП-подходе осуществляется анализ (распознавание) сигналов речи через синтез эталонных сигналов. Принципиально то, что в КДП-подходе сегментация речевого сигнала на части, соответствующие отдельным фонемам или словам, не предшествует распознаванию, а выполняется в едином взаимосвязанном с распознаванием процессе, при котором ответом распознавания непрерывного речевого сигнала является не только последовательность из дискретных элементов распознавания (например, слов), а и наиболее вероятные границы этих элементов в речевом сигнале.

2. КДП-подход основан на предположении о существовании сильных детерминированных закономерностей, связывающих разнообразные и изменяющиеся речевые сигналы классов, на том, что множества сигналов классов образуют маломерные многообразия в пространстве сигналов, на решающей роли математической модели речевого сигнала, ее детерминированной части, на возможности построения математической модели путем изучения свойств, допустимых преобразований и изменчивости речевого сигнала, обусловленных такими основными явлениями, как нелинейное изменение темпа и интенсивности произнесения, коартикуляция и редукция звуков, просодические характеристики речи, индивидуальные особенности голоса.

Существенно детерминированный характер математической модели речевого сигнала и вытекающая из этого возможность использования метода наибольшего правдоподобия для принятия решений составляют методологическую основу КДП-подхода — экономное задание множеств сигналов классов автоматными порождающими грамматиками и применение динамического программирования для принятия оптимального решения.

ГЛАВА 2

ПОЭЛЕМЕНТНЫЙ МЕТОД РАСПОЗНАВАНИЯ СЛОВ

В данной главе излагается метод распознавания отдельно произносимых слов (устных команд). Он основан на составлении эталонных сигналов слов из эталонных элементов, отдельно подбираемых для каждого слова, в соответствии с правилами автоматной порождающей грамматики, вытекающими из кусочно-постоянной математической модели речевого сигнала. Распознавание осуществляется с помощью динамического программирования.

Поэлементный метод распознавания слов был разработан в 1966—1968 гг. [1—4] и уточнен в 1969—1971 гг. [61—62], когда стала использоваться темпоральная транскрипция слова.

§ 2.1. ОПИСАНИЕ РЕЧЕВЫХ СИГНАЛОВ

При распознавании речевых сигналов, как правило, оперируют не с исходным речевым сигналом, получаемым на выходе микрофона, а с так называемым описанием речевого сигнала, экономно представляющим речевой сигнал и содержащим информацию о том, что говорится.

Обычно принято описывать (задавать) речевой сигнал последовательностью $\mathbf{X}_l = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l, \dots, \mathbf{x}_l)$ из элементов \mathbf{x}_l , которые являются отсчетами векторной функции $\mathbf{x}(t)$ в дискретные равноотстоящие моменты времени $t_i = i\Delta T$ с шагом ΔT , принимаемым равным, например, 15 мс. Тогда l — это длина речевого сигнала в дискретном равномерном времени с шагом ΔT . Вообще говоря, может быть использовано и неравномерное время с изменяющимся шагом ΔT , выбираемым, например, из диапазона 8—25 мс. В любом случае, однако, речевой сигнал будет представляться последовательностью \mathbf{X}_l из элементов \mathbf{x}_l .

Последовательности \mathbf{X}_l получают в результате предварительной обработки речевого сигнала на выходе микрофона, чем существенно сокращается объем информации. Так, исходный речевой сигнал, который характеризуется объемом 200 000 бит/с, как правило, описывается существенно меньшим объемом информации — от 9600 до 600 и менее бит/с, однако все еще сохраняющим существенную информацию о том, что говорится, чтобы по ней отвечать на вопросы о распознаваемом классе.

Вопросам предварительной обработки речевого сигнала посвящено огромное количество работ, в том числе ряд монографий [63—69]. Несмотря на многочисленность предложений все они сводятся к тому, что элементы речи описываются величинами, представляющими в той или иной форме мгновенные передаточную характеристику речевого тракта и параметры источников его возбуждения. Поскольку эти величины изменяют свои значения сравнительно медленно в процессе произнесения речи, то для подробного описания речевых сигналов вполне достаточно ограничиться временной дискретизацией элементов с шагом $\Delta T = 15$ мс.

Чаще всего элементами речи x_i выступают мгновенный амплитудный спектр речи или мгновенная автокорреляционная функция, мгновенный продольный профиль акустической трубы речевого тракта, мгновенные значения параметров линейной системы, представляющей речевой тракт, мгновенные значения системы двоичных признаков, характеризующих звуки по месту и способу образования и т. п. Для многих описаний речевого сигнала могут быть указаны взаимнооднозначные преобразования, позволяющие переходить от одного описания к другому.

Элементы x_i могут содержать компоненты, описываемые разнородными физическими величинами. Например, наряду с компонентами, представляющими форму амплитудного спектра речи или передаточную характеристику речевого тракта, могут быть компоненты, характеризующие интенсивность элемента, способ его образования (с участием голоса или только шума), относительную частоту основного тона и т. п.

Последовательности X_i элементов x_i получают, анализируя речевой сигнал на интервале (окне) анализа продолжительностью $\Delta T' \geq \Delta T$ и перемещая это окно вдоль оси времени с шагом ΔT . Таким образом, интервалы анализа либо соприкасаются, либо перекрываются.

Рассмотрим некоторые основные описания речевого сигнала, используемые в монографии, и укажем на их связь.

Пусть на текущем интервале анализа продолжительностью $\Delta T'$ наблюдается последовательность дискрет или отсчетов речевого сигнала f_n , $n = 1 : M$, $M = \left[\frac{\Delta T'}{\Delta t} \right]$, получаемых в результате аналогоцифрового преобразования речевого сигнала микрофона с частотой преобразований $v_{\text{пп}} = \frac{1}{\Delta t}$, Δt — расстояние между дискретами, например $\Delta t = 50$ мкс.

Тогда вычисляемыми компонентами элемента речи x для текущего интервала анализа могут быть отсчеты автокорреляционной функции речевого сигнала

$$B(s) = \sum_{n=1}^{M-s} f_n f_{n+s}, \quad s = 0 : m, \quad m < M, \quad (2.1.1)$$

где обычно полагается

$$m \leq \left[\frac{T_{0 \min}}{2\Delta t} \right] \quad (2.1.2)$$

и $T_{0\min}$ — минимально возможное значение периода основного тона речи, например, $T_{0\min} = 2$ мс. Выбор m по формуле (2.1.2) обусловлен желанием получить описание речевого сигнала, слабо зависящее от индивидуальности голоса, выраженной изменяющейся частотой основного тона [70], и сохраняющее сведения о передаточной характеристике речевого тракта и параметрах источников его возбуждения. Для типичных значений $\Delta t = 50$ мкс и $\Delta T' = 15$ мс получаем типовые значения $M = 300$ и $m = 20$.

Элемент-автокорреляция определяет некий сглаженный по частоте энергетический спектр речевого сигнала $G(p)$ (соответственно, амплитудный спектр $\mathcal{A}(p)$):

$$G(p) = B(0)g(0) + 2 \sum_{s=1}^m B(s)g(s) \cos ps\pi, \quad (2.1.3)$$

$$\mathcal{A}(p) = \sqrt{G(p)}, \quad (2.1.4)$$

где v — частота; $v_{rp} = \frac{1}{2\Delta t}$ — граничная частота; $p = \frac{v}{v_{rp}}$, $0 \leq p < 1$, — относительная частота; $g(s)$ — некоторая функция, обеспечивающая равномерную аппроксимацию спектров усеченными рядами Фурье, например, $g(s) = 1 - \frac{s}{m+1}$ или $g(s) = \frac{s\pi}{2(m+1)} \operatorname{ctg} \frac{s\pi}{2(m+1)}$ [71—73].

Сглаженный спектр $G(p)$ или $\mathcal{A}(p)$ передает огибающую спектра речевого сигнала, не содержащую осцилляций, обусловленных квазипериодическим характером голосового источника возбуждения речевого тракта.

Спектральный элемент x получается в результате вычисления спектра $G(p)$ или $\mathcal{A}(p)$ на дискретной сетке частот p_μ , $p_\mu > p_{\mu-1}$ и $\mu = 1 : Q$, причем Q может отличаться от m , например, $m < Q < M$.

Элемент-автокорреляция и спектральный элемент представляют соответственно автокорреляционную функцию и спектр всего речевого сигнала, рассматриваемого на текущем интервале анализа. Таким образом, эти элементы содержат в неразделенном виде информацию как о передаточной характеристике речевого тракта, так и о форме спектра источников его возбуждения.

Широкое распространение получило описание речевых сигналов с помощью параметров предсказания — параметров линейных систем авторегрессионного типа, моделирующих речеобразование [67, 74, 75]. Предполагается, что на интервале анализа параметры линейной системы $a = (a_1, a_2, \dots, a_s, \dots, a_m)$ не изменяются, а наблюдаемые отсчеты речевого сигнала f_n могут быть спрогнозированы по m предыдущим отсчетам через параметры a в соответствии с уравнением

$$f_n = - \sum_{s=1}^m a_s f_{n-s} + \epsilon_n. \quad (2.1.5)$$

В выражении (2.1.5) ϵ_n рассматривается как ошибка в прогнозе для отсчета f_n .

Параметры **a** линейной системы для текущего интервала анализа обычно оцениваются по отсчетам сигнала f_n , $n = 1 : M$, исходя из минимизации суммарной ошибки прогноза $\sum_{n=1}^M \epsilon_n^2$ или, что то же самое, исходя из метода наибольшего правдоподобия в предположении, что ϵ_n — отсчеты дискретного белого шума с нулевым средним и дисперсией σ . При этом, пользуясь уравнением (2.1.5), чаще всего принудительно полагают, что за пределами интервала анализа $f_{n-s} \equiv 0$ для $n - s < 1$ и $n - s > M$.

В этих предположениях элемент речи в виде **a**-параметров получают как решение системы уравнений

$$\sum_{s=1}^m a_s B(|s-v|) = -B(v), \quad v = 1 : m. \quad (2.1.6)$$

Эта система эффективно решается с помощью сходящегося за m итераций алгоритма Дурбина [55]:

$$\left. \begin{array}{l} E(0) = B(0), \\ k_s = -\left[B(s) + \sum_{v=1}^{s-1} a_v^{(s-1)} B(|s-v|) \right] / E(s-1), \\ a_s^{(s)} = k_s, \\ a_v^{(s)} = a_v^{(s-1)} + k_s a_{s-v}^{(s-1)}, \quad v = 1 : (s-1); \\ E(s) = (1 - k_s^2) E(s-1) \end{array} \right\} s = 1 : m, \quad (2.1.7)$$

причем полагаем

$$a_v = a_v^{(m)}, \quad v = 1 : m. \quad (2.1.8)$$

Максимально правдоподобная оценка дисперсии дискретного белого шума находится как

$$\sigma = \sqrt{\frac{1}{M} \left(B(0) + \sum_{v=1}^m a_v B(v) \right)} = \sqrt{\frac{1}{M} E(m)}. \quad (2.1.9)$$

Элемент речи в виде **a**-параметров определяет передаточную характеристику речевого тракта

$$H(z) = \sigma \left(1 + \sum_{i=1}^m a_i z^{-i} \right), \quad (2.1.10)$$

где $z = \exp(j\varphi)$ — переменная в z -преобразовании и $j^2 = -1$.

Можно отметить, что при выбранном способе оценивания **a**-параметров в передаточную характеристику речевого тракта (2.1.10) включена огибающая спектра реального источника возбуждения (голоса или шума), сам же сигнал возбуждения моделируется как дискретный белый шум с дисперсией σ или как квазипериодическая (с периодом основного тона) последовательность одиночных импульсов $\delta_n = 0$, $n = 1 : M$, кроме точек n_r , $r = 1, 2, \dots$, таких, что $(n_r - n_{r-1}) \times \Delta t = T_r$, где T_r — текущее значение периода основного тона.

В последнем случае амплитуда импульсов δ_n , выбирается, например, из условия

$$\sum_{r=1}^{\mathcal{P}} \delta_r^2 = M\sigma^2, \quad (2.1.11)$$

где \mathcal{P} — количество импульсов основного тона на интервале анализа.

Кроме **a**-параметров, элементы речи x могут представляться набором коэффициентов отражения $k = (k_1, k_2, \dots, k_s, \dots, k_m)$, которые получаются посредством алгоритма (2.1.7) в процессе решения системы уравнений (2.1.6).

Величины k_s названы коэффициентами отражения потому, что они однозначно определяют форму кусочно-постоянной акустической трубы, содержащей $(m + 1)$ цилиндрическую секцию фиксированной длины. Процессы в этой трубе — распространение плоской акустической волны — описываются тем же разностным уравнением (2.1.5), а площади C поперечных сечений соседних секций связаны коэффициентами отражения [67]

$$k_s = \frac{C_{s-1} - C_s}{C_{s-1} + C_s}, \quad s = 1 : m. \quad (2.1.12)$$

Интересно, что количество секций в трубе m , время дискретизации сигнала Δt и длина речевого тракта человека \mathcal{L} (среднее значение \mathcal{L} равно 17 см) связаны определенным соотношением

$$m = \frac{2\mathcal{L}}{c\Delta t}, \quad (2.1.13)$$

где c — скорость распространения звука в воздухе [67, 76]. Выбор m по (2.1.13) хорошо согласуется с m , определяемым формулой (2.1.2) [70].

Кроме **a**-параметров и коэффициентов отражения k в качестве элементов речи часто будет удобно использовать так называемые **b**-параметры, получаемые из **a** по следующим формулам:

$$\begin{aligned} \mathbf{b} &= (b_0, b_1, \dots, b_r, \dots, b_m), \\ b_0 &= \sum_{v=0}^m a_v^2, \quad b_r = 2 \sum_{v=0}^{m-r} a_v a_{v+r}, \quad r = 1 : m, \quad a_0 \equiv 1. \end{aligned} \quad (2.1.14)$$

Физически **b**-параметры интерпретируются как автокорреляционная функция импульсного отклика фильтра с передаточной характеристикой $1 + \sum_{t=1}^m a_t z^{-t}$ [77].

Располагая параметрами **b** и σ , нетрудно рассчитать энергетический спектр речевого сигнала [77], называемый авторегрессионным спектром!

$$G(p_\mu) = M\sigma^2 / \sum_{s=0}^m b_s \cos p_\mu s \pi, \quad \mu = 1 : Q. \quad (2.1.15)$$

Выражение (2.1.15) определяет еще один способ вычисления элемента речи x — авторегрессионного спектра.

Приведенные соотношения (2.1.1) — (2.1.15) показывают, что получаемые различными способами элементы речи связаны однозначными или взаимооднозначными преобразованиями. Как будет показано в дальнейшем, выбор того или иного описания будет зависеть от используемых мер сходства и стремления уменьшить объем вычислений при распознавании.

Отдельный класс описаний речевого сигнала составляют элементы с двоичными компонентами (признаками), принимающими значения 0 или 1. Это могут быть артикуляционные признаки, характеризующие звуки по способу и месту образования, например звук шумный или звонкий, гласный или согласный, носовой или неносовой, дрожащий или недрожащий, огубленный или неогубленный, заднеязычный или переднеязычный, фрикативный или аспиративный и т. п. Артикуляционные признаки относятся к категории трудновычисляемых по причинам принципиального характера.

В дальнейшем будет часто использоваться двоичное описание (элемент), имеющее смысл знака производной спектра $G(p)$ по частоте p на дискретной сетке частот p_μ , $\mu = 1 : Q$.

Пусть $x_i = (x_{i1}, x_{i2}, \dots, x_{i\mu}, \dots, x_{i(Q-1)})$ — текущий элемент речи для момента времени i и $x_{i\mu}$ ($\mu = 1 : (Q - 1)$) — компоненты этого элемента. Двоичное описание-элемент определим так:

$$x_{i\mu} = \begin{cases} 1, & \text{если } G(p_{\mu+1}) \geq G(p_\mu) \text{ и } G(p_{\mu+1}) \geq \Theta_\mu, \\ 0, & \text{в остальных случаях,} \end{cases} \quad (2.1.16)$$

где Θ_μ ($\mu = 1 : (Q - 1)$) — некоторые постоянные пороги, выбираемые экспериментально так, чтобы в стационарных условиях, когда на вход микрофона поступают только акустические помехи помещения,рабатывались в подавляющем большинстве случаев нулевые элементы-коды [78].

Легко убедиться, что двоичное описание (2.1.16) содержит информацию о местоположении формант, добротности полюсов речевого тракта и других параметрах речевого тракта, что позволяет сделать заключение о приемлемости двоичного описания для распознавания речи. Двоичное описание (2.1.16) слабо меняется с изменением громкости произнесения.

Еще более информативным является двоичное описание-элемент с компонентами

$$\text{Sign}(G(p_\mu) - G(p_v)), \quad \mu < v, \quad \mu, v = 1 : Q, \quad (2.1.17)$$

где

$$\text{Sign } \alpha = \begin{cases} 1, & \text{если } \alpha \geq 0, \\ 0, & \text{если } \alpha < 0. \end{cases} \quad (2.1.18)$$

Это двоичное описание-элемент из $\frac{1}{2} Q (Q - 1)$ компонент содержит информацию о форме спектра, определяет относительные амплитуды спектральных составляющих $G(p_\mu)$, $\mu = 1 : Q$, и не зависит от громкости произнесения. Предыдущее двоичное описание является по отношению к нему частным случаем.

Основные достоинства двоичных описаний — наглядность и возможность представления элемента речи в одной ячейке памяти ЭВМ.

Другие используемые описания речевого сигнала будут упоминаться по мере необходимости.

В заключение данного параграфа подчеркнем еще раз, что элемент речи это часто не только вектор спектра, автокорреляции, а-параметров, б-параметров, коэффициентов отражения, двоичных компонент, а и вектор, содержащий компоненты о громкости произнесения, например σ , признаке тон-шум и относительной частоте основного тона и т. п. Во всех случаях, однако, будем обозначать текущий элемент речи в момент i через x_i , каждый раз оговаривая содержание его компонент.

§ 2.2. КУСОЧНО-ПОСТОЯННАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РЕЧЕВЫХ СИГНАЛОВ

Исследование видеоспектограмм одного и того же слова в произнесении одного и того же диктора показывает, что все они (если не обращать особого внимания на интенсивность звуков в слове) могут быть приближенно преобразованы друг в друга путем нелинейного сжатия-растяжения временной оси с сохранением прямого хода времени. Этот факт позволяет рассматривать нелинейное растяжение-сжатие оси времени как основной фактор изменчивости речевых сигналов слова и положить его в основу построения кусочно-постоянной модели речевого сигнала слова [3, 79].

Из всех реализаций $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$, где l — длина реализации, слова с порядковым номером k , $k = 1 : K$ (K — объем слова-ря) выберем одну, которая соответствует наиболее быстрому и все еще четкому произнесению слова. Назовем такую последовательность элементов начальным (исходным) эталоном слова и обозначим ее $E_k = (e_{k1}, e_{k2}, \dots, e_{ks}, \dots, e_{kq_k})$, где k — номер слова, а q_k — длина начального эталона. Подчеркнем, что элементы e_{ks} , в отличие от наблюдаемых элементов x_l реализаций X_l , будут называться эталонными элементами и что элементы e_{ks} не обязательно имеют тот же физический смысл, что и x_l . Более определенно, что именно будет выбираться в качестве начального эталона слова и как его находить по обучающей выборке слова, будет сказано в следующей главе.

Далее вводится множество $\tau_k(l)$ возможных преобразований $v = (v_1, v_2, \dots, v_s, \dots, v_{q_k})$ исходного эталона E_k слова k :

$$\tau_k(l) = \left\{ v : \sum_{s=1}^{q_k} v_s = l, m_{ks} \leq v_s \leq M_{ks}, s = 1 : q_k \right\}, \quad (2.2.1)$$

которые приводят к образованию различных эталонных сигналов слова E_l длины $l \geq q_k$:

$$E_l = vE_k = (\underbrace{e_{k1}, \dots, e_{k1}}_{v_1 \text{ раз}}, \underbrace{e_{k2}, \dots, e_{k2}}_{v_2 \text{ раз}}, \dots, \underbrace{e_{ks}, \dots, e_{ks}}_{v_s \text{ раз}}, \dots, \underbrace{\dots, e_{kq_k}, \dots, e_{kq_k}}_{v_{q_k} \text{ раз}}) = (e_1, e_2, \dots, e_l, \dots, e_l). \quad (2.2.2)$$

Последовательность из пар (m_{ks}, M_{ks}) образует темпоральную транскрипцию слова

$$\tau_k = ((m_{k1}, M_{k1}), (m_{k2}, M_{k2}), \dots, (m_{ks}, M_{ks}), \dots, (m_{kq_k}, M_{kq_k})). \quad (2.2.3)$$

Эта транскрипция относительно каждого эталонного элемента e_{ks} слова указывает границы повторяемости элемента при нелинейном растяжении $v \in \tau_k(l)$ исходного эталона слова вдоль оси времени. Темпоральная транскрипция слова должна быть задана вместе с исходным эталоном слова. Она также вычисляется по обучающей выборке.

Напомним, что как наблюдаемые, так и эталонные элементы представляют речевой сигнал на интервале анализа продолжительностью $\Delta T' \geq \Delta T$ и ассоциируются с моментами дискретного времени $i\Delta T$. Элементы речи представляют фонемы, точнее их части, отдельные фазы.

Условимся, что в исходных эталонах слов первый e_{k1} и последний e_{kq_k} элементы являются элементами пауз (фона помещения) и что для них $m_{k1} = m_{kq_k} = 0$, $M_{k1} = M_{kq_k} = \infty$. Для всех других элементов $m_{ks} > 0$, а это означает, что при преобразованиях v пропускать эталонные элементы исходного эталона слова нельзя.

Эталонные сигналы vE_k , $v \in \tau_k(l)$, которые генерируются процессом (2.2.1) — (2.2.3), отличаются нелинейно изменяющимся темпом произнесения слова, а также различной длиной пауз, в том числе и нулевой, в начале и конце слова. Этот процесс также учитывает явление коартикуляции (взаимовлияния соседних звуков друг на друга), поскольку коартикулированными являются эталонные элементы e_{ks} в последовательностях E_k — исходных эталонах слов. Среди эталонных элементов e_{ks} есть и такие, для которых $m_{ks} = M_{ks} = 1$. Это переходные эталонные элементы, подпоследовательности из которых представляют переходные участки одних звуков в другие. Характерно, что длительность переходных участков для пары звуков изменяется сравнительно незначительно. При произнесении меняется в основном продолжительность стационарных фаз звуков.

Учтя основные факторы изменчивости сигналов слова посредством детерминированной модели vE_k , $v \in \tau_k(l)$, укажем на связь эталонных сигналов E_l с наблюдаемыми сигналами X_l .

Будем рассматривать реальные наблюдаемые сигналы $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$ как такие, которые происходят из эталонных сигналов E_l слов в результате действия аддитивного шума $R_l = (r_1, r_2, \dots, r_l, \dots, r_l)$:

$$X_l = E_l + R_l, \quad (2.2.4)$$

предполагая, что наблюдаемые элементы x_l являются результатом независимых искажений эталонных элементов $e_l = (vE_k)_l$:

$$x_l = e_l + r_l, \quad i = 1 : l. \quad (2.2.5)$$

Это вовсе не значит, что элементы x_l рассматриваются как независимые в последовательностях X_l . Наоборот, сильная детерминированная зависимость между элементами x_l учтена процессом составления эталонных сигналов vE_k , независимыми полагаются лишь отклонения наблюдаемых элементов от соответствующих им эталонных элементов.

Таким образом, математическая модель сигналов слова длины l задается условным распределением

$$p(\mathbf{X}_l / \mathbf{vE}_k) = \prod_{i=1}^l p(x_i / (\mathbf{vE}_k)_i), \quad \mathbf{v} \in \tau_k(l), \quad (2.2.6)$$

где $p(x_i / e_i)$ рассматривается как вероятность появления элемента x_i при условии эталонного элемента e_i .

Для полного описания модели необходимо еще задать вид распределения $p(x_i / e_i)$. Это уточнение будет сделано в дальнейшем.

Определяющая роль в модели (2.2.6), хотя она и вероятностная, принадлежит детерминированной части — множеству эталонных сигналов слова \mathbf{vE}_k , $\mathbf{v} \in \tau_k(l)$. Мы исходим из того, что при адекватной детерминированной части вид закона распределения $p(x_i / e_i)$ будет влиять несущественно на результаты распознавания, особенно при малой дисперсии этого распределения, и что, таким образом, его можно выбирать, исходя из удобства вычислений, например, полагая это распределение сферически симметричным.

Модель названа кусочно-постоянной, поскольку эталонные сигналы слова описываются кусочно-постоянными векторными функциями времени.

§ 2.3. ПОЭЛЕМЕНТНЫЙ МЕТОД РАСПОЗНАВАНИЯ

Воспользовавшись кусочно-постоянной моделью речевого сигнала (2.2.6) и методом наибольшего правдоподобия, сформулируем задачу распознавания предъявленного сигнала $\mathbf{X}_l = (x_1, x_2, \dots, x_l, \dots, x_l)$ как задачу отыскания для этого сигнала наиболее правдоподобного эталонного сигнала среди множества всех возможных эталонных сигналов \mathbf{vE}_k , $\mathbf{v} \in \tau_k(l)$, $k = 1 : K$ и указания номера слова $k(\mathbf{X}_l)$, которому этот наиболее правдоподобный эталонный сигнал принадлежит:

$$k(\mathbf{X}_l) = \operatorname{argmax}_k \max_{\mathbf{v} \in \tau_k(l)} p(\mathbf{X}_l / \mathbf{vE}_k). \quad (2.3.1)$$

С учетом независимости искажений элементов критерий распознавания (2.3.1) можно представить в виде аддитивного выражения:

$$k(\mathbf{X}_l) = \operatorname{argmax}_k \max_{\mathbf{v} \in \tau_k(l)} \sum_{i=1}^l g(x_i, (\mathbf{vE}_k)_i), \quad (2.3.2)$$

где

$$g(x_i, (\mathbf{vE}_k)_i) = \ln p(x_i / (\mathbf{vE}_k)_i) \leq 0 \quad (2.3.3)$$

интерпретируется как элементарная мера сходства наблюдаемого элемента речи x_i и эталонного элемента $e_i = (\mathbf{vE}_k)_i$. Величину же

$$G(\mathbf{X}_l, \mathbf{vE}_k) = \sum_{i=1}^l g(x_i, (\mathbf{vE}_k)_i) \leq 0 \quad (2.3.4)$$

естественно интерпретировать как интегральную меру сходства реализации \mathbf{X}_l и эталонного сигнала \mathbf{vE}_k [3, 4, 12, 13, 78, 79, 80, 96].

Рассмотрим примеры наиболее употребительных мер сходства.

1. $g(\mathbf{x}_t, \mathbf{e}_t) = -H(\mathbf{x}_t, \mathbf{e}_t)$, где $H(\mathbf{x}_t, \mathbf{e}_t)$ — хэммингово расстояние (количество несовпадающих компонент), если \mathbf{x}_t и \mathbf{e}_t имеют двоичные компоненты, например знаки разности энергий в соседних спектральных полосах [78, 80].

2. $g(\mathbf{x}_t, \mathbf{e}_t) = -\alpha(\mathbf{x}_t)(\mathbf{x}_t, \mathbf{e}_t)$, где $(\mathbf{x}_t, \mathbf{e}_t)$ — скалярное произведение векторов \mathbf{x}_t и \mathbf{e}_t , а $\alpha(\mathbf{x}_t)$ — некоторый скаляр, зависящий от \mathbf{x}_t . Эта мера сходства выводится, если речевые сигналы описывать посредством линейной авторегрессионной модели (см. § 2.1 и [67, 68, 77]). Тогда элемент \mathbf{x}_t имеет смысл элемента-автокорреляции, \mathbf{e}_t — смысл \mathbf{b} -параметров, а $\alpha(\mathbf{x}_t)$ может быть энергией элемента \mathbf{x}_t , взятой в степени $(-\frac{3}{4})$ [81].

3. $g(\mathbf{x}_t, \mathbf{e}_t) = -|\mathbf{x}_t - \mathbf{e}_t|^2$, где $|\mathbf{x}_t - \mathbf{e}_t|$ — евклидово расстояние между векторами.

4. $g(\mathbf{x}_t, \mathbf{e}_t) = -\sum_{v=1}^m \left(\frac{x_{tv} - e_{tv}}{\sigma_{tv}} \right)^2$, где x_{tv} и e_{tv} , $v = 1 : m$ — компоненты наблюдаемого \mathbf{x}_t и эталонного \mathbf{e}_t элементов, а σ_{tv} — дисперсия v -й компоненты эталонного элемента \mathbf{e}_t . Очевидно, что эта мера сходства удобна при использовании элементов, компоненты которых имеют различную физическую природу.

$$5. g(\mathbf{x}_t, \mathbf{e}_t) = \sum_{v=1}^m x_{tv} \ln \frac{p_{tv}}{1-p_{tv}} + c_t, \text{ где } c_t = \sum_{v=1}^m \ln(1-p_{tv}). \text{ Эта}$$

мера подобна рассмотренной в предыдущем пункте. Ее удобно использовать в случае, когда распознаваемые элементы \mathbf{x}_t состоят только из двоичных компонент 0 и 1, а эталонные элементы \mathbf{e}_t заданы частотами встречаемости p_{tv} значения 1 в v -й компоненте.

Выбор меры сходства $g(\mathbf{x}_t, \mathbf{e}_t)$ зависит от применяемого описания речевых сигналов и в значительной степени определяется удобствами вычислений. В дальнейшем вопросы выбора описаний речевого сигнала и мер сходства будут всесторонне рассматриваться. В частности, этим вопросам посвящена гл. 9.

Далее следует указать способ решения задачи распознавания (2.3.2).

С этой целью, отправляясь от того, что для аддитивной интегральной меры сходства на множестве эталонных сигналов слова выполняются условия оптимальности [82—84], величину

$$G_k(\mathbf{X}_t) = \max_{v \in \tau_k(t)} G(\mathbf{X}_t, v\mathbf{E}_k), \quad (2.3.5)$$

которая определяет сходство сигнала \mathbf{X}_t с наиболее вероятным эталонным сигналом k -го слова, вычисляют направленным перебором всех возможных эталонных сигналов с помощью динамического программирования [82—84], а задачу

$$k(\mathbf{X}_t) = \operatorname{argmax}_k G_k(\mathbf{X}_t) \quad (2.3.6)$$

решают полным перебором по k . Следовательно, если в словаре K слов, то каждый раз при распознавании одного произнесения понадобится решать K задач ДП.

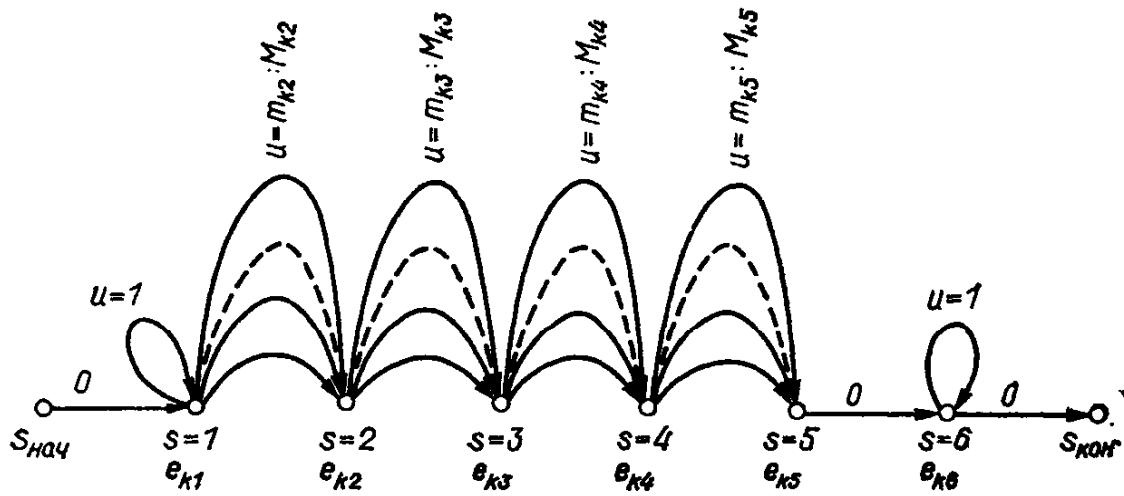


Рис. 2.1. Автоматная грамматика (граф), порождающая эталонные сигналы слова с номером k .

Для записи рекуррентных формул ДП и для удобства последующего изложения зададим процесс порождения эталонных сигналов (автоматную порождающую грамматику) с помощью графа, структура которого для случая $q_k = 6$ приведена на рис. 2.1. На графике всего q_k основных состояний $s = 1 : q_k$ и два вспомогательных $s_{\text{нач}}$ и $s_{\text{кон}}$. В состояние $s = 2 : (q_k - 1)$ входит $M_{ks} - m_{ks} + 1$ стрелок. Каждая из стрелок означает определенное количество $u = m_{ks} : M_{ks}$ повторений эталонного элемента e_{ks} , приписанного всем стрелкам, входящим в состояние s . Что касается первого $s = 1$ и последнего $s = q_k$ состояний графа слова, то в них входит по две стрелки, в том числе одна стрелка-петля $u = 1$, обозначающая одно повторение эталонного элемента e_{k1} или e_{kq_k} при движении по этой стрелке за один дискретный такт времени. Переходы по стрелкам $u = 0$ будем совершать за 0 тактов времени, при этом никакой эталонный элемент выбирать не будем.

В целом, при движении по какой-либо стрелке u , ведущей в состояние s за u тактов времени, будем выбирать подпоследовательность из u повторений эталонного элемента e_{ks} . Тогда очевидно, что при движении из $s_{\text{нач}}$ в $s_{\text{кон}}$ за l тактов времени будут генерироваться эталонные сигналы $E_l = vE_k$, $v \in \tau_k(l)$, длины l .

В дальнейшем график синтеза эталонных сигналов слова будем изображать схематически так, как показано на рис. 2.2, а. В этом схематическом графике одна жирная стрелка заменяет пучок стрелок $u = m_{ks} : M_{ks}$, входящих в состояние s .

Для записи рекуррентных формул ДП введем еще новые обозначения: $\Omega_{kj}(s)$ — множество эталонных сигналов, которые генерируются графиком k -го слова при условии, что в момент j процесс достиг состояния s ; $F_{kj}(s)$ — величина, которая выражает наилучшую интегральную меру сходства начальной части $X_j = (x_1, x_2, \dots, x_j)$ распознаваемого сигнала X_l на множестве эталонных сигналов $\Omega_{kj}(s)$.

Пусть $F_{kj}(s)$ уже вычислены для всех $k = 1 : K$, всех $s = 1 : q_k$ и всех моментов $j < i$, предшествующих моменту i . Тогда с приходом в момент i очередного текущего распознаваемого элемента x_i

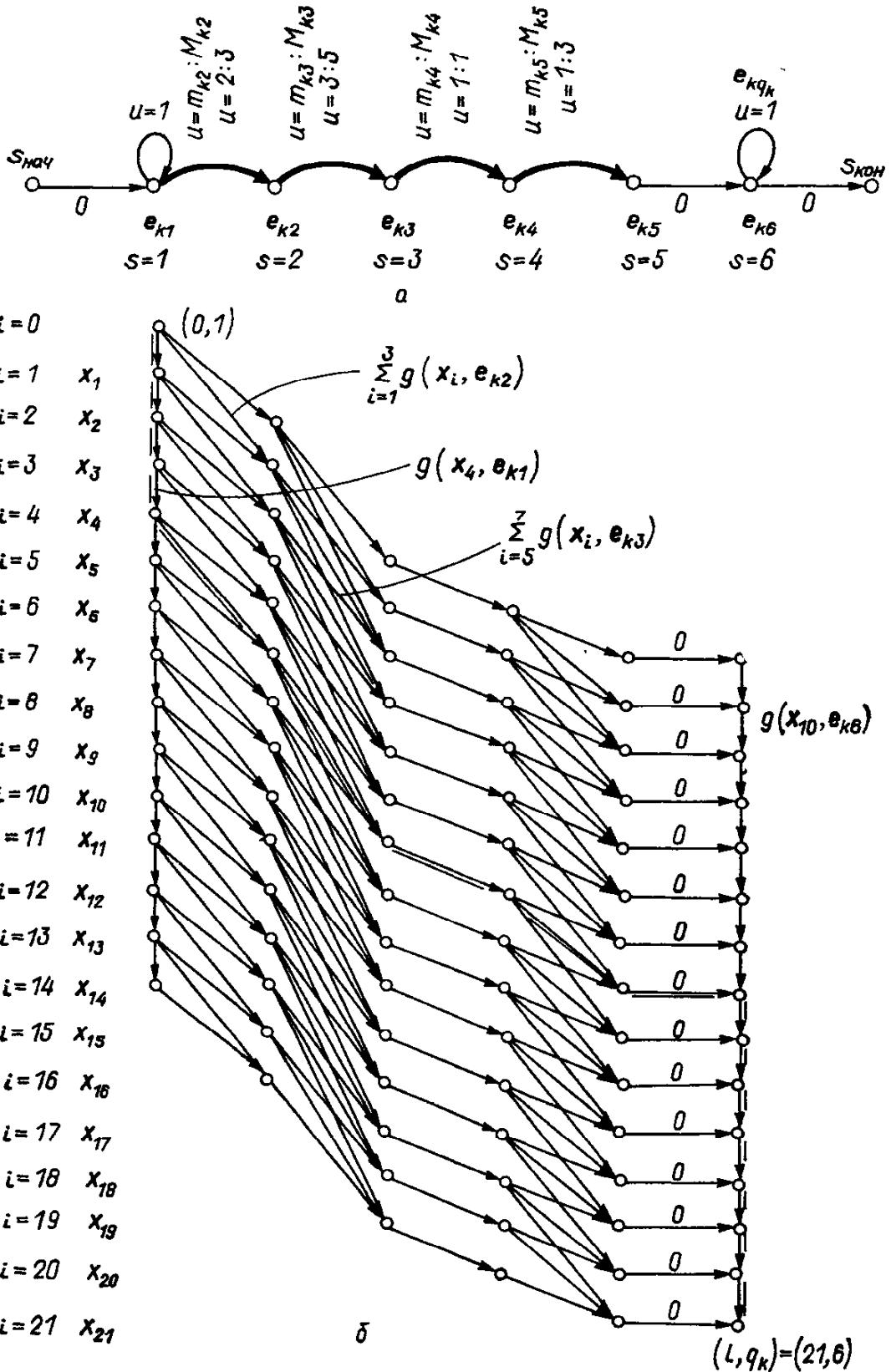


Рис. 2.2. Схематический (а) и развернутый (б) графы слова.

одновременно (если можно, то параллельно) для всех k и всех s по рекуррентным формулам ДП вычисляем (уточняем) новые значения $F_{kl}(s)$:

$$F_{kl}(1) = F_{k(l-1)}(1) + g(x_l, e_{kl}), \quad (2.3.7)$$

$$F_{kl}(s) = \max_{m_{ks} \leq u \leq M_{ks}} (F_{k(l-u)}(s-1) + G_{kl}(u, s)), \quad s = 2 : (q_k - 1), \quad (2.3.8)$$

где

$$G_{kl}(u, s) = \sum_{v=i-u+1}^l g(\mathbf{x}_v, \mathbf{e}_{ks}), \quad (2.3.9)$$

$$F_{kl}(q_k) = \max(F_{kl}(q_k - 1), F_{k(l-1)}(q_k) + g(\mathbf{x}_l, \mathbf{e}_{kq_k})). \quad (2.3.10)$$

Величина $F_{kl}(q_k)$, найденная по формулам (2.3.7) — (2.3.10) для момента $i = l$, как раз и будет равняться $G_k(\mathbf{X}_l)$.

При вычислениях по реккурентным формулам ДП все неопределенные в правых частях формул величины полагаются равными $-\infty$, кроме $F_{k0}(1) = 0$. Вычисления начинаем с момента $i = 1$. Так формально удается учитывать начальные и краевые условия процессов ДП на графах слов с линейной структурой (см. рис. 2.1 и 2.2, а).

Процесс ДП на графах слов может быть проиллюстрирован с помощью рис. 2.2, б. На нем представлен развернутый граф решения задачи распознавания для случая распознаваемой реализации \mathbf{X}_l длины $l = 21$. В отличие от обычного графа слова (см. рис. 2.2, а), который задает только сам процесс порождения эталонных сигналов слова произвольной длины l , развернутый граф, во-первых, задает все эти эталонные сигналы слова конкретной длины l , во-вторых, позволяет сравнить каждый эталонный сигнал слова с распознаваемой реализацией \mathbf{X}_l .

Развернутый граф — это прямоугольная сетка вершин (i, s) , соединяемых дугами. Структура развернутого графа полностью определяется соответствующим ему обычным графом. Две вершины (i_1, s_1) и (i_2, s_2) соединяются только тогда, когда состояния s_1 и s_2 являются соседними, а $u = i_2 - i_1 + 1$ — допустимая стрелка, идущая из состояния s_1 в s_2 . На развернутом графе часть вершин (i, s) исключается из рассмотрения, поскольку некоторые состояния s не достижимы из состояния $s_{\text{нач}}$ за i тактов времени либо из состояния s нельзя попасть в момент l в состояние $s_{\text{кон}}$ за $l - i$ оставшихся тактов времени. Это своеобразная форма выражения краевых условий процесса ДП.

Таким образом, дуга $((i_1, s_1), (i_2, s_2))$ на развернутом графе определяет $u = i_2 - i_1 + 1$ повторений эталонного элемента \mathbf{e}_{ks_2} , а допустимый путь из начальной вершины $(0, 1)$ в конечную $(l, q_k) = (21, 6)$ определяет соответствующий ему эталонный сигнал слова. Так, выделенному на рис. 2.2, б пути соответствует эталонный сигнал слова $\mathbf{E}_{21} = (\mathbf{e}_{k1}, \mathbf{e}_{k1}, \mathbf{e}_{k1}, \mathbf{e}_{k1}, \mathbf{e}_{k2}, \mathbf{e}_{k2}, \mathbf{e}_{k3}, \mathbf{e}_{k3}, \mathbf{e}_{k3}, \mathbf{e}_{k4}, \mathbf{e}_{k5}, \mathbf{e}_{k6}, \mathbf{e}_{k6}, \mathbf{e}_{k6}, \mathbf{e}_{k6}, \mathbf{e}_{k6})$.

Присвоим допустимым дугам, соединяющим вершины (i_1, s_1) и (i_2, s_2) , величину (длину) $G_{kl}(i_2 - i_1, s_2) = \sum_{v=i_1+1}^{i_2} q(\mathbf{x}_v, \mathbf{e}_{ks_2})$, характеризующую сходство сегмента (подпоследовательности) $(\mathbf{x}_{i_1+1}, \mathbf{x}_{i_1+2}, \dots, \mathbf{x}_{i_2})$ распознаваемого сигнала \mathbf{X}_l с эталонным элементом \mathbf{e}_{ks_2} . Тогда сумма длин дуг, лежащих на каком-либо пути из начальной в конечную вершину, определит интегральное сходство распознаваемого сигнала \mathbf{X}_l с эталонным сигналом, соответствующим выбранному пути.

Решить задачу нахождения наилучшего сходства распознаваемого сигнала на множество эталонных сигналов слова означает найти длину самого длинного пути на развернутом графе [3].

Использование развернутых графов облегчает запись рекуррентных формул ДП, способствует наглядному представлению задач, упрощает анализ и сравнение различных модификаций алгоритмов.

Рекуррентные формулы ДП (2.3.7) — (2.3.10) определяют некоторый параллельный вычислительный процесс-конвейер, однотипный как по словам k , так и по состояниям s внутри слов. Изменяются лишь значения параметров для типовых процессоров, зависящие от величин q_k , m_{ks} , M_{ks} . Этот процесс в равной мере может реализовываться и в традиционных (неймановских, непараллельных) машинах, если только их производительности достаточно, чтобы обработать целую группу слов или весь словарь. Для реального времени распознавания важно успеть вычислить все величины $F_{ki}(s)$, $k = 1 : K$, $s = 1 : q_k$ до прихода очередного распознаваемого элемента x_{i+1} , т. е. за отведенное время ΔT , равное, например, 15 мс.

Теперь заметим, что основная вычислительная формула (2.3.8) — (2.3.9) может быть существенно упрощена с точки зрения объема вычислений, если учесть внутренние особенности задачи [85]. Остановимся на этом несколько подробнее, поскольку формулы (2.3.8) — (2.3.9) будут использоваться и при распознавании, и при смысловой интерпретации слитной речи.

Сначала обратим внимание, что $G_{ki}(u, s)$ могут вычисляться рекуррентно:

$$G_{ki}(m_{ks}, s) = G_{k(i-1)}(m_{ks}, s) + g(x_i, e_{ks}) - g(x_{i-m_{ks}}, e_{ks}) \quad (2.3.11)$$

и

$$G_{ki}(u, s) = G_{k(i-1)}(u-1, s) + g(x_i, e_{ks}) \quad (2.3.12)$$

для $m_{ks} < u \leq M_{ks}$, что уменьшает объем вычислений. Далее, обозначив через $u_{ki}(s)$ величину

$$u_{ki}(s) = \underset{m_{ks} \leq u \leq M_{ks}}{\operatorname{argmax}} (F_{k(i-u)}(s-1) + G_{ki}(u, s)), \quad (2.3.13)$$

убеждаемся, что

$$F_{ki}(s) = \max(F_{k(i-m_{ks})}(s-1) + G_{ki}(m_{ks}, s), F_{k(i-1)}(s) + g(x_i, e_{ks})), \quad (2.3.14)$$

если только на предыдущем шаге оказывается, что $u_{k(i-1)}(s) \neq M_{ks}$.

Это позволяет (2.3.8) заменить более простым выражением

$$F_{ki}(s) = \begin{cases} \max(F_{k(i-m_{ks})}(s-1) + G_{ki}(m_{ks}, s), F_{k(i-1)}(s) + g(x_i, e_{ks})), & \text{если } u_{k(i-1)}(s) \neq M_{ks}; \\ \max_{m_{ks} \leq u \leq M_{ks}} (F_{k(i-u)}(s-1) + G_{ki}(u, s)), & \text{если } u_{k(i-1)}(s) = M_{ks}. \end{cases} \quad (2.3.15)$$

Несмотря на громоздкость записи формула (2.3.15) и сопутствующие ей формулы (2.3.11) и (2.3.12) гораздо проще для вычислений,

нежели (2.3.8), (2.3.9). Так, для типичных случаев, когда среднее значение величины $w = M_{ks} - m_{ks} + 1$ равно 10, и в предположении, что различные значения $u_{kl}(s)$ равновероятны и равны, таким образом, $1/w$, получим выигрыш в 3 раза по числу операций типа сложения. Этот эффект был подтвержден экспериментально.

На этом заканчивается изложение основного варианта метода поэлементного распознавания слов устной речи. Метод назван поэлементным, так как принятие решений о произнесенном слове производится на основе распознавания элементов речи — вычисления таблицы сходств наблюдаемых и эталонных элементов.

§ 2.4. АНАЛИЗ РАЗЛИЧНЫХ МОДИФИКАЦИЙ МЕТОДА

Метод поэлементного распознавания слов речи в первоначально сформулированном виде [1—4] не использовал темпоральную транскрипцию слова — ограничения на повторяемость эталонных элементов при нелинейном растяжении исходного эталона слова не накладывались. Формально это означало, что $m_{k1} = m_{kq_k} = 0$, $m_{ks} = 1$, $s = 2 : (q_k - 1)$ и $M_{ks} = \infty$, $s = 1 : q_k$.

Граф слова и соответствующий ему развернутый граф для этого случая представлены на рис. 2.3. Как и в случае рис. 2.2, рассматривается реализация X , длины $l = 21$. В каждую вершину развернутого графа входит не более двух дуг, причем всем дугам, входящим в вершину (i, s) , приписана длина $g(x_i, e_{ks})$. В этом случае рекуррентные формулы (ДП) (2.3.7) и (2.3.10) для состояний $s = 1$ и $s = q_k$ не меняются, а основная рекуррентная формула ДП (2.3.8) или (2.3.15) для состояний $s = 2 : (q_k - 1)$ упрощается к виду

$$F_{kl}(s) = \max(F_{k(l-1)}(s-1), F_{k(l-1)}(s)) + g(x_l, e_{ks}). \quad (2.4.1)$$

Необходимость и целесообразность введения темпоральных транскрипций была доказана в 1969—1971 гг. [61, 62].

В 1968 г. Г. Слуцкером [86] также был использован метод нелинейного сравнения слов с помощью ДП. Правда, при этом предполагалось, что начало и конец слова найдены каким-либо способом до распознавания.

В качестве исходного эталона слова использовалась какая-либо реализация этого слова.

Сравнение распознаваемой реализации X , с эталоном слова производилось с помощью графа, показанного на рис. 2.4. Всем диагональным дугам этого развернутого графа, входящим в вершину (i, s) , приписывалась величина $g(x_i, e_{ks})$, а всем вертикальным и горизонтальным дугам — величина $g \equiv 0$.

Точно такой же развернутый граф для распознавания слов использовали В. М. Величко и Н. Г. Загоруйко в 1969 г. [5], а затем Х. Сакоэ и С. Чиба в 1970 г. [7].

Можно убедиться, однако, что распознавание слов речи на основе развернутого графа типа рис. 2.4 обладает рядом существенных недостатков:

- 1) в процессе сравнения распознаваемой реализации с исходным

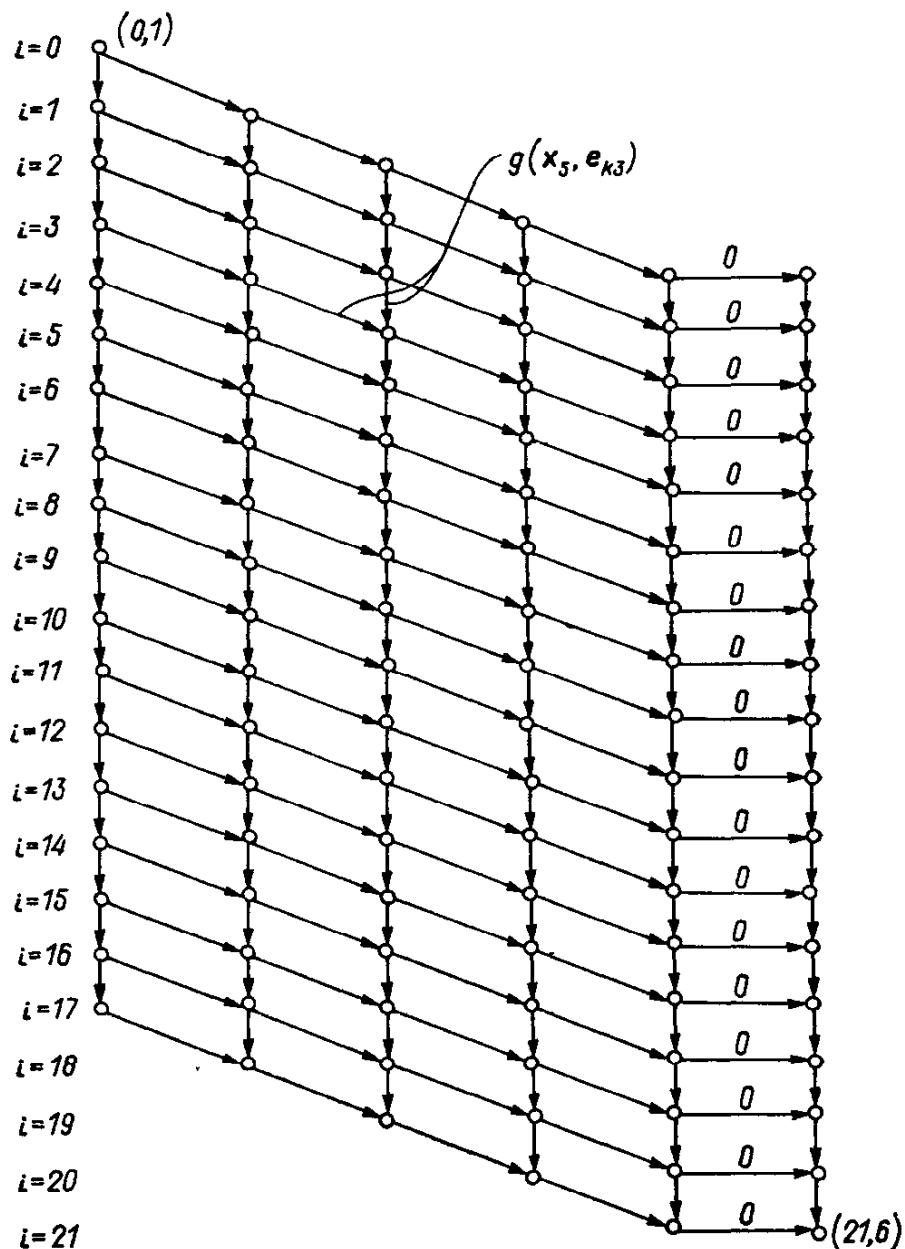
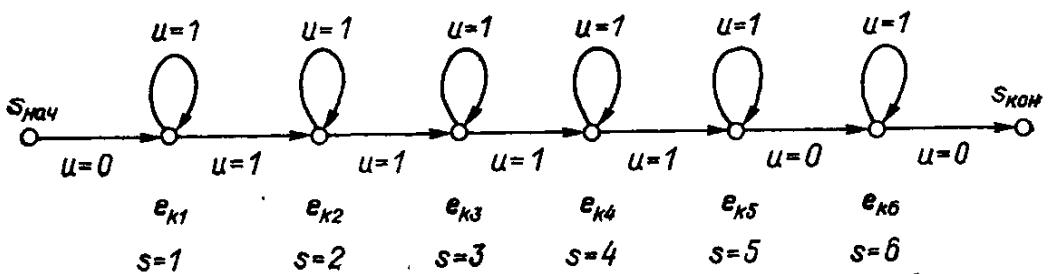


Рис. 2.3. Структура графов слова в случае, когда не используется темпоральная транскрипция.

эталоном слова пропускаются целые сегменты (цепочки из элементов) как в реализации, так и в эталонах, что приводит к взаимному перепутыванию пар слов, если только в этой паре одно слово может быть составлено из частей другого, например, ВОСЕМЬ и СЕМЬ, ТРИ и ЧЕТЫРЕ;

2) в процессе сравнения реализации с эталоном слова за счет выбрасывания сегментов происходит сокращение длин реализаций и эталона

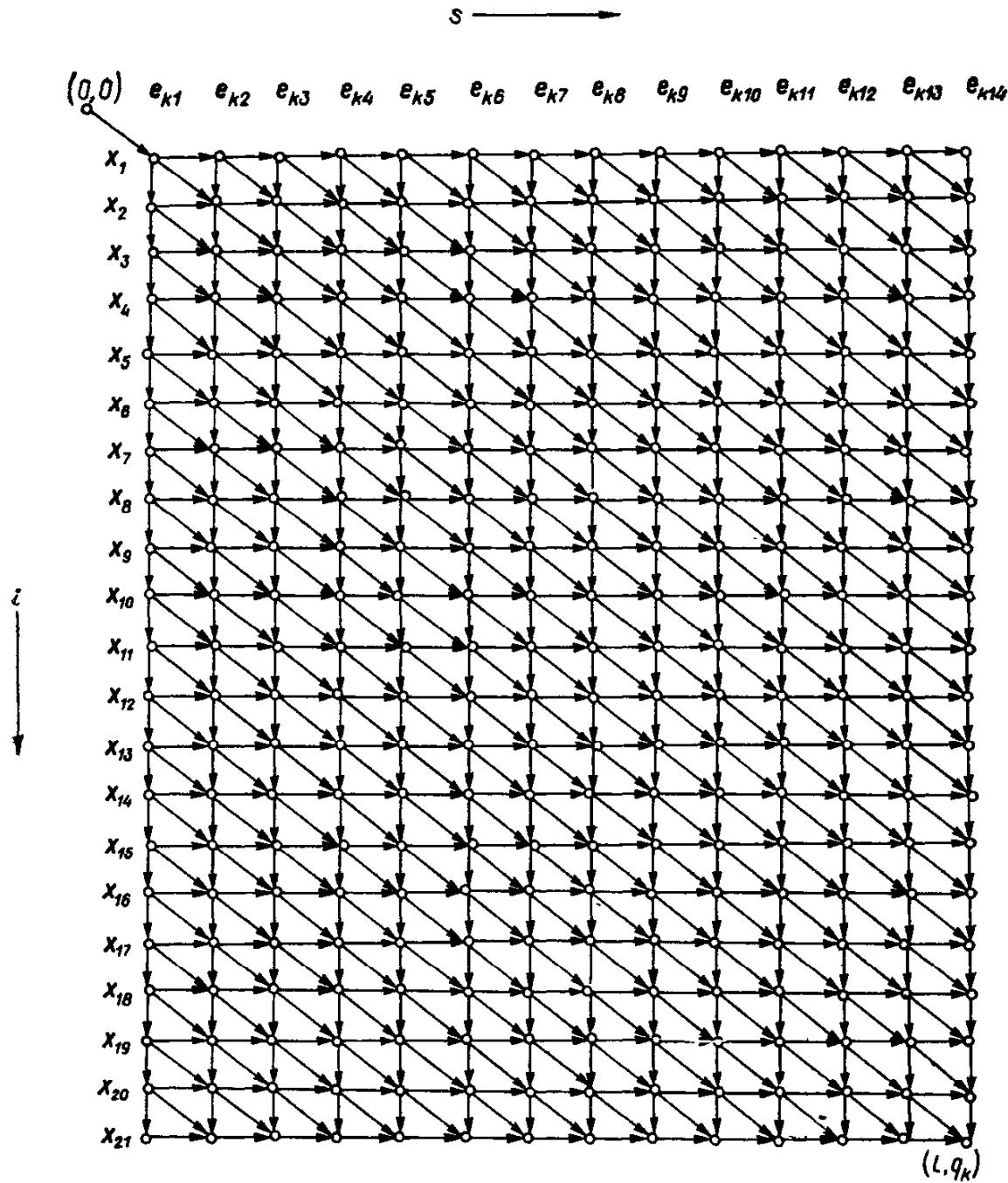


Рис. 2.4. Часто используемый развернутый граф слова.

(вплоть до одного элемента), что равносильно отказу от части информации, содержащейся как в исходном эталоне, так и в распознаваемой реализации, причем упомянутое сокращение происходит по-разному для разных эталонов.

Эти недостатки, которые служат причиной многих ошибок распознавания, отсутствуют в поэлементном методе — порождаемые эталонные сигналы имеют длину l , одинаковую с длиной распознаваемого сигнала, из распознаваемой реализации никакие элементы не выбрасываются, в эталонных же сигналах слова сохраняется порядок следования элементов исходного эталона слова, изменяется только их повторяемость.

Поэлементный метод оказывается более удобным с точки зрения объемов вычислений и памяти. Если в случае графа рис. 2.4 исходный

эталон слова имеет длину, в среднем равную длине распознаваемой реализации, т. е. в развернутом графе приблизительно l^2 вершин, то в поэлементном методе, поскольку $q_{k\text{среднее}}$ приблизительно в 4—5 раз меньше средней длины l , таких вершин соответственно в 4—5 раз меньше. Таким образом, достигается экономия в памяти на хранение исходных эталонов слова и в объеме вычислений. В последнем случае экономия даже больше, чем в 4—5 раз, поскольку вычисления по (2.4.1) проще вычислений по рекуррентным формулам для графа рис. 2.4:

$$F_{ki}(s) = \max(F_{k(i-1)}(s-1) + g(x_i, e_{ks}), F_{k(i-1)}(s), F_{ki}(s-1)). \quad (2.4.2)$$

Наконец, сопоставляемый метод обладает еще тем существенным недостатком, что предполагает поиск начала и конца слова до распознавания. В поэлементном же методе процедура поиска начала и конца слова автоматически выполняется в процессе распознавания.

Отмеченные недостатки объясняют, почему сопоставляемый метод не получил распространения. Начиная с 1973 г. сопоставляемый метод модифицировался. Сущность модификаций сводилась к следующему.

Всем дугам, входящим в состояние (i, s) , стали присваивать одну и ту же длину $g(x_i, e_{ks}) > 0$, в отдельных случаях диагональной дуге присыпали удвоенное значение $2g(x_i, e_{ks}) > 0$ [87, 88]. Такое нововведение уже не приводило к пропускам сегментов реализации и исходного эталона. Однако, как и раньше, порождаемые для X_i эталонные сигналы имели различную длину — от $\max(l, q_k)$ до $l + q_k - 1$, а распознаваемые элементы x_i искусственно повторялись с тем, чтобы длина преобразованной реализации X_i равнялась длине соответствующего эталонного сигнала. Удвоение длин диагональных дуг должно было, по замыслу авторов новшества, частично компенсировать изменение длин реализации и эталонных сигналов слова. Следующее нововведение заключалось в том, что с 1974 г. стали учитывать ограничения на возможный темп преобразования реализации и исходного эталона слова [87, 88]. Так, запрещалось двигаться по двум горизонтальным или двум вертикальным дугам подряд [88]. Формально это означало, что локальный темп «произнесения» распознаваемого или эталонного сигналов слова мог ускоряться или замедляться не более, чем в два раза. Учет этого локального ограничения приводил к изменению рекуррентных формул ДП. Допустимые переходы в вершину (i, s) теперь уже задавались, как на рис. 2.5, а вычисления $F_{ki}(s)$ вместо (2.4.2) были другими:

$$\begin{aligned} F_{ki}(s) = \max & (F_{k(i-1)}(s-2) + 2g(x_i, e_{k(s-1)}) + g(x_i, e_{ks}), \\ & F_{k(i-1)}(s-1) + 2g(x_i, e_{ks}), \\ & F_{k(i-2)}(s-1) + 2g(x_{i-1}, e_{ks}) + g(x_i, e_{ks})). \end{aligned} \quad (2.4.3)$$

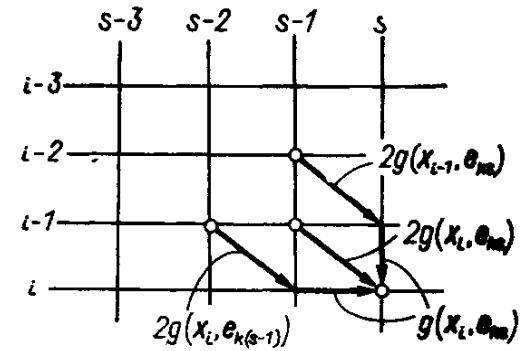


Рис. 2.5. Допустимые локальные переходы.

Убеждаемся, что применяемые ограничения на темп преобразования сигналов приводят к тому же эффекту, что и темпоральная транскрипция, введенная в поэлементный метод в 1969—1971 гг. [61, 62]. Кроме того, темпоральная транскрипция является более гибким и адекватным средством задания ограничений на темп «произнесения» эталонов, так как эти ограничения индивидуально задаются для каждого эталонного элемента слова, а не огульно — одинаково для всех элементов.

Различные варианты модифицированного метода распознавания слов, сопоставляемого поэлементному методу, применяются и в настоящее время.

Подытоживая, отметим, что поэлементный метод распознавания превосходит другие аналогичные методы распознавания слов, основанные на динамическом программировании, потому что:

1) является теоретически более ясным и обоснованным; оперирует с исходным эталоном слова, из которого составляются различные эталонные сигналы слова такой же длины, что и распознаваемая реализация; сама же распознаваемая реализация в процессе сравнения с эталонными сигналами не преобразуется; каждый наблюдаемый элемент речи участвует в интегральной мере сходства только один раз;

2) задача обучения распознаванию слов имеет четко выраженную постановку и алгоритм решения (см. гл. 3); при сравнении распознаваемой реализации с эталонными подвергается преобразованию только исходный эталон слова;

3) ограничения на темп преобразования исходного эталона слова удобно задаются с помощью темпоральной транскрипции слова, более гибко и адекватно передающей реальные множества сигналов слова;

4) применяемые в поэлементном методе короткие исходные эталоны слова позволяют в 4—5 раз уменьшить память на хранение эталонов слова и приблизительно во столько же раз уменьшить время на вычисление интегральных мер сходства (ср., например, формулы (2.3.15), (2.3.11) — (2.3.13) с формулой (2.4.3));

5) не нуждается в предварительном, до распознавания, нахождении начала и конца слова;

6) сравнительно просто обобщается на случаи распознавания и смысловой интерпретации слитной речи.

§ 2.5. ВОЗМОЖНЫЕ УСОВЕРШЕНСТВОВАНИЯ МЕТОДА

Дальнейшие усовершенствования поэлементного метода распознавания слов речи могут быть выполнены, если необходимо явно тем или иным способом учсть информацию о громкости произнесения звуков слова, способе их образования, тональных свойствах речевого сигнала слова. Бессспорно то, что эта дополнительная информация является полезной при распознавании.

Рассмотрим случай, когда и распознаваемые x_i , и эталонные элементы e_{ks} содержат информацию об интенсивности их произнесения. Пусть это будут спектральные или автокорреляционные элементы. Тогда имеет место выраженная тенденция к умножению элементов-векторов на множитель $\alpha > 0$ при изменении интенсивности произнесения. Эти изменения в последовательностях являются явно зависимыми.

Возникает необходимость учесть явление нелинейного изменения интенсивности элементов вдоль оси времени.

Первый способ учета изменений интенсивности произнесения элементов заключается в том, что при генерации эталонных сигналов слова, синтезируя $u = m_{ks}$: M_{ks} раз эталонный элемент e_{ks} с интенсивностью α_{ks} , т. е. выбрав элемент $\alpha_{ks}e_{ks}$, последующий элемент $e_{k(s+1)}$ синтезируют $u = m_{k(s+1)}$: $M_{k(s+1)}$ раз с интенсивностью $\alpha_{k(s+1)}$, т. е. выбирают элемент $\alpha_{k(s+1)}e_{k(s+1)}$, но при этом $\alpha_{k(s+1)}$ должно удовлетворять условию

$$\alpha_{ks} - \Theta \leq \alpha_{k(s+1)} \leq \alpha_{ks} + \Theta, \quad (2.5.1)$$

где Θ — некоторый параметр, ограничивающий скорость роста интенсивности для соседних эталонных элементов [61, 62, 79]. Задают возможный диапазон изменения интенсивности $\alpha_{\min} \leq \alpha_{ks} \leq \alpha_{\max}$, $k = 1 : K$, $s = 1 : q_k$ и вводят шкалу дискретных значений α_v , $v = 1 : N$, $\alpha_v > \alpha_{v-1}$, $\alpha_1 = \alpha_{\min}$, $\alpha_N = \alpha_{\max}$.

Соответствующая порождающая грамматика для $q_k = 6$, $N = 5$, $\Theta = \frac{\alpha_{\max} - \alpha_{\min}}{N}$ представлена в виде графа на рис. 2.6. Предполагается, что акустический фон помещения свою интенсивность не меняет. В этом случае порождающая грамматика задает двухразмерный процесс генерации эталонных сигналов слова, отличающихся нелинейно и независимо изменяющимися темпом и интенсивностью произнесения. При входе в состояние s уровня α_v по стрелке $u = m_{ks}$: M_{ks} выбирается u раз эталонный элемент $\alpha_v e_{ks}$.

Решение задачи распознавания для этого случая приводит к двухразмерному динамическому программированию. Вместо рекуррентных формул (2.3.7) — (2.3.15), применяемых для всех $k = 1 : K$, всех $s = 1 : q_k$ каждый раз в связи с обработкой очередного распознаваемого элемента x_t приходится пользоваться более громоздкими вычислениями — для всех $k = 1 : K$, всех $s = 1 : q_k$ и всех $v = 1 : N$ находить

$$F_{ki}(s, v) = \max_{\substack{m_{ks} \leq u \leq M_{ks}, \\ |v - \mu| \leq \omega}} \left(F_{k(t-u)}(s-1, v-\mu) + \sum_{r=t-u+1}^t g(x_r, \alpha_v e_{ks}) \right), \quad (2.5.2)$$

где $\omega = [\Theta N / (\alpha_{\max} - \alpha_{\min})]$.

Как это следует из анализа (2.5.2), вычисления по формуле (2.5.2) являются в $N\omega$ раз более громоздкими по сравнению с обычным поэлементным методом. Исследования показали [61, 62], что такой способ учета интенсивности элементов несущественно повышает надежность распознавания. Поэтому и из-за громоздкости вычислений использовать его не рекомендуется.

Гораздо более простой и эффективный второй способ учета интенсивности элементов при распознавании. Он предполагает, что сами элементы речи не несут информацию об интенсивности, например, элементы речи — это двоичное описание, а-параметры или б-параметры, коэффициенты отражения и т. п. Информацию же об интен-

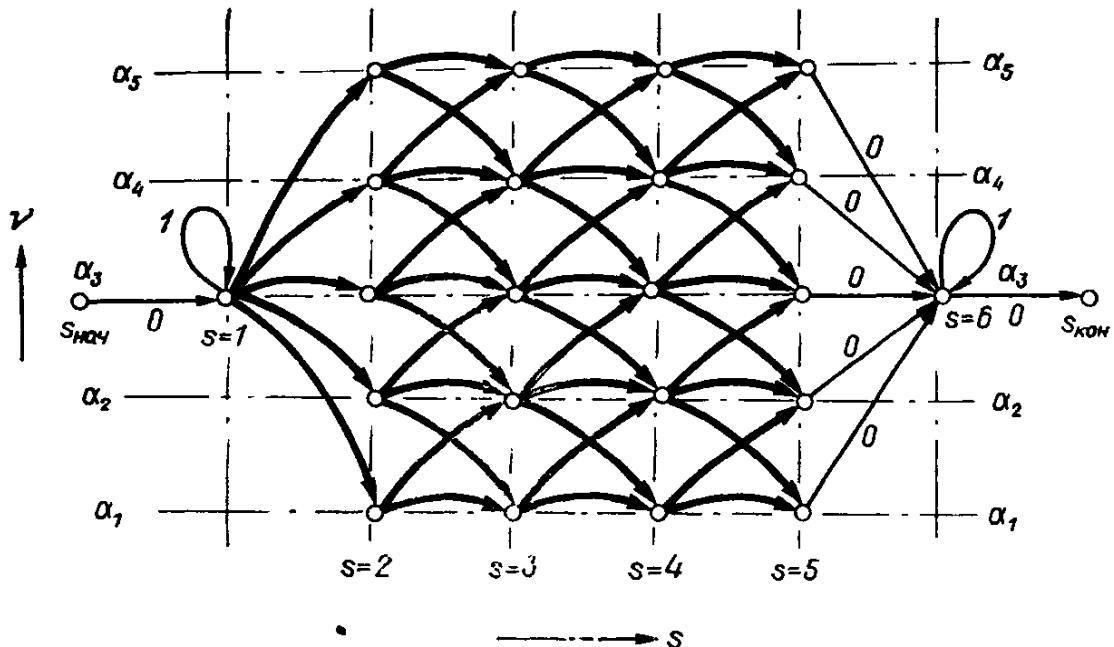


Рис. 2.6. Двухразмерная грамматика, порождающая эталонные сигналы слова с нелинейно изменяющимися темпом и интенсивностью произнесения.

сивности элемента предлагается передавать дополнительной компонентой. Пусть h_i , h_{ks} — компоненты интенсивности соответственно элемента x_i и эталонного элемента e_{ks} .

Аналогично введем еще одну дополнительную компоненту f_i для обозначения способа образования и тональности элемента x_i . Например, будем полагать $f_i = 0$, если элемент x_i шумный, т. е. образуется с помощью шума, и $0 < f_i \leq 1$, если элемент x_i тональный, т. е. образуется с помощью голоса. В этом последнем случае f_i можно трактовать как относительную частоту основного тона по отношению к максимально возможной на длине произнесения слова. Точно так же и эталонные элементы e_{ks} будем характеризовать тональностью f_{ks} .

Из сказанного вытекает, что каждое слово, помимо исходного эталона E_k и темпоральной транскрипции τ_k с одинаковой длиной q_k , должно быть задано еще своей громкостной транскрипцией

$$\mathbf{H}_k = (h_{k1}, h_{k2}, \dots, h_{ks}, \dots, h_{kq_k}) \quad (2.5.3)$$

и тональной

$$\mathbf{F}_k = (f_{k1}, f_{k2}, \dots, f_{ks}, \dots, f_{kq_k}) \quad (2.5.4)$$

той же длины q_k .

С учетом дополнительной информации об интенсивности и тональности элементов в поэлементном методе распознавания соответствующим образом уточняется элементарная мера сходства $g(x_i, e_{ks})$, например, путем добавления к $g(x_i, e_{ks})$ мер сходства интенсивностей $g_1(h_i, h_{ks})$ и тональностей $g_2(f_i, f_{ks})$. В остальном же поэлементный метод распознавания остается без изменений.

Отметим, что интенсивность элементов может передаваться по-разному, в том числе и в таком алфавите, как очень слабый, слабый, средний, сильный и очень сильный элемент.

В дальнейшем, говоря о поэлементном методе распознавания, будем подразумевать, что элементы x_i и e_{ks} содержат элементы интенсивности и тональности, что каждое слово представлено исходным эталоном и тремя транскрипциями — темпоральной, громкостной, тональной — и что элементы всех транскрипций используются при вычислениях меры сходства.

Случаи неиспользования информации об интенсивности и тональности элементов будут оговариваться особо.

§ 2.6. ЭКСПЕРИМЕНТЫ

Первые экспериментальные исследования поэлементного метода распознавания слов были выполнены в 1966—1967 гг. [1—4, 62].

Речевой сигнал с микрофона МД-55 вводился в ЭВМ М-50 с помощью 5-разрядного преобразователя аналог-код (частота дискретизации 10 кГц) и далее подвергался анализу и распознаванию с помощью ЭВМ.

Слова произносились диктором раздельно. Диктор нажимал кнопку, затем, спустя 0,2—0,5 с, произносил слово, после чего отпускал кнопку. Время нажатия кнопки определяло длину l введенного сигнала в дискретном времени с шагом $\Delta T = 20$ мс.

Уровень внешних акустических шумов и помех составлял 75 дБ.

Речевой сигнал подвергался текущему автокорреляционному анализу со взвешиванием — вычислялись $B(s)g(s)$, $g(s) = \frac{s\pi}{2(m+1)} \operatorname{ctg} \frac{s\pi}{2(m+1)}$, $s = 0 : m$, $m = 20$ на интервале анализа $\Delta T' = \Delta T = 20$ мс.

В качестве исходного эталона слова бралась реализация этого слова, соответствующая быстрому и четкому его произнесению.

В качестве меры сходства $g(x_i, e_{ks})$ использовалось скалярное произведение нормированных из условий $\|x_i\| = 1$ и $\|e_{ks}\| = 1$ элементов x_i и e_{ks} , представленных взвешенными автокорреляционными функциями. Таким образом, интенсивность произнесения элементов из рассмотрения исключалась.

Распознавались 12 слов: названия цифр 0—9 и слова ПЛЮС и МИНУС. На проверочной выборке из 600 произнесений слов (по 50 на слово) ошибок распознавания зарегистрировано не было. Длительности произнесений слов изменялись в пределах ± 60 % от средней продолжительности слова.

В течение 1969—1973 гг. эксперименты продолжались на БЭСМ-6. Условия и способ ввода сигналов в ЭВМ были аналогичны. Разница состояла лишь в том, что использовался 9-битовый преобразователь аналог — код и код — аналог, а также другой микрофон МК-61. Эксперименты проводились в условиях машинного зала, где работали две ЭВМ: БЭСМ-6 и М-220. Уровень помех и шумов — до 80 дБ. Частота дискретизации сигналов могла изменяться и составляла 40, 33, 25, 20, 16, 10, 8 или 6 кГц.

В экспериментах 1969—1971 гг. предварительная обработка заключалась в цифровой фильтрации речевого сигнала с помощью цифровых резонансных фильтров [62]. На интервалах анализа продолжи-

тельностью $\Delta T' = 18$ мс вычислялся текущий амплитудный спектр речевого сигнала. Размерность элементов речи равнялась 20 (20 различных резонансных фильтров). Шаг отсчета элементов $\Delta T = \Delta T' = 18$ мс. Исходная частота дискретизации речевого сигнала равнялась 16 кГц.

Перед использованием при обучении или распознавании реализации (последовательности элементов) $\mathbf{X}_l = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_l)$ нормировались из условия, что максимальный на длине реализации модуль элемента не превосходит 1: $\mathbf{x}_i = c\bar{\mathbf{x}}_i$, $i = 1 : l$, $c = 1/\max_i \|\mathbf{x}_i\|$. Такая нормировка сохраняет информацию об относительной интенсивности элементов в реализации и в то же время «стабилизирует» общую интенсивность произнесения слова.

Распознавались 111 слов. Список их и длина исходных эталонов слов даны в Приложении 1. Сначала выполнялось обучение распознаванию слов в соответствии с алгоритмом, описанным в гл. 3. Каждое слово произносилось 10 раз. В результате обучения распознаванию находились исходный эталон слова и его темпоральная транскрипция. Примеры исходного эталона и темпоральной транскрипции для слова ТОЧКА приведены на рис. 3.5.

При распознавании с целью сравнительного анализа исследовались три различных модификации алгоритма поэлементного распознавания слов.

В первой модификации не использовалась информация, содержащаяся в темпоральной транскрипции слова, и предполагалось, что относительная интенсивность элементов не меняется от произнесения к произнесению.

Во второй модификации дополнительно учитывалась темпоральная транскрипция слова.

Третья модификация рассматривалась как усовершенствованная вторая — в ней дополнительно учитывалось, что если эталонный элемент e_{ks} имеет относительную интенсивность α_{ks} , то последующий эталонный элемент $e_{k(s+1)}$ мог менять относительную интенсивность в интервале $[\alpha_{ks} - \epsilon, \alpha_{ks} + \epsilon]$. В экспериментах α_{ks} принимало $N = 20$ различных значений из диапазона $0,5 \leq \alpha \leq 1,5$, а ϵ равнялось 0,05.

На одной и той же проверочной выборке из 10 000 реализаций (около 100 произнесений на слово) были получены следующие результаты:

1-я модификация — 2 % ошибок, 5 % отказов от распознавания;

2-я модификация — менее 0,5 % ошибок, 2 % отказов от распознавания;

3-я модификация — менее 0,4 % ошибок и менее 1,5 % отказов от распознавания.

Отказ от распознавания вырабатывался при выполнении условия

$$|G_{k^*}(\mathbf{X}_l) - \max_{k \neq k^*} G_k(\mathbf{X}_l)| < \delta, \quad \delta = 0,0075, \quad (2.6.1)$$

где k^* — предполагаемый ответ распознавания; $G_k(\mathbf{X}_l)$ — величина, характеризующая принадлежность распознаваемой реализации \mathbf{X}_l .

к слову с номером k ; $\operatorname{argmax}_{k \neq k^*} G_k(X_i)$ определяет ближайший «чужой» класс.

Увеличением порога отказа δ можно было убрать ошибки распознавания, однако процент отказов при этом возрастал. Ошибки и отказы приходились на пары близких слов МИНУС — СИНУС, ЧИТАТЬ — ПИСАТЬ, ВОСЕМЬ — КОСИНУС, ВХОД — ВВОД и др. Отказ от распознавания осуществлялся также при выполнении условия $G_{k^*}(X_i) > 0,2$. Этим достигался отказ от распознавания многих слов, не вошедших в словарь.

Среднее время распознавания одной реализации составляло около 1 мин, причем время распределялось между тремя модификациями в пропорции 1 : 1,5 : 5.

Из сравнения видно, что предпочтение следует отдать второй модификации — основному варианту метода поэлементного распознавания. Таким образом, эксперименты, проводимые в 1969—1971 гг., подтвердили целесообразность учета ограничений на повторяемость эталонных элементов, т. е. целесообразность использования темпоральной транскрипции слова [61, 62].

Эти же эксперименты показали, что при распознавании слов, произнесенных другим диктором, по эталонам первого диктора количество ошибок увеличивалось до 10 % и более. Возрастание ошибок зависело от манеры и способа произношения диктора.

В 1971, 1972 гг. эксперименты с поэлементным методом распознавания слов продолжались. Словарь был увеличен до 200 слов (см. Приложение 1). При распознавании контрольной выборки из 6 000 реализаций, произнесенных одним диктором, было зарегистрировано 0,5 % ошибок и 2,5 % отказов от распознавания [16, 62, 89, 90].

В последующие годы поэлементный метод распознавания слов был существенно развит за счет введения принципа пофонемного распознавания. В дальнейшем усовершенствованный метод поэлементного распознавания слов стал называться пофонемным методом. Он оказался гораздо более эффективным с точки зрения используемой памяти и быстродействия.

Тем не менее методы распознавания слов речи, подобные поэлементному методу и основанные на динамическом программировании, все еще широко применяются. Однако, как это уже отмечалось в § 2.5, поэлементный метод продолжает превосходить их по надежности распознавания и существенно меньшим требованиям к объемам памяти и вычислений.

Поэлементный метод распознавания слов появился в результате попыток устранить недостатки алгоритмов распознавания слов, предполагающих так называемую временную нормализацию реализаций слова до распознавания, столь популярную в 60-е годы. Она заключалась в приведении реализации к некоторой стандартной длине, совпадающей с длиной эталона. Последующее распознавание сводилось к простому наложению реализации на эталоны слов и вычислению сходств. В конце концов подобные попытки привели к выводу, что никакой временной нормализацией до распознавания, какой бы изощ-

ренной она ни была, не удастся добиться адекватного совпадения элементов реализации с элементами эталонов слов. Выход был найден в результате включения в процесс распознавания временной нормализации в качестве обратной связи, что в конечном счете привело к необходимости генерации эталонных сигналов слова определенной длины с помощью порождающих грамматик и применению динамического программирования для сравнения распознаваемой реализации с генерируемыми эталонными сигналами [2—4, 72, 91—94].

ВЫВОДЫ

Разработан поэлементный метод распознавания слов речи, основанный на преобразовании исходного эталона слова в соответствии с его темпоральной транскрипцией и заключающийся в нахождении такого преобразованного эталонного сигнала слова, который наиболее похож на распознаваемый сигнал. Метод обеспечивает наилучшее соответствие элементов распознаваемого и исходного эталонного сигналов слова и устраняет принципиальные недостатки методов распознавания, предполагающих временную нормализацию описаний слов до распознавания.

Поэлементный метод учитывает разнообразие и изменчивость реализаций слова, обусловленные нелинейным изменением темпа произнесения и нелинейными изменениями интенсивности и тональности на длине слова. Метод не нуждается в предварительном нахождении начала и конца сигналов слов на оси времени.

Поэлементный метод эквивалентен такому способу распознавания, когда каждое слово задается огромным количеством эталонов, с которыми сравнивается распознаваемая реализация. Однако задание этого множества осуществляется экономным образом с помощью автоматной порождающей грамматики, а направленный поиск наиболее похожих эталонных сигналов выполняется посредством динамического программирования.

2. Поэлементный метод обеспечивает распознавание произвольных наборов слов с относительно высокой надежностью. Так, распознавание 200 слов характеризуется менее 0,5 % ошибок и 2,5 % отказов от распознавания. Метод предполагает предварительное обучение (настройку) на словарь и голос диктора. Он пригоден для практического использования.

3. Поэлементный метод распознавания слов выгодно отличается от появившихся позже аналогичных методов, основанных на динамическом программировании, тем, что требует в несколько раз меньших объемов памяти и вычислений.

4. Поэлементный метод распознавания слов сравнительно просто обобщается на распознавание слитной речи.

ГЛАВА 3

ОБУЧЕНИЕ ПОЭЛЕМЕНТНОМУ РАСПОЗНАВАНИЮ РЕЧИ

Поэлементный метод распознавания слов предполагает, что для каждого слова $k = 1 : K$ задан начальный (исходный) эталон слова E_k вместе с темпоральной τ_k , громкостной H_k и тональной F_k транскрипциями. Задать эти параметры вручную невозможно. К тому же они изменяются от диктора к диктору.

Возникает задача автоматического определения параметров слова по его обучающей выборке — совокупности реализаций слова, заговариваемых в микрофон в режиме обучения.

Алгоритм обучения распознаванию слов является составной частью поэлементного метода [79, 95—98].

§ 3.1. ПОСТАНОВКА И АНАЛИЗ ЗАДАЧИ ОБУЧЕНИЯ

Задача обучения распознаванию слов речи может рассматриваться как задача нахождения максимально правдоподобных оценок исходных эталонов слов и их транскрипций по обучающей выборке (OB), составленной из реализаций всех слов. Очевидно, что в такой постановке задача обучения распадается на K (по числу слов в словаре) независимых задач оценивания исходного эталона слова E_k и его транскрипций τ_k , H_k и F_k по OB k -го слова.

Пусть задана обучающая выборка для k -го слова, состоящая из \mathcal{U} реализаций X'_{l_r} :

$$X'_{l_r} = (x'_1, x'_2, \dots, x'_r, \dots, x'_{l_r}), \quad r = 1 : \mathcal{U}, \quad (3.1.1)$$

где l_r — длина r -й реализации в выборке.

Прежде чем записать критерий обучения, объединим исходный эталон слова E_k и транскрипции H_k и F_k в одну конструкцию, которую по-прежнему будем называть исходным эталоном слова и обозначать E_k . В самом деле, элементы h_{ks} и f_{ks} транскрипций H_k и F_k могут рассматриваться как компоненты эталонного элемента e_{ks} исходного эталона слова, поскольку элементы h_{ks} и f_{ks} используются точно так же, как и e_{ks} . Так, при нелинейных деформациях исходного эталона слова h_{ks} и f_{ks} повторяются синхронно с e_{ks} . Что же касается темпоральной транскрипции τ_k , то ее нет смысла объединять с E_k ,

поскольку она не влияет на элементарную меру сходства $g(x_i, e_{ks})$, в которую теперь будут входить в качестве слагаемого и $g_1(h_i, h_{ks})$, и $g_2(f_i, f_{ks})$, а задает лишь множество операторов в нелинейного растяжения исходного эталона.

В дальнейшем будем пользоваться в основном понятиями расширенного исходного эталона слова E_k и его темпоральной транскрипции τ_k .

Максимально правдоподобные оценки исходного эталона слова E_k (расширенного эталона) и его темпоральной транскрипции τ_k будем находить, максимизируя критерий обучения

$$\Phi(E_k, \tau_k, \{v'\}) = \sum_{r=1}^{\mathcal{U}} G(X'_{l_r}, v'E_k), \quad (3.1.2)$$

где $v' \in \tau_k (l_r)$ — оператор преобразования (растяжения) исходного эталона слова при сравнении с r -й реализацией. Максимизацию следует производить по тройке обобщенных переменных E_k, τ_k и $v', r = 1 : \mathcal{U}$.

Таким образом, в искомые при обучении распознаванию параметры, кроме E_k и τ_k , включены и мешающие параметры v' . Это сделано, во-первых, с целью уйти от необходимости интегрирования по мешающим параметрам с неизвестными и (или), главным образом, несуществующими распределениями. Во-вторых, включив v' в искомые параметры, можно будет дополнительно судить о качестве обучения на основании сегментации реализаций, задаваемых v' . Значит, v' можно рассматривать как контрольные параметры в обучении.

Итак, максимизируя (3.1.2) по тройке обобщенных переменных, вычисляем как искомые, так и контрольные параметры.

Подчеркнем еще, что в числе искомых параметров находится и длина q_k исходного эталона и транскрипций слова.

Теперь укажем на некоторые свойства критерия обучения Φ в зависимости от темпоральной транскрипции τ_k . Сначала условимся, что $m_{k1} = m_{kq_k} = 1$, а не $m_{k1} = m_{kq_k} = 0$, как было ранее. Такое ограничение не является принципиальным, однако оно упрощает последующее изложение. Физически это значит, что реализации ОВ не содержат пауз нулевой длительности в начале и конце слов, что легко выполнимо при накоплении ОВ.

В этих условиях нетрудно видеть, что для фиксированной длины q_k темпоральной транскрипции максимальное значение Φ достигается при такой темпоральной транскрипции $\tilde{\tau}_k$, в которой для всех $s = 2 : (q_k - 1)$ имеет место $m_{ks} = m$ ($m = 1$) и $M_{ks} = M$ ($M = \infty$ или $M = \max l_r$). При этом, как было установлено, всегда $m_{k1} = m_{kq_k} = 1$ и $M_{k1} = M_{kq_k} = M$.

Но точно такое же максимальное значение критерия будет, если все или отдельные элементы m_{ks} или M_{ks} , $s = 2 : (q_k - 1)$ транскрипции $\tilde{\tau}_k$ будем увеличивать начиная с m или уменьшать начиная с M соответственно. Найдется такая темпоральная транскрипция с $m_{ks} \geq m$ и $M_{ks} \leq M$, $m_{ks} \leq M_{ks}$, для которой значение критерия

обучения все еще остается максимально возможным, однако увеличение хотя бы одного m_{ks} или уменьшение хотя бы одного M_{ks} на единицу уже уменьшает критерий.

Это свойство критерия обучения позволяет решать задачу в два этапа: на первом этапе оценивать начальный эталон слова (вместе с громкостной и тональной транскрипциями), полагая темпоральную транскрипцию равной $\tilde{\tau}_k$; на втором этапе находить только темпоральную транскрипцию по результатам первого этапа.

Темпоральную транскрипцию следует выбирать с возможно большими m_{ks} и возможно меньшими M_{ks} , что обусловливается стремлением уменьшить «пересекаемость» множеств эталонных сигналов разных слов.

Итак, на первом этапе обучения необходимо оценить исходный эталон слова, максимизируя критерий

$$\Phi_1(\mathbf{E}_k, \{\mathbf{v}'\}) = \sum_{r=1}^{\mathcal{U}} G(\mathbf{X}'_{l_r}, \mathbf{v}'\mathbf{E}_k) = \sum_{r=1}^{\mathcal{U}} \sum_{i=1}^{l_r} g(x'_i, (\mathbf{v}'\mathbf{E}_k)_i) \quad (3.1.3)$$

в предположении, что $\mathbf{v}' \in \tau_k(l_r)$, а сама темпоральная транскрипция $\tilde{\tau}_k = \tau_k$ задана: $m_{k1} = m_{kq_k} = 1$; $M_{k1} = M_{kq_k} = M$; $m_{ks} = m$, например, $m = 1$ и $M_{ks} = M$ для $s = 2 : (q_k - 1)$.

Прежде чем изложить основные алгоритмы обучения распознаванию слов речи, решающие задачу (3.1.3), рассмотрим вспомогательные задачи оптимальной сегментации реализации и обучения по одной реализации ($\mathcal{U} = 1$), которые понадобятся в дальнейшем.

§ 3.2. ОПТИМАЛЬНАЯ СЕГМЕНТАЦИЯ РЕАЛИЗАЦИЙ. ОБУЧЕНИЕ ПО ОДНОЙ РЕАЛИЗАЦИИ

Исходный эталонный сигнал слова \mathbf{E}_k и его темпоральная транскрипция τ_k или $\tilde{\tau}_k$ для анализируемого сигнала $\mathbf{X}_l = (x_1, x_2, \dots, x_l, \dots, x_l)$ определяет (индуцирует) множество его разбиений (сегментаций) на q_k подпоследовательностей (сегментов) $\mathbf{X}_{w_{s-1}w_s}, s = 1 : q_k$:

$$\mathbf{X}_l = (\mathbf{X}_{w_0w_1}, \mathbf{X}_{w_1w_2}, \dots, \mathbf{X}_{w_{s-1}w_s}, \dots, \mathbf{X}_{w_{q_k-1}w_{q_k}}), \quad (3.2.1)$$

где $w_0 = 0$, $w_s > w_{s-1}$, $s = 1 : q_k$, $w_{q_k} = l$ и $\mathbf{X}_{w_{s-1}w_s} = (x_{w_{s-1}+1}, x_{w_{s-1}+2}, \dots, x_{w_s})$ является s -м сегментом реализации. Границы сегментов w_s взаимнооднозначно связаны с компонентами v_s операторов $\mathbf{v} = (v_1, v_2, \dots, v_s, \dots, v_{q_k})$, принадлежащих множествам $\tau_k(l)$ или $\tilde{\tau}_k(l)$:

$$v_s = w_s - w_{s-1}, \quad m_{ks} \leq v_s \leq M_{ks}, \quad s = 1 : q_k, \quad \sum_{s=1}^{q_k} v_s = l. \quad (3.2.2)$$

Таким образом, множество индуцируемых эталоном (\mathbf{E}_k, τ_k) сегментаций реализации \mathbf{X}_l определено так, как если бы \mathbf{X}_l была реализацией k -го слова.

Естественно назвать оптимальной сегментацией реализации \mathbf{X}_l относительно исходного эталона (\mathbf{E}_k, τ_k) ту сегментацию, которая

индуцируется оптимальным оператором $\mathbf{v}^* \in \tau_k(l)$:

$$\mathbf{v}^* = \underset{\mathbf{v} \in \tau_k(l)}{\operatorname{argmax}} G(\mathbf{X}_l, \mathbf{v}\mathbf{E}_k), \quad (3.2.3)$$

где

$$G(\mathbf{X}_l, \mathbf{v}\mathbf{E}_k) = \sum_{i=1}^l g(\mathbf{x}_i, (\mathbf{v}\mathbf{E}_k)_i) = \sum_{s=1}^{q_k} \sum_{v=\omega_{s-1}+1}^{\omega_s} g(\mathbf{x}_v, \mathbf{e}_{ks}). \quad (3.2.4)$$

Очевидно, что оптимальная сегментация однозначно определяется структурой наиболее похожего на \mathbf{X}_l эталонного сигнала k -го слова. В свою очередь, оптимальная сегментация позволяет интерпретировать структуру сигнала \mathbf{X}_l как реализации слова k .

Задача оптимальной сегментации (3.2.3) совпадает с задачей поэлементного распознавания (сравнения с исходным эталоном слова) по используемому критерию и отличается от нее тем, что необходимо указать не величину оптимального сходства, а то значение оператора $\mathbf{v} = \mathbf{v}^* \in \tau_k(l)$, при котором это максимальное сходство достигается.

С этой целью, наряду с вычислениями $F_{kl}(s)$, $s = 1 : q_k$ по рекуррентным формулам (2.3.7) — (2.3.15), одновременно будем запоминать и величины $u_{ki}(s)$, имеющие смысл потенциально-оптимальных длин сегментов:

$$F_{ki}(1) = F_{k(l-1)}(1) + g(\mathbf{x}_i, \mathbf{e}_{kl}), \quad (3.2.5)$$

$$u_{ki}(1) = u_{k(l-1)}(1) + 1; \quad (3.2.6)$$

$$u_{ki}(s) = \underset{m_{ks} \leqslant u \leqslant M_{ks}}{\operatorname{argmax}} \left(F_{k(l-u)}(s-1) + \sum_{v=i-u+1}^l g(\mathbf{x}_v, \mathbf{e}_{ks}) \right), \quad (3.2.7)$$

$$F_{kl}(s) = F_{k(l-u_{kl}(s))}(s-1) + \sum_{v=i-u_{kl}(s)+1}^l g(\mathbf{x}_v, \mathbf{e}_{ks}), \quad s = 2 : (q_k - 1), \quad (3.2.8)$$

$$F_{kl}(q_k) = \max(F_{kl}(q_k - 1), F_{k(l-1)}(q_k) + g(\mathbf{x}_i, \mathbf{e}_{kq_k})), \quad (3.2.9)$$

$$u_{kl}(q_k) = \begin{cases} 0, & \text{если } F_{kl}(q_k - 1) \geqslant F_{k(l-1)}(q_k) + g(\mathbf{x}_i, \mathbf{e}_{kq_k}), \\ u_{k(l-1)}(q_k) + 1 & \text{в противном случае.} \end{cases} \quad (3.2.10)$$

Вычисления начинаем, положив $F_{k0}(1) = 0$ и $u_{k0}(1) = 0$ и считая все неопределенные справа в формулах (3.2.5) — (3.2.10) величины $F_{k(l-u)}(s)$ равными $-\infty$.

После запоминания $u_{kl}(s)$ для всех $s = 1 : q_k$ и для всех $i = 1 : l$ находим оптимальные длины v_s и оптимальные границы w_s сегментов реализации \mathbf{X}_l , используя рекуррентные формулы выписывания:

$$\begin{aligned} w_{q_k} &= l, \quad v_s = u_{k\omega_s}(s), \\ w_{s-1} &= w_s - v_s, \quad s = q_k, q_{k-1}, q_{k-2}, \dots, 1. \end{aligned} \quad (3.2.11)$$

Оптимальную сегментацию реализации \mathbf{X}_l с помощью исходного эталона (\mathbf{E}_k , τ_k) слова k , которому \mathbf{X}_l принадлежит, будем называть для краткости сегментацией реализации \mathbf{X}_l , а соответствующий алгоритм оптимальной сегментации — алгоритмом сегментации.

С помощью алгоритма сегментации анализируемые реализации \mathbf{X}_l^r , $r = 1 : \mathcal{U}$, одного и того же слова могут быть просегментированы и для каждого эталонного элемента e_{ks} могут быть указаны s -е сегменты (части) реализаций, интерпретируемые как одни и те же звуки в различных реализациях слова, а точнее, как звуки, аппроксимируемые одним и тем же эталонным элементом e_{ks} исходного эталона слова. Сегментация реализаций позволяет найти соответствующие друг другу (одноименные) участки в разных реализациях слова.

Кроме сегментации понадобится еще и самосегментация реализации \mathbf{X}_l . Она осуществляется в тех же условиях, что и сегментация, с той существенной разницей, что исходный эталон слова E_k не задан, а известна только темпоральная транскрипция τ_k . Более того, чаще всего считается, что τ_k имеет стандартный вид $\tilde{\tau}_k$, т. е. $m_{ks} = m$, кроме $m_{k1} = m_{kq_k} = 1$ и $M_{ks} = M$, $s = 1 : q_k$, и тогда самосегментация реализации \mathbf{X}_l рассматривается как зависящая только от одного параметра $q_k = 1 : l$.

Целью оптимальной самосегментации реализации \mathbf{X}_l является такое ее разбиение на q_k сегментов согласно транскрипции τ_k , при котором достигает максимума критерий качества самосегментации:

$$\mathbf{v}^* = \underset{\mathbf{v} \in \tau_k^{(l)}}{\operatorname{argmax}} G(\mathbf{X}_l, \mathbf{v}), \quad (3.2.12)$$

где

$$G(\mathbf{X}_l, \mathbf{v}) = \max_{\mathbf{E}} G(\mathbf{X}_l, \mathbf{v}\mathbf{E}) = \sum_{s=1}^{q_k} \max_{\mathbf{e}} \sum_{v=w_{s-1}+1}^{w_s} g(\mathbf{x}_v, \mathbf{e}). \quad (3.2.13)$$

Как следует из критерия, при самосегментации находят наиболее однородные (по содержащимся в них элементам) сегменты, такие, что суммарный критерий однородности сегментов, вычисленный по всей реализации, принимает наибольшее значение. Самосегментация — это так же сегментация, но с возможностью свободного выбора наилучших эталонных элементов слова.

Задача самосегментации по постановке и методу решения аналогична сегментации.

Обозначим через $d_{ki}(u, s)$ величину:

$$d_{ki}(u, s) = \max_{\mathbf{e}} \sum_{v=i-u+1}^i g(\mathbf{x}_v, \mathbf{e}). \quad (3.2.14)$$

Тогда, подобно алгоритму сегментации (3.2.5) — (3.2.10), таблица потенциально-оптимальных длин сегментов $u_{ki}(s)$ заполняется с помощью следующих рекуррентных формул ДП:

$$u_{ki}(s) = \underset{m_{ks} \leq u \leq M_{ks}}{\operatorname{argmax}} (F_{k(i-u)}(s-1) + d_{ki}(u, s)), \\ F_{ki}(s) = F_{k(i-u_{ki}(s))}(s-1) + d_{ki}(u_{ki}(s), s), \quad s = 1 : q_k, \quad i = 1 : l, \quad (3.2.15)$$

причем, как и раньше, $F_{k0}(1) = 0$, $u_{k0}(1) = 0$, а все неопределенные справа величины $F_{k(i-u)}(s)$ равны $-\infty$.

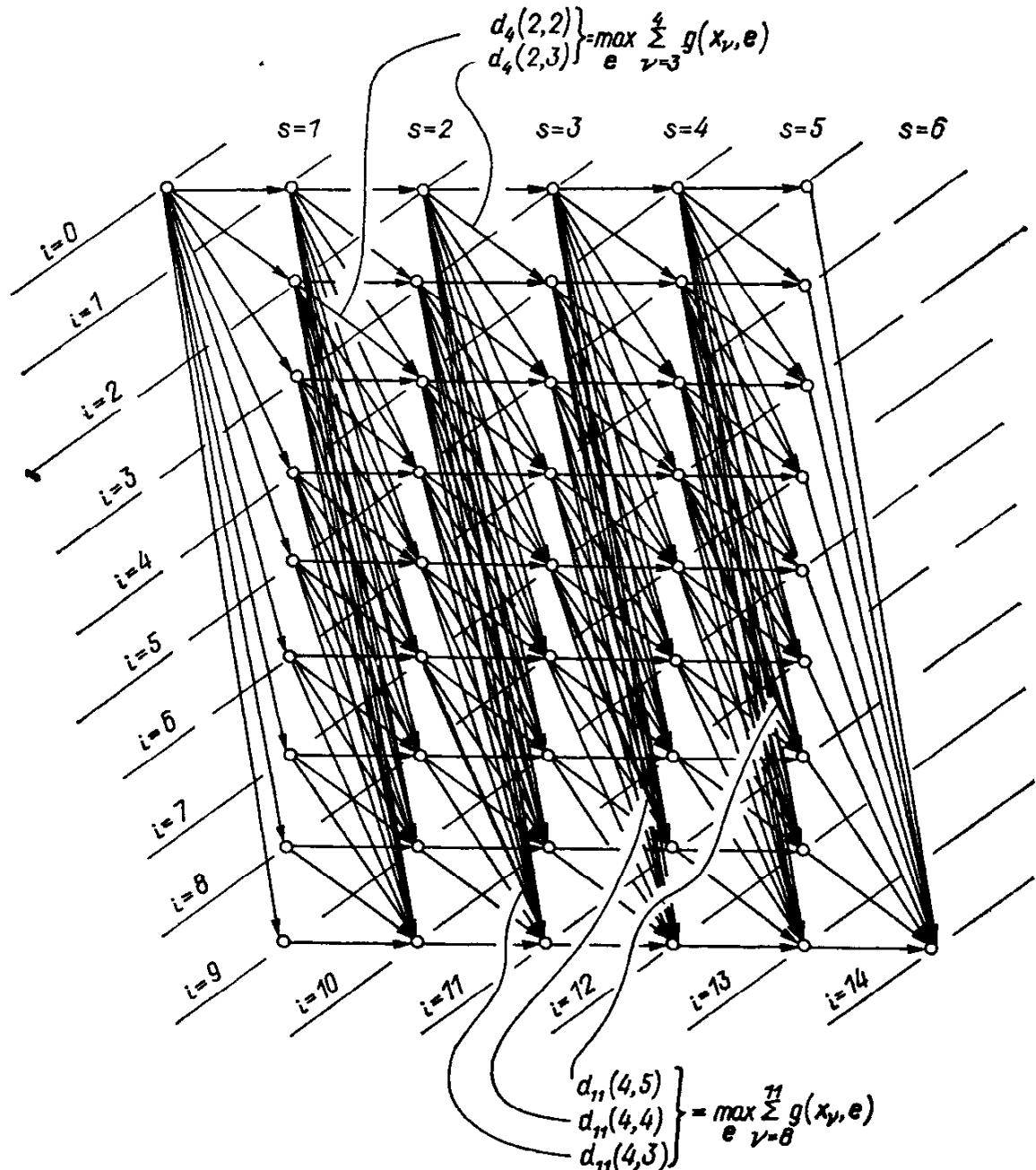


Рис. 3.1. Развернутый граф решения задачи самосегментации реализации.

Развернутый граф решения задачи самосегментации представлен на рис. 3.1 ($l = 14$, $\tau_k = \tilde{\tau}_k$, $m = 1$, $q_k = 6$).

Оптимальные границы самосегментации выписываем на основании u_{kl} (s) по той же формуле (3.2.11). Отправляясь от оптимальных границ w_s самосегментации, нетрудно найти оптимальный при данной темпоральной транскрипции τ_k , исходный эталон слова k :

$$e_{ks} = \operatorname{argmax}_e \sum_{v=w_{s-1}+1}^{w_s} g(x_v, e), \quad s = 1 : q_k. \quad (3.2.16)$$

Однако этот исходный эталон E_k слова является всего лишь решением задачи обучения (3.1.3) по одной реализации.

В отличие от сегментации самосегментация реализаций X'_k ОВ не позволяет находить одноименные (однотипные) участки во всех

реализациях слова, зато она дает возможность получить \mathcal{U} различных исходных эталонов слова (результатов обучения по одной реализации), которые можно будет использовать в качестве начального приближения в итерационных алгоритмах обучения.

Сравнивая сегментацию и самосегментацию реализации X_i , убеждаемся, что последняя зависит по существу только от длины q_k исходного эталона слова и его темпоральной транскрипции и параметра m . Имеет смысл рассматривать так называемую m -самосегментацию реализации, когда количество сегментов q_k не задано и его требуется найти в процессе m -самосегментации, а задано только ограничение m на длину сегмента.

Очевидно, что при m -самосегментации длина сегментов не может равняться или быть больше $2m$, поскольку в этом случае любой такой сегмент можно будет разбить на несколько сегментов длиной u , $m \leq u < 2m$.

При m -самосегментации рекуррентные формулы (3.2.15) модифицируются следующим образом:

$$\begin{aligned} u_i &= \operatorname{argmax}_{m \leq u < 2m} (F_{t-u} + d_i(u)), \\ F_i &= F_{t-u_i} + d_i(u_i), \\ q_i &= q_{t-u_i} + 1, \quad i = 1 : l, \end{aligned} \tag{3.2.17}$$

причем аналогично (3.2.14)

$$d_i(u) = \max_e \sum_{v=t-u+1}^t g(x_v, e) \tag{3.2.18}$$

и, как и раньше, $F_0 = 0$, $u_0 = 0$, $q_0 = 0$, а все неопределенные справа в (3.2.17) величины F_{t-u} равны $-\infty$.

Величина $q = q_i$ определит оптимальное количество сегментов m -самосегментации, а границы сегментов можно выписать по формуле (3.2.11), положив $q_k = q$, если до этого запомнить таблицу u_i , $i = 1 : l$, в вычислениях по (3.2.17).

Располагая границами ω_s , $s = 1 : q$, m -самосегментации, по формуле (3.2.16) можно найти исходный эталон слова (осуществить обучение по одной реализации) в условиях незаданной длины q_k исходного эталона слова.

§ 3.3. ТОЧНОЕ РЕШЕНИЕ ЗАДАЧИ ОБУЧЕНИЯ

Точное решение задачи (3.1.3) нахождения исходного эталона E_k по обучающей выборке (3.1.1) в принципе может быть получено с помощью динамического программирования.

Обозначим через $F(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ оптимальное значение критерия качества обучения (3.1.3) для случая, когда ОВ составляют начальные части $X'_{i_r} = (x'_1, x'_2, \dots, x'_{i_r})$, $i_r < l_r$, $r = 1 : \mathcal{U}$, исходных реализаций ОВ. Пусть $q(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ является оптимальным значением длины полученного таким образом исходного эталона.

Пусть величины $F(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ и $q(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ вычислялись в предположении темпоральной транскрипции $\tilde{\tau}_k$, такой, что для всех элементов $m_{ks} = m$ и $M_{ks} = M$, $s = 1 : q(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$, кроме $m_{kl} = 1$. И только для случаев выполнения хотя бы одного условия $i_r = l_r$, $r = 1 : \mathcal{U}$, будем полагать $\tilde{\tau}_k = \tilde{\tau}_k$, т. е. для всех элементов имеет место $m_{ks} = m$ и $M_{ks} = M$, кроме $m_{kl} = m_{kq(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})} = 1$.

Предположим, что для начальных длин реализаций, меньших i_r , уже найдены оптимальное значение критерия качества обучения и соответствующее ему значение длины исходного эталона. Тогда $F(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ и $q(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ для нового набора $(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ с увеличенной на единицу хотя бы одной компонентой i_r , вычисляются с помощью рекуррентных формул ДП:

$$F(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}) = \max_{v=(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}})} (F(i_1 - v_1, i_2 - v_2, \dots, \dots, i_r - v_r, \dots, i_{\mathcal{U}} - v_{\mathcal{U}}) + \\ + G_{i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}}(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}})), \quad (3.3.1)$$

где

$$G_{i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}}(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}}) = \max_e \sum_{r=1}^{\mathcal{U}} \sum_{\mu=i_r - v_r + 1}^{i_r} g(x_{\mu}, e). \quad (3.3.2)$$

Одновременно с вычислениями по формуле (3.3.1) будем запоминать потенциально-оптимальные длины сегментов реализаций

$$\mathbf{v}(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}) = \\ = \operatorname{argmax}_{v=(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}})} (F(i_1 - v_1, i_2 - v_2, \dots, i_r - v_r, \dots, i_{\mathcal{U}} - v_{\mathcal{U}}) + \\ + G_{i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}}(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}})). \quad (3.3.3)$$

Новое значение длины эталона теперь определяется формулой

$$q(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}) = \\ = q(i_1 - v_r^*, i_2 - v_r^*, \dots, i_r - v_r^*, \dots, i_{\mathcal{U}} - v_{\mathcal{U}}^*) + 1, \quad (3.3.4)$$

где v_r^* — r -я компонента вектора $\mathbf{v}(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$:

$$v_r^* = v_r(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}).$$

Вычисления по (3.3.1) — (3.3.4) выполняем последовательно для всех возможных возрастающих наборов $(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$, $i = 1 : l_r$, $r = 1 : \mathcal{U}$, начиная с набора $(1, 1, \dots, 1, \dots, 1)$. При этом полагаем, что $F(0, 0, \dots, 0, \dots, 0) = 0$, а все не определенные справа в (3.3.1) величины $F(i_1 - v_1, i_2 - v_2, \dots, i_r - v_r, \dots, i_{\mathcal{U}} - v_{\mathcal{U}})$ равны $-\infty$. Также полагаем, что $q(0, 0, \dots, 0, \dots, 0) = 0$.

Величина $F(l_1, l_2, \dots, l_r, \dots, l_{\mathcal{U}})$, найденная для заключительного набора $(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$, совпадет с оптимальным значением критерия качества обучения $\Phi_1(\mathbf{E}_k, \{\mathbf{v}^r\})$, а оптимальная длина q_k исходного эталона слова будет равна $q(l_1, l_2, \dots, l_r, \dots, l_{\mathcal{U}})$. Чтобы найти сам исходный эталон, требуется сначала по таблице $\mathbf{v}(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$ выписать согласованную оптимальную сегментацию реализаций \mathbf{X}'_i :

$$\begin{aligned} w'_{q_k} &= l_r, \quad r = 1 : \mathcal{U}; \quad v'_s = \mathbf{v}_r(w_s^1, w_s^2, \dots, w_s^{r'}, \dots, w_s^{\mathcal{U}}), \\ w'_{s-1} &= w'_s - v'_s, \quad r = 1 : \mathcal{U}; \\ s &= q_k, \quad q_k - 1, \quad q_k - 2, \quad \dots, \quad 1. \end{aligned} \quad (3.3.5)$$

Затем находим оптимальные эталонные элементы

$$\mathbf{e}_{ks} = \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{r=1}^{\mathcal{U}} \sum_{v=w'_{s-1}+1}^{w'_s} g(\mathbf{x}'_v, \mathbf{e}), \quad s = 1 : q_k, \quad (3.3.6)$$

что завершает решение первого этапа задачи обучения.

Однако алгоритм точного решения задачи обучения является весьма громоздким. Так, рекуррентные формулы (3.3.1) — (3.3.4) необходимо применять $\prod_{r=1}^{\mathcal{U}} l_r$ раз — именно столько будет различных наборов $(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}})$. Например, при средней длине реализаций $l_{\text{ср}} = 70$ и $\mathcal{U} = 10$ это число будет равно 70^{10} . Для запоминания потенциально-оптимальных длин сегментов также понадобится память на

$\prod_{r=1}^{\mathcal{U}} l_r$ чисел. В то же время даже однократное применение формул (3.3.1), (3.3.2) характеризуется значительной громоздкостью, которая практически не устраняется, если, воспользовавшись свойствами задачи, несколько (хотя и существенно) сузить область поиска решений по переменной \mathbf{v} . Так, оптимальность решений не изменится, если считать в (3.3.1) и (3.3.3) вектор $\mathbf{v} = (v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}})$ принадлежащим множеству $\Omega_{l_1, l_2, \dots, l_r, \dots, l_{\mathcal{U}}}(\mathbf{v})$, такому, что хотя бы для одной компоненты v_r выполняется условие $m \leq v_r < 2m$, в частности $v_r = 1$, если $m = 1$.

Для практического использования рекомендуется итерационный алгоритм обучения, называемый в дальнейшем основным алгоритмом обучения. Он использует как составную часть алгоритмы сегментации и самосегментации реализаций. Одна итерация итерационного алгоритма обучения содержит $\sum_{r=1}^{\mathcal{U}} l_r$ существенно более простых, чем в (3.3.1) — (3.3.4), операций, количество же итераций практически оказывается не превосходящим двух десятков.

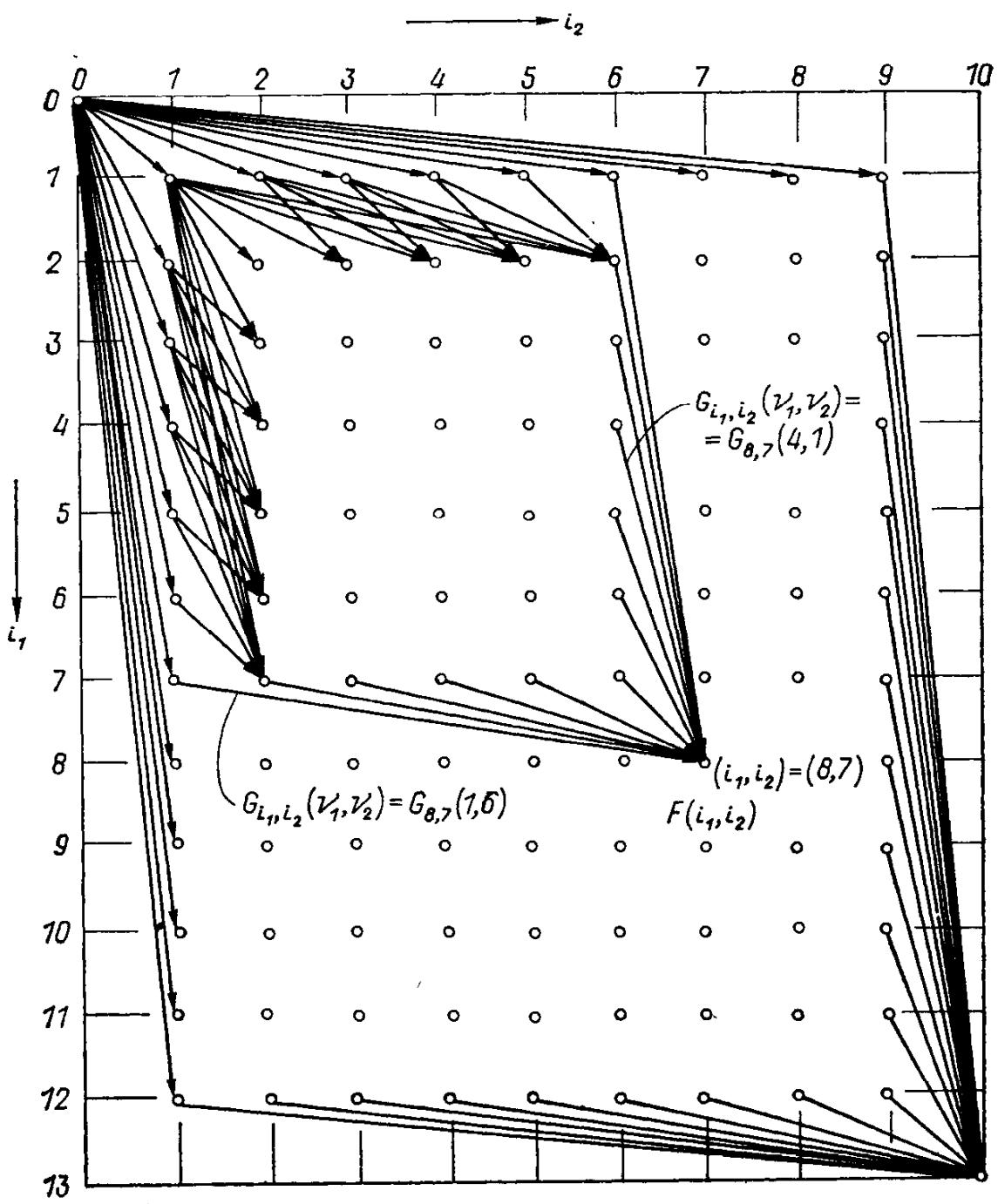


Рис. 3.2. Структура графа точного решения задачи обучения для случая $N = \mathcal{U}$, $l_1 = 13$, $l_2 = 10$ и $m = 1$.

Поэтому в противовес точному алгоритму решения задачи обучения с трудоемкостью, оцениваемой как $\prod_{r=1}^{\mathcal{U}} l_r$, основной алгоритм имеет существенно меньшую трудоемкость, оцениваемую примерно в тех же единицах как $\sum_{k=1}^{\mathcal{U}} l_k$.

Точное решение задачи обучения обращает внимание на то, что даже при $m = 1$ оптимальное значение длины q_k исходного эталона вовсе не обязательно равно $\min l_r$.

Точный алгоритм решения задачи обучения можно применять при небольшом количестве реализаций ОВ, например $\mathcal{U} = 2 - 4$.

На рис. 3.2 в качестве примера приведен граф точного решения задачи обучения для случая $\mathcal{U} = 2$, $l_1 = 13$, $l_2 = 10$ и $m = 1$. Показана только часть дуг: все дуги, входящие в вершины $(i_1, i_2) = (8,7)$, $(i_1, i_2) = (13,10)$, $(i_1, 2)$, $i_1 < 8$, $(2, i_2)$, $i_2 < 7$, и все дуги, выходящие из вершины $(i_1, i_2) = (0,0)$. Из этого рисунка следует, что формула ДП (3.3.1) при $m = 1$ и $\mathcal{U} = 2$ упрощается к виду

$$\begin{aligned} F(i_1, i_2) = \max & \left(\max_{1 \leq v_1 < i_1} (F(i_1 - v_1, i_2 - 1) + G_{i_1, i_2}(v_1, 1)), \right. \\ & \left. \max_{1 \leq v_2 < i_2} (F(i_1 - 1, i_2 - v_2) + G_{i_1, i_2}(1, v_2)) \right), \end{aligned} \quad (3.3.7)$$

а в случае произвольного $\mathcal{U} > 1$ та же формула (3.3.1) принимает вид

$$\begin{aligned} F(i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}) = & \\ = \max & \max_r \left(F(i_1 - v_1, i_2 - v_2, \dots, i_r - v_r, \dots, i_{\mathcal{U}} - v_{\mathcal{U}}) + \right. \\ & \left. + G_{i_1, i_2, \dots, i_r, \dots, i_{\mathcal{U}}}(v_1, v_2, \dots, v_r, \dots, v_{\mathcal{U}}) \right). \end{aligned} \quad (3.3.8)$$

Соответственно модифицируется и формула (3.3.3).

§ 3.4. ОСНОВНОЙ АЛГОРИТМ ОБУЧЕНИЯ

Основной алгоритм обучения является итерационным. Он реализует обобщенную покоординатную оптимизацию в пространстве двух обобщенных переменных \mathbf{E}_k и $\{\mathbf{v}^r\}$.

Рассмотрим сначала случай заданной длины эталона q_k .

Пусть в результате n -й итерации уже найден начальный эталон слова $\mathbf{E}_k^n = (\mathbf{e}_{k1}^n, \mathbf{e}_{k2}^n, \dots, \mathbf{e}_{ks}^n, \dots, \mathbf{e}_{kq_k}^n)$.

Тогда на $(n + 1)$ -й итерации последовательно выполняем два шага.

1-й шаг. При условии фиксированного \mathbf{E}_k^n решаем задачу нахождения оптимальной совокупности

$$\{\mathbf{v}^{(n+1)r}\} = \underset{\{\mathbf{v}^r\}}{\operatorname{argmax}} \Phi_1(\mathbf{E}_k^n, \{\mathbf{v}^r\}), \quad (3.4.1)$$

которая задает оптимальную сегментацию реализаций ОВ на q_k сегментов (подпоследовательностей).

Из сепарабельности критерия обучения (3.1.3) по переменным \mathbf{v}^r при условии заданного \mathbf{E}_k следует, что задача (3.4.1) распадается на \mathcal{U} независимых задач нахождения оптимальных операторов

$$\mathbf{v}^{(n+1)r} = \underset{\mathbf{v}^r \in \tilde{\tau}_k^{(l_r)}}{\operatorname{argmax}} G(\mathbf{X}'_{l_r}, \mathbf{v}^r \mathbf{E}_k^n), \quad r = 1 : \mathcal{U}. \quad (3.4.2)$$

Но каждая из задач (3.4.2) является задачей сегментации реализации \mathbf{X}'_{l_r} относительно исходного эталона $(\mathbf{E}_k^n, \tilde{\tau}_k)$, подробно рассмотренной в § 3.2.

Решив \mathcal{U} задач сегментации с помощью алгоритма сегментации (3.2.5) — (3.2.11), найдем границы всех $s = 1 : q_k$ сегментов всех

$r = 1 : \mathcal{U}$ реализаций ОВ:

$$\omega_0^{(n+1)r} = 0, \quad \omega_s^{(n+1)r}, \quad s = 1 : q_k, \quad \omega_{q_k}^{(n+1)r} = l_r, \quad r = 1 : \mathcal{U}. \quad (3.4.3)$$

Полученная информация о границах сегментов (3.4.3) позволяет приступить к выполнению второго шага итерации.

2-й шаг. При фиксированной совокупности $\{\mathbf{v}^{(n+1)r}\}$ оптимальных сегментаций реализаций ОВ решаем задачу нахождения начального эталона слова

$$\mathbf{E}_k^{n+1} = \operatorname{argmax}_{\mathbf{E}_k} \Phi_1(\mathbf{E}_k, \{\mathbf{v}^{(n+1)r}\}). \quad (3.4.4)$$

С этой целью критерий обучения (3.1.3) перепишем в виде

$$\Phi_1(\mathbf{E}_k, \{\mathbf{v}^r\}) = \sum_{r=1}^{\mathcal{U}} \sum_{s=1}^{q_k} \sum_{i=\omega_{s-1}^r + 1}^{\omega_s^r} g(\mathbf{x}_i^r, \mathbf{e}_{ks}), \quad (3.4.5)$$

что в силу сепарабельности критерия обучения по переменным \mathbf{e}_{ks} позволяет свести решение задачи (3.4.4) к решению q_k независимых задач нахождения эталонных элементов слова:

$$\mathbf{e}_{ks}^{n+1} = \operatorname{argmax}_{\mathbf{e}} \sum_{r=1}^{\mathcal{U}} \sum_{i=\omega_{s-1}^{(n+1)r} + 1}^{\omega_s^{(n+1)r}} g(\mathbf{x}_i^r, \mathbf{e}), \quad s = 1 : q_k. \quad (3.4.6)$$

В результате решения задач (3.4.6) будет получено новое значение исходного эталона слова \mathbf{E}_k^{n+1} — всех его эталонных элементов \mathbf{e}_{ks}^{n+1} вместе с компонентами громкости h_{ks}^{n+1} и тональности f_{ks}^{n+1} .

$$h_{ks}^{n+1} = \operatorname{argmax}_{h \in \Omega_1} \sum_{r=1}^{\mathcal{U}} \sum_{i=\omega_{s-1}^{(n+1)r} + 1}^{\omega_s^{(n+1)r}} g_1(h_i^r, h), \quad (3.4.7)$$

$$f_{ks}^{n+1} = \operatorname{argmax}_{f \in \Omega_2} \sum_{r=1}^{\mathcal{U}} \sum_{i=\omega_{s-1}^{(n+1)r} + 1}^{\omega_s^{(n+1)r}} g_2(f_i^r, f), \quad (3.4.8)$$

где Ω_1 и Ω_2 — множества допустимых значений громкости и тональности соответственно.

Условие останова итерационного алгоритма обучения запишется в виде

$$\mathbf{E}_k^{n+1} = \mathbf{E}_k^n. \quad (3.4.9)$$

Нетрудно убедиться, что если условие останова (3.4.9) не достигнуто, то от итерации к итерации критерий качества обучения увеличивает свое значение. В силу конечного количества возможных сегментаций ОВ на q_k сегментов предлагаемый алгоритм обучения является конечно-сходящимся. Для типичных элементарных мер сходства g , приведенных в § 2.3, и значений $\mathcal{U} = 10$ процесс обучения для заданного q_k сходился не более чем за 10 итераций во всех проведенных экспериментах по обучению поэлементному распознаванию слов.

/ Предлагаемый алгоритм обучения не гарантирует глобального решения задачи обучения для данного q_k . Однако, как показывает анализ результатов обучения, в частности, контрольных параметров обучения, он обеспечивает качественно хорошую, согласованную по всем реализациям сегментацию реализаций на квазистационарные участки, осуществляет адекватный поиск одинаковых (одноименных, аппроксимируемых одним и тем же эталонным элементом) сегментов в разных реализациях ОВ слова и, главное, обеспечивает высокую надежность распознавания слов речи.

Характерно, что в двухшаговой итерации алгоритма на первом шаге осуществляется поиск одноименных сегментов реализаций ОВ (назовем это согласованием по горизонтали), а на втором — выбор наилучшего эталонного элемента для всех одноименных сегментов, взятых из всех реализаций (назовем это согласованием по вертикали). Многократное же повторение итераций приводит к некоторому наилучшему согласованию одновременно и по горизонтали, и по вертикали. Во всяком случае, получаемое решение таково, что, меняя только одну из обобщенных переменных, улучшить значение критерия качества обучения уже не удается.

Для полного описания итерационного алгоритма обучения необходимо еще указать алгоритм выбора начальных условий. Очевидна также и та роль, которая отводится этим условиям в итерационных алгоритмах. В качестве «хороших» начальных условий предлагаются значения исходного эталона слова \mathbf{E}_k^{0r} , полученные в результате решения задач обучения по одной реализации \mathbf{X}'_l , $r = 1 : \mathcal{U}$:

$$\mathbf{E}_k^{0r} = \operatorname{argmax}_{\mathbf{E}_k} \max_{\mathbf{v} \in \tilde{\tau}_k(l)} G(\mathbf{X}'_l, \mathbf{v}\mathbf{E}_k), \quad r = 1 : \mathcal{U}. \quad (3.4.10)$$

Каждая из задач (3.4.10) решается с помощью алгоритма самосегментации, подробно изложенного в § 3.2. Решение это определяется формулами (3.2.11) — (3.2.16).

Можно осуществить \mathcal{U} запусков итерационного алгоритма с помощью начальных условий \mathbf{E}_k^{0r} , $r = 1 : \mathcal{U}$, и в качестве окончательного решения задачи обучения выбрать то, которое характеризуется абсолютно наибольшим значением критерия качества обучения.

Осталось выяснить, как при использовании основного алгоритма обучения следует выбирать значение длины q_k исходного эталона слова.

Еще в предыдущем параграфе подчеркивалось, что критерий качества обучения достигает своего максимального значения при некотором значении q_k^* , лежащем внутри интервала $[1, \min l_r]$. В качестве иллюстрации рассмотрим задачу обучения для простейшего случая, когда ОВ состоит из трех реализаций $\mathcal{U} = 3$, элементы содержат только одну компоненту с возможными значениями 0 и 1, элементарная мера сходства $g(\mathbf{x}_l, \mathbf{e}_l) = -|\mathbf{x}_l - \mathbf{e}_l|^2$ и эталонные элементы \mathbf{e}_l принимают значения из отрезка $0 \leq e_l \leq 1$. Пусть реализации ОВ имеют одинаковую длину $l_r = 7 : \mathbf{X}_7^1 = (0, 0, 0, 1, 1, 0, 0)$,

$\mathbf{X}_7^2 = (1, 0, 1, 1, 0, 0, 1)$, $\mathbf{X}_7^3 = (0, 1, 0, 0, 1, 1, 0)$. Решая задачу обучения для всех $q_k = 1 : 7$, получим качество обучения Φ_1 в зависимости от q_k :

$$\Phi_1(1) = 5,14; \Phi_1(2) = 4,48; \Phi_1(3) = 2,34; \Phi_1(4) = 1,33; \Phi_1(5) = 2,47; \Phi_1(6) = 2,91; \Phi_1(7) = 4,65.$$

Убеждаемся, что оптимальное значение длины исходного эталона равно $q_k^* = 4$.

При тех же условиях для выборки $\mathbf{X}_8^1 = (0, 0, 1, 1, 1, 1, 0, 0)$, $\mathbf{X}_8^2 = (0, 0, 0, 0, 0, 1, 1, 0)$, $\mathbf{X}_8^3 = (0, 1, 1, 1, 0, 0, 0, 0)$ получим, что $q_k = 3$ или $q_k^* = 4$, причем в обоих случаях $\Phi_1(3) = \Phi_1(4) = 0$.

Из приведенных примеров следует способ автоматического выбора длины q_k исходного эталона слова — необходимо применить основной алгоритм обучения для различных $q_k = q, q + 1, \dots, Q$; $q > 1$ и $Q < \min l_i$, и в качестве окончательного значения q_k выбрать то, для которого достигается абсолютно наибольшее значение критерия качества обучения.

Изучение зависимости критерия качества обучения от длины q_k начального эталона слова показывает, однако, что возможны гораздо более простые пути автоматического выбора q_k , существенно упрощающие основной алгоритм обучения.

§ 3.5. ВЫБОР ДЛИНЫ ИСХОДНОГО ЭТАЛОНА СЛОВА

На рис. 3.3 показан характерный график зависимости максимума критерия качества обучения Φ_1 от длины q_k исходного эталона слова. Эта зависимость получена для $U = 10$ реализаций слова ОДИН в произнесении диктора-мужчины. При этом интервал анализа $\Delta T'$ и шаг анализа ΔT были одинаковы и равны 18 мс. Компонентами элементов

речи были значения эффективных амплитуд сигналов на выходе цифровых резонансных фильтров. Количество фильтров (размерность элементов речи) равнялось 25. После предварительной обработки элементы x_i каждой реализации нормировались из условия $\max_{1 \leq i \leq l} |x_i| = 1$.

В качестве элементарной меры сходства использовалась функция $g(x_i, e_{ks}) = -|x_i - e_{ks}|^2$.

Как видно из рис. 3.3, наилучшее качество обучения достигает-

Рис. 3.3. Зависимость качества обучения от длины исходного эталона слова.

ся при $q_k^* = 20$. Однако и при $q_k = 9$ качество достаточно высокое, незначительно хуже наилучшего, но требуется память в два раза меньше для запоминания эталонных элементов слова.

Из рис. 3.3 видно, что существует достаточно широкий интервал значений q_k , для которых достигается близкое к оптимальному каче-

ство обучения. В этих условиях предпочтение следует отдать возможно меньшим значениям с целью экономии памяти на хранение исходных эталонов слова.

Некритичность к выбору значения q_k и последующие многочисленные эксперименты показали [61, 89, 95], что выбор значения q_k может делаться учителем одновременно с вводом текстового эквивалента слова либо осуществляться автоматически по тексту слова — сначала по тексту найти фонетическую транскрипцию слова, а затем положить длину q_k исходного эталона слова равной увеличенному в α раз ($\alpha = 1,5 - 2,0$) количеству символов в фонетической транскрипции плюс два символа на паузы в начале и конце слова.

Такой способ выбора q_k существенно упрощает задачу обучения в сравнении с другими способами автоматического выбора q_k , когда необходимо делать перебор возможных значений q_k .

§ 3.6. ФОРМИРОВАНИЕ ТЕМПОРАЛЬНОЙ ТРАНСКРИПЦИИ

После решения задачи (3.1.3) оценивания исходного эталона слова E_k , что делается на первом этапе, на втором этапе формируется темпоральная транскрипция слова τ_r .

Как отмечалось в § 3.1, темпоральную транскрипцию составляют, отправляясь от τ_k : стремятся выбрать такие максимально возможные m_{ks} и минимально возможные M_{ks} , чтобы критерий качества обучения не уменьшил своего значения. Эти рассуждения приводят к следующему приему нахождения элементов транскрипции τ_k по границам w_s' , оптимальных сегментаций реализаций ОВ, вычисленных точным алгоритмом (формулы (3.3.5)) или итерационным, на заключительной итерации, алгоритмом:

$$\begin{aligned} m_{ks} &= \min_s (w_s' - w_{s-1}') \geq m, \\ M_{ks} &= \max_s (w_s' - w_{s-1}') \leq M, \\ s &= 2 : (q_k - 1). \end{aligned} \quad (3.6.1)$$

Что же касается m_{k1} , m_{kq_k} , M_{k1} и M_{kq_k} , то принудительно полагаем $m_{k1} = m_{kq_k} = 0$ и $M_{k1} = M_{kq_k} = \infty$.

Заметим только, что выбором ограничений m и M , влияющих на решение первого этапа задачи обучения, можно изменять структуру множества эталонных сигналов. В частности, если положить $m = 3$ и $M = 5$ для $s = 2 : (q_k - 1)$, то эталонные сигналы слов (без пауз на концах) будут в 3—5 раз длиннее исходного эталона (также без пауз на концах).

§ 3.7. ПРИМЕРЫ РЕШЕНИЙ ЗАДАЧИ ОБУЧЕНИЯ

На рис. 3.4 приведен пример исходного эталона слова ОДИН, полученный в тех же условиях, при которых снималась зависимость качества обучения от длины исходного эталона (см. § 3.5 и рис. 3.3).

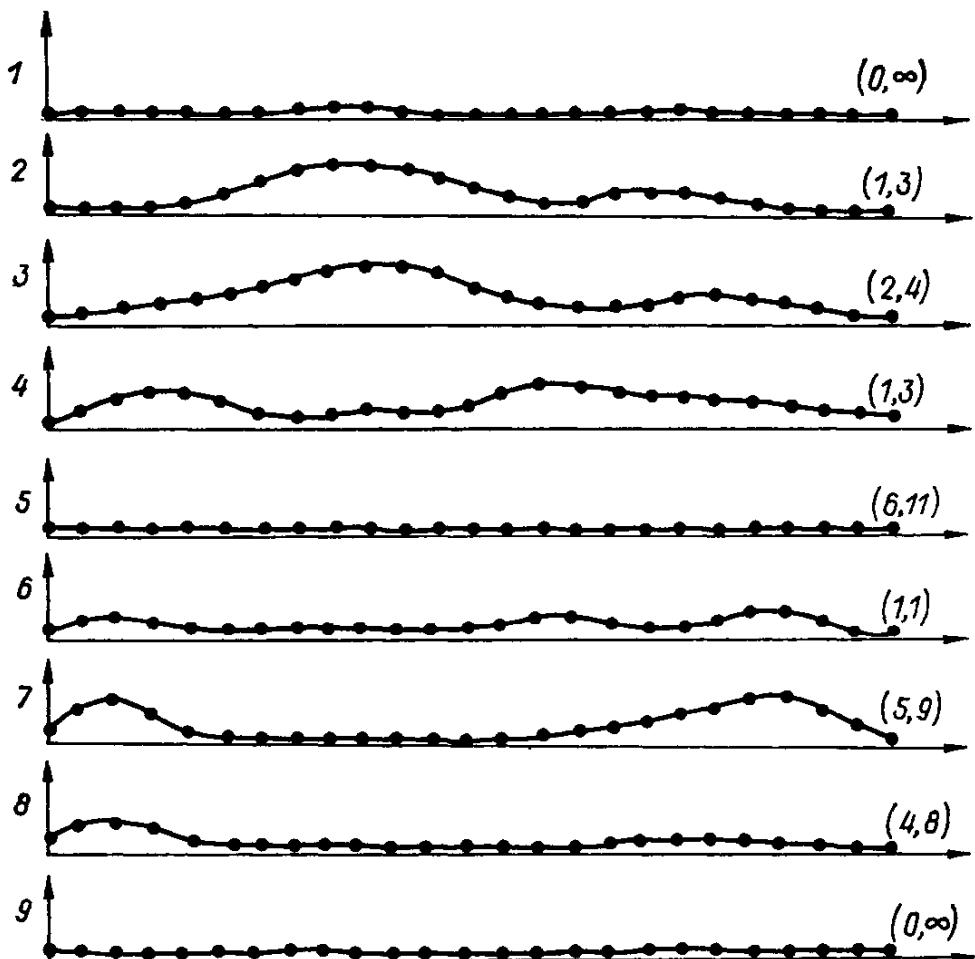


Рис. 3.4. Исходный эталонный сигнал слова ОДИН.

Обучающая выборка состояла из $\mathcal{U} = 10$ реализаций. Длина исходного эталона слова q_k равна 9. По осям абсцисс отложена частота (всего 25 дискретных значений), по оси ординат — спектральная амплитуда. Цифры у начал координат указывают порядковый номер эталонного элемента, а пара чисел у оси абсцисс является парой (m_{ks}, M_{ks}) соответствующего элемента темпоральной транскрипции. Первый и девятый элементы — эталонные элементы пауз в начале и конце слова соответственно. Элемент 5 отвечает за смычку внутри слова ОДИН. Шестой эталонный элемент — переходной от смычки к звуку И, поскольку для него $m_{ks} = M_{ks} = 1$. Элементы 2, 3 и 4 представляют различные фазы звука А, а элементы 7 и 8 соответствуют звукам И и Н.

На рис. 3.5 приведен пример исходного эталона слова ТОЧКА. Он получен в совершенно аналогичных условиях по $\mathcal{U} = 10$ реализациям диктора-мужчины. Разница лишь в том, что размерность элементов равнялась 20. У этого слова шесть переходных элементов, в том числе переходный сегмент из трех элементов.

Элементы громкостной и тональной транскрипций на рис. 3.4, 3.5 не показаны.

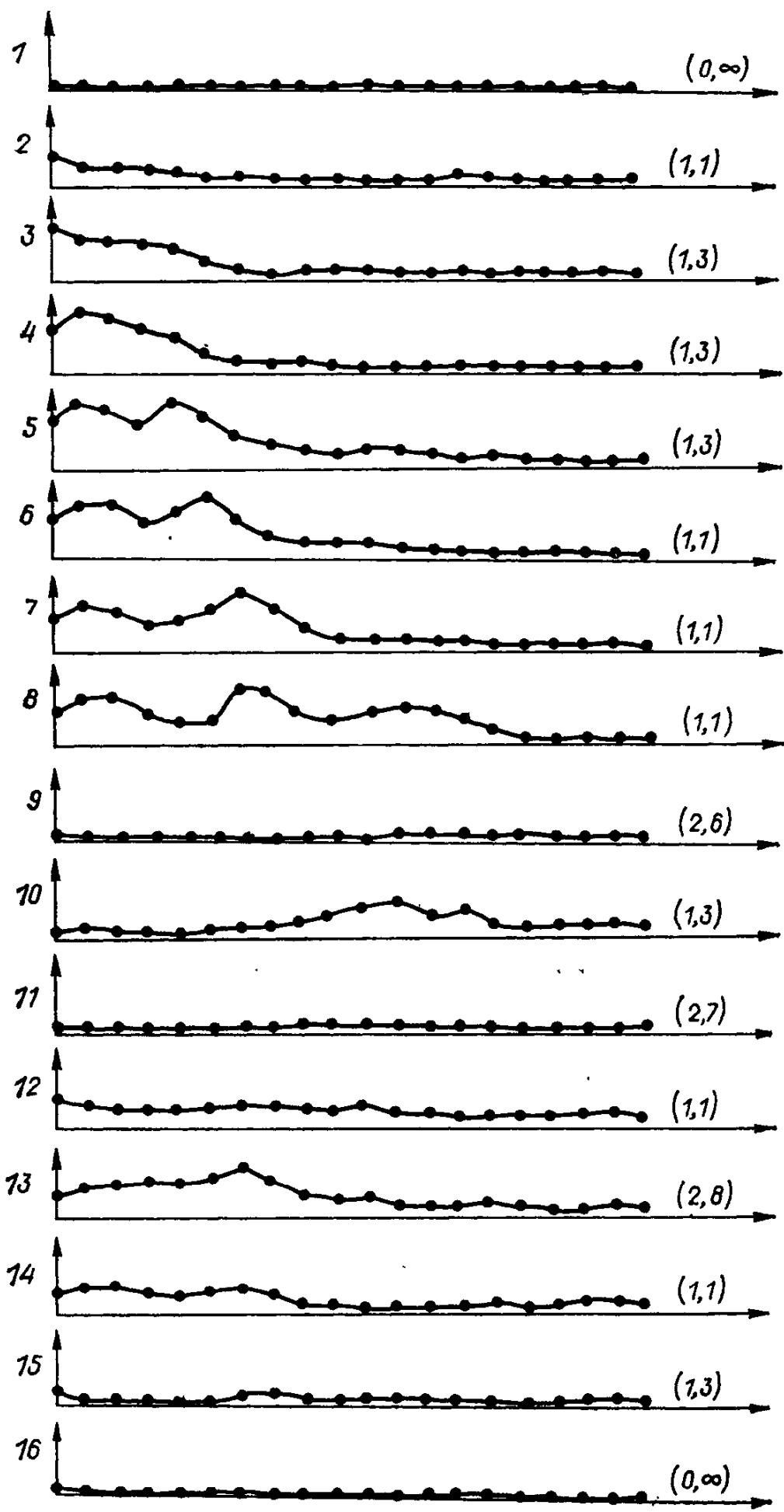


Рис. 3.5. Исходный эталонный сигнал слова ТОЧКА.

§ 3.8. ДРУГИЕ ВАРИАНТЫ ОРГАНИЗАЦИИ ПРОЦЕДУРЫ ОБУЧЕНИЯ

Обучение поэлементному распознаванию слов речи можно организовать несколько иначе, если воспользоваться дополнительной априорной информацией о свойствах речевых сигналов.

Известно, что темп речи в нормальных условиях произнесения изменяется в пределах $\pm 60\%$ от некоторого среднего темпа. В таких же пределах варьируется и длительность произнесения одного и того же слова.

Все это дает основание задаваться стандартным элементом темпоральной транскрипции, например элементом $(m_{ks}, M_{ks}) = (1, 2)$ или $(m_{ks}, M_{ks}) = (2, 3)$, или $(m_{ks}, M_{ks}) = (3, 5)$, и набирать темпоральные транскрипции слов из одного стандартного элемента.

Условимся, что темпоральные транскрипции слов имеют стандартный вид — все элементы, кроме первого и последнего, равны $(t, 2t - 1)$, а первый и последний (элементы пауз в начале и конце слова) равны $(0, \infty)$. При такой структуре темпоральной транскрипции слова минимальная и максимальная собственные (без пауз в начале и конце) длины слова будут соотноситься как $t : (2t - 1)$, например $2 : 3$ или $3 : 5$, или $4 : 7$. В такой же пропорции будут находиться и изменения локального темпа произнесения отдельных звуков слова.

Итак, пусть темпоральная транскрипция слова имеет стандартный вид. Не известна только ее длина. Теперь задача обучения распознаванию формулируется как задача оценивания по ОВ слова только исходного эталона слова, одновременно с указанием его длины.

Далее обучение распознаванию организуем так.

Применив m -самосегментацию (см. § 3.2) реализаций X'_{l_r} , $r = 1 : \mathcal{U}$, ОВ, получим \mathcal{U} результатов обучения по одной реализации с указанием количества эталонных элементов для каждого результата обучения. Используя эти результаты обучения и соответствующие им длины исходного эталона слова в качестве начальных условий для итерационного алгоритма обучения, далее находим \mathcal{U} результатов полного обучения и в качестве окончательного решения задачи обучения выбираем тот полученный при обучении исходный эталонный сигнал слова (вместе с соответствующей длиной), который характеризуется абсолютно наибольшим значением критерия качества обучения.

Таким образом, применение стандартной темпоральной транскрипции позволяет простыми средствами в процессе обучения распознаванию автоматически находить длину исходного эталона и темпоральной транскрипции слова. Применение такой транскрипции с параметром m гарантирует то, что длина исходного эталона слова (без учета пауз на концах слова) будет не менее, чем в m раз короче самой короткой реализации (без учета пауз на концах) слова в ОВ.

Использование стандартной темпоральной транскрипции при распознавании автоматически гарантирует поддержание темпа «произнесения» эталонных сигналов слов в определенных границах. Тем самым, как следствие, учитывается априорная информация о длительности сигналов слов.

Использование стандартной темпоральной транскрипции также регуляризует (делает более однородной) структуру вычислительных средств, реализующих поэлементный метод распознавания.

§ 3.9. ОБЩИЕ ЗАМЕЧАНИЯ

Благодаря обучению осуществляется настройка системы распознавания речи на словарь и голос диктора одновременно. Оказывается, что исходные эталоны и транскрипции слов зависят от индивидуальных особенностей голоса. Эта зависимость проявляется по-разному для различных описаний элементов речи. Однако в целом, только учитывая индивидуальные особенности голоса в процессе обучения распознаванию, удается обеспечить высокую надежность распознавания речи. Благодаря обучению система распознавания становится гибкой — в любой момент времени можно пополнить словарь, заменить отдельные слова, настроиться на голос нового оператора.

Процесс обучения в целом легко распараллеливается. Он сводится к решению однотипных задач, которые возникают на различных этапах обучения в силу сепарабельности критерия качества обучения по варьируемым параметрам. Процедуры решения этих задач подобны используемым в распознавании и во многих случаях от них мало отличаются.

Как и в распознавании, процесс обучения распараллеливается по словам и по отдельным элементам (сегментам) внутри слов. Дополнительно, однако, возможно распараллеливание по отдельным реализациям ОВ. В отличие от распознавания, параллельные процессы обучения выполняются только на отдельных шагах итерационного (последовательного) алгоритма обучения, а весь процесс обучения представляет собой последовательное (пошаговое) согласование параллельных решений, получаемых на предыдущих шагах.

Следует отметить, что если распараллеливание вычислений при распознавании необходимо для обеспечения реального времени распознавания, то оно не обязательно при обучении, которое предшествует распознаванию, выполняется сравнительно редко и может быть выполнено заранее, в том числе и в замедленном масштабе.

Опыт работы показывает, что исходный этalon слова «не стареет». Эталоны трех- и пятигодичной давности обеспечивали примерно ту же надежность распознавания, что и новые эталоны.

Таким образом, при обучении следует ограничиться распараллеливанием только тех вычислений, которые подобны выполняемым при распознавании и все равно реализуются в распознающей системе, как в системе реального времени.

ВЫВОДЫ

1. Разработан метод обучения поэлементному распознаванию речи, позволяющий находить исходный эталон слова (вместе с его темпоральной, громкостной и тональной транскрипциями) по ОВ этого слова.

Обучение заключается в нахождении максимально правдоподобных оценок искомых параметров и параметров преобразований исходного

эталона, при которых достигается наилучшее суммарное сходство преобразованных эталонных сигналов на реализациях ОВ.

Задача обучения точно решается с помощью динамического программирования. Для практического применения рекомендуется итерационный алгоритм обучения, заключающийся в многократном повторении сегментации (распознавания) реализаций на одноименные участки (сегменты) и в поиске наилучшего эталонного элемента для одноименных сегментов. Итерационный алгоритм реализует покоординатную оптимизацию в пространстве двух обобщенных переменных: искомых параметров и параметров сегментации (преобразования исходного эталона).

Составными частями алгоритмов обучения являются оптимальная сегментация и самосегментация реализаций.

Показано, что задачи обучения поэлементному распознаванию являются по существу задачами самообучения.

2. Метод обучения поэлементному распознаванию является составной частью поэлементного метода распознавания слов и слитной речи.

3. Обучение поэлементному распознаванию обеспечивает высокую надежность распознавания.

4. Благодаря обучению системы поэлементного распознавания речи становятся гибкими — в любой момент можно пополнить словарь, заменить отдельные слова, настроиться на голос нового оператора, полностью сменить словарь.

ГЛАВА 4

ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ СЛОВ РЕЧИ

В данной главе покажем, как реализация пофонемного принципа распознавания в поэлементном методе распознавания слов может привести к существенному уменьшению объемов памяти и вычислений. Полученный таким образом метод назван нами простым пофонемным методом распознавания.

Основные положения пофонемного метода распознавания слов и слитной речи были сформулированы в 1971—1973 гг. [89, 99—102]. С 1973 г. в разработке метода и создании системы пофонемного распознавания слов речи принял активное участие А. Г. Шинкажд [103].

Пофонемный метод распознавания слов в современном виде представлен в публикациях [17, 78, 80, 81, 89, 102, 104—108].

§ 4.1. ПЕРЕХОД ОТ ПОЭЛЕМЕНТНОГО К ПОФОНЕМНОМУ РАСПОЗНАВАНИЮ

Особенностью поэлементного метода является то, что начальный эталон слова составляется из эталонных элементов, характерных только данному слову. Совпадение же эталонных элементов слова или разных слов могло быть чисто случайным. Возникает вопрос, почему бы эталонные элементы слов не выбирать из общей для всех слов совокупности эталонных элементов, не зависящей от состава и объема словаря [89]. Принципиально такая возможность вытекает из того, что одни и те же звуки речи встречаются в разных словах. Так, в [105, 106] было показано, что наблюдаемые 48-мерные элементы-коды x_i (см. § 2.1) двухсот слов (по 5 реализаций на слово, всего около 60 000 элементов) аппроксимируются со средней точностью 6 % всего-навсего 80-ю эталонными элементами.

Пусть имеется совокупность E эталонных элементов $e(j) \in E$. Пусть в этой совокупности всего J элементов, например $J = 128$ или $J = 256$. Символом $j = 1 : J$ обозначим имя (порядковый номер) эталонного элемента $e(j) \in E$. Этalonные элементы $e(j)$ будем интерпретировать как представляющие фонемы или, точнее, как части (отдельные фазы) фонем, которые ответственны за элементарные участки речевого сигнала продолжительностью $\Delta T'$, например $\Delta T' = 20$ мс.

Если начальные эталоны слов в поэлементном методе распознавания составлять из общей для всех слов совокупности эталонных элементов, то поэлементный метод распознавания переходит в пофонемный.

Как будет показано в дальнейшем, такое нововведение в поэлементный метод, не изменяя его сущности, приводит к значительному уменьшению объемов памяти и вычислений. Правда, в пофонемном методе гораздо более сложной становится задача обучения.

Использование общей для всех слов совокупности эталонных элементов при составлении исходных эталонов слов — это только первый шаг в направлении пофонемного распознавания. Более глубокая реализация принципов пофонемного распознавания будет дана в гл. 6.

§ 4.2. РОЛЬ ТРАНСКРИПЦИИ СЛОВА

Как было отмечено, в пофонемном методе распознавания слов должна быть задана общая для всех слов совокупность E эталонных элементов $e(j) \in E$, где $j = 1 : J$ — имя (порядковый номер) элемента.

Каждое слово $k = 1 : K$ должно быть задано четверкой транскрипций: темпоральной τ_k , громкостной H_k , тональной F_k и акустической транскрипцией

$$R_k = (j_{k1}, j_{k2}, \dots, j_{ks}, \dots, j_{kq_k}), \quad (4.2.1)$$

указывающей последовательность имен элементов $e(j) \in E$, из которых может быть составлен исходный эталонный сигнал k -го слова

$$E_k = R_k E = (e(j_{k1}), e(j_{k2}), \dots, e(j_{ks}), \dots, e(j_{kq_k})). \quad (4.2.2)$$

Таким образом, при пофонемном распознавании исходный эталон слова набирается из E по акустической транскрипции слова R_k .

С помощью R_k исходный эталон слова задается гораздо более экономными средствами, чем в поэлементном методе.

Для удобства в дальнейшем будем называть Q_k -транскрипцией слова k тройку транскрипций (R_k, H_k, F_k). Транскрипция Q_k имеет длину q_k , а ее s -й элемент z_{ks} составляет тройка элементов: $z_{ks} = (j_{ks}, h_{ks}, f_{ks})$. Запись $E_k = Q_k E$ (иногда $R_k E$) будет означать, что в исходном эталоне слова длины q_k эталонный элемент $e(j_{ks})$ характеризуется громкостью h_{ks} и тональностью f_{ks} .

Пофонемное распознавание слов осуществляется по тем же рекуррентным формулам (2.3.7) — (2.3.15), что и в поэлементном методе, однако в них вместо E_k следует положить $R_k E$ или $Q_k E$, вместо $g(x_i, (vE_k)_i)$ — $g(x_i, (vR_k E)_i)$ или $g(x_i, (vQ_k E)_i)$, а вместо $g(x_i, e_{ks})$ — $g(x_i, e(j_{ks}))$ или $g(x_i, e(j_{ks})) + g_1(h_i, h_{ks}) + g_2(f_i, f_{ks})$. Сохраняются, таким образом, все свойства поэлементного метода, в том числе и возможные способы распаралеливания вычислений.

В целом, в пофонемном методе для каждого слова k , исходя из общей для всех слов совокупности эталонных элементов и Q_k -транскрипции, составляется исходный эталон, который далее подвергается преобразованиям согласно темпоральной транскрипции τ_k ; преобразованные эталонные сигналы слов сравниваются с распознаваемым сигналом X_i ; распознаваемый сигнал X_i относится к тому классу (слову), пре-

образованный эталонный сигнал которого оказался наиболее похожим на распознаваемый.

Таким образом, в пофонемном методе априорная информация используется по иерархическому принципу: сначала выбираем транскрипции слова, затем строим исходный эталон слова, после чего запускается процесс порождения эталонных сигналов слова.

§ 4.3. ПРЕИМУЩЕСТВА ПОФОНЕМНОГО МЕТОДА

В пофонемном методе распознавания, по сравнению с поэлементным, достигается значительная экономия памяти и вычислений.

Так, на хранение исходных эталонов слов в поэлементном методе нужна память на $N_s = 2n \sum_{k=1}^K q_k$ байт, если n — размерность элементов речи и на одну компоненту элемента отводить 2 байта. В то же время в пофонемном методе эта память существенно меньше:

$$N_\Phi = N_{\Phi_1} + N_{\Phi_2} = \left(2nJ + \frac{\log_2 J}{8} \sum_{k=1}^K q_k \right) \text{байт},$$

причем на хранение одного элемента акустической транскрипции слова требуется всего $\log_2 J$ бит или 1 байт памяти. При $K = 500$, $n = 20$, $J = 256$ и среднем значении q_k , равном 12, получаем выигрыш по памяти в

$$\eta_p = \frac{N_s}{N_\Phi} = \frac{2 \cdot 20 \cdot 500 \cdot 12}{2 \cdot 20 \cdot 256 + 500 \cdot 12} \approx 18 \text{ раз.}$$

Очевидно, что затраты памяти на хранение темпоральной, громкостной и тональной транскрипций в обоих методах одинаковы.

В пофонемном методе, как и в поэлементном, затраты памяти хотя и растут линейно с увеличением объема словаря, однако с существенно меньшим (в $2n$ раз) коэффициентом пропорциональности.

Экономия в вычислениях достигается за счет значительного уменьшения количества вычисляемых мер сходства $g(\mathbf{x}_i, \mathbf{e}_{ks})$: $\sum_{k=1}^K q_k$ мер на один элемент \mathbf{x}_i в поэлементном методе и J мер в пофонемном. Значит, экономия в вычислениях более ощутима: при $K = 500$ и $J = 128$ получаем выигрыш в

$$\eta_v = \sum_{k=1}^K q_k / J = \frac{500 \cdot 12}{128} \approx 46 \text{ раз.}$$

Пофонемный метод имеет и другие преимущества: большая наглядность, удобство в исследовании, возможность фонетической интерпретации эталонных элементов $e(j)$.

Есть еще одно достоинство, которое лучше проявляется в глубоком пофонемном распознавании (см. гл. 6). Оно заключается в повышении надежности распознавания.

§ 4.4. ПОСТАНОВКА И АНАЛИЗ ЗАДАЧИ ОБУЧЕНИЯ И САМООБУЧЕНИЯ ПОФОНЕМНОМУ РАСПОЗНАВАНИЮ

При пофонемном методе усложняется задача обучения, поскольку, кроме транскрипций слова, по ОВ требуется найти общую для всех слов совокупность эталонных элементов E . При этом, чтобы обеспечить именно пофонемное распознавание, задают ограничения на количество эталонных элементов $J \leq J_0$, $J_0 = 128$ или 256 .

Пусть задана ОВ

$$(\mathbf{X}'_{l_r}, k(r)), \quad r = 1 : \mathcal{U}, \quad (4.4.1)$$

где $\mathbf{X}'_{l_r} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_i, \dots, \mathbf{x}'_{l_r})$ — реализация с номером r в ОВ; l_r — длина r -й реализации; $k(r)$ — функция, принимающая значение на множестве целых чисел $k = 1 : K$ (K — количество слов в словаре) и указывающая, какому слову принадлежит реализация с номером r ; \mathcal{U} — количество реализаций в ОВ.

При обучении пофонемному распознаванию слов требуется на основании ОВ найти совокупность E , содержащую $J \leq J_0$ эталонных элементов и совокупность (\mathbf{Q}_k, τ_k) , $k = 1 : K$, транскрипций всех слов, которые вместе с совокупностью \mathbf{v}' , $r = 1 : \mathcal{U}$, операторов преобразования $\mathbf{v}' \in \tau_{k(r)}$ (l_r) исходных эталонов $\mathbf{Q}_k E$ доставляют максимум критерию качества обучения:

$$\Phi(E, \{(\mathbf{Q}_k, \tau_k)\}, \{\mathbf{v}'\}) = \sum_{r=1}^{\mathcal{U}} G(\mathbf{X}'_{l_r}, \mathbf{v}' \mathbf{Q}_{k(r)} E). \quad (4.4.2)$$

Постановка задачи обучения (4.4.2) (эта задача в равной мере может быть названа и задачей самообучения) предполагает нахождение максимально правдоподобных оценок как искомых параметров: совокупности E и совокупности $(\mathbf{Q}_k, \tau_k) = (\mathbf{R}_k, \mathbf{H}_k, \mathbf{F}_k, \tau_k)$, $k = 1 : K$, из акустических, громкостных, тональных и темпоральных транскрипций слов, так и мешающих (в рассматриваемом случае — контрольных) параметров \mathbf{v}' , $r = 1 : \mathcal{U}$.

Постановка задачи обучения пофонемному распознаванию слов предполагает также оценивание и длин q_k транскрипций всех слов.

В отличие от обучения поэлементному распознаванию слов в пофонемном методе задача обучения не распадается на K подзадач (по числу слов), однако, как и раньше, существенно целочисленная задача обучения решается с помощью обобщенной покоординатной оптимизации с использованием свойств сепарабельности критерия качества обучения.

Критерий качества обучения обладает тем свойством, что достигает максимума при $J = J_0$ и при темпоральных транскрипциях $\tau_k = \tau_k : m_{k1} = m_{kq_k} = 1$, $m_{ks} = m$ ($m = 1$) для $s = 2 : (q_k - 1)$ и $M_{ks} = \infty$ (или $M_{ks} = \min l_r$) для $s = 1 : q_k$.

Это свойство позволяет решать задачу обучения в два этапа.

На первом этапе, полагая $J = J_0$ и $\tau_k = \tau_k$, находят оптимальную

совокупность E эталонных элементов $e(j) \in E$ и оптимальные транскрипции \mathbf{Q}_k всех слов $k = 1 : K$.

На втором этапе из совокупности E исключаются одинаковые элементы. Тем самым находится оптимальное количество элементов $J^* \leq J_0$. Далее, с учетом сокращения количества элементов в E осуществляется перекодировка транскрипций \mathbf{Q}_k , полученных на первом этапе. Завершается второй этап формированием темпоральных транскрипций τ_k .

§ 4.5. ИТЕРАЦИОННЫЙ АЛГОРИТМ ОБУЧЕНИЯ

Рассмотрим сначала случай заданных длин транскрипций q_k . Как отмечалось в § 3.5, длины q_k могут задаваться вручную учителем в процессе накопления ОВ либо выбираться автоматически по тексту слова, который также вводится в систему распознавания речи при формировании словаря.

Случай заданных длин транскрипций q_k (вручную или автоматически по тексту слова) является основным, чаще всего используемым в обучении пофонемному распознаванию речи. Обобщение итерационного алгоритма обучения на случай незаданных длин транскрипций будет приведено в следующем параграфе.

На первом этапе обучения решается задача оценивания совокупности E из J_0 эталонных элементов и совокупности \mathbf{Q}_k -транскрипций слов длины q_k . Темпоральные транскрипции полагаются равными τ_k .

Итерационный алгоритм обучения осуществляет покоординатную оптимизацию критерия качества обучения Φ (см. (4.4.2)) в пространстве трех обобщенных переменных E , $\{\mathbf{Q}_k\}$, $\{\mathbf{v}^r\}$.

Пусть в результате n -й итерации обучения найдены: а) E^n — совокупность из эталонных элементов $e^n(j) \in E^n$, $j = 1 : J_0$; б) \mathbf{Q}_k^n , $k = 1 : K$ — совокупность \mathbf{Q}_k -транскрипций всех слов словаря. Тогда на $(n + 1)$ -й итерации последовательно выполняются три шага.

1-й шаг. При фиксированных E^n и $\{\mathbf{Q}_k^n\}$ решаем задачу нахождения совокупности

$$\{\mathbf{v}^{(n+1)r}\} = \underset{\{\mathbf{v}^r\}}{\operatorname{argmax}} \Phi(E^n, \{\mathbf{Q}_k^n\}, \{\mathbf{v}^r\}). \quad (4.5.1)$$

Из сепарабельности критерия (4.4.2) по переменным \mathbf{v}^r следует, что задача (4.5.1) распадается на \mathcal{U} независимых задач нахождения оптимальной сегментации реализаций:

$$\mathbf{v}^{(n+1)r} = \underset{\mathbf{v} \in \tau_{k(r)}(l_r)}{\operatorname{argmax}} G(\mathbf{X}'_{l_r}, \mathbf{v} \mathbf{Q}_{k(r)}^n E^n), \quad r = 1 : \mathcal{U}. \quad (4.5.2)$$

Для их решения следует воспользоваться алгоритмом сегментации, изложенным в § 3.2.

Напомним только, что, пользуясь формулами (3.2.5) — (3.2.11), в данном случае под величиной $g(x_i, e_{ks})$ следует подразумевать величину

$$g(x_i, z_{ks}) = g(x_i, e(j_{ks})) + g_1(h_i, h_{ks}) + g_2(f_i, f_{ks}). \quad (4.5.3)$$

Решив \mathcal{U} задач сегментации (4.5.2), получим оптимальные границы $w_s^{(n+1)r}$, $r = 1 : \mathcal{U}$, $s = 1 : q_{k(r)}$, сегментов $(w_{s-1}^{(n+1)r}, w_s^{(n+1)r}]$ всех реализаций, определяющих посредством $v_s^{(n+1)r} = w_s^{(n+1)r} - w_{s-1}^{(n+1)r}$ компоненты оптимальных операторов сегментации $\mathbf{v}^{(n+1)r}$.

Полученные результаты позволяют приступить к выполнению 2-го шага алгоритма обучения.

2-й шаг. При фиксированных E^n и $\{\mathbf{v}^{(n+1)r}\}$ решаем задачу нахождения совокупности \mathbf{Q}_k -транскрипций слов:

$$\{\mathbf{Q}_k^{n+1}\} = \operatorname{argmax}_{\{\mathbf{Q}_k\}} \Phi(E^n, \{\mathbf{Q}_k\}, \{\mathbf{v}^{(n+1)r}\}). \quad (4.5.4)$$

Критерий обучения Φ сформулирован по \mathbf{Q}_k -транскрипциям (по отдельным словам), так как представляет собой сумму из K слагаемых $B_k^{n+1}(\mathbf{Q}_k)$:

$$\Phi(E^n, \{\mathbf{Q}_k\}, \{\mathbf{v}^{(n+1)r}\}) = \sum_{k=1}^K B_k^{n+1}(\mathbf{Q}_k), \quad (4.5.5)$$

где

$$B_k^{n+1}(\mathbf{Q}_k) = \sum_{r: k(r)=k} G(\mathbf{X}_l^r, \mathbf{v}^{(n+1)r} \mathbf{Q}_{k(r)} E^n). \quad (4.5.6)$$

Поэтому задача (4.5.4) распадается на K независимых задач нахождения \mathbf{Q}_k -транскрипций:

$$\mathbf{Q}_k^{n+1} = \operatorname{argmax}_{\mathbf{Q}_k} B_k^{n+1}(\mathbf{Q}_k), \quad k = 1 : K. \quad (4.5.7)$$

Для решения задач (4.5.7) преобразуем функцию $B_k^{n+1}(\mathbf{Q}_k)$:

$$\begin{aligned} B_k^{n+1}(\mathbf{Q}_k) &= \sum_{r: k(r)=k} \sum_{s=1}^{q_{k(r)}} \sum_{i=w_{s-1}^{(n+1)r}+1}^{w_s^{(n+1)r}} g(\mathbf{x}_i, \mathbf{z}_{ks}) = \\ &= \sum_{s=1}^{q_k} \sum_{r: k(r)=k} \sum_{i=w_{s-1}^{(n+1)r}+1}^{w_s^{(n+1)r}} (g(\mathbf{x}_i, \mathbf{e}^n(j_{ks})) + g_1(h_i^r, h_{ks}) + \\ &\quad + g_2(f_i^r, f_{ks})) = \sum_{s=1}^{q_k} (D_{ks}^{n+1}(j_{ks}) + D_{1ks}^{n+1}(h_{ks}) + D_{2ks}^{n+1}(f_{ks})), \end{aligned} \quad (4.5.8)$$

где

$$D_{ks}^{n+1}(j_{ks}) = \sum_{r: k(r)=k} \sum_{i=w_{s-1}^{(n+1)r}+1}^{w_s^{(n+1)r}} g(\mathbf{x}_i, \mathbf{e}^n(j_{ks})), \quad (4.5.9)$$

$$D_{1ks}^{n+1}(h_{ks}) = \sum_{r: k(r)=k} \sum_{i=w_{s-1}^{(n+1)r}+1}^{w_s^{(n+1)r}} g_1(h_i^r, h_{ks}), \quad (4.5.10)$$

$$D_{2ks}^{n+1}(f_{ks}) = \sum_{r: k(r)=k} \sum_{i=w_{s-1}^{(n+1)r}+1}^{w_s^{(n+1)r}} g_2(f_i^r, f_{ks}). \quad (4.5.11)$$

Из анализа (4.5.8) — (4.5.11) следует, что каждое из слагаемых в (4.5.8) зависит только от всех s -х сегментов всех реализаций k -го слова.

Из (4.5.8) также следует, что функция $B_k^{n+1}(\mathbf{Q}_k)$ сепарабельна по переменным j_{ks} , h_{ks} и f_{ks} транскрипций $\mathbf{Q}_k = (\mathbf{R}_k, \mathbf{H}_k, \mathbf{F}_k)$. Поэтому каждая из задач (4.5.7), в свою очередь, распадается на $3q_k$ независимых задач нахождения компонент j_{ks}^{n+1} , h_{ks}^{n+1} и f_{ks}^{n+1} транскрипций \mathbf{Q}_k^{n+1} .

Введем понятие условного аргмаксимума функции $a(i)$ по значению j аргумента i :

$$\operatorname{argmax}_i a(i) | j = \begin{cases} j, & \text{если } j \in I; \\ \operatorname{argmax}_i a(i), & \text{если } j \notin I, \end{cases} \quad (4.5.12)$$

где

$$I = \{i : a(i) = \max_v a(v)\}. \quad (4.5.13)$$

Воспользовавшись понятием условного аргмаксимума, новые элементы j_{ks}^{n+1} , h_{ks}^{n+1} и f_{ks}^{n+1} акустической, громкостной и тональной транскрипций будем находить по формулам

$$j_{ks}^{n+1} = \operatorname{argmax}_{i=1:J_0} D_{ks}^{n+1}(j) | j_{ks}^n, \quad (4.5.14)$$

$$h_{ks}^{n+1} = \operatorname{argmax}_{h \in \Omega} D_{1ks}^{n+1}(h) | h_{ks}^n, \quad (4.5.15)$$

$$f_{ks}^{n+1} = \operatorname{argmax}_{f \in \Omega} D_{2ks}^{n+1}(f) | f_{ks}^n, \quad (4.5.16)$$

где Ω_1 и Ω_2 — множества дискретных значений громкости h и тональности f соответственно.

По формулам (4.5.14) — (4.5.16) производится консервативный выбор новых элементов транскрипций, гарантирующий строгое возрастание функций $D_{ks}^{n+1}(j)$, $D_{1ks}^{n+1}(h)$, $D_{2ks}^{n+1}(f)$, а значит, и всего критерия обучения Φ , если только элементы j_{ks}^n , h_{ks}^n и f_{ks}^n заменяются неравными им элементами j_{ks}^{n+1} , h_{ks}^{n+1} и f_{ks}^{n+1} соответственно.

Решив задачи (4.5.14) — (4.5.16) для всех $s = 1 : q_k$, получим решение задачи (4.5.7), а решив задачу (4.5.7) для всех слов $k = 1 : K$, получим решение задачи (4.5.4).

Таким образом, на втором шаге алгоритма обучения по фиксированным сегментациям реализаций ОВ и по фиксированной совокупности эталонных элементов находятся новые \mathbf{Q} -транскрипции слов.

3-й шаг. При фиксированных \mathbf{Q} -транскрипциях $\{\mathbf{Q}_k^{n+1}\}$ и сегментациях $\{\mathbf{v}^{(n+1)r}\}$ решаем задачу нахождения новой совокупности E^{n+1} эталонных элементов $e^{n+1}(j)$, $j = 1 : J_0$:

$$E^{n+1} = \operatorname{argmax}_E \Phi(E, \{\mathbf{Q}_k^{n+1}\}, \{\mathbf{v}^{(n+1)r}\}). \quad (4.5.17)$$

Для этого случая критерий обучения Φ запишем в виде

$$\Phi(E, \{\mathbf{Q}_k^{n+1}\}, \{\mathbf{v}^{(n+1)r}\}) = \sum_{r=1}^R \sum_{s=1}^{q_{k(r)}} \sum_{i=w_{s-1}^{(n+1)r} + 1}^{w_s^{(n+1)r}} g(\mathbf{x}'_i, \mathbf{e}(j_{k(r)s}^{n+1})) + c, \quad (4.5.18)$$

где c — добавка, не зависящая от E . Сгруппируем все слагаемые выражения (4.5.18) с одинаковым именем $j_{k(r)s}^{n+1} = j$ эталонного элемента $\mathbf{e}(j_{k(r)s}^{n+1})$. Содержательно это означает сборку всех таких сегментов всех реализаций всех слов ОВ, которые аппроксимируются одним и тем же эталонным элементом $j_{k(r)s}^{n+1} = j$. Такая группировка (сборка) позволит записать критерий обучения Φ в виде

$$\Phi(E, \{\mathbf{Q}_k^{n+1}\}, \{\mathbf{v}^{(n+1)r}\}) = \sum_{j=1}^{J_0} C_j^{n+1}(\mathbf{e}(j)) + c, \quad (4.5.19)$$

где

$$C_j^{n+1}(\mathbf{e}(j)) = \sum_{(r,s) \in I^{n+1}(j)} \sum_{i=w_{s-1}^{(n+1)r} + 1}^{w_s^{(n+1)r}} g(\mathbf{x}'_i, \mathbf{e}(j_{k(r)s}^{n+1})), \quad (4.5.20)$$

$$I^{n+1}(j) = \{(r, s) : j_{k(r)s}^{n+1} = j\}. \quad (4.5.21)$$

Из сепарабельности критерия (4.5.19) по переменным $\mathbf{e}(j)$ следует, что задача (4.5.17) распадается на J_0 независимых задач нахождения новых элементов $\mathbf{e}^{n+1}(j)$, которые будем вычислять по формуле консервативного выбора

$$\mathbf{e}^{n+1}(j) = \begin{cases} \underset{\mathbf{e}}{\operatorname{argmax}} C_j^{n+1}(\mathbf{e}) | \mathbf{e}^n(j), & \text{если } I^{n+1}(j) \neq \emptyset; \\ \mathbf{e}^n(j), & \text{если } I^{n+1}(j) = \emptyset, \\ j = 1 : J_0. & \end{cases} \quad (4.5.22)$$

По формуле (4.5.22) гарантируется строгое возрастание критерия обучения Φ , если элемент $\mathbf{e}^n(j)$ заменяется неравным ему элементом $\mathbf{e}^{n+1}(j)$. Определив по формуле (4.5.22) элементы $\mathbf{e}^{n+1}(j)$ для всех $j = 1 : J_0$, получим совокупность E^{n+1} , содержащую J_0 элементов и являющуюся решением задачи (4.5.17).

На этом заканчивается выполнение $(n + 1)$ -й итерации алгоритма обучения.

Если условие останова алгоритма

$$\begin{cases} E^{n+1} = E^n, \\ \{\mathbf{Q}_k^{n+1}\} = \{\mathbf{Q}_k^n\} \end{cases} \quad (4.5.23)$$

не выполнилось, приступаем к выполнению следующей итерации.

Если условие останова выполнилось, совершают заключительную итерацию алгоритма. Она составляет содержание второго этапа обучения.

Сначала отсеиваем лишние элементы из совокупности E^{n+1} . Таковыми являются элементы с номерами j , для которых $I^{n+1}(j) = \emptyset$,

поскольку в этом случае элемент $e^{n+1}(j)$ не присутствует ни в одном исходном сигнале $R_k^{n+1}E^{n+1}$. Далее, в совокупности E^{n+1} могут оказаться одинаковые элементы. Оставим один из них, считая остальные лишними. Исключив из E^{n+1} лишние элементы и упорядочив оставшиеся, получим искомую совокупность E эталонных элементов $e(j) \in E$, содержащую $J \leq J_0$ элементов. С учетом изменившейся нумерации элементов в совокупности E производим перекодировку транскрипций R_k^{n+1} . В результате получим искомую совокупность $\{R_k\}$ транскрипций слов. Что касается громкостной H_k и тональной F_k транскрипций, то они остаются без изменений: $H_k = H_k^{n+1}$, $F_k = F_k^{n+1}$.

Далее формируются темпоральные транскрипции слов τ_k . Искомые значения элементов (m_{ks}, M_{ks}) темпоральных транскрипций слов выбираются из условия таких максимально возможных m_{ks} и минимально возможных M_{ks} , чтобы критерий качества обучения еще сохранял свое максимально возможное значение, достигаемое при $\tau_k = \tilde{\tau}_k$. Такой способ выбора τ_k обусловлен стремлением уменьшить «пересекаемость» классов (слов).

Таким образом, элементы (m_{ks}, M_{ks}) транскрипций τ_k находим по формулам

$$\begin{aligned} m_{ks} &= \min_{r: k(r)=k} (w_s^{(n+1)r} - w_{s-1}^{(n+1)r}), \\ M_{ks} &= \max_{r: k(r)=k} (w_s^{(n+1)r} - w_{s-1}^{(n+1)r}), \\ k &= 1 : K, \quad s = 2 : (q_k - 1). \end{aligned} \quad (4.5.24)$$

Что касается первого и последнего элементов транскрипций всех слов, то для них полагаем $m_{k1} = m_{kq_k} = 0$, $M_{k1} = M_{kq_k} = \infty$, а сам эталонный элемент паузы (ему присвоим имя $j = 1$) получаем путем «усреднения» элементов всех первых и последних сегментов всех реализаций ОВ:

$$e(1) = \underset{e}{\operatorname{argmax}} \sum_{r=1}^{\omega_1^{(n+1)r}} \left(\sum_{i=1}^{l_r} g(x_i^r, e) + \sum_{i=w_{q_k(r)-1}^{(n+1)r}+1}^{l_r} g(x_i^r, e) \right), \quad (4.5.25)$$

присвоив ему громкость $h(1)$, $h(1) = h_{k1} = h_{kq_k}$:

$$h(1) = \underset{h \in \Omega_1}{\operatorname{argmax}} \sum_{r=1}^{\omega_1^{(n+1)r}} \left(\sum_{i=1}^{l_r} g_1(h_i^r, h) + \sum_{i=w_{q_k(r)-1}^{(n+1)r}+1}^{l_r} g_1(h_i^r, h) \right) \quad (4.5.26)$$

и тональность $f(1)$, $f(1) = f_{k1} = f_{kq_k}$:

$$f(1) = \underset{f \in \Omega_2}{\operatorname{argmax}} \sum_{r=1}^{\omega_1^{(n+1)r}} \left(\sum_{i=1}^{l_r} g_2(f_i^r, f) + \sum_{i=w_{q_k(r)-1}^{(n+1)r}+1}^{l_r} g_2(f_i^r, f) \right). \quad (4.5.27)$$

На этом алгоритм обучения пофонемному распознаванию слов заканчивает свою работу.

Для запуска итерационного алгоритма необходимо задать начальные условия. Таковыми может служить любая из пар совокупностей:

- 1) $E^0, \{\mathbf{Q}_k^0\}$; 2) $\{\mathbf{Q}_k^0\}, \{\mathbf{v}^{0r}\}$; 3) $E^0, \{\mathbf{v}^{0r}\}$.

Один из рекомендуемых и используемых вариантов заключается в том, что сначала разбивают всю ОВ на подвыборки отдельных слов, реализации r которых удовлетворяют условию $k(r) = k$. Далее решают K задач обучения поэлементному распознаванию слов на основании подвыборок отдельных слов, считая длины q_k исходного эталона и транскрипций слов заданными (см. гл. 3). В результате решения этих задач будут найдены одноименные сегменты реализаций ОВ одного и того же слова k , интерпретируемые как один таксон, как проявление одного и того же эталонного элемента. Обозначим сборку элементов \mathbf{x}'_i одноименных сегментов как множество пар индексов (r, i) этих элементов:

$$T(k, s) = \{(r, i) : k(r) = k, i \in (w_{s-1}', w_s']\}. \quad (4.5.28)$$

Для одного слова k будет q_k сборок $T(k, s), s = 1 : q_k$, причем $w'_0 = 0$ и $w'_{q_k(r)} = l_r$. Всего же получим $I = \sum_{k=1}^K q_k$ различных сборок $T(k, s), k = 1 : K, s = 1 : q_k$; соответственно столько же различных таксонов.

Будем соединять (собирать) все образовавшиеся сборки так, чтобы получилось $J_0 \ll I$ новых сборок, каждую из которых будем также интерпретировать как один таксон, как проявление какого-то одного эталонного элемента $\mathbf{e}(j) \in E, j = 1 : J_0$. Найдем оптимальную сборку сборок $T(k, s)$, чтобы получилось наилучшее суммарное сходство таксонов (эталонных элементов $\mathbf{e}(j) \in E$) на все сборки $T(k, s)$:

$$E^0 = \operatorname{argmax}_E \sum_{k=1}^K \sum_{s=1}^{q_k} \max_{\mathbf{e}(j) \in E} \sum_{(r, i) \in T(k, s)} g(\mathbf{x}'_i, \mathbf{e}(j)). \quad (4.5.29)$$

Задача (4.5.29) является одной из задач самообучения (таксономии, развода на кучи), широко распространенных в распознавании образов [91, 109—113]. Решаются эти задачи, чаще всего, итерационными алгоритмами. На первом шаге, полагая E^0 заданной, находят все сборки $T(k, s)$, аппроксимируемые одним и тем же эталонным элементом $\mathbf{e}^0(j) \in E^0$, а затем, на втором шаге, вычисляют новые элементы $\mathbf{e}^0(j)$ на основании тех сборок $T(k, s)$, которые аппроксимировались старым элементом $\mathbf{e}^0(j)$.

Другой рекомендуемый и апробированный способ решения задачи (4.5.29) заключается в последовательном соединении (сборке) наиболее похожих сборок.

Величиной, характеризующей сходство сборок T_1 и T_2 , естественно назвать

$$G(T_1, T_2) = \max_{\mathbf{e}} \sum_{(r, i) \in T_1 \cup T_2} g(\mathbf{x}'_i, \mathbf{e}). \quad (4.5.30)$$

Из всех I первоначальных сборок $T(k, s)$ найдем самую похожую пару сборок и заменим ее новой объединенной сборкой. Получим $I - 1$ сборок. Среди них опять найдем самую похожую пару сборок. Заменив ее одной объединенной сборкой, получим $I - 2$ сборки. Действуя аналогично $I - J_0$ раз, получим J_0 сборок T_j , для каждой из которых затем находим наиболее похожий эталонный элемент

$$\mathbf{e}^0(j) = \operatorname{argmax}_{\mathbf{e}} \sum_{(r, i) \in T_j} g(\mathbf{x}'_i, \mathbf{e}). \quad (4.5.31)$$

Чтобы ускорить процесс объединения сборок, необходимо при очередном полном просмотре образовавшегося массива сборок выбирать не одну, а $N > 1$ наиболее похожих пар сборок. Выделенные таким способом не более $2N$ сборок заменяются, например, N сборками путем объединения наиболее близких сборок.

Этот прием ускорения вычислений позволяет за один просмотр массива сборок уменьшить их количество на N . Суммарное количество просмотров будет не $I - J_0$, а существенно меньше, например в N раз.

Чтобы получить ровно J_0 эталонных элементов $\mathbf{e}^0(j) \in E^0$, необходимо по мере уменьшения количества объединенных сборок уменьшать число N , например полагать $N = 1$, если количество объединенных сборок стало меньше $2J_0$.

Другой вариант алгоритма нахождения E^0 , подобный алгоритму объединения сборок, описан в [114].

Заметим, что точное решение задачи (4.5.29) может быть получено алгоритмом ДП, который оказывается весьма громоздким. Предлагаемые же алгоритмы нахождения E^0 реализуются сравнительно просто. «Локальность» и «эвристичность» получаемых решений оправдываются тем, что находится всего лишь начальное значение E^0 совокупности эталонных элементов, которое затем будет «улучшено» итерационным алгоритмом обучения. Нетрудно видеть, однако, что способы получения начального значения E^0 учитывают свойства и структуру речевых сигналов, а это позволяет надеяться на получение хорошего начального значения E^0 .

Далее, полагая, что E^0 вычислена, находят значение второго начального условия $\{\mathbf{Q}_k^0\}$.

Для этого возвращаемся к сборкам $T(k, s)$ элементов \mathbf{x}'_i одноименных сегментов реализаций отдельных слов k . Напомним, что эти сборки определены формулой (4.5.28) по результатам обучения поэлементному распознаванию слов.

Начальные значения элементов $\mathbf{z}_{ks}^0 = (j_{ks}^0, h_{ks}^0, f_{ks}^0)$ транскрипций \mathbf{Q}_k^0 находим по $T(k, s)$, $k = 1 : K$, $s = 1 : q_k$:

$$j_{ks}^0 = \operatorname{argmax}_i \sum_{(r, i) \in T(k, s)} g(\mathbf{x}'_i, \mathbf{e}^0(j)), \quad (4.5.32)$$

$$h_{ks}^0 = \operatorname{argmax}_{h \in \Omega_i} \sum_{(r, i) \in T(k, s)} g_1(h'_i, h), \quad (4.5.33)$$

$$f_{ks}^0 = \operatorname{argmax}_{f \in \Omega_s} \sum_{(r, i) \in T(k, s)} g_2(f'_i, f). \quad (4.5.34)$$

На этом заканчивается формирование начальных условий E^0 , $\{\mathbf{Q}_k^0\}$ для итерационного алгоритма обучения пофонемному распознаванию слов речи.

Нетрудно убедиться, что если условие останова итерационного алгоритма (4.5.23) не выполняется, то от итерации к итерации гарантируется монотонное возрастание критерия качества обучения Φ . Это обеспечивается правилами консервативного выбора новых параметров на каждой итерации. Кроме того, конечность множества возможных сегментаций $\{\mathbf{v}'\}$ и конечность множества возможных транскрипций $\{\mathbf{Q}_k\}$, которые в принципе могут быть «просмотрены» алгоритмом, позволяют сделать заключение о сходимости алгоритма за конечное число итераций. Так как критерий качества обучения ограничен ($\Phi \leq 0$) и монотонно растет от итерации к итерации, то можно рассчитывать на сходимость за небольшое число итераций. Практически для $K = 200$ слов и $\mathcal{U} = 1000$ (по пять реализаций на слово) количество итераций не превышало 12. При этом значительный рост критерия качества обучения наблюдался лишь на первых двух итерациях.

Итерационный алгоритм обучения не гарантирует нахождение глобального максимума критерия качества обучения. Относительно полученного решения можно лишь утверждать, что изменение одной из трех обобщенных переменных, при фиксированных других, не приводит к росту критерия.

Практически же оказалось, что алгоритм обучения пофонемному распознаванию обеспечивает высокую надежность распознавания слов. Однако достигается эта надежность при существенно меньших затратах памяти и вычислений, чем в поэлементном методе.

§ 4.6. СЛУЧАЙ НЕЗАДАННЫХ ДЛИН ТРАНСКРИПЦИЙ

Основной алгоритм обучения пофонемному распознаванию слов предполагает заданными длины q_k транскрипций. Они задаются вручную учителем либо вычисляются автоматически по тексту слова, т. е. находятся до обучения.

Представляют интерес, однако, и такие постановки и решения задач, которые предполагают автоматическое нахождение длин q_k транскрипций в процессе обучения. При этом дополнительно задают ограничения на минимально q'_k и максимально q''_k возможные длины транскрипций всех слов:

$$1 < q'_k \leq q_k \leq q''_k < \min_{r: k(r)=k} l_r, \quad k = 1 : K. \quad (4.6.1)$$

Постановка задачи обучения для случая незаданных длин транскрипций предполагает, таким образом, существование оптимальной длины транскрипции, меньшей, чем минимально возможная длина слова. Подробнее об этом было сказано в § 3.5.

Для решения задачи обучения пофонемному распознаванию в новой постановке может быть предложен итерационный алгоритм, яв-

ляющийся дальнейшим обобщением итерационного алгоритма обучения в случае заданных длин транскрипций.

Пусть $\{q_k^n\}$ — длины транскрипций слов, полученные на n -й итерации.

Тогда на $(n+1)$ -й итерации, как и в основном алгоритме, последовательно выполняем три шага.

Отличие от основного алгоритма в том, что находимые на 1-м шаге границы сегментов $\omega_s^{(n+1)r}$, $s = 1 : q_{k(r)}$, $r = 1 : \mathcal{U}$ и на 2-м шаге элементы j_{ks}^{n+1} , h_{ks}^{n+1} и f_{ks}^{n+1} транскрипций \mathbf{Q}_k^{n+1} теперь становятся функциями параметра q_k . Поэтому названные величины должны быть вычислены для всех возможных значений параметра q_k .

Функциями параметра q_k становятся и слагаемые $B_k^{n+1}(\mathbf{Q}_k)$ в выражении критерия качества обучения Φ в виде (4.5.5). Это представление критерия Φ позволяет указать способ нахождения оптимальных длин транскрипций $\{q_k^{n+1}\}$ на $(n+1)$ -й итерации.

Как следует из сепарабельности критерия Φ в виде (4.5.5) по переменным k , для этого необходимо решить K независимых задач нахождения величин q_k^{n+1} :

$$q_k^{n+1} = \underset{q_k' \leq q_k \leq q_k''}{\operatorname{argmax}} B_k^{n+1}(\mathbf{Q}_k(q_k)) | q_k^n, \quad k = 1 : K, \quad (4.6.2)$$

где q_k^n — оптимальная длина транскрипции на n -й итерации.

3-й шаг алгоритма не меняется. Лишь полагается, что длины транскрипций слов равны значениям q_k^{n+1} , которые вычислены по формуле (4.6.2).

Останов алгоритма определяется условием

$$\begin{cases} E^{n+1} = E^n, \\ \mathbf{Q}_k^{n+1}(q_k) = \mathbf{Q}_k^n(q_k), \quad q_k = q_k' : q_k'', \quad k = 1 : K. \end{cases} \quad (4.6.3)$$

Если условие останова достигнуто, выполняем заключительную итерацию, полагая оптимальные длины транскрипций равными $q_k = q_k^{n+1}$.

Свойства итерационного алгоритма обучения для случая незаданных длин транскрипций определяются следующими утверждениями, приводимыми без доказательств.

Утверждение 1. От итерации к итерации критерий качества обучения не уменьшается:

$$\Phi^{n+1} \geq \Phi^n; \quad (4.6.4)$$

здесь

$$\Phi^n = \Phi(E^n, \{\mathbf{Q}_k^n(q_k^n)\}, \{\mathbf{v}'(q_{k(r)}^n)\}). \quad (4.6.5)$$

Утверждение 2. Если на n -й итерации не выполнилось условие останова (4.6.3), то найдется такое натуральное число μ , что на $(n+\mu)$ -й итерации:

1) либо условие останова выполнится;

2) либо будет справедливо неравенство

$$\Phi^{n+\mu} > \Phi^n. \quad (4.6.6)$$

Утверждение 3. Равенство в (4.6.4) имеет место, когда

$$\begin{cases} E^{n+1} = E^n, \\ q_k^{n+1} = q_k^n, \\ \mathbf{Q}_k^{n+1}(q_k^{n+1}) = \mathbf{Q}_k^n(q_k^n), \end{cases} \quad (4.6.7)$$

но находятся такие k и $q_k \neq q_k^{n+1}$, что

$$\mathbf{Q}_k^{n+1}(q_k) \neq \mathbf{Q}_k^n(q_k). \quad (4.6.8)$$

Утверждение 4. Алгоритм обучения сходится за конечное количество итераций.

Другой способ автоматического нахождения длин транскрипций слов связан с применением стандартной темпоральной транскрипции слова или m -темпоральной транскрипции слова (об этом см. § 3.2).

Если использовать m -темпоральные транскрипции, то уже на этапе нахождения начальных приближений для итерационного алгоритма обучения пофонемному распознаванию будут найдены длины транскрипций слов. Далее обучение выполняем, используя итерационный алгоритм, полагая известными длины транскрипций слов и стандартные m -темпоральные транскрипции слов.

§ 4.7. ЗАДАЧА ДООБУЧЕНИЯ РАСПОЗНАВАНИЮ РЕЧИ

В пофонемном методе распознавания речи можно ожидать, что по мере увеличения объема словаря все меньше и меньше будет изменяться общая для всех слов совокупность эталонных элементов $e(j) \in E$. Эта особенность метода была подтверждена экспериментами.

Если это так, то нет необходимости выполнять сразу обучение на весь словарь. Достаточно обучиться на распознавание какого-то минимального (по количеству слов) словаря, т. е. оценить совокупность E эталонных элементов $e(j) \in E$, общих для всей речи, и транскрипции (\mathbf{Q}_k, τ_k) всех слов этого минимального словаря. Дальнейшее пополнение словаря можно вести, зафиксировав E и оценивая только транскрипции (\mathbf{Q}_k, τ_k) новых слов по ОВ этих слов, что можно уже делать раздельно для каждого слова.

Задачу оценивания транскрипций (\mathbf{Q}_k, τ_k) слова по ОВ этого слова при условии фиксированной общей для всех слов совокупности E эталонных элементов $e(j) \in E$ назовем задачей дообучения. Задачу дообучения будем отличать от задачи полного обучения пофонемному распознаванию (или просто обучения), когда по ОВ всех слов оцениваются и совокупность E , и транскрипции (\mathbf{Q}_k, τ_k) всех слов, составляющих ОВ.

Решение задачи обучения пофонемному распознаванию слов речи в два приема (сначала полное обучение на минимальный словарь и затем серия задач дообучения с целью расширения словаря) в целом направлено на уменьшение необходимых объемов памяти для запоми-

нания ОВ и на ускорение процесса обучения. Благодаря дообучению возможна быстрая замена отдельных слов в словаре, пополнение словаря новыми словами.

Дообучение пофонемному распознаванию слов подобно обучению поэлементному распознаванию слов. Разница лишь в том, что совокупность, из которой можно выбирать эталонные элементы, не свободна, а задана и равна E . По этой же причине метод (алгоритм) дообучения может быть получен как частный случай основного итерационного алгоритма обучения пофонемному распознаванию, когда совокупность E эталонных элементов $e(j) \in E$ задана, что приводит к распаду задачи обучения на независимые задачи нахождения транскрипций (Q_k, τ_k) отдельных слов.

Сначала накапливаем ОВ из \mathcal{U} реализаций слова, на которое будем дообучаться. Затем, полагая $\tau_k = \tilde{\tau}_k$, решаем задачу оценивания транскрипции Q_k слова. Предварительно находим длину q_k транскрипций слова по тексту слова, вводимому одновременно с накоплением ОВ.

Транскрипция Q_k оценивается с помощью итерационного алгоритма, каждая итерация которого содержит два шага. На первом шаге, отправляясь от E и Q_k^n , осуществляют согласованную (совместную) сегментацию реализаций ОВ — находят решение задач (4.5.2) с помощью алгоритма сегментации (см. § 3.2). Затем на втором шаге по найденным границам сегментов $w_s^{(n+1)r}$, $s = 1 : q_k$, $r = 1 : \mathcal{U}$, $w_0^{(n+1)r} = 0$, $w_{q_k}^{(n+1)r} = l$, и используя E , вычисляют транскрипцию Q_k^{n+1} по формулам (4.5.9) — (4.5.16). Если выполнилось условие останова $Q_k^{n+1} = Q_k^n$, то переходят к заключительной итерации, в результате чего по формулам (4.5.24) находят темпоральную транскрипцию слова τ_k .

Для запуска итерационного алгоритма дообучения необходимо указать начальное значение Q_k^0 транскрипции Q_k слова. С этой целью решают \mathcal{U} задач нахождения Q_k^{0r} -транскрипций слова по одной реализации слова, $r = 1 : \mathcal{U}$. Другими словами, решают \mathcal{U} задач E -самосегментации реализаций X'_i , ОВ. E -самосегментация реализации отличается от самосегментации (см. § 3.2) только тем, что в формуле (3.2.14) e выбирается не свободно, а из совокупности E , т. е. $e \in E$. Поэтому в соответствии с (3.2.16) элементы Q_k^{0r} -транскрипций должны выбираться так:

$$j_{ks}^{0r} = \operatorname{argmax}_{j=1:J} \sum_{i=w_{s-1}^{0r}+1}^{w_s^{0r}} g(x_i^r, e(j)), \quad (4.7.1)$$

$$h_{ks}^{0r} = \operatorname{argmax}_{h \in \Omega} \sum_{i=w_{s-1}^{0r}+1}^{w_s^{0r}} g_1(h_i^r, h), \quad (4.7.2)$$

$$f_{ks}^{0r} = \operatorname{argmax}_{f \in \Omega} \sum_{i=w_{s-1}^{0r}+1}^{w_s^{0r}} g_2(f_i^r, f), \quad (4.7.3)$$

где w'_s , $s = 1 : q_k$, $w'_0 = 0$, $w'_{q_k} = l_r$ — оптимальные граници E -самосегментации r -й реализации ОВ.

Запустив \mathcal{U} раз итерационный алгоритм дообучения начальными условиями \mathbf{Q}_k^{0r} , получим \mathcal{U} решений, из которых выбираем то, которое обеспечивает максимум критерия качества дообучения.

Если же использовать стандартные темпоральные транскрипции слов, то при дообучении темпоральную транскрипцию слова считаем заданной, а для запуска итерационного алгоритма дообучения решаем \mathcal{U} задач m -самосегментации в условиях заданной совокупности E эталонных элементов. Возникающая задача (m, E) -самосегментации отличается от задачи m -самосегментации (см. § 3.2) только тем, что в формуле (3.2.18) е выбирается не свободно, а из совокупности E , т. е. $e \in E$.

Решив \mathcal{U} задач (m, E) -самосегментации реализаций ОВ, находим \mathbf{Q}_k^{0r} -транскрипции слова ($r = 1 : \mathcal{U}$) с длинами транскрипций q'_k . Для этого используем все те же формулы (4.7.1) — (4.7.3). Далее запускаем \mathcal{U} раз итерационный алгоритм дообучения начальными условиями $(\mathbf{Q}_k^{0r}, q'_k)$ и выбираем в качестве результата дообучения ту \mathbf{Q}_k -транскрипцию с соответствующим значением q_k , при которой критерий качества дообучения принимает наибольшее значение.

В связи с задачей дообучения пофонемному распознаванию слов возникает предложение максимально упростить процедуру обучения пофонемному распознаванию речи. Суть этого предложения сводится к тому, что задача решается в два приема: сначала по одной ОВ с помощью того или иного метода самообучения (таксономии, развода на кучи) находится общая для всей речи данного диктора совокупность из J_0 эталонных элементов $e(j) \in E$; затем для включения того или иного слова в словарь распознаваемых слов накапливается ОВ этого слова и решается столько задач дообучения, сколько слов предполагается иметь в словаре. Естественно при этом стремление упростить и сами алгоритмы независимого оценивания E . Так, вместо алгоритма нахождения E , хорошо отражающего специфику речевых сигналов (см. § 4.5), но предполагающего накопление большой ОВ, можно воспользоваться и универсальными последовательными алгоритмами самообучения, не предполагающими накопление ОВ [112].

Необходимость упрощения методов обучения пофонемному распознаванию речи диктуется условиями реализуемости систем пофонемного распознавания.

§ 4.8. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

Эксперименты с пофонемным распознаванием слов проводились начиная с 1973 г. В качестве описания речевого сигнала использовалось двоичное описание (§ 2.1). Элементами x_i были 48-разрядные двоичные коды, представляющие знак производной спектра по частоте на дискретной сетке из 49 частот [78]. Позже в качестве описания стали использовать автокорреляционную функцию речевого сигнала и параметры предсказания [115].

С использованием двоичных кодов была выполнена серия экспери-

ментов по ручному, полуавтоматическому и автоматическому транскрибированию слов речи.

Первый эксперимент состоял в том, что в качестве транскрипции R_k слова бралась квазифонетическая транскрипция. Она задавалась исследователем вручную. Ее отличие от фонетической транскрипции состояло лишь в том, что некоторые звуки рассматривались как состоящие из простых звуков. Правила ручного квазифонетического транскрибирования легко прослеживаются на примерах:

ОДИН	— «АД'ИН»
ДВА	— «ДВА»
ТРИ	— «ТР'И»
ЧЕТЫРЕ	— «ЧЭТЫР'Ь»
ПЯТЬ	— «П'ААТ'»
ШЕСТЬ	— «ШЭС'Т'»
СЕМЬ	— «С'ЭЭМ»
ВОСЕМЬ	— «ВОС'ЬМ»
ДЕВЯТЬ	— «Д'ЕВ'ЬТ'»
НОЛЬ	— «НОЛ'»
РАЗДЕЛИТЬ	— «РАЗД'ЭЛ'ЙТ'»
СЛОЖИТЬ	— «СЛАЖЫТ'»
ВЫЧЕСТЬ	— «ВЫЧ'С'Т'»
ПРОБЕЛ	— «ПРАБ'ЭЭЛ»
ТОЧКА	— «ТОЧКЪ»
ОТКРЫТЬ	— «АТ'КРЫТ'»
ЗАКРЫТЬ	— «ЗАКРЫТ'»
УМНОЖИТЬ	— «УМНОЖЬТ'»
ЗАПЯТАЯ	— «ЗАП'АТАА»
РАВНО	— «РАВНО»
НЕРАВНО	— «Н'ЭРАВНО»
КАВЫЧКА	— «КАВЫЧКЪ»
ДВОЕТОЧИЕ	— «ДВАЭТОЧЪ»
ДАННЫЕ	— «ДАНЫ»
ПЕРЕЙТИ	— «П'ЭР'ЭИТ'»
ЕСЛИ	— «ЈЭЭСЛ'»
ЦИКЛ	— «ЦЫКЛ»
СТОП	— «СТОП»
ЧИТАТЬ	— «ЧИТАТ'»
ПЕЧАТАТЬ	— «П'ЭЧААТАТ'»
ПРОБИТЬ	— «ПРАБ'ЙТ'»
ПИСАТЬ	— «П'ИСАТ'»
КОНЕЦ	— «КАН'ЭЭЦ»
ПРОСТРАНСТВО	— «ПРАСТРАНСТВЪ»
ПЕРЕМОТАТЬ	— «П'ЭР'ЭМАТАТ'»
НАЗАД	— «НАЗАТ'»
ИСТИННО	— «ЙС'Т'ИНЬ»
ЛОЖНО	— «ЛОЖНЬ»

ОПИСАНИЕ
 ИНДЕКС
 ОПЕРАТОР
 НЕТ
 ИЛИ
 ТАКЖЕ
 ПОМЕТИТЬ
 БОЛЬШЕ
 МЕНЬШЕ
 ПРОГРАММА
 ПОДПРОГРАММА
 ФУНКЦИЯ

— ЦАЦП'ИСАН'И'Ь
 — ЦИНДЪЦКС
 — ЦАЦП'ЭРÁЦТЬ
 — ЦН'ЭЭЦТ
 — ЦЙЛ'И
 — ЦТАЦКЖЬ
 — ЦПАМ'ЕЦТ'ИЦТ'
 — ЦБОЛ'ШЬ
 — ЦМ'ЕН'ШЬ
 — ЦПРАЦГРАМЬ
 — ЦПАЦПРАЦГРАМЬ
 — ЦФУНЦКЦЦ'Ь

Вручную было затранскрибировано 200 слов. Затем была накоплена ОВ — 600 реализаций (по три реализации на слово).

Далее с помощью основного алгоритма обучения пофонемному распознаванию слов на основании ОВ оценивалась общая для всех слов совокупность E эталонных элементов $e(j) \in E$. Задача обучения упрощалась тем, что транскрипции слов R_k были заданы. В качестве начального значения для E была взята совокупность эталонных элементов (двоичных кодов), подобранная вручную по видеоспектrogramмам речи для каждого символа из алфавита квазифонетических символов.

Темпоральные транскрипции τ_k определялись автоматически по правилу

$$m_{ks} = m(j_{ks}), \quad M_{ks} = M(j_{ks}), \quad s = 1 : q_k, \quad (4.8.1)$$

а сами величины $m(j)$ и $M(j)$ задавались вручную по видеоспектrogramмам речи. Так, для символов С, С', Ш пара ($m(j)$, $M(j)$) равнялась (6, 20), для Т', Д', Ц, Ч — (2, 8), а для твердых и мягких вариантов В, Л, М, Н, Р, Ф, Х, Й полагалось ($m(j)$, $M(j)$) = (3, 10).

Шаг анализа был равен $\Delta T = 15$ мс.

В качестве элементарной меры сходства была выбрана величина $g(x_i, e(j)) = -H(x_i, e(j))$, где $H(x_i, e(j))$ — хэммингово расстояние между кодами x_i и $e(j)$.

На контрольной выборке из 1000 реализаций (по пять реализаций на слово) было зарегистрировано 15 % ошибок.

Первый эксперимент проводился с целью выяснения применимости простых правил транскрибирования слов, близких к обычному фонетическому транскрибированию. Этот эксперимент, однако, показал, что квазифонетическое транскрибирование плохо отражает акустические явления в словах, в частности не отражает коартикуляцию звуков. Как следствие этого, эталонные сигналы, порождаемые квазифонетическими транскрипциями, плохо аппроксимировали распознаваемые реализации. Обнаружилось также, что в квазифонетических транскрипциях на один звук в слове приходится один элемент транскрипции, а не хотя бы два, как должно быть. Далее стало ясно, что элементы темпоральных транскрипций должны выбираться отдельно для каждого слова.

Первый эксперимент позволил сделать заключение, что квазифонетические транскрипции не отражают акустические образы реали-

заций слов и что необходимо существенно уточнить процесс порождения эталонных сигналов, чтобы улучшить качество аппроксимации распознаваемых сигналов.

Второй эксперимент состоял в том, что составляемые вручную квазифонетические транскрипции R_k слов теперь записывались с учетом акустических явлений в реализациях, с учетом коартикуляции звуков и с большей степенью подробности. Получаемые транскрипции уже слабо напоминали фонетические. Они были названы акустофонетическими (АФ-транскрипциями) [78].

Алфавит символов транскрипций был радикально изменен. Использовалось всего 86 символов ($J = 86$). Гласные и их фазы были представлены весьма подробно. В то же время взрывные (твердые и мягкие) П, П', К, К', Б, Б', Д, Г, Г' оказались весьма невыразительными в кодовом описании. Поэтому символы этих звуков не использовались. Влияние же согласных на гласные представлялось весьма тщательно. Символы гласных размечались вспомогательными знаками огубленности (^), мягкости (') и (‘), переднеязычности (—), назальности (~) и другими признаками, характеризующими коартикуляцию звуков по способу и месту образования. Некоторые суждения об АФ-транскрипциях слов и порядке их составления можно получить из следующих примеров:

ОДИН	— ӦАА·ӦД'ЙИН
ДВА	— ӦВӨААЬ
ТРИ	— ӦР'ЙИ
ЧЕТЫРЕ	— ӦЧЭЦЫЫЭ
ПЯТЬ	— Ӧ·ААА·ӦТ'
ШЕСТЬ	— ӦШЭЭ·ӦЕС'ӦТ'
СЕМЬ	— ӦС'Е·ӦЭЭМ
ВОСЕМЬ	— ӦВӨО·ӦС'ЭМ
ДЕВЯТЬ	— ӦД'ЕЁЕВ'ИӦТ'
НОЛЬ	— ӦНӨО·ӦЛ'
РАЗДЕЛИТЬ	— ӦРАЗ·ӦД'ЕЛ'·ӦЙИӦТ'
СЛОЖИТЬ	— ӦСЛАЖЫ·ӦЙИӦТ'
ВЫЧЕСТЬ	— ӦВЫЫЫ·ӦЧЕС'ӦТ'
ПРОБЕЛ	— ӦРАА·ӦЕ·ӦЭЭЛ
ТОЧКА	— ӦӨӨО·ӦЧ·ААЬ
ОТКРЫТЬ	— ӦАА·ӦРЫЫ·ӦТ'
ЗАКРЫТЬ	— ӦЗАА·ӦРЫЫ·ӦТ'
УМНОЖИТЬ	— ӦҮМНӨБӨОЖЫ·ӦТ'
ЗАПЯТАЯ	— ӦЗАА·ӦАА·ААА·ӦЬ
РАВНО	— ӦРААВ·ӦНӨОЬ
НЕРАВНО	— ӦН'Е·ӦРААВ·ӦНӨОЬ
КАВЫЧКА	— ӦААВЫЫ·ӦЧ·ААЬ

ДВОЕТОЧИЕ	— ےВØÀÀ·ЭЛØØ·ØЧИЭ
ДАННЫЕ	— ےÀÀÁННЫ·Э
ПЕРЕЙТИ	— ےЕР'ЕИ·Т'И
ЕСЛИ	— ےЈЕЁЕС'Л'ЫИ
ЦИКЛ	— ےЦЫЫ·Л
СТОП	— ےСЛØØ
ЧИТАТЬ	— ےЧИ ےÀÀ·Л'
ПЕЧАТАТЬ	— ےЕØЧ·ÀÀ·ЛÀ·Л'
ПРОБИТЬ	— ےРÀÀ·Л'И·Л'
ПИСАТЬ	— ےИ·ЫСÀÀ·Л'
КОНЕЦ	— ےÀÀ·Н'Е·ЭЭØЦ
ПРОСТРАНСТВО	— ےРÀС·РÀÀ·НС·ВÀ·
ПЕРЕМОТАТЬ	— ےЕР'Е·ЭМÀÀ·ЛÀ·Л'
НАЗАД	— ےНАЗАÀÀ·Л
ИСТИННО	— ےЛ'ИС·Л'И·ЫНН·
ЛОЖНО	— ےЛØØЖНА·
ОПИСАНИЕ	— ےÀÀ·Л·И·ЫСÀÀ·НИ·
ИНДЕКС	— ےЛ'ИН·Л'Э
ОПЕРАТОР	— ےÀÀ·Л·ЭРÀÀ·ЛÀ·
НЕТ	— ےН'Е·ЭЭ·Л'
ИЛИ	— ےЛ'ИЛ'И
ТАКЖЕ	— ےÀÀ·Л·Ж·
ПОМЕТИТЬ	— ےÀÀ·М'ЕЁ·Л'И·Л'
БОЛЬШЕ	— ےЛØØ·Л'Ø·Ш·
МЕНЬШЕ	— ےМ'ЕЁ·Н'Ø·Ш·
ПРОГРАММА	— ےРÀС·РÀÀ·ЛÀ·
ПОДПРОГРАММА	— ےÀ·Л·РÀС·РÀÀ·ЛÀ·
ФУНКЦИЯ	— ےФÙÙ·ЦЫ·Л.

АФ-транскрипции составлялись по видеоспектrogramмам слов.

Предполагалось, что в процессе ручного АФ-транскрибирования человек научится составлять АФ-транскрипции новых слов, не глядя на их видеоспектrogramмы. После этого опыта считалось возможным сформулировать алгоритм автоматического АФ-транскрибирования слова по его тексту.

Легко видеть, однако, что АФ-транскрибирование зависит от выбранного способа описания речевых сигналов. В таком случае трудно представить себе такой алгоритм автоматического АФ-транскрибирования, который учитывал бы используемое описание речевого сигнала.

Тем не менее трудная работа по ручному АФ-транскрибированию 200 слов была выполнена. Как и в первом эксперименте, в процессе автоматического обучения оценивалась только совокупность E эталонных элементов. Темпоральные транскрипции по-прежнему задавались вручную, точно так же, как и в первом эксперименте.

Затем была распознана контрольная выборка из 3000 реализаций (по 15 реализаций на слово). Были получены следующие результаты:

0,5 % ошибок и 3 % отказов от распознавания. Отказ от распознавания вырабатывался при выполнении условия

$$G_{k^*}(\mathbf{X}_I) - \max_{k \neq k^*} G_k(\mathbf{X}_I) < 10, \quad (4.8.2)$$

где k^* — ответ распознавания.

Далее объем словаря был увеличен до 300 слов. АФ-транскрипции дополнительных 100 слов были записаны вручную без анализа видеоспектрограмм этих слов. Шесть транскрипций из ста потом все-таки пришлось скорректировать по видеоспектрограммам.

Окончательно была достигнута 98-процентная надежность распознавания 300 слов.

Работа с АФ-транскрипциями поставила остро вопрос об автоматизации процесса задания АФ-транскрипций и темпоральных транскрипций слов. В результате был разработан алгоритм обучения пофонемному распознаванию слов в таком виде, как он представлен в § 4.4—4.7. Этот алгоритм обучения стал применяться начиная с 1975 г.

Интересно отметить, что полученные во втором эксперименте АФ-транскрипции 300 слов и совокупность E эталонных элементов были использованы в качестве начальных условий для итерационного алгоритма обучения (§ 4.5). Полученные в результате обучения новые R_k -транскрипции заметно отличались от заданных вручную АФ-транскрипций: они уже не поддавались фонетической интерпретации и потому были названы акустическими.

Любопытно, что автоматически полученные R_k -транскрипции не ухудшили в целом надежность распознавания контрольной выборки. Это дало основание применять метод автоматического транскрибирования слов на практике [80, 105].

Перейдя к полной автоматизации процесса обучения пофонемному распознаванию слов, пришлось значительно ослабить действие принципов пофонемного распознавания. В самом деле, если при ручном АФ-транскрибировании, например, слов СЕГМЕНТ и ОБМЕН транскрипции этих слов имели общие части, то при автоматическом транскрибировании подобное пересечение транскрипций слов уже не гарантируется.

Это потери, обусловленные переходом к автоматическому транскрибированию. Как увязать и согласовать акустические и фонетические транскрипции слов и тем самым обеспечить глубокое пофонемное распознавание, будет сказано в гл. 6.

Последующие эксперименты проводились только с применением алгоритма обучения пофонемному распознаванию слов. В одном из экспериментов была сначала выполнена процедура полного обучения пофонемному распознаванию 200 слов. Обучающая выборка состояла из 1000 реализаций — по пять реализаций на слово. Затем с помощью дообучения словарь был увеличен до 500 слов [17] и до 1000 слов [116]. Используемый словарь приведен в Приложении 2.

В совокупности E было всего $J = 80$ эталонных элементов $e(j) \in E$.

Время обучения на распознавание 200 слов составило не более 2 ч машинного времени БЭСМ-6, включая накопление ОВ. Время дообу-

чения на одно слово при пяти реализациях ОВ равнялось в среднем 15 с.

Надежность пофонемного распознавания слов для словарей из 200, 500 и 1000 слов составила соответственно 99,5, 98 и 96 %. При этом распознавание есть выбор одной гипотезы из 200, 500 или 1000 гипотез.

Запаздывание ответа распознавания после окончания произнесения слова составило 1, 4 и 8 с для словаря из 200, 500 и 1000 слов соответственно.

Оценки надежности распознавания слов были получены на больших контрольных выборках — от 20 до 50 реализаций на слово.

В Приложении 3 приведены акустические транскрипции первых пятидесяти слов словаря из Приложения 2.

На рис. 4.1—4.3 представлены реализации отдельных слов (левая часть рисунков) и соответствующие им оптимальные эталонные сигналы vR_kE (правая часть). Элементы-коды (их компоненты 0 и 1) изображаются с помощью символов «—» и «ж». Сопоставляя левые и правые части рисунков, можно убедиться, что процесс порождения эталонных сигналов слов vR_kE , $v \in \tau_k(l)$, основанный на использовании транскрипций слова и общей для всех слов совокупности эталонных элементов, достаточно хорошо отображает реальное разнообразие речевых сигналов.

Приведенные результаты по обучению и распознаванию слов речи с помощью пофонемного метода были многократно воспроизведены.

Аналогичные данные по быстродействию и надежности распознавания были получены для элементов-автокорреляций и элементов, представленных параметрами предсказания [81, 108, 115]. Размерность векторов x и e равна 11. На хранение эталонных элементов $e(i) \in E$, которые представлялись 11-мерными b -векторами (§ 2.1), теперь уходило хотя и в 11 раз больше памяти, однако всего лишь 80×11 ячеек. Как и раньше, акустическая R_k и темпоральная τ_k транскрипции одного слова задавались с помощью не более чем 36 байт. Используемая мера сходства $g(x_i, e_i)$ основывалась на скалярном произведении элемента-автокорреляции x_i и эталонного элемента e_i , представленного b -параметрами (§ 2.3, [108]).

Выполненные экспериментальные исследования продемонстрировали эффективность пофонемного метода распознавания слов, подтвердили его преимущества в сравнении с поэлементным методом.

ВЫВОДЫ

1. Разработан метод пофонемного распознавания слов речи, основанный на составлении эталонных сигналов слов из общей для всех слов совокупности небольшого количества эталонных элементов по правилам, определяемым акустическими, темпоральными, громкостными и тональными транскрипциями слов.

Пофонемный метод является частным случаем поэлементного метода, который переходит в пофонемный, если выбирать эталонные элементы слова из общей для всех слов речи совокупности небольшого

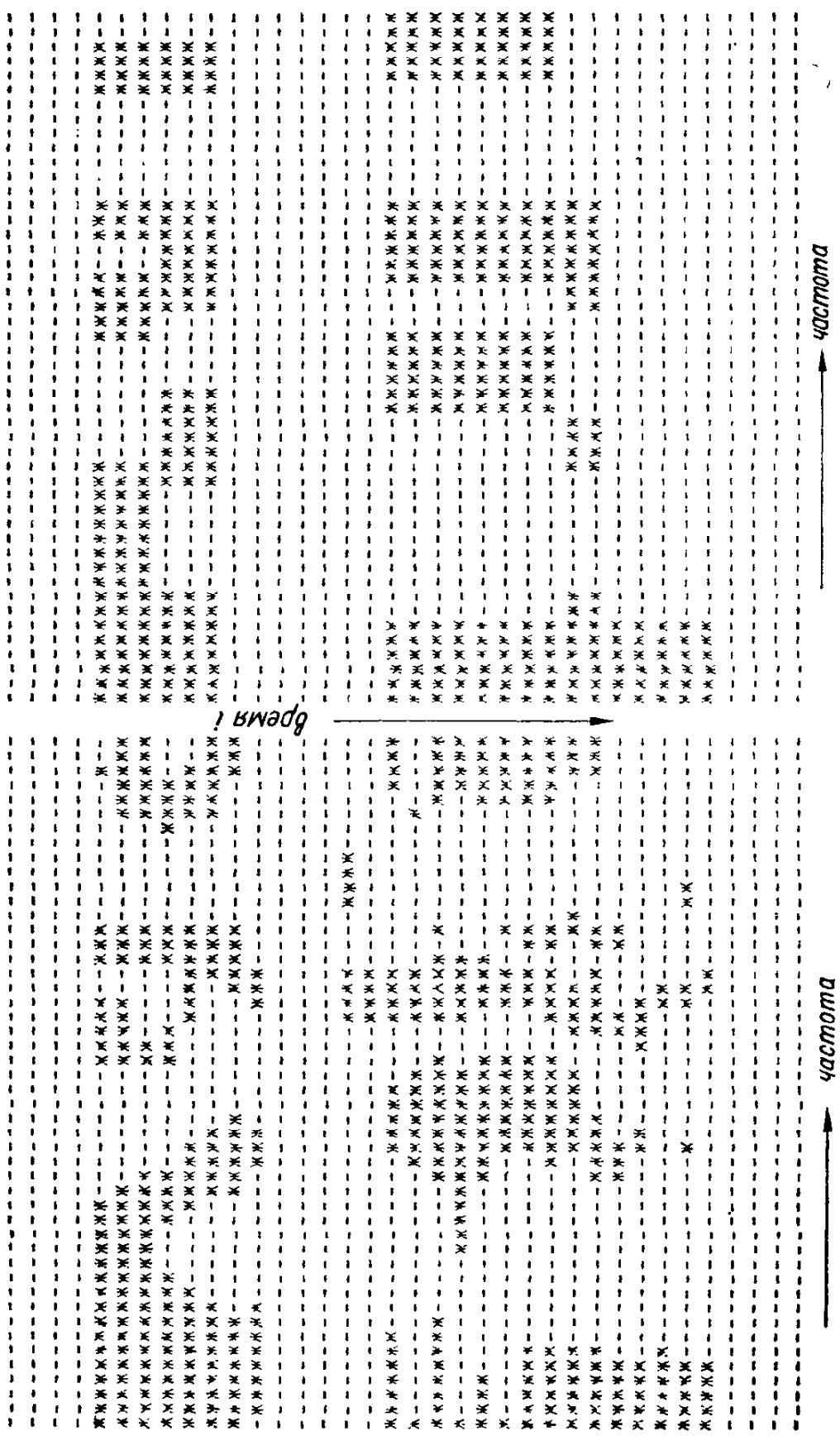


Рис. 4.1, Распознаваемая реализация и соответствующий ей оптимальный эталонный сигнал слова ОДИН.

частота

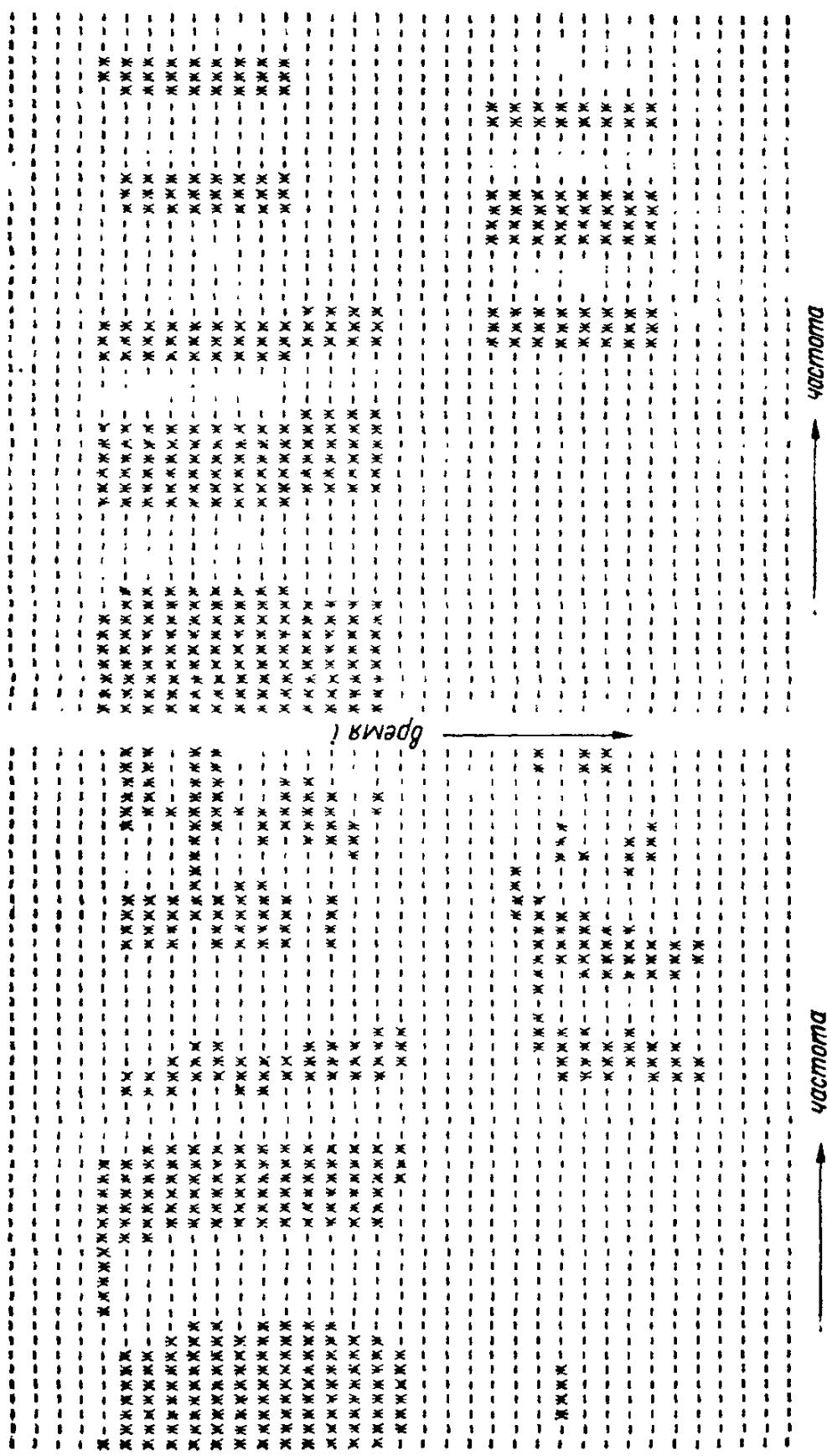


Рис. 4.2. Распознаваемая реализация и соответствующий ей оптимальный эталонный сигнал слова ПЯТЬ.

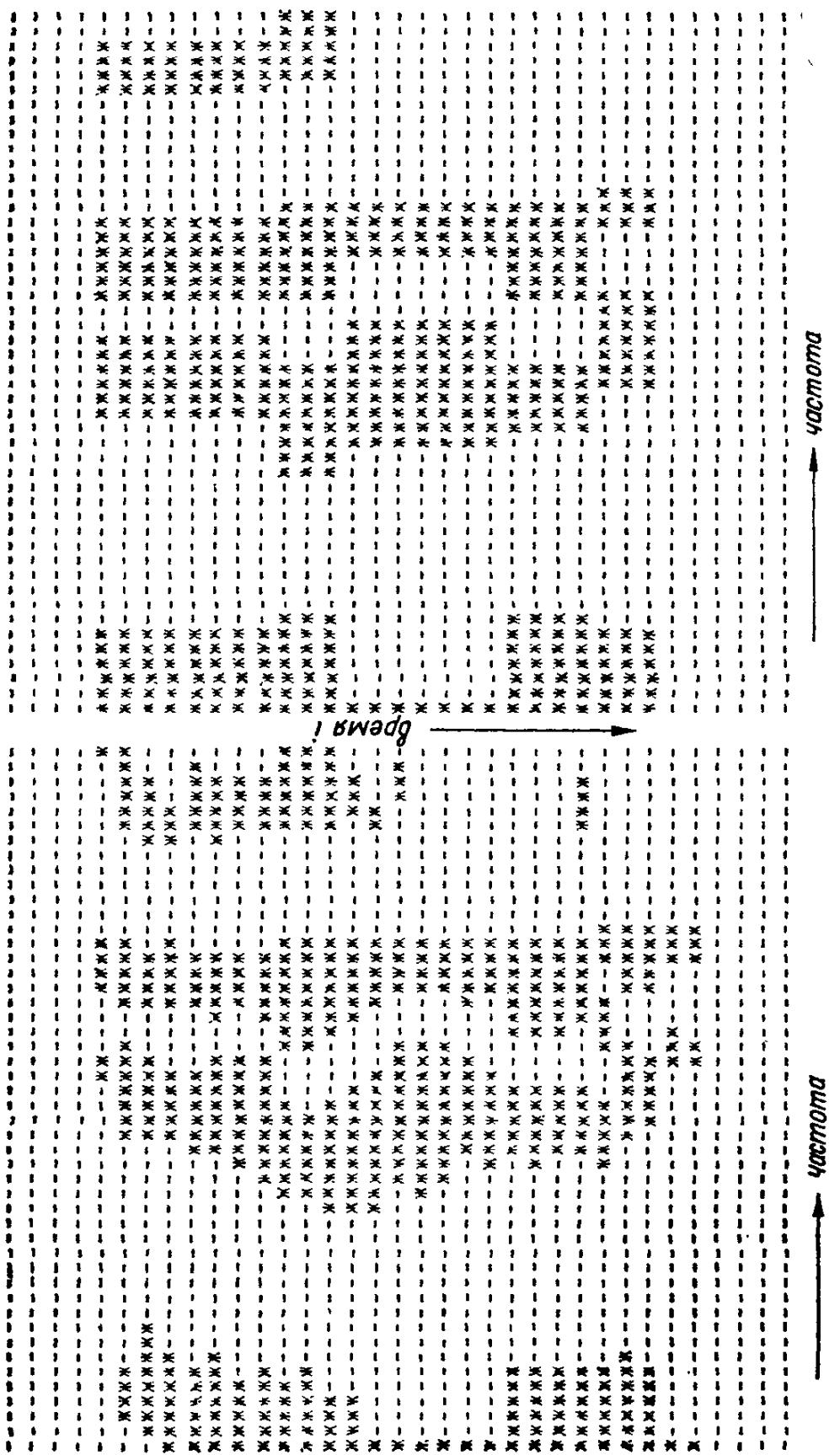


Рис. 4.3. Распознаваемая реализация и соответствующий ей оптимальный эталонный сигнал слова ИЛИ.

количества эталонных элементов. Возникающее в этих условиях понятие акустической транскрипции слова отражает акустические свойства реализаций слов, обусловленные коартикуляцией звуков, и является компромиссом, отражающим как речевые особенности сигналов, так и языковые свойства слов, выражаемые фонетическими транскрипциями слов.

Процедурно метод пофонемного распознавания хотя и совпадает с поэлементным, однако отличается от него существенно меньшими затратами памяти и вычислений.

2. При пофонемном распознавании усложняется задача обучения распознаванию. Она уже не распадается на K (по числу слов в словаре) независимых задач, как в поэлементном методе, а сводится к согласованному по всем словам выбору общей совокупности эталонных элементов и транскрипций слов. Предлагаемые алгоритмы обучения пофонемному распознаванию являются итерационными и заключаются в многократном согласовании сегментации реализаций, транскрипций слов и совокупности эталонных элементов.

По мере увеличения объема словаря уменьшается изменяемость совокупности эталонных элементов, что позволяет фиксировать эту совокупность при достижении некоторого объема словаря, а последующее увеличение или замену словаря вести в режиме дообучения, когда по обучающей выборке отдельного слова оцениваются только транскрипции слова.

3. Пофонемный метод, включающий в себя обучение и дообучение распознаванию, обеспечивает высокую надежность распознавания. Он апробирован на словарях до 1000 слов. Метод ориентирован и рекомендуется для реализации в технических системах.

ГЛАВА 5

РАСПОЗНАВАНИЕ СЛИТНОЙ РЕЧИ, СОСТАВЛЯЕМОЙ ИЗ СЛОВ ВЫБРАННОГО СЛОВАРЯ

Одно из достоинств КДП-подхода к распознаванию речи — возможность перехода от распознавания слов к распознаванию слитной речи, составляемой из слов выбранного словаря. В данной главе будет в основном рассмотрен случай свободного порядка следования слов в предложениях.

Основы метода распознавания слитной речи, составляемой из слов выбранного словаря, были разработаны в 1969—1971 гг. [12, 13, 89]. Сначала было изучено поэлементное распознавание, затем сформулирован метод пофонемного распознавания слитной речи как вариант поэлементного метода для случая составления исходных эталонов слов из общей совокупности небольшого количества эталонных элементов. Как и в случае распознавания отдельно произносимых слов, при пофонемном распознавании слитной речи, в сравнении с поэлементным, существенно уменьшаются необходимые объемы памяти и вычислений. В разработке средств распознавания слитной речи наиболее активное участие приняла О. Н. Гаврилюк.

§ 5.1. ПОСТАНОВКА ЗАДАЧИ ПОЭЛЕМЕНТНОГО РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ.

ПОРОЖДАЮЩИЕ ГРАММАТИКИ И ГРАФЫ СЛИТНОЙ РЕЧИ

Пусть в словаре K слов и пусть каждое слово $k = 1 : K$ задано своим исходным эталонным сигналом E_k и темпоральной транскрипцией τ_k . Условимся, что компонентами эталонного элемента e_{ks} являются также элементы h_{ks} и f_{ks} громкостной H_k и тональной F_k транскрипций соответственно.

Совокупность vE_k , $v \in \tau_k(l)$, составляет множество эталонных сигналов k -го слова, которые имеют длину l (§ 2.2). Сигналы vE_k , $v \in \tau_k(l)$, отличаются нелинейно изменяющимся во времени темпом произнесения, являются коартикулированными, характеризуются различной длиной пауз в начале и конце слова.

Стыкуя (объединяя) эталонные сигналы слов в последовательности, получим эталонные сигналы слитной речи.

Пусть предъявленный для распознавания сигнал слитной речи $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$ имеет длину l . С целью распознавания

сигнала X_l будем составлять эталонные сигналы слитной речи длины l из эталонных сигналов отдельных слов и сравнивать их с распознаваемым сигналом.

Рассмотрим произвольную последовательность слов \mathcal{K}_n , состоящую из произвольного количества n слов:

$$\mathcal{K}_n = (k_1, k_2, \dots, k_r, \dots, k_n). \quad (5.1.1)$$

Произвольному слову k_r с порядковым номером r в этой последовательности можно сопоставить множество $v^r E_{k_r}$, $v^r \in \tau_{k_r}(l_r)$ эталонных сигналов слова k_r , произвольной длины l_r . Всей же последовательности слов \mathcal{K}_n можно сопоставить множество эталонных сигналов слитной речи $E_l(\mathcal{K}_n)$:

$$E_l(\mathcal{K}_n) = (v^1 E_{k_1}, v^2 E_{k_2}, \dots, v^r E_{k_r}, \dots, v^n E_{k_n}), \quad (5.1.2)$$

таких, что

$$v^r \in \tau_{k_r}(l_r), \quad r = 1 : n \quad (5.1.3)$$

и

$$\sum_{r=1}^n l_r = l. \quad (5.1.4)$$

Последовательность

$$E_{\mathcal{K}_n} = (E_{k_1}, E_{k_2}, \dots, E_{k_r}, \dots, E_{k_n}) \quad (5.1.5)$$

можно интерпретировать как исходный эталонный сигнал фразы \mathcal{K}_n , а последовательность

$$v = (v^1, v^2, \dots, v^r, \dots, v^n) — \quad (5.1.6)$$

как оператор преобразования исходного эталона фразы \mathcal{K}_n , такой, что можно записать эталонный сигнал слитной речи в виде

$$E_l(\mathcal{K}_n) = v E_{\mathcal{K}_n}, \quad v \in \tau_{\mathcal{K}_n}(l), \quad (5.1.7)$$

где

$$\tau_{\mathcal{K}_n}(l) = \left\{ v : v^r \in \tau_{k_r}(l_r), r = 1 : n, \sum_{r=1}^n l_r = l \right\}. \quad (5.1.8)$$

Порождаемые с помощью правил (5.1.1) — (5.1.8) эталонные сигналы слитной речи E_l отличаются тем, что соответствуют различным последовательностям слов со свободным порядком их следования, а в случае фиксированной последовательности слов отличаются нелинейно изменяющимся темпом произнесения как всей фразы, так и отдельных слов во фразе, различной длиной пауз между словами, различной длиной слов и всей фразы. Подчеркнем, что сигналы E_l слитной речи — это коартикулированные сигналы.

Задача распознавания сигнала X_l слитной речи далее формулируется как задача отыскания для него наиболее правдоподобного эталонного сигнала слитной речи такой же длины l среди множества всех эталонных сигналов слитной речи, которые генерируются процес-

сом (5.1.1) — (5.1.8), и указания той последовательности слов, из эталонных сигналов которых этот наиболее правдоподобный эталонный сигнал слитной речи состоит. Такая постановка задачи распознавания слитной речи X_i предполагает и автоматическое нахождение количества слов, содержащихся в предъявленном сигнале X_i .

Из постановки задачи следует, что в искомые параметры при распознавании кроме последовательности слов включены еще контрольные параметры — оператор v , определяющий как границы между словами, так и границы звуков в отдельных словах фразы.

Как и в случае распознавания отдельно произносимых слов, отправляясь от того, что наблюдаемые элементы x_i являются результатом независимых искажений соответствующих эталонных элементов $e_i = (vE_{\mathcal{K}_n})_i$, критерий распознавания слитной речи представим в различной форме записи:

$$\begin{aligned}
 \mathcal{K}(X_i) &= \operatorname{argmax}_{\mathcal{K}} \max_{v \in \tau_{\mathcal{K}^{(l)}}} G(X_i, vE_{\mathcal{K}}) = \\
 &= \operatorname{argmax}_{(\mathcal{K}_n, n)} \max_{v \in \tau_{\mathcal{K}_n^{(l)}}} G(X_i, vE_{\mathcal{K}_n}) = \\
 &= \operatorname{argmax}_{(\mathcal{K}_n, n)} \max_{\{w_r\}} \sum_{r=1}^n \max_{v \in \tau_{k_r}(w_r - w_{r-1})} G(X_{w_{r-1} w_r}, vE_{k_r}) = \\
 &= \operatorname{argmax}_{(\mathcal{K}_n, n)} \max_{v \in \tau_{\mathcal{K}_n^{(l)}}} \sum_{i=1}^l g(x_i, (vE_{\mathcal{K}_n})_i) = \\
 &= \operatorname{argmax}_{\mathcal{K}} \max_{v \in \tau_{\mathcal{K}^{(l)}}} \sum_{i=1}^l g(x_i, (vE_{\mathcal{K}})_i). \tag{5.1.9}
 \end{aligned}$$

В формулах (5.1.9) полагается, что $X_{w_{r-1} w_r} = (x_{w_{r-1}+1}, x_{w_{r-1}+2}, \dots, x_{w_r})$ — подпоследовательность (сегмент) распознаваемого сигнала, интерпретируемая как одно слово, причем $w_0 = 0$, $w_n = l$, $w_{r-1} < w_r$, $r = 1 : n$.

По аналогии с распознаванием отдельных слов величина $G(X_{w_{r-1} w_r}, vE_{k_r})$ интерпретируется как интегральное сходство сегмента $X_{w_{r-1} w_r}$ с эталонным сигналом vE_{k_r} , $v \in \tau_{k_r}$, ($w_r - w_{r-1}$), слова k_r , а $G(X_i, vE_{\mathcal{K}_n})$ или $G(X_i, vE_{\mathcal{K}})$ — как интегральное сходство распознаваемого сигнала X_i с эталонным сигналом $vE_{\mathcal{K}_n}$ или $vE_{\mathcal{K}}$ слитной речи, соответствующим последовательности слов \mathcal{K}_n или \mathcal{K} .

В соответствии с записью (5.1.9) задача распознавания слитной речи еще может быть проинтерпретирована как нахождение такого разбиения распознаваемого сигнала X_i на такое количество сегментов, соответствующих отдельным словам, что при интерпретации каждого сегмента как реализации наиболее похожего эталонного сигнала наиболее вероятного слова достигается наилучшая суммарная аппроксимация распознаваемого сигнала последовательностью из эталонных сигналов отдельных слов.

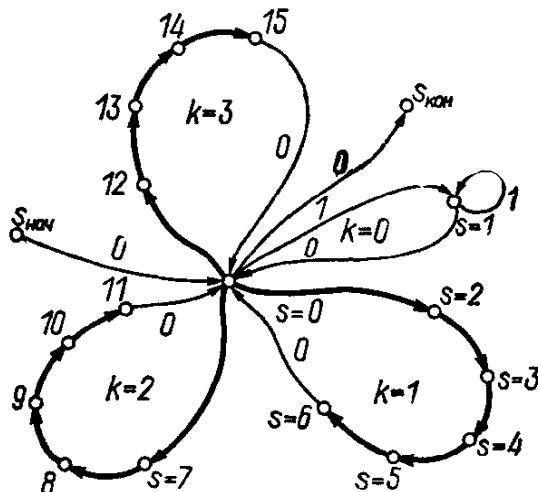


Рис. 5.1. Граф слитной речи.

п рождающую эталонные сигналы слитной речи.

Для этого объединим графы отдельных слов (рис. 2.2, а) в один граф слитной речи (рис. 5.1). С этой целью совместим начальные $s = s_{\text{нач}}$ и конечные $s = s_{\text{кон}}$ состояния графов всех слов в одно состояние $s = 0$. Затем совместим первые $s = 1$ и последние $s = q_k$ состояния графов слов в одно общее состояние паузы $s = 1$. Так получим граф слитной речи, представленный на рис. 5.1, для случая, когда в словаре всего четыре слова (одно из них слово-пауза $k = 0$). Один лепесток графа представляет одно слово $k = 0 : 3$.

Введем сквозную нумерацию состояний s для графа слитной речи, такую, чтобы она отражала порядок прохождения состояний внутри лепестков-слов (см. рис. 5.1).

Как и в случае графов слов, 0-стрелка будет означать переход за 0 тактов времени, а каждая жирная стрелка, входящая в состояние s , будет обозначать веер стрелок $u = m(s) : M(s)$, где $(m(s), M(s))$ — элемент темпоральной транскрипции слова $k(s)$, которому принадлежит состояние s . Переход по стрелке u , входящей в состояние s , будет совершаться за u тактов времени и при этом u раз будет выбираться эталонный элемент $e(s)$ исходного эталона слова $k(s)$, определяемый порядковым номером $\mu(s)$ состояния s внутри слова $k(s) : e(s) = e_{k(s)\mu(s)}$. Напомним, что за счет того, что первые и последние состояния графов всех слов объединены в одно состояние $s = 1$, количество состояний внутри каждого слова и длина исходных эталонов слов уменьшились на две единицы. Это значит, что первые состояния внутри слов на графике слитной речи совпадают со вторыми состояниями графов отдельных слов. Аналогично, последние состояния внутри слов на графике слитной речи совпадают с предпоследними состояниями графов отдельных слов. Как и раньше, длину исходного эталона слова обозначаем q_k , подразумевая, что в это количество уже не включены элементы пауз в начале и конце слова.

Граф слова-паузы $k = 0$ имеет только одно состояние $s = 1$. Переход по стрелкам, входящим в это состояние, осуществляется за один такт времени и сопровождается выбором одного эталонного элемента паузы $e(1)$.

Сформулированная переборная задача распознавания слитной речи в силу специфики процесса порождения эталонных сигналов слитной речи, задаваемого (5.1.1) — (5.1.8), и свойств критерия распознавания (5.1.9) эффективно решается с помощью специальной схемы динамического программирования, учитывающей особенности задачи.

Для записи рекуррентных формул ДП, определяющих решение задачи распознавания слитной речи, представим автоматную грамматику, в виде графа слитной речи.

Для этого объединим графы отдельных слов (рис. 2.2, а) в один граф слитной речи (рис. 5.1). С этой целью совместим начальные $s = s_{\text{нач}}$ и конечные $s = s_{\text{кон}}$ состояния графов всех слов в одно состояние $s = 0$. Затем совместим первые $s = 1$ и последние $s = q_k$ состояния графов слов в одно общее состояние паузы $s = 1$. Так получим граф слитной речи, представленный на рис. 5.1, для случая, когда в словаре всего четыре слова (одно из них слово-пауза $k = 0$). Один лепесток графа представляет одно слово $k = 0 : 3$.

Введем сквозную нумерацию состояний s для графа слитной речи, такую, чтобы она отражала порядок прохождения состояний внутри лепестков-слов (см. рис. 5.1).

Как и в случае графов слов, 0-стрелка будет означать переход за 0 тактов времени, а каждая жирная стрелка, входящая в состояние s , будет обозначать веер стрелок $u = m(s) : M(s)$, где $(m(s), M(s))$ — элемент темпоральной транскрипции слова $k(s)$, которому принадлежит состояние s . Переход по стрелке u , входящей в состояние s , будет совершаться за u тактов времени и при этом u раз будет выбираться эталонный элемент $e(s)$ исходного эталона слова $k(s)$, определяемый порядковым номером $\mu(s)$ состояния s внутри слова $k(s) : e(s) = e_{k(s)\mu(s)}$. Напомним, что за счет того, что первые и последние состояния графов всех слов объединены в одно состояние $s = 1$, количество состояний внутри каждого слова и длина исходных эталонов слов уменьшилось на две единицы. Это значит, что первые состояния внутри слов на графике слитной речи совпадают со вторыми состояниями графов отдельных слов. Аналогично, последние состояния внутри слов на графике слитной речи совпадают с предпоследними состояниями графов отдельных слов. Как и раньше, длину исходного эталона слова обозначаем q_k , подразумевая, что в это количество уже не включены элементы пауз в начале и конце слова.

Граф слова-паузы $k = 0$ имеет только одно состояние $s = 1$. Переход по стрелкам, входящим в это состояние, осуществляется за один такт времени и сопровождается выбором одного эталонного элемента паузы $e(1)$.

Очевидно, что при движении из начального состояния $s_{\text{нач}}$ в конечное $s_{\text{кон}}$ за l тактов времени граф слитной речи будет порождать разнообразные эталонные сигналы слитной речи длины l .

§ 5.2. АЛГОРИТМ РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ. ОСНОВНЫЕ СВОЙСТВА

Естественно предположить, что предъявленный для распознавания сигнал слитной речи $\mathbf{X}_l = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l, \dots, \mathbf{x}_l)$ начинается и кончается какими-либо словами. В таком случае решить задачу распознавания слитной речи означает найти для распознаваемого сигнала оптимальную траекторию из $s_{\text{нач}}$ в $s_{\text{кон}}$ на графе слитной речи рис. 5.1, которая и определит наиболее похожий эталонный сигнал слитной речи.

Подчеркнем, однако, что целью распознавания является последовательность слов, соответствующая наиболее похожему эталонному сигналу слитной речи, а не сам этот сигнал во всех его подробностях. Это значит, что в конечном счете нас будет интересовать не сама оптимальная траектория на графике слитной речи, а только то, по каким лепесткам-словам и в какой последовательности лепестков-слов эта оптимальная траектория пройдет. Именно эта особенность задачи определяет специфический алгоритм распознавания слитной речи [13], основанный на использовании динамического программирования.

Обозначим через $\Omega_i(s)$ множество эталонных сигналов слитной речи длины i , которые генерируются графиком слитной речи при движении из $s_{\text{нач}}$ в состояние s за i тактов времени. Через $F_i(s)$ обозначим наибольшую интегральную меру сходства для сигнала $\mathbf{X}_i = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i)$ на множестве эталонных сигналов $\Omega_i(s)$, а через $v_i(s)$ обозначим потенциально-оптимальный индекс [13], указывающий момент начала последнего слова в наиболее похожем на \mathbf{X}_i эталонном сигнале из множества $\Omega_i(s)$. Это последнее слово обозначим $k_i(s)$, поскольку оно определяется тем словом, которому состояние s принадлежит.

Пусть $F_v(s)$, $v_v(s)$, $k_v(s)$ уже вычислены для всех состояний s графа слитной речи и для всех моментов времени $v < i$, предшествующих i . Тогда с приходом очередного распознаваемого элемента \mathbf{x}_i одновременно (параллельно) для всех s по рекуррентным формулам ДП за время ΔT до прихода очередного элемента \mathbf{x}_{i+1} вычисляем новые значения $F_i(s)$, $v_i(s)$, $k_i(s)$:

$$F_i(s) = \max_{m(s) \leq u \leq M(s)} (F_{i-u}(\mu) + G_i(u, s)), \quad (5.2.1)$$

причем

$$G_i(u, s) = \sum_{v=i-u+1}^i g(\mathbf{x}_v, \mathbf{e}(s)), \quad (5.2.2)$$

$$u_i(s) = \operatorname{argmax}_{m(s) \leq u \leq M(s)} (F_{i-u}(\mu) + G_i(u, s)), \quad (5.2.3)$$

$$v_i(s) = v_{i-u_i(s)}(\mu), \quad (5.2.4)$$

где $\mathbf{e}(s)$, $m(s)$ и $M(s)$ — соответственно эталонный элемент и ограничения на повторяемость эталонного элемента, приписанного стрелкам,

входящим в состояние s (кроме 0-стрелок). В формулах (5.2.1) — (5.2.4) следует положить $\mu = s - 1$ для всех состояний s , кроме первых состояний слов (на рис. 5.1 это $s = 2, 7, 12$), состояния $s = 1$ слова-паузы $k = 0$ и главного состояния $s = 0$. Для первых состояний слов в формулах (5.2.1) — (5.2.3) необходимо принять $\mu = 0$, а $v_t(s)$ вычислять по формуле

$$v_t(s) = i - u_t(s). \quad (5.2.5)$$

Для состояния $s = 1$ слова-паузы $F_t(s)$ и $v_t(s)$ вычисляем по формулам

$$F_t(1) = \max_{\mu=0,1} F_{t-1}(\mu) + g(x_t, e(1)), \quad (5.2.6)$$

$$u_t(1) = \operatorname{argmax}_{\mu=0,1} F_{t-1}(\mu), \quad (5.2.7)$$

$$v_t(1) = \begin{cases} i - 1, & \text{если } u_t(1) = 0; \\ v_{t-1}(1), & \text{если } u_t(1) = 1. \end{cases} \quad (5.2.8)$$

Наконец, для главного состояния $s = 0$ вычисления производим по особым формулам

$$F_t(0) = \max_{k=0:K} F_t(s_k), \quad (5.2.9)$$

$$k_t(0) = \operatorname{argmax}_{k=0:K} F_t(s_k), \quad (5.2.10)$$

$$v_t(0) = v_t(s_{k_t(0)}), \quad (5.2.11)$$

где s_k — последнее состояние слова k на графе слитной речи (для графа рис. 5.1 последние состояния слов $s_0 = 1, s_1 = 6, s_2 = 11, s_3 = 15$).

В процессе вычислений по формулам (5.2.1) — (5.2.11) понадобится запоминать тройку величин $F_t(0), k_t(0)$ и $v_t(0)$ для всех $i = 1 : l$. Еще раз подчеркнем, что величина $k_t(0)$ указывает потенциально-оптимальное слово, которое закончилось в момент i , а величина $v_t(0)$ — момент начала этого слова.

Поскольку предполагалось, что слитная речь $X_l = (x_1, x_2, \dots, x_l)$ начинается с какого-либо слова (в том числе и со слова-паузы) и кончается каким-либо словом (в том числе словом-паузой), то очевидно, что $k^* = k_l(0)$ укажет последнее слово, содержащееся в сигнале X_l , а $v^* = v_l(0)$ — момент его начала. Точно так же $k^{**} = k_{v^*}(0)$ определит предпоследнее слово в распознаваемом сигнале X_l , а величина $v^{**} = v_{v^*}(0)$ — его начало. Тогда предпредпоследним словом будет $k^{***} = k_{v^{**}}(0)$, а его началом $v^{***} = v_{v^{**}}(0)$. Действуя таким образом до тех пор, пока не достигнем очередного v со звездочками, равного 0, получим в обратном порядке последовательность слов, которая содержится в предъявленном для распознавания сигнале X_l . Индексы v^*, v^{**} и т. д. укажут границы между словами.

Алгоритм формирования ответа распознавания по массиву $F_t(0), k_t(0), v_t(0)$ формально записывается так:

$$\begin{cases} v_1^* = l, \\ k_r^* = k_{v_r^*}(0), \quad v_{r+1}^* = v_{v_r^*}(0), \\ r = 1, 2, 3, \dots \text{ до достижения } v_{r+1}^* = 0. \end{cases} \quad (5.2.12)$$

Слова k_r , $r = 1, 2, 3, \dots$, составят ответ распознавания в виде обратной последовательности слов, а величины v_r , $r = 1, 2, 3, \dots$, зададут соответствующие границы между словами.

Алгоритм распознавания слитной речи начинает свою работу с того, что $F_0(0)$ полагается равным нулю, а все неопределенные в правых частях формул (5.2.1) — (5.2.10) величины $F_t(s)$ — равными — ∞ .

Как и в случае распознавания слов, под $g(x_i, e(s))$ в (5.2.2) и (5.2.6) подразумевается сумма трех элементарных сходств, включая сходство громкостей $g_1(h_i, h(s))$ и сходство тональностей $g_2(f_i, f(s))$. Разумеется, отдельные слагаемые g_1 или g_2 или оба одновременно могут отсутствовать, если информация о громкости и (или) о тональности звуков не используется.

Точно так же, как и в случае слов, при распознавании слитной речи вместо основных формул (5.2.1), (5.2.2) для вычисления $F_t(s)$ следует пользоваться формулами, аналогичными (2.3.11) — (2.3.15), которые существенно экономят вычисления:

$$F_t(s) = \begin{cases} \max(F_{t-m(s)}(\mu) + G_t(m(s), s), \\ F_{t-1}(s) + g(x_i, e(s))), \\ \text{если } u_{t-1}(s) \neq M(s), \\ \max_{m(s) \leq u \leq M(s)} (F_{t-u}(\mu) + G_t(u, s)), \\ \text{если } u_{t-1}(s) = M(s). \end{cases} \quad (5.2.13)$$

В связи с заменой формулы для вычисления $F_t(s)$ формула для $u_t(s)$ также изменяется:

$$u_t(s) = \begin{cases} m(s), \text{ если } F_{t-m(s)}(\mu) + G_t(m(s), s) \geq \\ \geq F_{t-1}(s) + g(x_i, e(s)) \\ \text{и } u_{t-1}(s) \neq M(s), \\ \text{в противном случае:} \\ u_{t-1}(s) + 1 \text{ при } u_{t-1}(s) \neq M(s), \\ \operatorname{argmax}_{m(s) \leq u \leq M(s)} (F_{t-u}(\mu) + G_t(u, s)) \\ \text{при } u_{t-1}(s) = M(s). \end{cases} \quad (5.2.14)$$

Из анализа алгоритма распознавания слитной речи следует, что в пересчете на секунду речи требуемые объемы памяти и вычислений несущественно увеличиваются по сравнению с распознаванием отдельных слов. Добавляются лишь вычисления потенциально-оптимальных индексов, однако они сводятся в основном к операциям пересылки чисел. Требуется и дополнительная память из $3l$ ячеек на хранение массива чисел $F_t(0)$, $k_t(0)$, $v_t(0)$, $t = 1 : l$. Сам же процесс формирования ответа распознавания по этому массиву также требует дополнительно незначительного количества пересылок чисел.

На рис. 5.2 представлены схематический (а) и развернутый (б) графы решения задачи распознавания слитной речи. Рассмотрен случай $l = 17$. Для простоты изображения выбраны малые значения

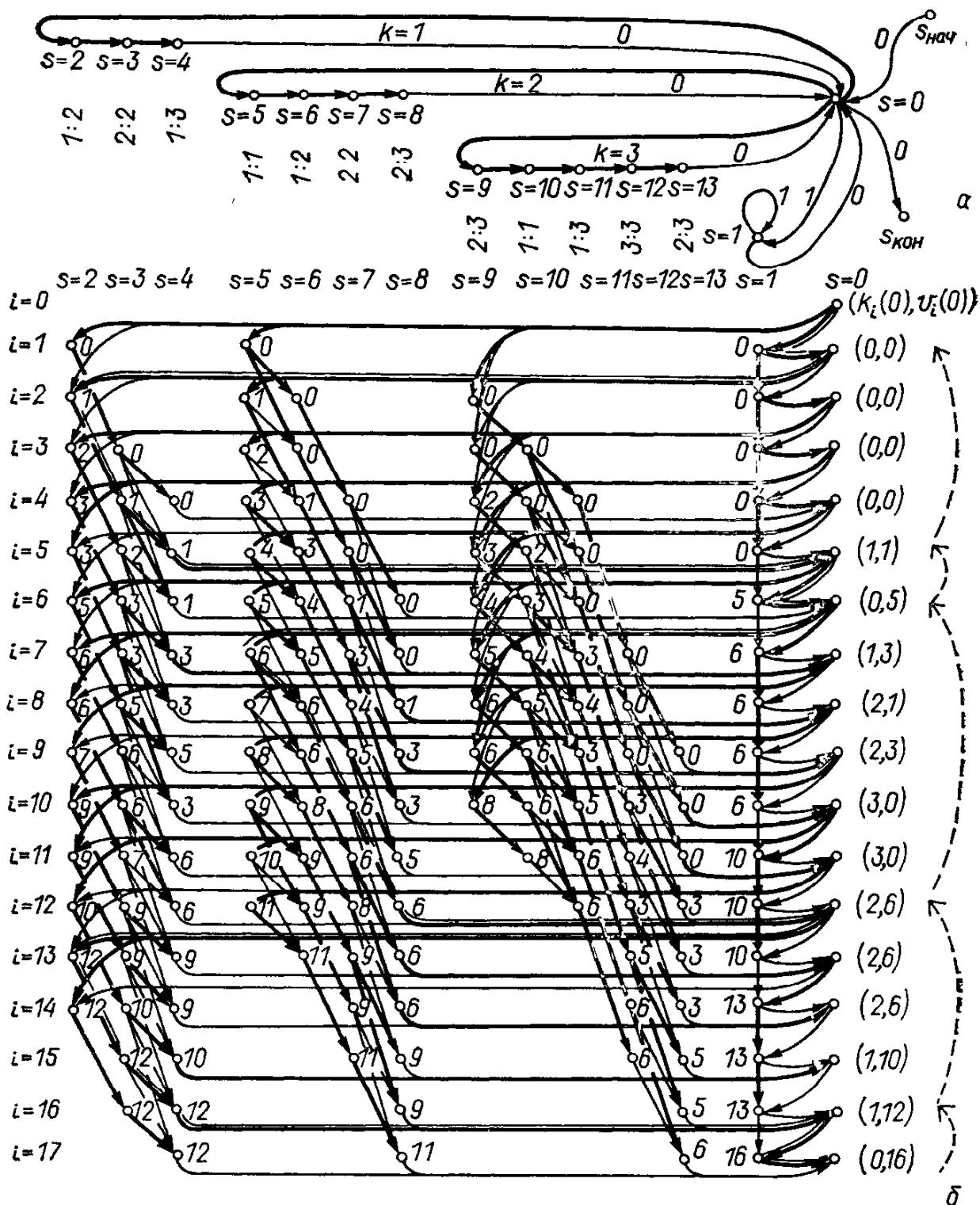


Рис. 5.2. Схематический (а) и соответствующий ему развернутый (б) графы решения задачи распознавания слитной речи.

параметров q_k , $k = 0 : 3$, m (с) и M (с). На развернутом графе потенциально-оптимальные переходы выделены жирными стрелками, а потенциально-оптимальные индексы указаны у вершин графа. Оптимальная траектория и формирование ответа распознавания обозначены соответственно двойной линией и стрелками.

Изложенный алгоритм распознавания слитной речи [13] во много раз (в 5–20 раз) эффективнее других алгоритмов по требуемым объемам вычислений и памяти (см., например, [117]; сравнение с [117] дано в [118]).

Так, в [117] решение задачи распознавания слитной речи организуется таким образом, что для распознаваемого сигнала $\mathbf{X}_l = (x_1, x_2, \dots, x_i, \dots, x_l)$ сначала рассматриваются все возможные сегменты $\mathbf{X}_{jl} = (x_{j+1}, x_{j+2}, \dots, x_l)$, $j < i$, $T_{\min} \leq i - j \leq T_{\max}$, $i, j = 1 : l$, где T_{\min} и T_{\max} — ограничения на минимальную и максимальную длительности слова соответственно. Далее эти сегменты распознаются как реализации отдельных слов из словаря в K слов, а сам процесс распознавания слитной речи в используемых нами обозначениях задается рекуррентными формулами ДП

$$F_l(0) = \max_{\{j: T_{\min} \leq i-j \leq T_{\max}\}} (F_j(0) + \max_k \max_{v \in \tau_k(i-j)} G(\mathbf{X}_{jl}, vE_k)), \quad (5.2.15)$$

$$k_l(0) = \operatorname{argmax}_k \max_{\{j: T_{\min} \leq i-j \leq T_{\max}\}} (F_j(0) + \max_{v \in \tau_k(i-j)} G(\mathbf{X}_{jl}, vE_k)), \quad (5.2.16)$$

$$v_l(0) = \operatorname{argmax}_{\{j: T_{\min} \leq i-j \leq T_{\max}\}} (F_j(0) + \max_k \max_{v \in \tau_k(i-j)} G(\mathbf{X}_{jl}, vE_k)), \quad (5.2.17)$$

посредством которых заполняются массивы $F_l(0)$, $k_l(0)$, $v_l(0)$, $i = 1 : l$, необходимые для формирования ответа распознавания по формуле выписывания (5.2.12). По существу эти формулы отражают тот факт, что сегменты \mathbf{X}_{jl} должны не перекрываться, а покрывать весь сигнал \mathbf{X}_l таким образом, чтобы конец одного сегмента совпадал с началом другого.

Формулы (5.2.15) — (5.2.17) определяют двухступенчатый ДП-алгоритм распознавания слитной речи, поскольку величины $G_k(\mathbf{X}_{jl}) = \max_{v \in \tau_k(i-j)} G(\mathbf{X}_{jl}, vE_k)$ в этих формулах в свою очередь вычисляются с помощью алгоритма ДП.

Предлагаемый нами одноступенчатый ДП-алгоритм распознавания слитной речи эффективно объединяет оба процесса в один ДП-процесс за счет той особенности задачи, что величины $G_k(\mathbf{X}_{jl})$ сразу для всех допустимых значений j и i можно сравнительно просто и быстро вычислять, если выразить их через элементарные сходства распознаваемых и эталонных элементов на структурах развернутых графов (см. рис. 2.2, б и 5.2, б).

Как бы там ни было, можно показать, что в двухступенчатом ДП-методе каждый раз при распознавании сигнала \mathbf{X}_l вычисляется всего $(T_{\max} - T_{\min} + 1) Kl$ сходств $G_k(\mathbf{X}_{jl})$ на развернутых графах размером $(i - j) \times q_k$, тогда как в предлагаемом одноступенчатом ДП-алгоритме (5.2.1) — (5.2.14) фактически таких сходств вычисляется существенно меньше — всего-навсего Kl , правда, на развернутых графах с несколько большими размерами $l \times q_k$. Отсюда получаем выигрыш по быстродействию в $(T_{\max} - T_{\min} + 1)$ раз (более чем в 100 раз при $\Delta T = 15$ мс, $T_{\min} = 300$ мс и $T_{\max} = 2000$ мс). С учетом, однако, того обстоятельства, что размеры развернутых графов различны и необходимо вычислять и пересыпать потенциально-оптимальные индексы, одноступенчатый алгоритм распознавания оказывается в целом в 10—20 раз экономнее по объему вычислений.

§ 5.3. ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ СЛИТНОЙ РЕЧИ

Как уже отмечалось, алгоритм пофонемного распознавания слитной речи, составляемой из слов выбранного словаря, по своей сути не отличается от алгоритма поэлементного распознавания. Используются все те же формулы (5.2.1) — (5.2.14). Однако в них в силу того, что теперь исходный эталон слова E_k составляется из общей для всех слов совокупности эталонных элементов E в соответствии с Q_k -транскрипцией слова, необходимо произвести ряд замен, не влияющих на порядок, характер и объем вычислений.

Поскольку $E_k = Q_k E$, то в (5.1.2) — (5.1.9) вместо $v'E_k$ необходимо писать $v'Q_k E$, а вместо $E_{\mathcal{K}} - E_{\mathcal{K}} = Q_{\mathcal{K}} E$ или $E_{\mathcal{K}_n} = Q_{\mathcal{K}_n} E$, где $Q_{\mathcal{K}}$ или $Q_{\mathcal{K}_n}$ следует рассматривать как Q -транскрипцию фразы \mathcal{K} или \mathcal{K}_n (см. § 5.1). Q -транскрипция фразы образуется объединением в последовательность транскрипций слов, составляющих эту фразу:

$$Q_{\mathcal{K}_n} = (Q_{k_1}, Q_{k_2}, \dots, Q_{k_r}, \dots, Q_{k_n}). \quad (5.3.1)$$

Темпоральная транскрипция $\tau_{\mathcal{K}}$ и $\tau_{\mathcal{K}_n}$ фразы была определена ранее формулами (5.1.6) — (5.1.8).

При переходе к пофонемному распознаванию в алгоритме § 5.2 делаем только одну замену. Вместо $g(x_i, e(s))$ следует писать $g(x_i, e(j(s)) + g_1(h_i, h(s)) + g_2(f_i, f(s))$, где $z(s) = (j(s), h(s), f(s))$ — элемент Q -транскрипции слова $k(s)$, приписанный состоянию s графа слитной речи.

Как и в случае слов, с переходом к пофонемному распознаванию слитной речи существенно экономится память, необходимая для запоминания исходных эталонов слов, а также существенно уменьшается объем вычислений, необходимый для нахождения значений элементарных сходств g . Во всем остальном — полное совпадение с поэлементным методом распознавания слитной речи.

Таким образом, при пофонемном методе генерация эталонных сигналов слитной речи организуется по следующему иерархическому принципу: сначала, отправляясь от общей для всех слов совокупности E небольшого количества эталонных элементов и пользуясь Q -транскрипциями слов, составляются исходные эталонные сигналы слов; затем эти сигналы подвергаются нелинейным деформациям вдоль оси времени в соответствии с темпоральными транскрипциями слов, так что образуются различные эталонные сигналы слов; наконец, последние, объединяясь в последовательности, образуют разнообразные эталонные сигналы слитной речи, которые затем сравниваются с распознаваемым сигналом слитной речи.

§ 5.4. ОБУЧЕНИЕ РАСПОЗНАВАНИЮ СЛИТНОЙ РЕЧИ

Метод распознавания слитной речи, составляемой из слов выбранного словаря, предполагает заданными исходные эталоны и транскрипции слов: (E_k, τ_k) , $k = 1 : K$ — в случае поэлементного распознавания или совокупность эталонных элементов E , общую для всех слов, и транскрипции (Q_k, τ_k) , $k = 1 : K$ — при пофонемном распознавании.

Перечисленные параметры находятся в режиме обучения распоз-

наванию отдельно произносимых слов (гл. 3 и 4). В предположении, что сигналы слов в слитной речи не очень отличаются от сигналов отдельно произносимых слов, результаты обучения по распознаванию слов в равной мере могут быть использованы и при распознавании слитной речи.

Именно так и рекомендуется поступать при распознавании слитной речи, составляемой из слов выбранного словаря. Как будет показано далее, это приводит к вполне приемлемым для практики результатам по надежности распознавания слитной речи.

Все же сигналы слов в слитной речи подвержены дополнительной изменчивости, обусловленной влиянием соседних слов, положением в синтагме, фразовым ударением, типом предложения и т. п.

Поэтому естественно стремление определять параметры слов по реализации слов в слитной речи. Необходимо с этой целью научиться автоматически выделять реализации отдельных слов из потока слитной речи. При наличии такой процедуры обучение распознаванию слитной речи не будет отличаться от обучения распознаванию отдельно произносимых слов, однако будет уже выполнено по слитной речи.

Разработанные средства обучения и распознавания речи позволяют организовать обучение распознаванию слитной речи по ОВ слитной речи.

Обучим систему сначала распознаванию отдельно произносимых слов (гл. 3 и 4). Далее, отправляясь от результатов обучения на слова и исходя из того, что каждая реализация ОВ слитной речи сопровождена указанием учителя о содержащейся в ней последовательности слов, составим для каждой реализации (фразы) ОВ слитной речи исходный эталон фразы по формулам (5.1.1) и (5.1.5) или Q-транскрипцию этой фразы по формулам (5.1.1) и (5.3.1).

При условии фиксированного исходного эталона или транскрипции любая реализация слова или фразы, что все равно, разбивается оптимальным образом на сегменты (подпоследовательности), соответствующие отдельным эталонным элементам исходного эталона или элементам транскрипции. Для этого необходимо воспользоваться алгоритмом сегментации (см. § 3.2). В результате применения алгоритма сегментации все реализации ОВ слитной речи будут разбиты на сегменты, соответствующие отдельным словам (с указанием номера слова, которое представляет сегмент).

Составим новую ОВ как из отдельно произнесенных слов, так и из реализаций слов в слитной речи, выделенных в результате сегментации, и к вновь полученной ОВ применим алгоритм обучения распознаванию отдельно произносимых слов. Получим новые результаты обучения, но уже скорректированные реализациями ОВ слитной речи.

Вновь полученные результаты обучения необходимо опять использовать для оптимальной сегментации реализаций ОВ слитной речи с тем, чтобы уточнить параметры обучения. Далее аналогичную процедуру многократно повторяем.

Реализуя эту итерационную процедуру обучения распознаванию слитной речи, за конечное число итераций достигнем положения равновесия, когда параметры обучения уже меняться не будут.

Для обучения распознаванию слитной речи по слитной речи необходимо многократно использовать всю ОВ как отдельно произносимых слов, так и слитной речи. Возможны, конечно, и упрощенные процедуры, не требующие накопления реализаций ОВ.

Обучение по слитной речи способствует повышению надежности распознавания слитной речи.

§ 5.5. ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ МЕТОДА

Из рекуррентных формул (5.2.1) — (5.2.14) следует, что алгоритм распознавания слитной речи очевидным образом распараллеливается как по отдельным словам, так и по отдельным состояниям внутри слов — точно так же, как и при распознавании отдельно произносимых слов. Эти параллельные процессы реализуются в однотипных вычислителях, в том числе и таких, которые образуют структуру с общим потоком команд. Необходимо несколько видов типовых процессоров для обработки отдельно внутренних и начальных состояний слов. По одному разу должны быть использованы процессор для обработки слова-паузы (состояния паузы $s = 1$) и процессор главного состояния $s = 0$.

Одно из замечательных свойств предлагаемого метода распознавания слитной речи — приблизительно одинаковые в сравнении с распознаванием отдельных слов затраты по объемам памяти и вычислений в пересчете на секунду речи.

Предлагаемый алгоритм распознавания слитной речи обладает еще одной особенностью — для формирования ответа распознавания необходимо получение и запоминание всей последовательности $\mathbf{W}_t = ((F_1(0), k_1(0), v_1(0)), (F_2(0), k_2(0), v_2(0)), \dots, (F_t(0), k_t(0), v_t(0)), \dots, (F_{t+1}(0), k_{t+1}(0), v_{t+1}(0)))$, сколь бы протяженным во времени ни оказался распознаваемый сигнал \mathbf{X}_t . В то же время интуитивно ясно, что после истечения определенного времени с начала произнесения фразы или текста накопленной информации уже может быть достаточно для уверенного указания, какое слово было сказано первым, вторым и т. д.

Возникает задача нахождения некоторых условий, по которым на основании уже накопленной информации можно было бы утверждать, что в некоторый момент времени i^* закончилось некоторое слово $k_{i^*}(0)$, при этом будучи уверенным, что это наше суждение не изменится при анализе всего последующего после i^* речевого сигнала.

Сформулируем некоторые очевидные достаточные условия того, что в момент i^* заканчивается слово $k_{i^*}(0)$.

Утверждение. Если для всех состояний s графа слитной речи, достижимых в момент времени i , имеет место одинаковое значение потенциально-оптимальных индексов, т. е.

$$v_i(s) = c, \quad c = \text{const} \text{ для всех } s, \quad (5.5.1)$$

то слово $k_{i^*}(0)$, где $i^* = c$, является оптимальным словом, которое закончилось в момент времени $i^* = c$.

Из приведенного утверждения вытекает дополнение к алгоритму распознавания слитной речи (§ 5.2).

Одновременно с вычислениями по формуле (5.2.1) — (5.2.11), (5.2.13), (5.2.14) осуществляем проверку условия (5.5.1). При выполнении условия (5.5.1) полагаем $l = i^*$ и по формулам (5.2.12) находим начальную часть ответа распознавания. После этого нет необходимости запоминать начальную часть \mathbf{W}_{i^*} последовательности \mathbf{W}_i . Далее запоминаем число i^* и продолжаем вычисления по формулам (5.2.1) — (5.2.11), (5.2.13) — (5.2.14) с проверкой условия (5.5.1); при выполнении последнего полагаем $l = i^{**}$ и формируем вторую начальную часть ответа распознавания. Далее вторую начальную часть от i^* до i^{**} последовательности \mathbf{W}_i забываем. Действуя далее аналогично, будем формировать ответ распознавания по частям, не дожидаясь всего речевого сигнала \mathbf{X}_t .

Для реализации вычислений по рекуррентным формулам (5.2.1) — (5.2.14), как и в случае распознавания слов, нужна память для хранения промежуточных значений $F_t(s)$, $v_t(s)$. Объем этой памяти (количество чисел) оценивается величиной N :

$$N = 2 \sum_{k=1}^K \sum_{s=1}^{q_k} M_{ks}. \quad (5.5.2)$$

В формуле (5.5.2) еще не учтены состояния $s = 0$ и $s = 1$ графа слитной речи. Так, для слова-паузы $k = 0$ (состояние $s = 1$) необходимо хранить только два числа $F_t(1)$ и $v_t(1)$.

Чтобы учесть все разнообразие проявлений сигналов слова в слитной речи под влиянием соседних слов, в зависимости от положений в синтагме, фразового ударения, типа предложения и т. п., следует задавать слово не одним исходным эталоном или не одной **Q**-транскрипцией, а заводить несколько исходных эталонов или несколько **Q**-транскрипций на слово. Граф слова или подграф слова в графе слитной речи теперь будет представляться либо несколькими отдельными графиками (подграфами-лепестками) с раздельными входами и выходами, либо в виде одного графа (подграфа) с частью общих состояний и раздельными входами и выходами. Такое расширение исходных эталонов, транскрипций или графов (подграфов) слова не изменяет метода распознавания слитной речи. Оно только приводит к росту объемов вычислений и памяти и делает проблему распараллеливания вычислений еще более актуальной.

Количество исходных эталонов или **Q**-транскрипций слова может определяться автоматически в режиме обучения распознаванию слитной речи исходя из требования достижения необходимой точности аппроксимации сигналов ОВ.

При распознавании слитной речи целесообразно ввести и синтаксические ограничения на возможный порядок следования слов. Учет этой априорной информации должен способствовать повышению надежности распознавания. Метод решения задачи распознавания слитной речи для случая синтаксиса языка, задаваемого автоматной грамматикой, когда известны подсловари начальных и конечных слов фраз и для каждого слова словаря указан подсловарь слов, которые могут за ним следовать, подробно рассмотрен в [119]. В этом методе эталонные сигналы слитной речи получаются объединением в последова-

тельность эталонных сигналов слов с учетом допустимого порядка их следования. Задача распознавания слитной речи и в этом случае по-прежнему решается с привлечением динамического программирования и характеризуется примерно тем же объемом вычислений, что и в случае свободного порядка следования слов. Однако за учет ограничений на порядок слов приходится платить увеличением в K раз объема памяти на хранение вспомогательных величин, аналогичных последовательности \mathbf{W}_l , и определяющих ответ распознавания ($3Kl$ чисел вместо $3l$). Процесс вычислений при этом дополнительно распараллеливается еще и по подсловарям.

Метод распознавания слитной речи с учетом синтаксиса языка является частным проявлением более широкого метода смысловой интерпретации слитной речи, когда учитывается не только синтаксис, а и семантика, и прагматика языка устного диалога человека и ЭВМ. Этот метод излагается в гл. 7.

§ 5.6. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Первые экспериментальные исследования поэлементного метода распознавания слитной речи относятся к 1972, 1973 гг. [16, 90]. В условиях, полностью совпадающих с экспериментами 1969—1971 гг. по распознаванию слов, распознавалась слитная речь, составляемая из словаря в 10, 20 и 66 слов.

Предварительная обработка заключалась в цифровой фильтрации речевого сигнала с помощью цифровых резонансных фильтров. Размерность элементов речи \mathbf{x}_i (количество фильтров) равнялось 20 или 25. Длина интервала анализа $\Delta T' = 18$ мс, шаг анализа $\Delta T = 18$ мс.

Распознаваемые реализации слитной речи $\mathbf{X}_i = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_l)$, прежде чем быть распознанными, подвергались нормировке из условия, что максимальный на длине реализации модуль элемента не превосходит единицы: $\mathbf{x}_i = c\tilde{\mathbf{x}}_i$, $i = 1 : l$, $c = 1/\max_i |\mathbf{x}_i|$.

Слитная речь распознавалась по результатам обучения распознаванию отдельно произносимых слов.

При этом с целью сравнения, как и в случае раздельно произносимых слов (§ 2.6), исследовались три различных модификации алгоритма поэлементного распознавания слитной речи.

В первой модификации не накладывались ограничения на повторяемость эталонных элементов, т. е. темпоральные транскрипции слов τ_k полагались равными $\tilde{\tau}_k$.

Во второй модификации учитывалась темпоральная транскрипция слов τ_k .

В третьей модификации дополнительно по отношению ко второй осуществлялось локально-оптимальное сложение за изменением интенсивности элементов. Это сложение было введено с целью устранить принципиальные недостатки применявшейся нормировки элементов по интенсивности. Если для распознавания отдельно произносимых слов эта нормировка обеспечивала хорошее совпадение реализаций со «своими» эталонами, то в слитной речи относительная интенсивность

реализаций слов все время меняется. Суть локально-оптимального слежения за интенсивностью заключалась в том, что если при вычислении элементарного сходства элементов \mathbf{x}_i и \mathbf{e}_i эталонный элемент \mathbf{e}_i брался с некоторой интенсивностью $\alpha_i^* \mathbf{e}_i$, то при вычислении элементарного сходства элементов \mathbf{x}_{i+1} и \mathbf{e}_{i+1} последний уже брался с оптимальной интенсивностью α_{i+1}^* , выбираемой из интервала $[\alpha_i^* - \varepsilon, \alpha_i^* + \varepsilon]$, $\varepsilon = 0,03$:

если $g(\mathbf{x}_i, \mathbf{e}_i) = -|\mathbf{x}_i - \alpha_i^* \mathbf{e}_i|^2$, то

$$g(\mathbf{x}_{i+1}, \mathbf{e}_{i+1}) = \max_{\alpha_{i+1}^* - \varepsilon \leq \alpha_{i+1} \leq \alpha_{i+1}^* + \varepsilon} -|\mathbf{x}_{i+1} - \alpha_{i+1} \mathbf{e}_{i+1}|^2. \quad (5.6.1)$$

Интенсивность первого элемента α_1^* выбираем равной единице, поскольку этот элемент считается совпадающим с элементом паузы.

Локальность слежения обусловливалась тем, что решение о величине α_{i+1}^* выбиралось как оптимальное при условии фиксированного решения на предыдущем шаге.

Локальное слежение за интенсивностью позволяло сравнительно просто, не переходя к двухразмерному динамическому программированию (§ 2.5), учесть нелинейное изменение интенсивности слов в сигналах слитной речи и тем самым устранить недостатки применявшейся нормировки по интенсивности.

Таким образом, элементарной мерой сходства служила величина $g(\mathbf{x}_i, \mathbf{e}_i) = -|\mathbf{x}_i - \mathbf{e}_i|^2$, а в третьей модификации — $g(\mathbf{x}_i, \mathbf{e}_i) = -|\mathbf{x}_i - \alpha_i^* \mathbf{e}_i|^2$. В описываемых экспериментах громкостная и тональная транскрипции слов не использовались.

Были введены также отказы от распознавания для отдельных слов в слитной речи. Пусть v_r^* и v_{r-1}^* — границы слова в потоке слитной речи, определяемые формулами (5.2.12). Сегмент слитной речи с границами v_r^* и v_{r-1}^* , $v_r^* < v_{r-1}^*$, выдается в отказ (объявляется как неопознанное слово), если выполняется условие

$$|F_{v_r^*}(0) - F_{v_{r-1}^*}(0) + \max_{k \neq k_{r-1}^*} G_k(X_{v_r^* v_{r-1}^*})| < \delta \wedge k_{r-1}^* \neq 0. \quad (5.6.2)$$

В формуле (5.6.2) $G_k(X_{v_r^* v_{r-1}^*})$ — величина, характеризующая принадлежность сегмента $X_{v_r^* v_{r-1}^*}$ распознаваемого сигнала X_i к слову с номером k . При этом по аналогии с распознаванием слов (§ 2.6) полагалось $\delta = 0,0075$.

Результаты распознавания слитной речи сведены в табл. 5.1. Количество слов в слитных фразах изменялось от 0 до 8. В таблице приведены результаты экспериментов для двух случаев: а) в словаре 20 слов; б) в словаре 66 слов. Это первые 20 и 66 слов из словаря, приведенного в Приложении 1. Очевидно, что в случае свободного порядка следования слов всего можно составить

$$N = \sum_{i=0}^8 K^i \quad (5.6.3)$$

Таблица 5.1. Результаты экспериментов по поэлементному распознаванию слитной речи

Модификации алгоритма	Объем словаря	Ошибки, %	Ошибки (%) за счет		Отказы, %
			расщеплений слов	слияний слов	
I	20	5	1	1	4
	66	15	3	6	8
II	20	2	0,5	0	2
	66	8	1	2	5
III	20	0,5	0	0	2
	66	9	0	0	3

различных фраз. Даже для случая небольшого словаря $K = 20$ число N достигает весьма больших значений ($N = 25 \cdot 10^9$). Это значит, что в реальных экспериментах по распознаванию слитной речи могут использоваться только отдельные фразы. В случае словаря из 20 слов было распознано 1000 фраз, отдельные фразы произносились по несколько раз. В случае словаря из 66 слов было обработано 300 различных фраз.

Малые объемы экспериментов по распознаванию слитной речи в поэлементном методе были обусловлены замедленным масштабом времени распознавания.

Начиная с 1974 г. выполнялись эксперименты с пофонемным распознаванием слитной речи, сначала в замедленном масштабе времени, затем в квазиреальном времени.

В табл. 5.2 приведены результаты экспериментов с пофонемным распознаванием слитной речи, составляемой из первых 10, 50 и 200 слов словаря, приведенного в Приложении 2.

Условия эксперимента полностью идентичны описанным в § 4.8. Элементы речи представлялись 48-разрядными двоичными кодами. Интервал анализа и шаг анализа $\Delta T' = \Delta T = 18$ мс. Используемая элементарная мера сходства — хэммингово расстояние между кодами x_i и e_i , взятое со знаком минус.

Ошибки распознавания имели следующий типичный характер: слова во фразах распознаются как другое слово или распадаются на несколько слов, чаще всего на два слова. Слова-вставки появляются редко [85].

Таблица 5.2. Результаты экспериментов с пофонемным распознаванием слитной речи

Объем словаря	Количество распознанных фраз	Общее количество слов во фразах	Процент ошибок для слов во фразах	Процент отказов от распознавания отдельных слов
10	3000	9000	0,8	1,7
50	500	2000	2,5	5
200	1000	2000	3	7

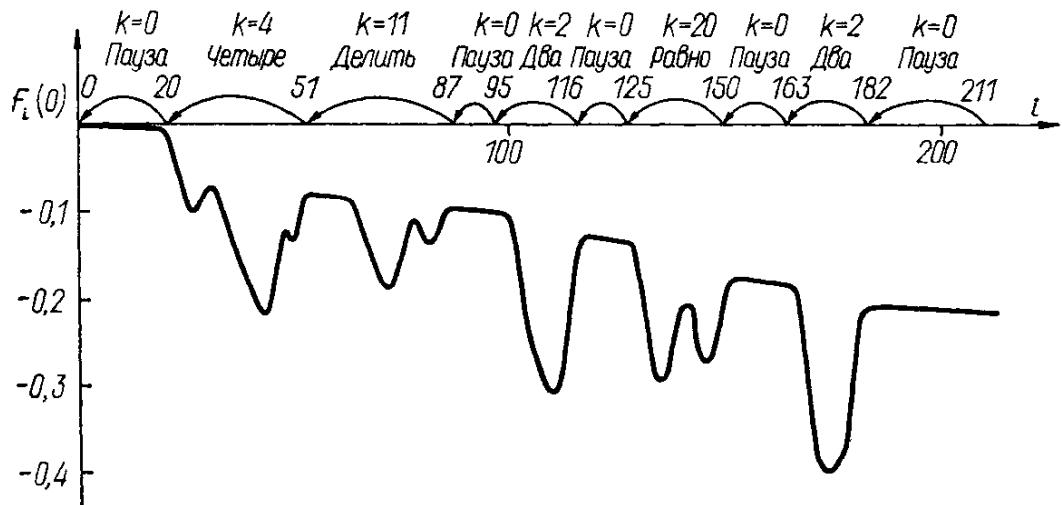


Рис. 5.3. График функции $F_t(0)$ для фразы ЧЕТЫРЕ ДЕЛИТЬ ДВА РАВНО ДВА.

Аналогичные результаты были получены при распознавании слитной речи по параметрам предсказания [81].

С переходом к квазиреальному времени распознавания (как в случае описания речевых сигналов двоичными элементами-кодами, так и в случае использования автокорреляций или параметров предсказания) во многочисленных экспериментах были получены приблизительно одинаковые результаты по надежности распознавания слитной речи, составляемой из слов выбранного словаря. Эта надежность в пересчете на распознавание слов составляет 93 % при словаре в 200 слов [17, 108].

Дальнейшее повышение надежности распознавания слитной речи должно осуществляться путем использования априорной информации о возможном порядке следования слов, выраженной синтаксисом, семантикой и pragmatикой языков устного диалога.

На рис. 5.3 показана работа алгоритма распознавания слитной речи. По оси абсцисс отложено время i , по оси ординат — функция $F_t(0)$. На оси абсцисс указаны оптимальные границы слов v_i^* и сами оптимальные слова k_i .

ВЫВОДЫ

1. Разработаны методы поэлементного и пофонемного распознавания слитной речи, составляемой из слов выбранного словаря. Основаны они на составлении эталонных сигналов слитной речи из эталонных сигналов отдельных слов путем объединения последних в последовательности. При этом сами эталонные сигналы слов образуются путем преобразования исходных эталонных сигналов слова в соответствии с его темпоральными транскрипциями (случай поэлементного распознавания) либо посредством образования исходных эталонных сигналов слов из общей для всех слов совокупности небольшого количества эталонных элементов (согласно акустическим, громкостным и тональным транскрипциям слов) с последующим преобразованием полученных исходных эталонных сигналов слов по темпоральным

транскрипциям этих слов (случай пофонемного распознавания). Распознавание слитной речи заключается в нахождении для предъявленного сигнала наиболее похожего эталонного сигнала слитной речи и в указании последовательности слов, из эталонных сигналов которых составлен этот наиболее похожий эталонный сигнал. Задача распознавания слитной речи эффективно решается с помощью специальной вычислительной схемы одноступенчатого динамического программирования, использующей понятия о потенциально-оптимальном индексе и потенциально-оптимальном слове. Метод распознавания не требует предварительного членения речевого сигнала на участки, соответствующие отдельным словам. Более того, эти участки указываются автоматически в результате решения задачи распознавания слитной речи.

Составной частью методов распознавания слитной речи является предлагаемый метод обучения распознаванию слитной речи.

Методы распознавания слитной речи учитывают основные факторы изменчивости и разнообразия речевых сигналов слитной речи: коартикуляцию звуков, нелинейные изменения темпа и интенсивности произнесения, изменения длительности пауз между словами и т. п.

2. Методы распознавания слитной речи обеспечивают распознавание слитной речи с практически приемлемой надежностью — 93 % в пересчете на надежность распознавания слов при словаре в 200 слов. Метод необходимо использовать при условии обучения на словарь и настройки на голос пользователя. Рекомендуется осуществлять обучение распознаванию по слитной речи. Дальнейшее повышение надежности распознавания слитной речи должно осуществляться путем использования априорной информации о синтаксисе, семантике и прагматике языков устного диалога.

3. В пересчете на обработку одной секунды речевого сигнала предлагаемые методы распознавания слитной речи характеризуются несущественно большей, чем в случае распознавания отдельно произносимых слов, трудоемкостью. Это с учетом ранее отмеченных достоинств предлагаемых методов распознавания отдельно произносимых слов позволяет заключить, что предлагаемые методы распознавания слитной речи во много раз эффективнее других ДП-методов как по объемам используемой памяти, так и по быстродействию.

ГЛАВА 6

ГЛУБОКОЕ ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ РЕЧИ

В задачах пофонемного распознавания слов и слитной речи, рассмотренных в гл. 4 и 5, принцип пофонемного распознавания был реализован поверхностно. Используемые акустические транскрипции слов хотя и позволяли задавать исходные эталонные сигналы слов из общей для всех слов совокупности небольшого количества эталонных элементов, однако выбирались эти транскрипции независимо для каждого слова, вне связи с его фонетической транскрипцией.

Между тем принципиально необходимо увязать выбор акустических транскрипций слов с их фонетическими транскрипциями. Речь идет, таким образом, о более глубокой реализации принципа пофонемного распознавания.

В данной главе в рамках КДП-подхода рассматривается глубокое пофонемное распознавание, использующее понятие фонетической транскрипции слова, а также обсуждаются достигаемые при этом преимущества.

§ 6.1. ПРОЯВЛЕНИЕ ФОНЕМНОСТИ

Слабое проявление фонемного распознавания в методах, изложенных в гл. 4 и 5, иллюстрируется следующим.

Во-первых, для слов, содержащих на фонетическом и акустическом уровнях одинаковые части, никак не гарантируется совпадение соответствующих частей акустических транскрипций. Так, слова ВЕСТЬ и ШЕСТЬ должны были бы отличаться только начальными элементами. Для надежного же распознавания слов этой пары и во всех других аналогичных случаях совпадение частей транскрипций надо гарантировать. А это возможно сделать, согласовав выбор акустических транскрипций слов с их фонетическими транскрипциями.

Во-вторых, поскольку акустическая транскрипция одного слова выбирается вне связи с акустическими транскрипциями других слов, то для включения нового слова в используемый словарь принципиально необходимы накопление ОВ нового слова и решение задачи дообучения, что затрудняет применение метода. Наоборот, зависимый

выбор акустических транскрипций слов, например, в соответствии с их фонетическими транскрипциями принципиально позволяет задавать акустическую транскрипцию нового слова по его фонетической транскрипции, не прибегая к накоплению ОВ и решению задачи дообучения. Последнее обстоятельство приводит к упрощению системы распознавания речи и делает эту систему удобной в использовании.

Таким образом, переход к глубокому пофонемному распознаванию, т. е. к использованию фонетических транскрипций слов, принципиально должен способствовать повышению надежности распознавания и упрощению процедуры обучения (настройки) и использования системы распознавания.

§ 6.2. ОСВОБОЖДЕНИЕ ОТ ЛЕКСИЧЕСКИХ ОГРАНИЧЕНИЙ. ОБЩИЙ ФОНЕМНЫЙ ГРАФ

Глубокое пофонемное распознавание можно реализовать, отправляясь от общего фонемного графа (ОФГ) [89, 99—101], позволяющего генерировать эталонные сигналы слитной речи для каких угодно последовательностей фонем данного языка. Графы же отдельных слов (соответственно акустическая, темпоральная, громкостная и тональная транскрипции) получаются из ОФГ путем вырезки траектории слова в соответствии с его фонетической транскрипцией.

Этот подход подробно рассмотрен в работах [89, 99—101, 120]. Изложим его в дифонной интерпретации, напомнив, что под дифоном подразумевается переход одного звука в другой. В дальнейшем понятия дифона и фонемы будут уточнены с помощью вспомогательных понятий эталонных сигналов дифона и фонемы.

Общая идея подхода будет заключаться в том, чтобы, освободившись сначала от лексических ограничений и учитывая инерционные свойства речеобразующего аппарата и фонетику языка, построить некоторую автоматную грамматику (ОФГ), порождающую все возможные эталонные сигналы слитной речи для каких угодно последовательностей фонем. Эта грамматика должна быть такой, чтобы генерируемые эталонные сигналы слитной речи отражали явления коартикуляции звуков, нелинейного изменения темпа и интенсивности произнесения, удовлетворяли статистике речи, например, касающейся длительности фонем и их фаз и т. п. Существенное ограничение, накладываемое на ОФГ, будет заключаться в том, что он должен иметь такую структуру, чтобы движение по допустимой цепочке состояний этого графа можно было бы трактовать как генерацию эталонного сигнала фонемы в том или ином ее окружении. Это позволит сформулировать задачу фонемного распознавания слитной речи как задачу отыскания для предъявленного речевого сигнала наиболее правдоподобного эталонного сигнала, генерируемого общим фонемным графиком, и указания последовательности фонем, из эталонных сигналов которых этот наиболее правдоподобный эталонный сигнал состоит. Чтобы учесть лексические ограничения, теперь достаточно будет, воспользовавшись фонетическими транскрипциями слов, вырезать из ОФГ линейные

цепочки состояний — графы отдельных слов (подобные приведенным на рис. 2.2, а) и составить из них граф слитной речи (рис. 5.1).

Нетрудно убедиться, что таким образом полученный граф слитной речи будет отражать глубокую фонемность. Его же использование в распознавании слитной речи не будет отличаться от описанного в гл. 5.

Перейдем к более конкретному описанию дифонного варианта.

Известно [121, 122], что в русском языке 43 фонемы (включая паузу): ˘ (пауза), А, О, У, Э, И, Ы, Й, Б, БЬ, В, ВЬ, Г, ГЬ, Д, ДЬ, З, ЗЬ, Ж, К, КЬ, Л, ЛЬ, М, МЬ, Н, НЬ, П, ПЬ, Р, РЬ, С, СЬ, Т, ТЬ, Ф, ФЬ, Х, ХЬ, Ш, ШЬ, Ц, ЧЬ. Соответственно возможно образование 43² дифонов. Ряд дифонов, однако, в русском языке недопустим. В свою очередь, для некоторых дифонов будем различать варианты.

Среди допустимых выделим дифоны ˘˘, А*А* (* — символ удара), АА, О*О*, ОО, У*У*, УУ, Э*Э*, ЭЭ, И*И*, ИИ, Ы*Ы*, ЫЫ, ЙЙ, ВВ, ВЬВЬ, ЗЗ, ЗЬЗЬ, ЖЖ, ЛЛ, ЛЬЛЬ, ММ, МЬМЬ, НН, НЬНЬ, РР, РЬРЬ, СС, СЬСЬ, ФФ, ФЬФЬ, ХХ, ХЬХЬ, ШШ, ШЬШЬ, представляющие стационарные части ударных и безударных гласных и твердых и мягких согласных ˘, А*, А, О*, О, У*, У, Э*, Э, И*, И, Ы*, Ы, Й, В, ВЬ, З, ЗЬ, Ж, Л, ЛЬ, М, МЬ, Н, НЬ, Р, РЬ, С, СЬ, Ф, ФЬ, Х, ХЬ, Ш, ШЬ соответственно. Все гласные фонемы (ударные и безударные) с каждой согласной будут образовывать по два дифона, например, БА* и БА, БЬА* и БЬА, ПА* и ПА, ПЬА* и ПЬА. Аналогично будут образовываться обратные дифоны: А*З и АЗ, А*ЗЬ и АЗЬ, А*Т и АТЬ и т. д.

Особое место занимают дифоны типа согласный-согласный, среди них выделим дифоны типа взрывной согласный-взрывной согласный.

Всего рассматривается около 1 500 дифонов.

От фонемной транскрипции слова нетрудно перейти к его дифонной транскрипции. Так, слову РЕВОЛЮЦИЯ соответствует фонетическая транскрипция ˘РЬЭВАЛЬУЦЫЙА˘, а соответствующая последовательность дифонов имеет вид

˘˘ ˘РЬ РЬРЬ РЬЭ ЭЭ ЭВ ВВ ВА АА АЛЬ ЛЬЛЬ ЛЬУ* У*У*
У*Ц ЦЫ ЫЫ ЫЙ ЙЙ ЙА АА А˘˘.

Для словосочетания БЫСТРЫЙ ВЗДОХ имеем фонетическую транскрипцию ˘БЫСТРЫЙ ВЗДОХ˘ и последовательность дифонов

˘˘ ˘Б БЫ* Ы*Ы* Ы*С СС СТ ТР РР РЫ ЫЫ ЫЙ ЙЙ
Й˘˘ ˘В ВВ ВЗ ЗЗ ЗД ДО*О*О* О*Х ХХ Х˘˘.

Пусть задана общая для всех слов речи и, таким образом, для всех дифонов совокупность E эталонных элементов $e(j) \in E$, $j = 1 : J$, j — имя (порядковый номер) эталонного элемента в E . Как и раньше, пусть эталонные элементы $e(j)$ представляют участки речевого сигнала продолжительностью $\Delta T'$, например, $\Delta T' = 20$ мс, наблюдаемые с шагом $\Delta T = 15$ мс.

Пусть количество J эталонных элементов в совокупности E равно 128, 256, 512 или 1024.

Пусть также каждый дифон d задан парой транскрипций (Q_d , τ_d): Q -транскрипцией и темпоральной. Напомним, что, как и в случае слов,

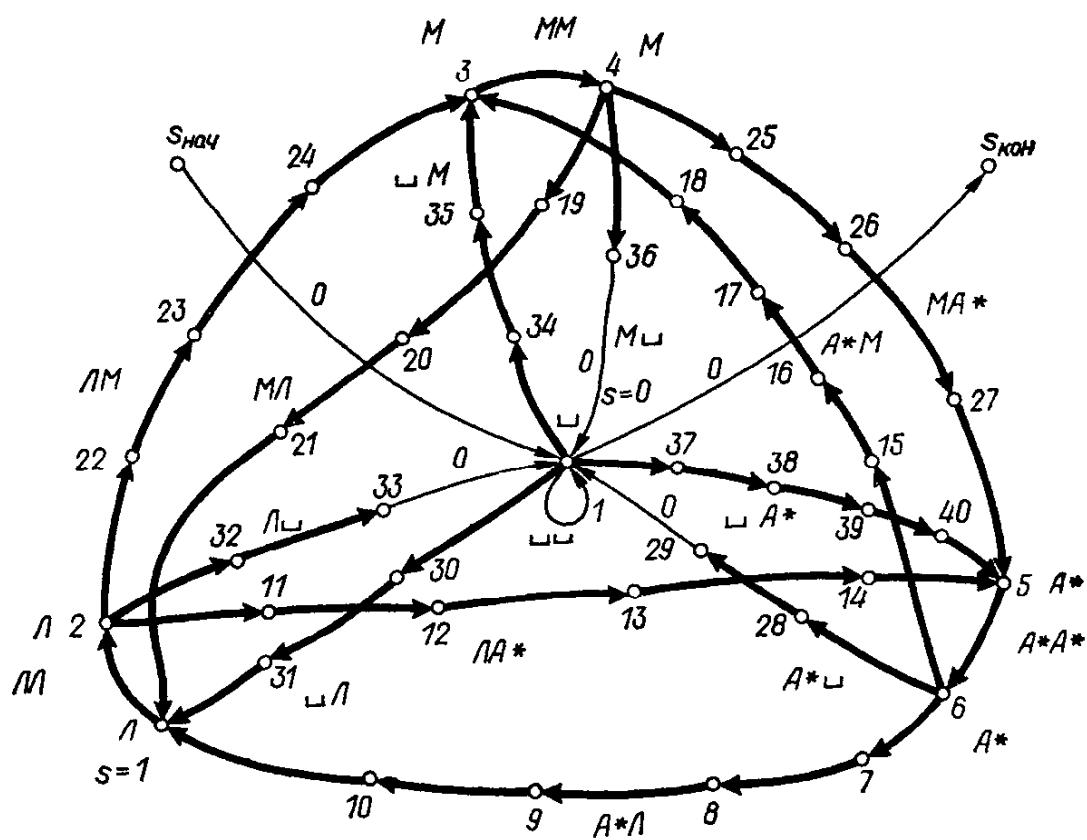


Рис. 6.1. Пример общего фонемного графа.

Q-транскрипция дифона — это тройка, состоящая из акустической R_d , громкостной H_d и тональной F_d транскрипций дифона. Акустическая транскрипция дифона (или **Q**-транскрипция) позволяет составить из E исходный эталонный сигнал дифона, поскольку она указывает последовательность имен элементов $e(j)$ из E для данного дифона d .

Пара транскрипций (Q_d, τ_d) полностью определяет граф дифона, который, подобно графу слова, имеет линейную структуру, аналогичную приведенной на рис. 2.2, а.

Отличие будет состоять лишь в том, что первый и последний элементы транскрипций уже не обязательно будут представлять паузу.

Каждый дифон имеет двойное имя, определяемое парой символов дифона: входное (первый символ дифона) и выходное (второй символ дифона). Например, для дифона АСЬ входное имя А, а выходное — СЬ; для дифона АА входное и выходное имя совпадают.

Соединяя графы дифонов в соответствии с их входными и выходными именами, получим граф слитной речи, названный нами ОФГ. При соединении графов дифонов в ОФГ будем совмещать последнее состояние графа одного дифона с первым состоянием графа второго дифона, если только выходное имя первого дифона и входное имя второго дифона совпадают.

Отметим наиболее характерные состояния s ОФГ.

Прежде всего, это первые или последние состояния графов отдельных дифонов. Поскольку на ОФГ последние состояния одних дифонов совпадают с первыми состояниями других, то естественно эти состояния называть основными и именовать их соответствующими

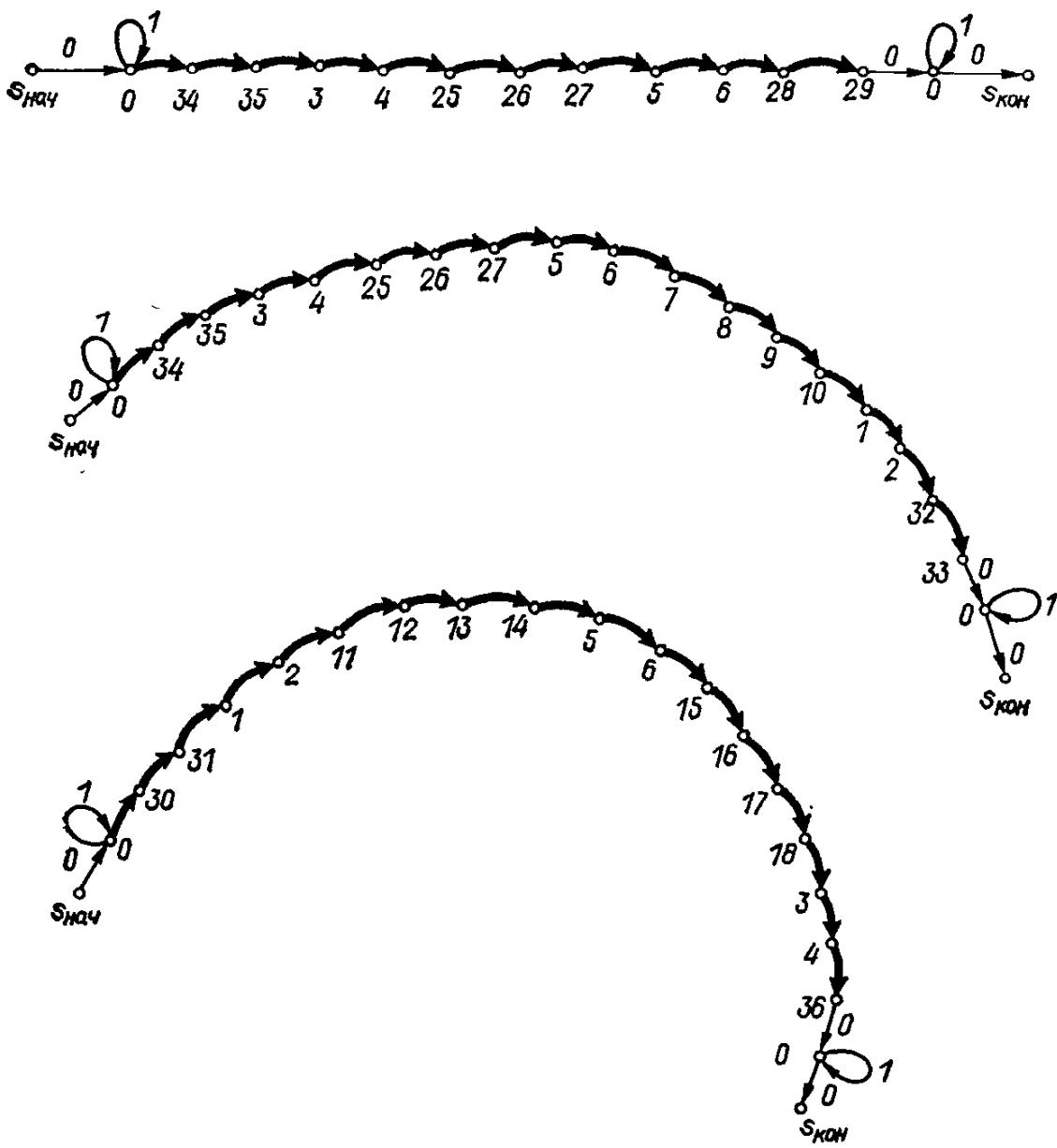


Рис. 6.2. Графы слов МА, МАЛ и ЛАМ, полученные из общего фонемного графа рис. 6.1.

именами дифонов или фонем. Легко убедиться, что на ОФГ будет всего 83 основных состояния.

Среди основных состояний графа ОФГ выделим главное $s = 0$ с именем \square (пауза). В это состояние можно попасть и из начального состояния $s_{\text{нач}}$, а также выйти из него в конечное состояние $s_{\text{кон}}$.

Структура ОФГ представлена на рис. 6.1. Для простоты рассмотрен случай, когда в алфавите всего 4 фонемы \square , A^* , M , L . В качестве основных состояний взяты последние состояния всех дифонов, т. е. $s = 0, 1, 2, 3, 4, 5, 6$. Дифоны помечены двойными именами. Внутри дифонов состояния s нумеруются в порядке возможного их прохождения. Как и в случае графов отдельных слов или графа слитной речи, жирные стрелки на ОФГ обозначают веер стрелок $u = m(s) : M(s)$, определяющих u -кратное повторение эталонного элемента $e(s) \in E$, приписанного стрелкам, входящим в состояние s . Элемент $e(s)$ и ограничива-

ния (m (с), M (с)) однозначно определяются \mathbf{Q}_d - и τ_d -транскрипциями дифона d (с), которому состояние s принадлежит.

Двигаясь вдоль стрелок ОФГ из $s_{\text{иач}}$ в $s_{\text{кон}}$ за l тактов времени, будут генерироваться все возможные эталонные сигналы слитной речи, соответствующие произвольным последовательностям фонем в данном языке и отличающиеся нелинейно изменяющимся темпом произнесения.

Пользуясь ОФГ или, что то же самое, набором графов всех дифонов, нетрудно, отправляясь от фонетической транскрипции слова, вырезать из ОФГ или составить из графов дифонов граф данного слова, удовлетворяющий требованиям глубокой фонемности. Для этого достаточно от фонемной транскрипции слова перейти к его дифонной транскрипции, а затем, в соответствии с последней, соединить графы отдельных дифонов в цепочку графов, которая и составит граф слова.

Примеры графов отдельных слов, полученных из ОФГ рис. 6.1, приведены на рис. 6.2.

Из общего фонемного графа вытекает, что, в отличие от эталонного сигнала дифона и границ дифона, понятия эталонного сигнала фонемы и границ фонемы не имеют четкой выраженности. Однако это обстоятельство не будет препятствовать решению задачи распознавания произвольных последовательностей фонем на основе общего фонемного графа.

§ 6.3. РАСПОЗНАВАНИЕ ПРОИЗВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ФОНЕМ

Используя ОФГ, можно сформулировать задачу распознавания произвольных последовательностей фонем данного языка. Она будет заключаться в том, чтобы для распознаваемого сигнала $\mathbf{X}_t = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t, \dots, \mathbf{x}_f)$ сначала найти наиболее правдоподобный эталонный сигнал слитной речи среди множества всех эталонных сигналов слитной речи, генерируемых ОФГ, и затем на основании этого наиболее правдоподобного эталонного сигнала указать последовательность фонем, которая передается распознаваемым сигналом \mathbf{X}_t .

Такая постановка задачи распознавания является традиционной для КДП-подхода. Особенностью ее является то, что никакая другая априорная информация о речи и языке, кроме акустики речи, фонологических правил ее образования, фонетических знаний о языке, закономерностей генерации речи, включая информацию о длительности фонем, способе и месте их образования, не используется. Попытки же привлечь дополнительную априорную информацию о речи и языке, например, содержащуюся в лексике, синтаксисе, семантике и прагматике языка устного диалога, приводят, что уже было частично показано в предыдущих главах, к несколько другим постановкам задач распознавания речи, однако по-прежнему разрешимых в рамках КДП-подхода.

Итак, в процессе распознавания произвольных последовательностей фонем осуществляется переход от исходного речевого сигнала \mathbf{X}_t к последовательности фонем, которая этим сигналом передается.

Как и в случае распознавания слитной речи, составляемой из слов выбранного словаря, для формирования ответа распознавания в виде последовательности фонем достаточно знать не сам наиболее правдоподобный эталонный сигнал слитной речи во всех его подробностях, а только последовательность имен дифонов, графы которых порождают этот наиболее правдоподобный эталонный сигнал слитной речи. Эта особенность задачи позволяет сформулировать следующий алгоритм распознавания слитной речи на основе ОФГ.

Обозначим через $\Omega_j(s)$ множество эталонных сигналов слитной речи, генерируемых ОФГ при движении из $s_{\text{ нач}}$ в состояние s за j тактов времени, а через $F_j(s)$ — максимальное сходство, которое может быть достигнуто сигналом $\mathbf{X}_j = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j)$ на множестве эталонных сигналов $\Omega_j(s)$.

Потенциально-оптимальным индексом $v_j(s)$ для неосновного состояния s в момент времени j назовем ближайший к j потенциально-оптимальный предшествующий момент времени $v_j(s) < j$ начала дифона $d(s)$, которому принадлежит состояние s . Значение $v_j(s)$, таким образом, определяется наиболее правдоподобным эталонным сигналом из множества $\Omega_j(s)$.

Для основных состояний ОФГ будем определять потенциально-оптимальную фонему $k_j(s)$ и потенциально-оптимальный индекс $v_j(s)$. Поскольку состояние s является основным, то при достижении его в момент времени j будет обязательно закончен какой-то дифон, выходное имя которого совпадет с именем фонемы, приписанной состоянию s . При этом можно утверждать, что фонема, определяемая входным именем этого дифона, уже наверняка закончилась. Поэтому потенциально-оптимальной фонемой $k_j(s)$ основного состояния s , достижимого в момент j , будем называть выходное имя дифона, который в момент j заканчивается в состоянии s . Таким образом, потенциально-оптимальная фонема $k_j(s)$ определяется входным именем дифона, которым заканчивается наиболее правдоподобный эталонный сигнал из множества $\Omega_j(s)$. Потенциально-оптимальный момент начала этого дифона естественно обозначить $v_j(s)$.

Очевидно, что если s — основное состояние и $\mu(s)$ — имя фонемы, приписанной этому состоянию, то имя $(k_j(s), \mu(s))$ определяет дифон, который заканчивается в состоянии s в момент j ; причем ему предшествует дифон с выходным именем $k_j(s)$ и заканчивающийся в момент $v_j(s)$.

Пусть $F_j(s)$, $v_j(s)$, $k_j(s)$ вычислены для всех моментов времени $j < i$ и всех состояний s , достижимых в этот момент.

Тогда с приходом очередного распознаваемого элемента \mathbf{x}_i вычисляем новые значения $F_i(s)$, $v_i(s)$, $k_i(s)$ для всех состояний ОФГ, пользуясь приведенными ниже рекуррентными формулами динамического программирования.

Для всех неосновных состояний s

$$F_i(s) = \max_{m(s) \leq u \leq M(s)} (F_{i-u}(z(s)) + G_i(u, s)), \quad (6.3.1)$$

где

$$G_i(u, s) = \sum_{v=i-u+1}^i g(\mathbf{x}_v, \mathbf{e}(s)); \quad (6.3.2)$$

$$u_i(s) = \operatorname{argmax}_{m(s) \leq u \leq M(s)} (F_{i-u}(z(s)) + G_i(u, s)), \quad (6.3.3)$$

$$v_i(s) = \begin{cases} v_{i-u_i(s)}(z(s)), & \text{если } z(s) = s - 1; \\ i - u_i(s) & \text{в противном случае.} \end{cases} \quad (6.3.4)$$

В формулах (6.3.1) — (6.3.4) функция $z(s)$ определяет предшествующее состояние ОФГ, из которого стрелки непосредственно ведут в состояние s . Например, для первых неосновных состояний дифонов имеем $z(7) = 6$, $z(15) = 6$, $z(28) = 6$, $z(19) = 4$, $z(22) = 2$ (см. рис. 6.1), для всех же непервых неосновных состояний всегда $z(s) = s - 1$.

Для основных состояний ОФГ рекуррентные формулы ДП имеют другой вид

$$F_i(s) = \max_{z \in \mathcal{A}(s)} \max_{m(s) \leq u \leq M(s)} (F_{i-u}(z) + G_i(u, s)), \quad (6.3.5)$$

где $\mathcal{A}(s)$ — множество состояний ОФГ, из которых непосредственно можно попасть в состояние s , например, $\mathcal{A}(3) = \{18, 24, 25\}$, $\mathcal{A}(4) = \{3\}$, $\mathcal{A}(5) = \{14, 27, 40\}$ (см. рис. 6.1);

$$z_i(s) = \operatorname{argmax}_{z \in \mathcal{A}(s)} \max_{m(s) \leq u \leq M(s)} (F_{i-u}(z) + G_i(u, s)), \quad (6.3.6)$$

$$u_i(s) = \operatorname{argmax}_{m(s) \leq u \leq M(s)} (F_{i-u}(z_i(s)) + G_i(u, s)), \quad (6.3.7)$$

$$v_i(s) = v_{i-u_i(s)}(z_i(s)), \quad (6.3.8)$$

$$k_i(s) = k(z_i(s)), \quad (6.3.9)$$

где $k(z)$ — функция, указывающая входное имя (имя фонемы) дифона, которому состояние z принадлежит. Так, $k(31) = \sqcup$, $k(24) = L$, $k(4) = M$, $k(10) = A^*$ и т. п. (см. рис. 6.1).

Для главного состояния ОФГ $s = 0$:

$$F_i(0) = \max (\max_{z \in \mathcal{A}(0)} F_i(z), F_{i-1}(0) + g(\mathbf{x}_i, \mathbf{e}(1))), \quad (6.3.10)$$

причем $z = 0$ не включено в $\mathcal{A}(0)$;

$$z_i(0) = \begin{cases} 0, & \text{если } \max_{z \in \mathcal{A}(0)} F_i(z) < F_{i-1}(0) + g(\mathbf{x}_i, \mathbf{e}(1)); \\ \operatorname{argmax}_{z \in \mathcal{A}(0)} F_i(z) & \text{в противном случае} \end{cases} \quad (6.3.11)$$

$$v_i(0) = \begin{cases} i - 1, & \text{если } z_i(0) = 0; \\ v_i(z_i(0)), & \text{если } z_i(0) \neq 0, \end{cases} \quad (6.3.12)$$

$$k_i(0) = \begin{cases} \sqcup, & \text{если } z_i(0) = 0; \\ k(z_i(0)), & \text{если } z_i(0) \neq 0. \end{cases} \quad (6.3.13)$$

В процессе вычислений по формулам (6.3.1) — (6.3.13) понадобится запоминать величины $k_i(s)$ и $v_i(s)$ для всех основных состояний s ОФГ и всех моментов времени $i = 1 : l$. Подчеркнем, что если $k_i(s)$ указывает входное имя дифона (имя фонемы), который заканчивается в момент i в состоянии s (выходное имя этого дифона (имя фонемы) равно $k(s)$), то $v_i(s)$ определяет момент начала дифона ($k_i(s)$, $k(s)$).

Условившись, что распознаваемый сигнал X_t начинается и кончается паузой, нетрудно убедиться, что по таблице $k_i(s)$, $v_i(s)$, $i = 1 : l$, для всех основных состояний s можно сформировать ответ распознавания в виде последовательности фонем.

В самом деле, последней фонемой в распознаваемом сигнале всегда будет пауза. На это укажет $k_l(0)$, если пауза в конце слитной речи X_t имеет ненулевую длину, в противном случае $k_l(0)$ является первым именем дифона, второе имя которого является паузой по определению.

Таким образом, $k^* = k_l(0)$ задаст имя предпоследней, перед паузой, фонемы. Найдем на ОФГ основное состояние s , имя которого $k(s)$ совпадает с k^* . Таких состояний может быть одно или два. Если их два, тогда выберем то, которое является выходным для дифонов типа WW . Это дифоны A^*A^* , $СЬСЬ$, $ММ$, $ЛЬЛЬ$ и т. п. На графике рис. 6.1, таким образом, будем выбирать состояния $s = 0, 2, 4, 6$, а не $s = 1, 3, 5$ соответственно. Пусть найденное состояние s будет равно s^* . Положим $i^* = v_l(0)$. Тогда $k^{**} = k_{l*}(s^*)$ будет предпоследней фонемой. Далее положим $i^{**} = v_{l*}(s^*)$ и займемся поиском состояния s^{**} , имя которого равно k^{**} . Это будет опять одно или два состояния. Если два, то выберем выходное состояние дифона типа WW , если только $s^{**} \neq s^*$, или входное состояние этого дифона, чтобы обязательно оказалось $s^{**} \neq s^*$. Тогда можно будет выписать имя предпредпоследней фонемы $k^{***} = k_{l**}(s^{**})$ и индекс $i^{***} = v_{l**}(s^{**})$. Действуя далее аналогично, последовательно будем выписывать фонемы k со звездочками, начиная с конца сигнала, вместе с границами i со звездочками, являющимися границами фонем (точнее, дифонов). Выписывание заканчиваем при достижении очередного i со звездочками, равного 0.

Алгоритм распознавания последовательности фонем начинает свою работу с того, что $F_0(0)$ полагается равным нулю, а все неопределенные справа в формулах (6.3.1) — (6.3.13) величины $F_i(s)$ и $G_i(u, s)$ — равными — ∞ .

Заметим, что получаемая в результате выписывания ответа распознавания последовательность фонем может содержать пары одинаковых соседних фонем, если это не фонема \perp , или более чем две подряд фонемы \perp . Совпадающие соседние фонемы следует заменить одним символом. В этом заключается редактирование ответа распознавания.

Как и в случае распознавания слов и слитной речи, составляемой из слов выбранного словаря, вместо основных формул (6.3.1) — (6.3.3), (6.3.5) — (6.3.7), (6.3.10), (6.3.11) можно пользоваться формулами, аналогичными (2.3.11) — (2.3.15), которые существенно экономят вычисления.

Алгоритм распознавания последовательностей фонем легко распараллеливается на однотипные вычисления. В частности, распараллеливание можно вести по дифонам.

В формулах (6.3.2) и (6.3.10) под $g(x_i, e(s))$ подразумевается сумма трех элементарных сходств, включая сходство громкостей $g_1(h_i, h(s))$ и сходство тональностей $g_2(f_i, f(s))$. Однако, как и раньше, отдельные слагаемые в $g(x_i, e(s))$ могут изменяться или опускаться. В частности, в силу значительной вариативности основного тона в слитной речи

в качестве сходства тональностей $g_2(f_i, f(s))$ может использоваться только сходство признаков тон-шум элементов x_i и $e(s)$.

В целом, распознавание произвольных последовательностей фонем в потоке слитной речи следует рассматривать все же как вспомогательную задачу в общей проблематике распознавания и смысловой интерпретации слитной речи.

§ 6.4. ИСПОЛЬЗОВАНИЕ ФОНЕТИЧЕСКОЙ ТРАНСКРИПЦИИ СЛОВА. ПЕРЕХОД К ГЛУБОКОМУ ПОФОНЕМНОМУ РАСПОЗНАВАНИЮ СЛОВ И СЛИТНОЙ РЕЧИ

В § 6.2 было показано, как, пользуясь общим фонемным графом или транскрипциями дифонов, следует составлять графы слов или (Q_k, τ_k) -транскрипции слов по их фонетической транскрипции. Для этого достаточно от фонетической транскрипции слова перейти к последовательности его дифонов и далее, объединяя графы или транскрипции дифонов согласно последовательности дифонов слова, получить график слова или (Q_k, τ_k) -транскрипцию слова.

В дальнейшем для распознавания отдельно произносимых слов речи или слитной речи, составляемой из слов выбранного словаря, следует без изменений применить метод пофонемного распознавания речи, изложенный в гл. 4 и 5. Однако в этом случае уже будет достигаться глубокое пофонемное распознавание, поскольку исходные эталоны слов будут выбираться согласно фонетическим транскрипциям слова.

Таким образом, глубокое пофонемное распознавание достигается теми же средствами обработки информации и вычислений, что и простое. Отличие лишь в способе выбора транскрипций, более полном использовании априорной информации о структуре и свойствах речевых сигналов, фонетическом составе слова, способе обучения пофонемному распознаванию.

При глубоком пофонемном распознавании процесс образования эталонных сигналов речи приобретает явно выраженный иерархический характер. Имеются общая совокупность E эталонных элементов и общая совокупность (Q_d, τ_d) -транскрипций всех дифонов данного языка. Каждое слово задано своей фонетической транскрипцией. По ней на основании (Q_d, τ_d) -транскрипций дифонов составляется (Q_k, τ_k) -транскрипция слова. Она же задает способ составления всех возможных эталонных сигналов слова из эталонных элементов совокупности E . Эталонные сигналы слов, объединяясь в последовательности, образуют эталонные сигналы слитной речи, которые далее сравниваются с распознаваемым сигналом.

Глубокое пофонемное распознавание слов и слитной речи использует сильную априорную информацию о лексике языка. Пофонемное же распознавание произвольных последовательностей фонем подобную информацию не использует. Существуют предложения в последнем случае привлечь дополнительную априорную информацию о вероятностях (частотах) встречаемости пар или триад фонем, что должно повысить надежность распознавания произвольных последовательностей

фонем. Соответствующие изменения в КДП-метод распознавания произвольных последовательностей фонем, учитывающие вероятности пар или триад фонем, даны в [100—101]. В целом же, однако, априорная информация о частотах встречаемости пар и триад фонем является не очень существенной. Гораздо более существенной и сильной априорной информацией о языке и речи является информация, выраженная лексикой предметной области. Именно по этой причине в данной главе и уделено основное внимание глубокому пофонемному распознаванию слов и слитной речи.

§ 6.5. ОБУЧЕНИЕ ГЛУБОКОМУ ПОФОНЕМНОМУ РАСПОЗНАВАНИЮ

Обучение глубокому пофонемному распознаванию речи может быть организовано по-разному. Один из рекомендуемых способов предполагает решение задачи в несколько этапов. Сначала на первом этапе по обучающей выборке решается задача самообучения (таксономии, разбала на кучи) — нахождения совокупности E из заданного количества $J = 128, 256, 512$ или 1024 эталонных элементов, наилучшим образом аппроксимирующих все наблюдаемые элементы x_i ОВ. Возможные постановки и способы решения этой задачи рассматривались в § 4.5 и 4.7.

На втором этапе по ОВ (реализациям слов или отрезкам слитной речи, сопровождаемым указаниями учителя о передаваемых последовательностях фонем или дифонов) определяются максимально правдоподобные оценки (Q_d, τ_d) -транскрипций всех дифонов d . При этом предполагается, что совокупность E эталонных элементов уже задана (она была найдена на первом этапе).

Для решения задачи второго этапа исходят из того, что длины q_d транскрипций дифонов заданы. Рекомендуется эти параметры задать вручную, что должно быть сделано опытным исследователем. Предполагается, что длины q_d транскрипций дифонов для выбранного шага анализа ΔT одинаковы для всех дикторов. Поэтому задание длин транскрипций дифонов делается один раз.

Задача второго этапа решается с помощью итерационного алгоритма [100—101]. На первом шаге итерации, предполагая Q_d -транскрипции дифонов известными, находят границы дифонов в реализациях ОВ и границы отдельных сегментов дифонов, аппроксимируемых одним и тем же элементом транскрипции дифона. Этот шаг итерации совпадает с оптимальной сегментацией реализаций ОВ по заданным их дифонным транскрипциям и при условии заданного E . Здесь имеет место полная аналогия с шагом 1 итерационного алгоритма обучения пофонемному распознаванию слов (см. § 4.5). Затем на втором шаге при фиксированных границах сегментов дифонов определяют новые элементы транскрипций дифонов (полная аналогия с шагом 2 названного выше алгоритма из § 4.5). Первоначальное же задание транскрипций Q_d^0 дифонов d при заданной совокупности E может осуществляться по отдельным реализациям дифонов с помощью алгоритма транскрибирования по одной реализации (§ 4.7). Разметка границ отдельных дифонов в слитной речи может быть сделана вручную.

В результате второго этапа обучения все реализации слов и слитной речи, входящие в обучающую выборку, окажутся разбитыми на участки, соответствующие отдельным дифонам. Рассматривая эти участки как реализации дифонов, составим новую ОВ как из реализаций отдельных дифонов, так и из дифонов, выделенных из слов и слитной речи. Далее, рассматривая полученные в конце второго этапа Q_d -транскрипции дифонов и полученную в конце первого этапа совокупность E эталонных элементов в качестве начальных условий, приступаем к третьему этапу обучения — совместному оцениванию совокупности E эталонных элементов и совокупности $\{Q_d\}$ всех транскрипций дифонов.

Задача совместного оценивания этих параметров полностью совпадает с задачей обучения пофонемному распознаванию отдельно произносимых слов речи с той лишь разницей, что первые и последние элементы слов (в данном случае дифонов) должны выбираться теперь независимо и не обязательно должны быть равны паузе.

Четвертый этап обучения заключается в многократном последовательном повторении двух шагов, аналогичных второму и третьему этапам соответственно. Сначала, располагая максимально правдоподобными оценками E и $\{Q_d\}$, полученными в результате шага, аналогичного третьему этапу, и используя эти оценки как начальное условие, находим новые максимально правдоподобные границы дифонов во всех реализациях ОВ и составляем новую ОВ из реализаций дифонов. Это — оптимальная сегментация слов и слитной речи. Затем на следующем шаге, аналогичном третьему этапу, занимаемся совместным оцениванием E и $\{Q_d\}$ по новой ОВ дифонов, пользуясь новыми границами дифонов и передаваемыми от шага к шагу начальными условиями относительно E и $\{Q_d\}$.

Подобный итерационный процесс согласования решений о сегментации слитной речи, совокупности E и транскрипциях Q дифонов за конечное количество итераций достигает такого состояния, когда улучшить оценки E или $\{Q_d\}$, изменяя одну из обобщенных переменных, уже не удается.

Подчеркнем, что обучение глубокому пофонемному распознаванию хотя и громоздко, однако осуществляется один раз для одного диктора. Изменение и пополнение словаря сводятся лишь к формированию по тексту слова его фонетической транскрипции и составлению по ней графа слова согласно дифонной транскрипции слова. Это преимущество глубокого пофонемного распознавания достигается тем, что совокупность E эталонных элементов и совокупность $\{Q_d\}$ транскрипций дифонов остаются неизменными.

С появлением описаний речевого сигнала, слабо зависящих от диктора, обучение глубокому пофонемному распознаванию становится абсолютно разовым и должно быть тщательно осуществлено в лабораторных условиях.

§ 6.6. ВЗАИМОСВЯЗЬ ЗАДАЧ РАСПОЗНАВАНИЯ И СИНТЕЗА РЕЧИ

На примере глубокого пофонемного распознавания речи наиболее ярко проявляется связь развивающегося подхода к распознаванию и смысловой интерпретации речи с синтезом речи.

Прежде всего, этот подход основан на идее синтеза (генерации) эталонных сигналов речи из общей для всех слов совокупности эталонных элементов и в соответствии с акустическими, громкостными, тональными и темпоральными транскрипциями слов и дифонов, составляемыми по их фонетическим транскрипциям. Далее, в этом подходе синтез речи (генерация эталонных сигналов слитной речи) включен в обратную связь по отношению к процессу распознавания.

Все это позволяет трактовать проблемы автоматического распознавания и синтеза речи с единых позиций, в частности, рассматривать синтез речи как компоненту процесса распознавания речи.

В рамках КДП-подхода наиболее удобно изложить синтез речи, отправляясь от общего фонемного графа или совокупности графов дифонов.

Произвольный исходный текст, предназначенный для озвучивания, сначала подвергается автоматическому фонемному транскрибированию (см., например, [123—124]), причем этот процесс сопровождается выделением ударений в словах, синтагматических и фразовых ударений, вычислением типов интонации (перечисления, завершенности, обращения, утверждения, вопроса и т. п.). Далее последовательность фонем заменяется последовательностью дифонов. Затем составляются **Q**-транскрипция и темпоральная **τ**-транскрипция синтезируемого текста путем выписывания в последовательность **Q**- и **τ**-транскрипций дифонов.

Выберем некоторый средний темп произнесения, т. е. будем повторять *s*-й элемент транскрипции **Q_d** дифона *d* среднее число раз $v_{ds} = \frac{1}{2} (m_{ds} + M_{ds})$, где (m_{ds}, M_{ds}) — *s*-й элемент темпоральной транскрипции дифона τ_d . Тогда **Q**- и **τ**-транскрипции текста позволят записать последовательность эталонных элементов (эталонный сигнал слитной речи для анализируемого текста), «произнесенную» в среднем темпе.

Поскольку эталонные элементы характеризуют передаточную характеристику речевого тракта и параметры источников его возбуждения, то полученная последовательность эталонных элементов определяет динамику управляющих параметров для модели речеобразования (собственно синтезатора речи — линейной системы с изменяемыми, управляемыми параметрами). Однако, прежде чем использовать громкостную, тональную и темпоральную транскрипции синтезируемого текста, в них рекомендуется внести, что вовсе не обязательно, определенные корректизы в соответствии с синтагматическими и фразовыми ударениями и типом интонации. Сами же эталонные элементы в акустических транскрипциях дифонов не корректируются, поскольку коартикуляция и редукция звуков уже учтены в самих этих транскрипциях.

В целом, синтез речи при КДП-подходе идет встык — дифон за дифоном, согласно дифонной последовательности текста, с незначительной корректировкой громкостной, тональной и темпоральной транскрипций дифонов под влиянием текста.

Идея синтеза речи в рамках КДП-подхода хорошо согласуется с фонемно-формантным методом синтеза речи, развитым в работах Б. М. Лобанова [125—127] и реализованным в системе речевого диалога (СРД) «Речь-1» [128].

ВЫВОДЫ

1. Разработан метод глубокого пофонемного распознавания слов и слитной речи, составляемой из слов выбранного словаря. Его отличие от метода пофонемного распознавания, описанного в гл. 4 и 5, состоит в том, что акустическая, темпоральная, громкостная и тональная транскрипция слов теперь выбираются в соответствии с их фонетическими транскрипциями. Сам же процесс распознавания не меняется.

С переходом к глубокому пофонемному распознаванию обучение распознаванию речи становится разовым. Для осуществления полной или частичной замены словаря, пополнения словаря новыми словами достаточно ввести только тексты слов, автоматически получить их фонетические транскрипции и затем составить **Q**-транскрипции и темпоральные транскрипции слов, не прибегая к обучению или дообучению.

Кроме удобств в использовании, глубокое пофонемное распознавание обеспечивает повышение надежности распознавания речи.

2. Разработан метод распознавания произвольных последовательностей фонем в потоке слитной речи. Он основан на использовании общего фонемного графа, позволяющего генерировать все возможные эталонные сигналы слитной речи для произвольных последовательностей фонем. Эти сигналы являются коартикулированными и отличаются нелинейно изменяющимся темпом произнесения. Общий фонемный граф, таким образом, учитывает инерционные свойства речеобразующего тракта, фонемный состав языка, статистику речевых сигналов и другую априорную информацию об акустике и фонетике речи. Распознавание речевого сигнала заключается в нахождении для него наиболее правдоподобного эталонного сигнала слитной речи и указании последовательности фонем, эталонные сигналы которых составляют этот наиболее правдоподобный эталонный сигнал. Распознавание эффективно реализуется с помощью специальной схемы одноступенчатого динамического программирования, где используются понятия о потенциально-оптимальной фонеме и потенциально-оптимальном индексе.

Обучение распознаванию произвольных последовательностей фонем заключается в оценивании параметров общего фонемного графа на основании обучающей выборки. Реализуется с помощью метода обобщенной покоординатной оптимизации.

Распознавание произвольных последовательностей фонем в потоке слитной речи является вспомогательной задачей в общей проблематике распознавания и смысловой интерпретации речевых сигналов.

ГЛАВА 7

СМЫСЛОВАЯ ИНТЕРПРЕТАЦИЯ СЛИТНОЙ РЕЧИ

В настоящей главе показано, как в рамках КДП-подхода могут решаться задачи смысловой интерпретации слитной речи. Эти задачи будут рассмотрены в связи с устным диалогом человека и ЭВМ на формализованных или усеченных естественных языках, ориентированных на ту или иную предметную область или тему.

Проблема смысловой интерпретации слитной речи имеет наибольшую значимость, так как с ее решением достигается наибольшая эффективность речевого ввода информации.

При КДП-подходе задачи смысловой интерпретации слитной речи рассматриваются как дальнейшее обобщение задач распознавания слитной речи, когда дополнительно учитываются априорные сведения, выраженные синтаксисом, семантикой и прагматикой языков устного диалога. Однако, как и раньше, смысловая интерпретация предъявленного речевого сигнала будет заключаться в поиске для него наиболее правдоподобного эталонного сигнала слитной речи среди множества всех сигналов, генерируемых порождающей грамматикой в условиях дополнительных априорных ограничений, и в лексико-семантико-синтаксическом разборе последнего.

Работы по смысловой интерпретации слитной речи были начаты в 1975 г. [18, 104, 129—131]. С 1981 г. в этих работах принял активное участие К. М. Биатов.

§ 7.1. ВЗАИМОСВЯЗЬ ЗАДАЧ РАСПОЗНАВАНИЯ И СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ РЕЧИ. ГЕНЕРАТИВНАЯ МОДЕЛЬ СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ

Смысловая интерпретация слитной речи заключается в записи смысла, передаваемого речевым сигналом, в определенной канонической форме, удобной для непосредственного использования с целью вызова в автоматической системе действий, угодных говорящему человеку. Таким образом, при смысловой интерпретации речи разные предложения, передающие один и тот же смысл, должны отображаться в один и тот же результат, т. е. ответ распознавания (последовательность слов) не должен противоречить синтаксису, семантике и прагматике предметной области.

Задача смысловой интерпретации слитной речи существенно сложнее задачи распознавания, поскольку для ее решения необходимо дополнительно учитывать априорную информацию, выраженную синтаксисом, семантикой и прагматикой предметной области. Нельзя рассчитывать на успех, если задачу смысловой интерпретации решать по упрощенной схеме в следующие два этапа: сначала распознать слитную речь, используя только лексику и предположение о свободном порядке слов, а затем с помощью синтаксиса, семантики и прагматики отфильтровать возможные ошибки в распознавании слов и по скорректированной последовательности выработать каноническую форму высказывания. К сожалению, ни теоретически, ни из опыта не следует возможность надежного распознавания слитной речи вне учета синтаксиса, семантики и прагматики языка. Представляется, что задачи распознавания и смысловой интерпретации слитной речи должны решаться в едином взаимосвязанном процессе, в котором собственно распознавание управляет со стороны семантико-синтаксического уровня так, что достигается самая высокая надежность как распознавания, так и смысловой интерпретации.

Этот подход к распознаванию и смысловой интерпретации слитной речи может быть реализован в генеративной модели понимания (смысловой интерпретации) слитной речи [129].

Пусть необходимо указать, какое смысловое высказывание из заданного конечного множества смысловых высказываний содержится в предъявленном речевом сигнале.

В дальнейшем задачу распознавания смыслового высказывания из заданного множества смысловых высказываний будем считать основной при смысловой интерпретации речи. Развивающийся диалог будем рассматривать как постепенную (подчиненную определенным правилам) смену возможных множеств смысловых высказываний.

Каждое смысловое высказывание зададим в канонической форме, записанной на некотором семантическом языке (формальном математическом языке, выражающем понятия и отношения между ними). Далее введем преобразования канонической формы, не разрушающие смысл высказывания. Эти преобразования зададим с помощью генератора семантически эквивалентных предложений (ГСЭП), на вход которого поступает каноническая форма. Таким образом, ГСЭП порождает все возможные тексты с одинаковым смыслом, определяемым канонической формой. Этот генератор осуществляет переход от смысла к тексту и учитывает лексику и грамматику конкретного языка (русского, английского).

Далее введем преобразования, порождающие все возможные эталонные сигналы слитной речи для каждого предложения, сгенерированного ГСЭП, например, посредством фонетического транскрибирования предложения и использования ОФГ (см. гл. 6). Эти эталонные сигналы являются коартикулированными и отличаются друг от друга нелинейно изменяющимися темпом и интенсивностью произнесения. Модель синтеза эталонных сигналов для автоматического понимания речи приведена на рис. 7.1.

Автоматическое понимание (смысловая интерпретация) предъяв-

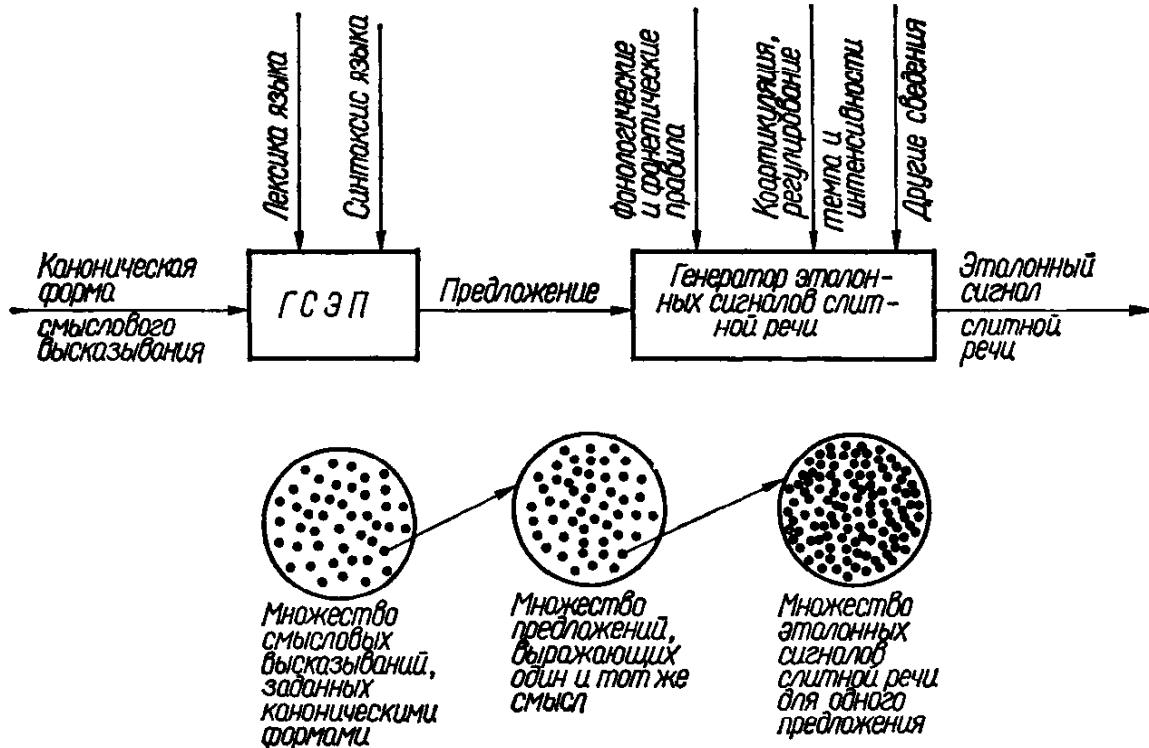


Рис. 7.1. Модель синтеза эталонных сигналов слитной речи для смысловой интерпретации.

ленного речевого сигнала с помощью генеративной модели (см. рис. 7.1) будет заключаться в том, чтобы сначала для анализируемого сигнала найти наиболее правдоподобный эталонный сигнал речи среди всех сигналов, порождаемых генеративной моделью, а затем определить каноническую форму того смыслового высказывания, предложение которого соответствует наиболее правдоподобному эталонному сигналу.

Последующее изложение будет посвящено конкретной реализации генеративной модели смысловой интерпретации слитной речи в рамках КДП-подхода. Будут указаны конструктивные приемы задания всех возможных предложений языка, выражающих один и тот же смысл, генерации и поиска наиболее правдоподобных эталонных сигналов слитной речи, семантико-синтаксического разбора предложений.

Имея в виду устный диалог человека и ЭВМ на формализованных или усеченных естественных языках предметных областей, в рамках КПД-подхода предложим некоторый универсальный конструктивный прием решения задач смысловой интерпретации слитной речи [104, 129—131].

Он основан на разработке средств экономного задания всех возможных предложений языка диалога, передающих один и тот же смысл. В соответствии с требованиями КДП-подхода эти средства должны позволять осуществлять как направленный поиск эталонных сигналов слитной речи, так и формирование результатов смысловой интерпретации.

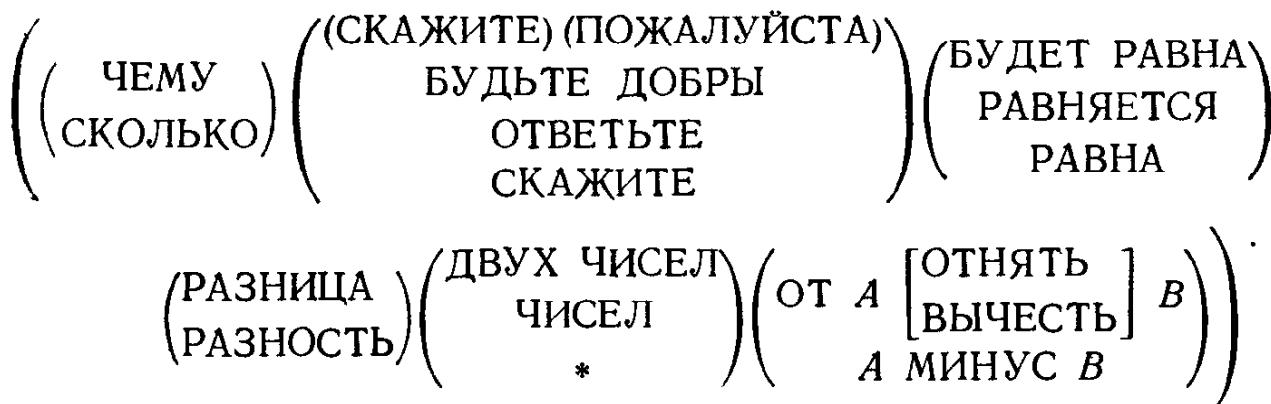
§ 7.2. ЗАДАНИЕ ИНФОРМАЦИИ О ЯЗЫКАХ УСТНОГО ДИАЛОГА

Будем задавать все возможные предложения языка устного диалога с помощью так называемой ориентированной семантической сети, опуская термины синтаксис и прагматика, поскольку они гораздо слабее влияют на порядок слов, чем семантика. Для построения ориентированной семантической сети (ОСС) будем использовать понятия «тип смысла» и «тип предложения».

Все мыслимые предложения языка диалога разобьем по передаваемому смыслу на типы смыслов. Например, применительно к информационно-справочной службе аэропорта типы смыслов выражаются вопросами о прилете и отлете самолетов, наличии свободных мест в самолете, маршруте, расположении служб аэропорта и т. п. Каждой предметной области соответствует не так уж много типов смысла.

Каждому типу смысла соответствует конечное множество типов предложений. Тип предложения — это конструкция, экономно задающая множество предложений, получающихся из одного предложения независимыми допустимой заменой и допустимой перестановкой отдельных слов и словосочетаний. Основным элементом типа предложений является подсловарь. Подсловари именуются по их отношению к предметной области.

Пример типа предложения для вопроса о разности двух чисел (тип смысла — разность двух чисел):



В скобках () указаны переставляемые подсловари, а в [] — непереставляемые. Переставлять подсловари можно только внутри старших скобок (). Тип предложения, как правило, параметрический. В рассматриваемом примере параметрами являются числа-операнды A и B . Символ * означает пустое слово (отсутствие слова).

Нетрудно убедиться, что, даже если не считать возможные варианты значений operandов A и B , приведенный тип предложений задает всего $6! \cdot 2 \cdot 5 \cdot 3 \cdot 2 \cdot 3 \cdot 3 = 388\,800$ различных предложений, допустимых в языке диалога и выражающих один и тот же смысл о разности двух чисел. Среди этих предложений находятся, например, и такие:

ЧЕМУ СКАЖИТЕ РАВНЯЕТСЯ РАЗНОСТЬ ЧИСЕЛ A МИНУС B ,

ЧЕМУ РАВНА РАЗНОСТЬ A МИНУС B СКАЖИТЕ ПОЖАЛУЙСТА,

ЧЕМУ ОТ A ОТНЯТЬ B БУДЕТ РАВНА РАЗНИЦА ДВУХ ЧИСЕЛ
ОТВЕТЬТЕ,

ОТ A ВЫЧЕСТЬ B ЧЕМУ РАВНА РАЗНОСТЬ СКАЖИТЕ,

ОТ A ОТНЯТЬ B ЧЕМУ РАВНА РАЗНОСТЬ СКАЖИТЕ,

СКОЛЬКО ОТВЕТЬТЕ БУДЕТ РАВНА A МИНУС B РАЗНОСТЬ ДВУХ ЧИСЕЛ.

Эти примеры иллюстрируют, что в данный тип предложений включены многие синтаксически допустимые предложения разговорной речи.

Пример другого типа предложения для семантического вопроса о разности двух чисел:

$\alpha \qquad \beta \qquad \gamma$
 $\left(((\text{УМЕНЬШАЕМОЕ } A) (\text{ВЫЧИТАЕМОЕ } B)) \right) \left(\begin{array}{c} \text{ЧЕМУ} \\ \text{СКОЛЬКО} \end{array} \right)$
 $\delta \qquad \varepsilon \qquad \xi$
 $\left(\begin{array}{c} \text{БУДЕТ РАВНА} \\ \text{РАВНЕТСЯ} \\ \text{РАВНА} \end{array} \right) \left(\begin{array}{c} \text{РАЗНОСТЬ} \\ \text{РАЗНИЦА} \end{array} \right) \left(\begin{array}{c} \text{ДВУХ ЧИСЕЛ} \\ \text{ЧИСЕЛ} \\ * \end{array} \right) \right).$

Для этого примера выполним именование подсловарей: α — уменьшаемое (операнд 1), β — вычитаемое (операнд 2), γ — вопросительное слово, δ — действие, ε — операция, ξ — объекты действий.

Каждому типу смысла соответствует не так уж много типов предложений.

Все типы предложений для данного типа смысла можно задавать, используя языки списочных структур, например язык ЛИСП. Очевидно, что данный тип смысла всегда можно пополнить новыми типами предложений, если в этом появится необходимость.

Таким образом, все предложения языка диалога можно задавать с помощью типов смысла и соответствующих им типов предложений.

Эквивалентный способ задания всех возможных предложений в языке диалога — ориентированная семантическая сеть (ОСС). Она имеет состояния y , среди них — одно начальное $y_{\text{нач}}$ и одно конечное $y_{\text{кон}}$ (рис. 7.2). Состояния соединены стрелками. Каждой стрелке, соединяющей состояния $y = \mu$ и $y = \nu$, приписан подсловарь $z_{\mu\nu}$ (на рисунке подсловари изображены прямоугольниками). Подсловари описаны (поименованы) по их отношению к предметной области с точки зрения семантики этой области. Слова в подсловарях отмечены жирными отрезками. Одно и то же слово может принадлежать разным подсловарям. Отдельные подсловари могут быть отмечены как базовые (ключевые). Пословари (точнее, прямоугольники) перенумерованы, номер подсловаря соответствует его месту в ОСС. Условимся, что, двигаясь по стрелке $\mu\nu$, непосредственно соединяющей состояния μ и ν , будем выбирать одно из слов $k \in z_{\mu\nu}$. Будем строить ОСС так, чтобы при движении из $y_{\text{нач}}$ в $y_{\text{кон}}$ образовывались только допустимые в языке

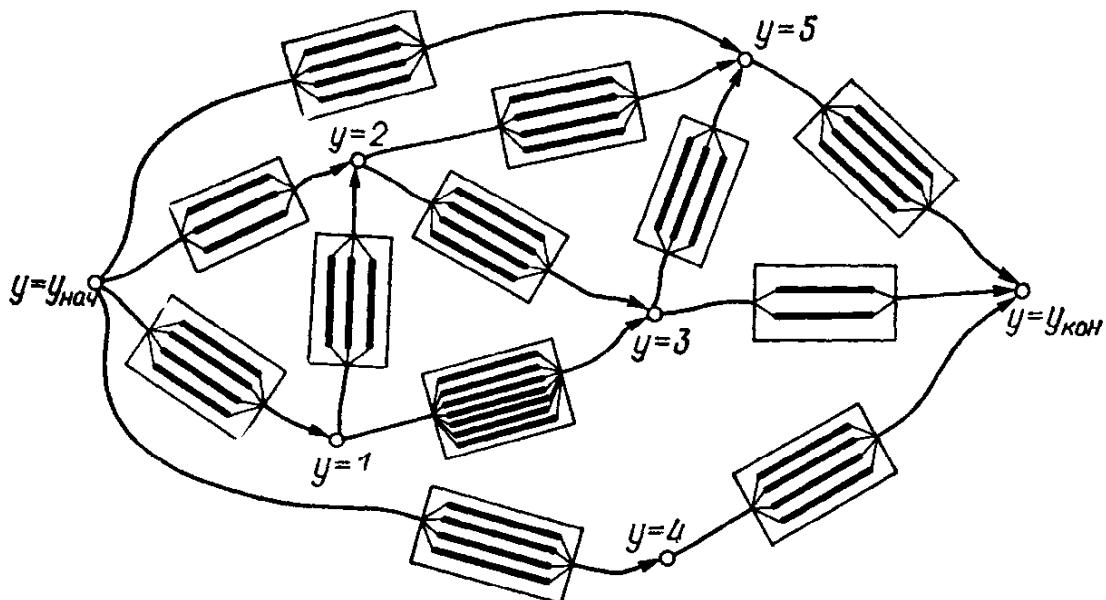


Рис. 7.2. Структура ориентированной семантической сети.

диалога предложения, т. е. удовлетворяющие синтаксису, семантике и прагматике предметной области.

ОСС удобно строить, отправляясь от типов смысла и типов предложений. Каждому типу предложения может быть сопоставлена своя частная ОСС. На рис. 7.3 представлена такая сеть, соответствующая второму примеру типа предложений (использованы имена подсловарей).

Объединив сети ОСС отдельных типов предложений, соответствующих одному смыслу, получим ОСС для данного типа смысла. При таком объединении частных ОСС будем совмещать лишь начальные и конечные состояния соответственно. Аналогично, объединив сети ОСС отдельных типов смысла, получим сеть ОСС для данной предметной области.

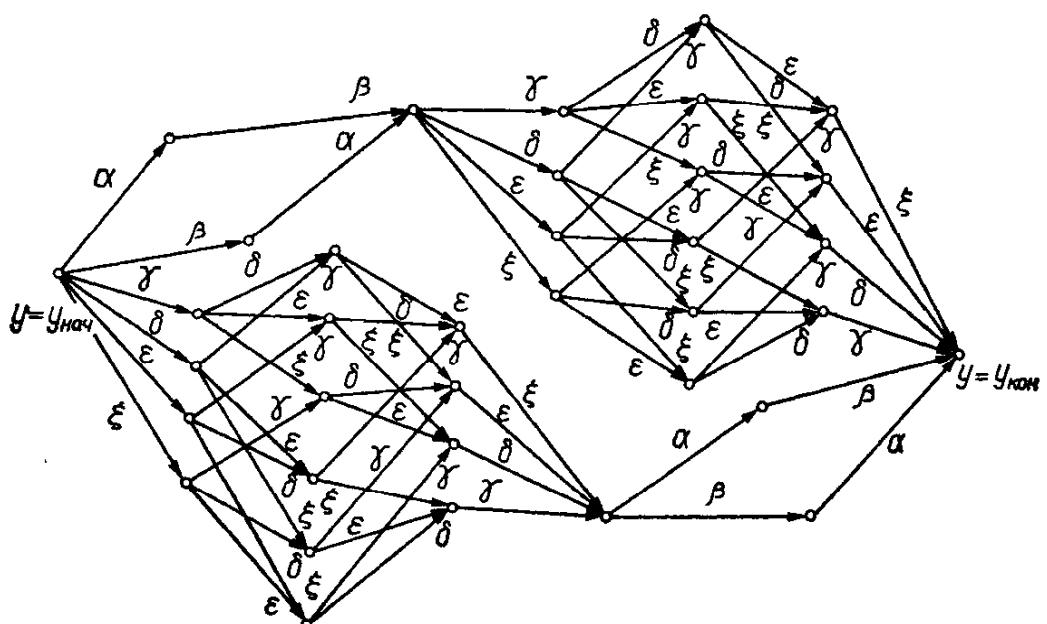


Рис. 7.3. Частная ориентированная семантическая сеть.

ОСС желательно строить так, чтобы в ней было как можно меньше состояний.

Легко убедиться, что если тип предложения экономно указывает лишь способ образования различных предложений, передающих один и тот же смысл, то ОСС для данного типа предложения экономно перечисляет (задает) все эти предложения. ОСС более громоздка, однако и более удобна при смысловой интерпретации слитной речи.

§ 7.3. АЛЬТЕРНАТИВНЫЕ ПУТИ РЕШЕНИЯ ЗАДАЧИ СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ

После того как мы научились экономно задавать все возможные допустимые предложения языка диалога с помощью ОСС или типов смысла и типов предложений, возникает проблема использования этих структур при смысловой интерпретации слитной речи.

В рамках КДП-подхода и генеративной модели смысловой интерпретации слитной речи могут быть указаны три основных пути использования введенных структур.

Первый путь состоит в акустической детализации ОСС [104, 130]. Заменяется каждое слово $k \in z_{\mu v}$ в подсловаре $z_{\mu v}$ его графом (см. рис. 2.2, a) так, чтобы начальное состояние слова совпадало с состоянием $y = \mu$, а конечное с состоянием $y = v$. В результате такого замещения сеть ОСС перейдет в граф СРОСС для генерирования эталонных сигналов слитной речи (СР) согласно ОСС. Фрагмент графа СРОСС для ОСС рис. 7.2 приведен на рис. 7.4.

Нетрудно убедиться, что, отправляясь из $y = y_{\text{нач}}$ и двигаясь по сети СРОСС в $y = y_{\text{кон}}$ за l тактов времени, будут генерироваться коартикулированные эталонные сигналы слитной речи, отличающиеся

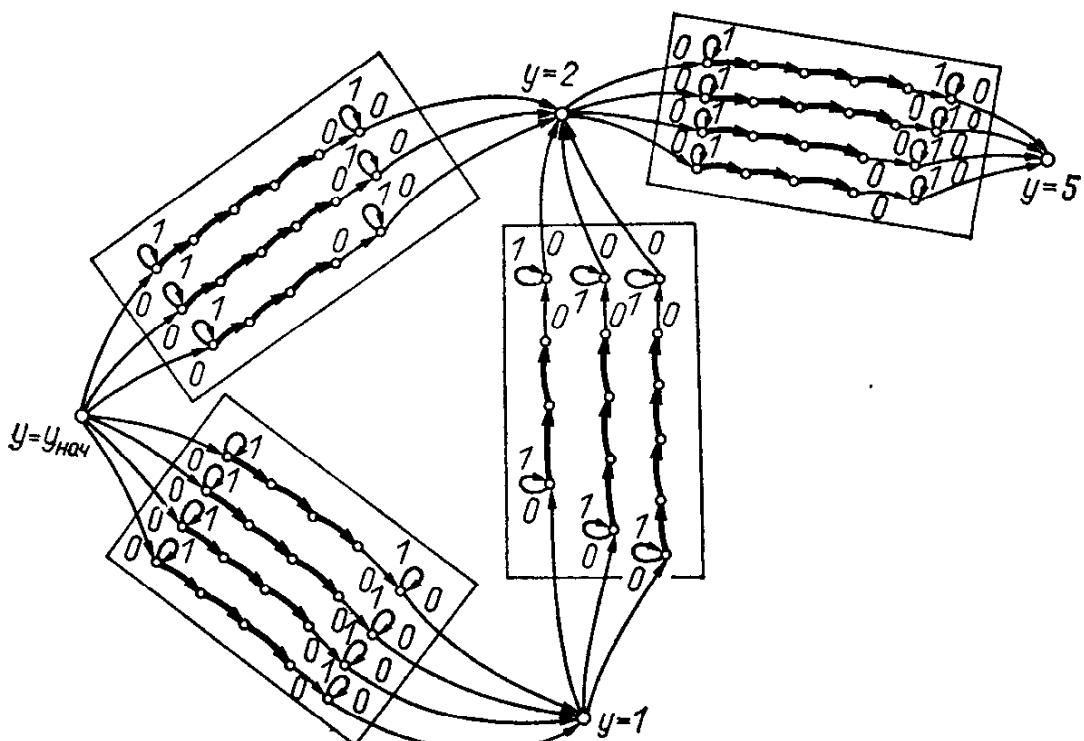


Рис. 7.4. Фрагмент графа СРОСС для ОСС рис. 7.2.

темпом и интенсивностью произнесения, длиной пауз между словами, однако эти эталонные сигналы будут соответствовать только предложениям, допустимым с точки зрения синтаксиса, семантики и прагматики предметной области.

Теперь задача распознавания и смысловой интерпретации слитной речи может быть сформулирована как задача нахождения для распознаваемого сигнала $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$, наиболее правдоподобного эталонного сигнала слитной речи длины l среди множества всех эталонных сигналов слитной речи, которые генерируются сетью СРОСС, и как задача анализа (разбора) последнего, в результате которого указывается последовательность слов и каноническая форма переданного сигналом X_l смысла.

Сформулированная задача решается с помощью одноступенчатого динамического программирования, рекуррентные формулы которого записываются на основании графа СРОСС. Алгоритм решения этой задачи назван алгоритмом речевой ориентированной семантической сети и излагается в следующем параграфе [130].

Второй путь использования сети ОСС или типов смысла и типов предложений заключается в разбиении решения задачи смысловой интерпретации слитной речи на два этапа. На первом этапе решается так называемая обобщенная задача распознавания слитной речи [104, 131, 132], которая состоит в том, что исходя из предположения о свободном порядке следования слов для распознаваемого сигнала $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$ находят $N > 1$ наиболее вероятных разных последовательностей слов, которые ранжируют по убыванию вероятности (правдоподобия или сходства). Затем, на втором этапе, последовательно подряд анализируют эти N последовательностей слов до тех пор, пока очередная просматриваемая последовательность слов не окажется принадлежащей одному из типов предложений и соответственно одному из типов смысла. Эта последовательность объявляется результатом распознавания, а каноническая форма передаваемого смысла составляется с помощью ОСС или с помощью типов смысла и типов предложений.

Таким образом, второй путь основан на использовании многозначных решений для распознавания слитной речи. Метод решения обобщенной задачи распознавания слитной речи, являющийся дальнейшим расширением обычного метода распознавания слитной речи (гл. 5), будет излагаться в § 7.6.

Третий путь является промежуточным. Он устраняет недостатки и объединяет достоинства двух предыдущих способов.

Этот путь по существу состоит в том, что в процессе распознавания слитной речи используются семантико-синтаксические ограничения. Этот путь заключается в решении обобщенной задачи распознавания слитной речи в условиях, когда рассматриваемые $N > 1$ последовательностей слов (соответственно $N > 1$ подпоследовательностей слов) предполагаются допустимыми в языке диалога, т. е. удовлетворяющими синтаксису, семантике и прагматике языка диалога. Теперь для текущего момента времени i будем находить только $N > 1$ допустимых в языке диалога начальных последовательностей слов, а в последующие

моменты времени рассматривать только такие слова, которые могут составить допустимые в языке диалога продолжения последовательностей слов, накопленных в предшествующие моменты времени. Таким образом, следуя третьему пути, в любой момент времени имеем дело с текущей многозначной смысловой интерпретацией слитной речи.

Алгоритм многозначной смысловой интерпретации слитной речи, в отличие от алгоритма речевой ориентированной сети, не гарантирует оптимальное решение. Однако использование многозначных решений и текущий отбор только допустимых в диалоге последовательностей слов, задаваемых ОСС или типами смысла и типами предложений, позволяют надеяться на то, что при надлежащем выборе $N > 1$ оптимальное решение не будет теряться и что таким образом будет получена достаточно высокая надежность распознавания и смысловой интерпретации слитной речи с гораздо меньшими затратами вычислений и памяти, чем в случае графа СРОСС.

В отличие от двухэтапного алгоритма, в котором требуется задаваться достаточно большим числом N , чтобы гарантировать выбор хотя бы одного допустимого в языке диалога предложения, в алгоритме многозначной смысловой интерпретации все отбираемые N последовательностей слов являются допустимыми. Поэтому нет необходимости задаваться большими $N > 1$, а достаточно ограничиться существенно меньшими N , чем в случае двухэтапного алгоритма. Значит, использование алгоритма многозначной смысловой интерпретации заметно снижает затраты на память и вычисления также и в сравнении с двухэтапным алгоритмом.

Далее рассмотрим подробнее все три алгоритма смысловой интерпретации слитной речи.

§ 7.4. АЛГОРИТМ СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ, ОСНОВАННЫЙ НА РЕЧЕВОЙ ОРИЕНТИРОВАННОЙ СЕМАНТИЧЕСКОЙ СЕТИ

Данный алгоритм использует рекуррентные формулы ДП, подобные приведенным в гл. 5. Для записи этих формул следует воспользоваться графиком СРОСС и ранее введенными понятиями о потенциально-оптимальном индексе и потенциально-оптимальном слове.

Записываемые формулы составят трехступенчатое динамическое программирование, в котором каждый из трех процессов оптимизации (по сходству отрезков речевого сигнала на слова, по границам между словами и по допустимым последовательностям слов), реализуясь с помощью ДП, вкладывается друг в друга так, что образуется некоторый единый одноступенчатый процесс динамического программирования [130].

Обозначим тройкой (z, k, s) состояние сети СРОСС для k -го слова из подсловаря z , z — порядковый номер подсловаря (прямоугольника), точнее, номер или место на сети ОСС, s — порядковый номер состояния графа слова: $s = 1 : q_k$. Состояния (z, k, s) будем отличать от игрековых состояний u сети СРОСС, включая $u_{\text{ нач}}$ и $u_{\text{ кон}}$.

Через $\Omega_i(z, k, s)$ и $\Omega_i(y)$ обозначим множества траекторий на графе СРОСС, начинающихся из состояния $y_{\text{нач}}$ и заканчивающихся в момент времени i в состоянии (z, k, s) или y соответственно. Очевидно, что $\Omega_i(y_{\text{кон}})$ задает множество всех возможных эталонных сигналов слитной речи длины l , а $\Omega_i(z, k, s)$ и $\Omega_i(y)$ — начальные части длины i этих эталонных сигналов из множества $\Omega_l(y_{\text{кон}})$.

Пусть $X_i = (x_1, x_2, \dots, x_i, \dots, x_l)$ — распознаваемый сигнал слитной речи, а $X_i = (x_1, x_2, \dots, x_i)$ — его начальная часть длины $i < l$.

Обозначим через $F_i(z, k, s)$ и $F_i(y)$ наибольшие сходства сигнала X_i на множествах эталонных сигналов $\Omega_i(z, k, s)$ и $\Omega_i(y)$ соответственно.

Введем потенциально-оптимальные индексы $v_i(z, k, s)$, указывающие момент начала последнего слова k в наиболее похожем на X_i эталонном сигнале из $\Omega_i(z, k, s)$, а также потенциально-оптимальные индексы $v_i(y)$, указывающие момент начала последнего слова $k_i(y)$ в наиболее правдоподобном для X_i эталонном сигнале из $\Omega_l(y)$. Это последнее слово $k_i(y)$ будем называть потенциально-оптимальным словом для состояния y в момент i .

Отметим, что в графе СРОСС отдельное слово-пауза не выделено и граф каждого слова в СРОСС содержит паузные элементы и в начале, и в конце слова.

Еще обозначим через $y_i(y)$ потенциально-оптимальное игрековое состояние, которое непосредственно предшествует состоянию y и из которого начинается слово $k_i(y)$ в наиболее правдоподобном для X_i эталонном сигнале из $\Omega_i(y)$.

Рассмотрим массив троек $w_i(y) = (y_i(y), v_i(y), k_i(y))$ для всех y и всех $i = 1 : l$; очевидно, что слово $k_i(y)$ начинается в момент времени $v_i(y) < i$ и в момент $v_i(y)$ выходит из игрекового состояния $y_i(y)$.

Покажем сначала, что массив троек $w_i(y)$ для всех y и всех $i = 1 : l$ определяет последовательность слов в распознаваемом сигнале X_i и соответствующую ей каноническую форму передаваемого сигналом X_i смысла.

В самом деле, пусть в момент i оптимальная траектория на графике СРОСС проходит через состояние y . Тогда оптимальным предшествующим игрековым состоянием будет $y^* = y_i(y)$; оптимальным словом, заканчивающимся в момент i , будет $k_i(y)$, причем оно начинается в момент $i^* = v_i(y)$. В таком случае $k_{i^*}(y^*)$ определит оптимальное слово, предшествующее слову $k_i(y)$, $v_{i^*}(y^*)$ — его начало на оси времени, а $y_{i^*}(y^*)$ — ближайшее предшествующее игрековое состояние. Полагая $i^{**} = v_{i^*}(y^*)$ и $y^{**} = y_{i^*}(y^*)$, далее выписываем оптимальное предшествующее слово $k_{i^{**}}(y^{**})$, его начало $v_{i^{**}}(y^{**})$, а также игрековое состояние $y^{***} = y_{i^{**}}(y^{**})$ и т. д. Действуя так, дойдем до момента времени i со звездочками, равного нулю. Иначе говоря, оптимальная и допустимая в предметной области последовательность слов для сигнала X_i может быть указана просто, если известен массив $w_i(y)$ для всех y и всех $i = 1 : l$ и если будет известно заключительное игрековое состояние для момента $i = l$. Но по определению в момент $i = l$ процесс на графике СРОСС должен достичь состояния $y = y_{\text{кон}}$, т. е. выписывание ответа распознавания начинается с $y = y_{\text{кон}}$.

Таким образом, алгоритм формирования ответа распознавания по массиву $w_i(y)$ для всех y и всех $i = 1 : l$ записывается так.

Полагая $y_0^* = y_{\text{кон}}$ и $v_0^* = l$, последовательно выписываем тройки

$$y_{\mu+1}^* = y_{v_{\mu}^*}(y_{\mu}^*), \quad (7.4.1)$$

$$v_{\mu+1}^* = v_{v_{\mu}^*}(y_{\mu}^*), \quad (7.4.2)$$

$$k_{\mu+1}^* = k_{v_{\mu}^*}(y_{\mu}^*) \quad (7.4.3)$$

для $\mu = 0, 1, 2, \dots$ и до тех пор, пока не достигнем $v_{\mu+1}^* = 0$.

Ответ распознавания составят слова k_{μ}^* , $\mu = 1, 2, 3, \dots$, расположенные в распознаваемом сигнале в обратном порядке. Границами слов будут v_{μ}^* , $\mu = 1, 2, 3, \dots$

Поскольку найденная последовательность слов k_{μ}^* , $\mu = 1, 2, 3, \dots$, удовлетворяет синтаксису, семантике и прагматике предметной области, то не представляет особого труда, используя ОСС или типы смыслов и типы предложений, сформировать каноническую форму переданного сигналом X_i смысла. Так, последовательность y_{μ}^* , $\mu = 1, 2, 3, \dots$, выписанная согласно (7.4.1) — (7.4.3), определяет подсловари z_{μ}^* , $\mu = 1, 2, 3, \dots$, которым принадлежат слова k_{μ}^* , $\mu = 1, 2, 3, \dots$, соответственно. Имена (описания) этих подсловарей позволяют выписать объекты-слова и найти отношения между ними, выписать ключевые слова и т. п. Совсем просто определяется передаваемый сигналом X_i смысл с помощью типов смысла и типов предложений: достаточно указать, в какой тип смысла входит тип предложений, которому принадлежит последовательность слов k_{μ}^* , $\mu = 1, 2, 3, \dots$

Теперь остается указать способ вычисления массива $w_i(y)$ для всех y и всех $i = 1 : l$.

Для этого понадобится по мере поступления распознаваемого сигнала $X_i = (x_1, x_2, \dots, x_i, \dots, x_l)$ наряду с $w_i(y)$ для всех y и каждого текущего i вычислять $F_i(z, k, s)$ и $v_i(z, k, s)$ для всех подсловарей z , всех слов $k \in z$ и всех состояний s слов $k \in z$.

Итак, с каждым вновь поступившим распознаваемым элементом x_i значения $F_i(y)$, $F_i(z, k, s)$, $y_i(y)$, $k_i(y)$, $v_i(y)$, $v_i(z, k, s)$ будем вычислять по значениям этих же функций в предыдущие моменты времени в соответствии с рекуррентными формулами ДП, приведенными ниже и учитывающими специфику решаемой задачи. Последовательно для $i = 1 : l$ с каждым текущим предъявленным элементом x_i за время ΔT до прихода следующего элемента x_{i+1} выполняются подряд пункты 1, 2 и 3:

1) одновременно (параллельно) для всех подсловарей z , всех слов $k \in z$ и всех состояний $s = 2 : q_k$ всех слов $k \in z$:

для $s = 2 : (q_k - 1)$

$$F_i(z, k, s) = \max_{m_{ks} \leq u \leq M_{ks}} (F_{i-u}(z, k, s-1) + G_i(z, k, s, u)), \quad (7.4.4)$$

где

$$G_t(z, k, s, u) = \sum_{v=i-u+1}^t g(\mathbf{x}_v, \mathbf{e}(j_{ks})); \quad (7.4.5)$$

$$u_t(z, k, s) = \underset{m_{ks} \leq u \leq M_{ks}}{\operatorname{argmax}} (F_{t-u}(z, k, s-1) + G_t(z, k, s, u)), \quad (7.4.6)$$

$$v_t(z, k, s) = v_{t-u_t(z, k, s)}(z, k, s-1); \quad (7.4.7)$$

для $s = q_k$

$$F_t(z, k, q_k) = \max (F_{t-1}(z, k, q_k) + g(\mathbf{x}_t, \mathbf{e}(j_{kq_k})), F_t(z, k, q_k - 1)), \quad (7.4.8)$$

$$u_t(z, k, q_k) =$$

$$= \begin{cases} 0, & \text{если } F_t(z, k, q_k - 1) > F_{t-1}(z, k, q_k) + g(\mathbf{x}_t, \mathbf{e}(j_{kq_k})); \\ 1 & \text{в противном случае,} \end{cases} \quad (7.4.9)$$

$$v_t(z, k, q_k) = v_{t-u_t(z, k, q_k)}(z, k, q_k - 1 + u_t(z, k, q_k)); \quad (7.4.10)$$

2) одновременно (параллельно) для всех y , кроме $y = y_{\text{нач}}$:

$$F_t(y) = \max_{z \in Z(y)} \max_{k \in z} F_t(z, k, q_k), \quad (7.4.11)$$

где $Z(y)$ — множество подсловарей z , непосредственно входящих в состояние y ;

$$k_t(y) = \underset{k \in z}{\operatorname{argmax}} \max_{z \in Z(y)} F_t(z, k, q_k), \quad (7.4.12)$$

$$z_t(y) = \underset{z \in Z(y)}{\operatorname{argmax}} \max_{k \in z} F_t(z, k, q_k), \quad (7.4.13)$$

$$v_t(y) = v_t(z_t(y), k_t(y), q_{k_t(y)}), \quad (7.4.14)$$

$$y_t(y) = \mathcal{L}(z_t(y)), \quad (7.4.15)$$

где $\mathcal{L}(z(y))$ — функция, указывающая, из какого игрекового состояния выходит подсловарь z , чтобы затем непосредственно входить в состояние y ;

3) одновременно (параллельно) для всех z , всех $k \in z$ и всех $s = 1$:

$$F_t(z, k, 1) = \max (F_{t-1}(z, k, 1) + g(\mathbf{x}_t, \mathbf{e}(j_{k1})), F_t(\mathcal{L}(z))), \quad (7.4.16)$$

$$u_t(z, k, 1) = \begin{cases} 0, & \text{если } F_t(\mathcal{L}(z)) > F_{t-1}(z, k, 1) + g(\mathbf{x}_t, \mathbf{e}(j_{k1})); \\ 1 & \text{в противном случае,} \end{cases} \quad (7.4.17)$$

$$v_t(z, k, 1) = \begin{cases} i, & \text{если } u_t(z, k, 1) = 0; \\ v_{t-1}(z, k, 1), & \text{если } u_t(z, k, 1) = 1. \end{cases} \quad (7.4.18)$$

Как и раньше, для всех неопределенных справа величин F и G полагаем, что они равны $-\infty$, кроме $F_0(z, k, 1) = 0$ и $v_0(z, k, 1) = 0$ для всех z и $k \in z$, таких, что $\mathcal{L}(z) = y_{\text{нач}}$. Кроме того, первоначальные значения $F_t(y) = -\infty$ для всех y и всех $i = 1 : l$ выставляются каждый раз перед началом распознавания и смысловой интерпретации очередной реализации слитной речи.

Как видно из формул (7.4.4) — (7.4.18), специфические вычисления для первых и последних состояний всех слов $k \in z$ всех подсловарей

z обусловлены тем, что длины пауз в начале и конце слов могут быть любыми, в том числе и равными нулю. Особая же обработка игрековых состояний обусловлена тем, что вход в них и выход из них осуществляются за нуль тактов времени (переходы по 0-стрелкам).

Из приведенного алгоритма следует, что в процессе обработки речевого сигнала с целью распознавания и смысловой интерпретации запоминаем лишь массив $w_i(y)$ для всех y и всех $i = 1 : l$. Остальные же величины, а именно $F_i(z, k, s)$ и $v_i(z, k, s)$, кратковременно запоминаются не более чем на $\max_{k=1}^{s=2(q_k-1)} M_{ks}$ тактов времени.

Как и раньше, в элементарное сходство g в формулах (7.4.5), (7.4.8) — (7.4.9), (7.4.16) — (7.4.17) входят добавки сходств $g_1(h_i, h_{ks})$ и $g_2(f_i, f_{ks})$ за счет элементов громкостной и тональной транскрипций.

Из анализа рекуррентных формул (7.4.4) — (7.4.18) вытекает, что процесс распознавания и смысловой интерпретации слитной речи распараллеливается по подсловарям z , по словам $k \in z$ внутри подсловарей z и по состояниям s внутри слов $k \in z$. Среди типовых процессоров могут быть названы вычислитель элементарной меры сходства $g(x_i, e(j))$, вычислитель меры сходства $G_l(z, k, s, u)$ сегментов (отрезков) речи с эталонными элементами $e(j_{ks})$, вычислители первых, внутренних и последних состояний (z, k, s) слов, вычислитель игрековых состояний.

Алгоритм речевой ОСС можно также реализовать, перебирая частные СРОСС, соответствующие отдельным типам смысла или типам предложений, и формируя ответ распознавания и смысловой интерпретации по той частной СРОСС, которая обеспечила абсолютно наибольшее сходство эталонного сигнала с распознаваемым. Этот прием решения задачи уменьшает требования к объемам памяти, но увеличивает объем вычислений.

Описываемый метод распознавания и смысловой интерпретации слитной речи, основанный на сети ОСС и графе СРОСС, является достаточно универсальным. Как частный случай, из этого метода может быть получен алгоритм распознавания слитной речи, учитывающий автоматную грамматику следования слов. Здесь задаются начальным $V_{\text{нач}} \subset V$ и конечным $V_{\text{кон}} \subset V$ словарями и для каждого слова k указывается подсловарь $V(k) \subset K$ слов, которые могут следовать после слова k (V — полный словарь) [119].

Чтобы получить алгоритм решения этой задачи, необходимо на сети ОСС завести K подсловарей z (K — количество слов в словаре V) — по одному слову k , $k = 1 : K$, в каждом подсловаре $z = k$. Затем подсоединить стрелками подсловари $z = k \in V_{\text{нач}}$ к начальному игрековому состоянию $y = y_{\text{нач}}$, подсловари $z = k \in V_{\text{кон}}$ — к конечному $y = y_{\text{кон}}$. Далее игрековое состояние на выходе подсловаря-слова $z = k$ соединим со входами подсловарей-слов $k^* \in V(k)$. Поступив так со всеми подсловарями-словами, получим ОСС для случая автоматного синтаксиса языка. Эта сеть будет генерировать только допустимые автоматным синтаксисом предложения. Далее детализируем СРОСС с учетом того, что в каждом подсловаре $z = k$ только одно слово k .

Алгоритм распознавания слитной речи для этого случая полностью определяется рекуррентными формулами (7.4.1) — (7.4.18) с учетом того, что $z = k$ и $k = 1 : K$. В этом случае для хранения массива $w_i(y)$ для всех y и всех $i = 1 : l$ потребуется память на $3Kl$ чисел, вместо $3l$ чисел при простом распознавании слитной речи [119].

§ 7.5. ДВУХЭТАПНЫЙ АЛГОРИТМ СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ

Двухэтапный алгоритм смысловой интерпретации слитной речи (см. § 7.3) предполагает решение задачи в два этапа. На первом этапе решается обобщенная задача распознавания слитной речи, когда в предположении свободного порядка слов для распознаваемого сигнала находят $N > 1$ наиболее вероятных, ранжированных по убыванию правдоподобия или сходства, последовательностей слов. Затем, на втором этапе, анализируя подряд эти N последовательностей слов, с помощью ОСС или типов смысла и типов предложений находят такую одну последовательность слов, которая одновременно и наиболее вероятна и допустима в языке диалога. Если в N -ке последовательностей слов нет допустимой последовательности слов, то выдается отказ от распознавания и смысловой интерпретации. В противном случае, найденная наиболее вероятная допустимая последовательность слов объявляется ответом распознавания, а каноническая форма переданного смысла находится с помощью ОСС или типов смысла и типов предложений. Например, распознаваемым сигналом передается смысл, определяемый тем типом смысла, которому принадлежит тот тип предложений, в который попадает выделенная из N -ки допустимая последовательность слов.

Двухэтапный метод решения задачи распознавания и смысловой интерпретации слитной речи предпочтительнее метода речевой ориентированной семантической сети в смысле разделения работ: на первом этапе решается сугубо распознавальская задача, на втором — чисто лингвистическая. Недостаток метода состоит, в частности, в том, что нельзя априори указать, каким минимальным следует выбрать число N претендентов в ответ распознавания, чтобы среди них оказался хотя бы один допустимый в сети ОСС.

Для описания двухэтапного метода необходимо указать алгоритм решения обобщенной задачи распознавания.

§ 7.6. ОБОБЩЕННАЯ ЗАДАЧА РАСПОЗНАВАНИЯ СЛИТНОЙ РЕЧИ

Обобщенная задача является обобщением обычной задачи распознавания слитной речи, составляемой из слов выбранного словаря в условиях свободного порядка следования слов (см. гл. 5). Отличие в постановке задачи лишь в том, что надо указать не одну, наиболее вероятную последовательность слов, передаваемую распознаваемым сигналом $X_l = (x_1, x_2, \dots, x_l, \dots, x_l)$, а $N > 1$ наиболее вероятных и различных, ранжированных по убыванию правдоподобия или сходства последовательностей слов [104, 131, 132].

Алгоритм решения обобщенной задачи распознавания слитной речи [132] получаем в результате дальнейшего обобщения обычного метода распознавания слитной речи (гл. 5). Для формулировки алгоритма воспользуемся определениями, рисунками, графиком слитной речи и частью обозначений главы 5.

Пусть для всех моментов времени $v < i$, всех начальных частей $X_v = (x_1, x_2, \dots, x_v)$ распознаваемого сигнала X_i и всех состояний s графа слитной речи (рис. 5.1) указаны (вычислены) N -ки величин:

$$(F'_v(s), \mathcal{K}'_v(s)), \quad r = 1 : N, \quad (7.6.1)$$

где $\mathcal{K}'_v(s)$, $r = 1 : N$, — наиболее вероятные последовательности слов, которые отличаются друг от друга и соответствуют наиболее вероятным по отношению к X_v эталонным сигналам слитной речи из множества $\Omega_v(s)$; $F'_v(s)$ — соответствующая $\mathcal{K}'_v(s)$ наилучшая интегральная мера сходства (значение выражения правдоподобия).

Пусть в N -ке (7.6.1) пары величин $(F'_v(s), \mathcal{K}'_v(s))$ ранжированы по убыванию сходства (правдоподобия): $F'_v^1(s) \geq F'_v^2(s) \geq \dots \geq F'_v^r(s) \geq \dots \geq F'_v^N(s)$.

Далее с приходом в текущий момент i очередного распознаваемого элемента x_i одновременно (параллельно) для всех состояний s графа слитной речи рекуррентно будем вычислять новые N -ки $(F'_i(s), \mathcal{K}'_i(s))$, $r = 1 : N$. Вычисления будем выполнять за время ΔT с тем, чтобы успеть подготовиться к приему и обработке очередного, $(i + 1)$ -го, элемента x_{i+1} .

Для вычисления новых N -ок будем составлять для всех состояний s , кроме $s = 1$ и главного состояния $s = 0$, все возможные суммы $\Phi_u^\omega(s)$ и последовательности слов $\mathcal{L}_u^\omega(s)$:

$$\begin{aligned} \Phi_u^\omega(s) &= F_{i-u}^\omega(\mu) + G_i(u, s), \\ \mathcal{L}_u^\omega(s) &= \mathcal{K}_{i-u}^\omega(\mu), \\ u &= m(s) : M(s), \quad \omega = 1 : N, \end{aligned} \quad (7.6.2)$$

где

$$G_i(u, s) = \sum_{v=i-u+1}^i g(x_v, e(j(s))), \quad (7.6.3)$$

и полагается $\mu = s - 1$, если s не совпадает с первыми состояниями слов, или $\mu = 0$ — в противном случае. Для фиксированного s выберем такую N -ку сумм $\Phi_u^\omega(s)$, чтобы все $\Phi_u^\omega(s)$ были наибольшими, а соответствующие им $\mathcal{L}_u^\omega(s)$ — разными. Такие N пар $(\Phi_u^\omega(s), \mathcal{L}_u^\omega(s))$, будучи ранжированными по ω в порядке убывания величины $\Phi_u^\omega(s)$, и составят новую N -ку $(F'_i(s), \mathcal{K}'_i(s))$, $r = 1 : N$, для состояния s и момента времени i .

Что касается состояния $s = 1$ для слова-паузы $k = 0$, то для него из массива чисел

$$F_{i-1}^\omega(\mu), \quad \mu = 0, 1; \quad \omega = 1 : N, \quad (7.6.4)$$

ищем N -ку наибольших и таких, чтобы соответствующие им $\mathcal{K}_{i-1}^w (\mu)$ были все разными. Ранжируя полученную N -ку пар по убыванию сходства и прибавляя ко всем сходствам N -ки величину $g(x_i, e(1))$, получаем новую N -ку $(F'_i(1), \mathcal{K}'_i(1)), r = 1 : N$, для состояния $s = 1$ и момента времени i .

Наконец, для главного состояния $s = 0$ вычисления N -ок будем вести по особым формулам. Именно здесь происходит дописывание новых слов в N -ку. Будем рассматривать массив пар

$$(F'_i(s_k), \mathcal{K}'_i(s_k) \downarrow k), \quad w = 1 : N, \quad k = 0 : K, \quad (7.6.5)$$

где s_k — последнее состояние k -го слова на графе слитной речи; \downarrow — символ дописывания справа слова k к последовательности слов $\mathcal{K}'_i(s_k)$. Среди этих пар выберем N с наибольшими $F'_i(s_k)$, для которых все $(\mathcal{K}'_i(s_k) \downarrow k)$ разные. Ранжировав полученные пары по убыванию сходства, получим новую N -ку $(F'_i(0), \mathcal{K}'_i(0)), r = 1 : N$, для главного состояния $s = 0$ и момента времени i .

Очевидно, что N -ка $(F'_i(0), \mathcal{K}'_i(0)), r = 1 : N$, найденная по рекуррентным формулам для главного состояния $s = 0$ в заключительный момент времени $i = l$, определит обобщенный ответ распознавания.

Все неопределенные величины сходств в формулах (7.6.1) — (7.6.5) полагаются равными $-\infty$. Алгоритм решения обобщенной задачи распознавания слитной речи начинает свою работу с того, что полагается $(F'_0(0) = 0, \mathcal{K}'_0(0) = \emptyset), r = 1 : N$.

Напоминаем, что в элементарное сходство $g(x_i, e(j(s)))$ включаются добавки сходств громкости $g_1(h_i, h(s))$ и тональности $g_2(f_i, f(s))$.

Из рекуррентного алгоритма решения задачи обобщенного распознавания следует, что и при обобщенном распознавании слитной речи процесс обработки информации по-прежнему распараллеливается по словам и состояниям внутри слов.

Что же касается объемов вычислений и объемов памяти, необходимых для вычисления и хранения промежуточных N -ок, то в обобщенном алгоритме они возрастают в N раз по сравнению с обычным алгоритмом распознавания слитной речи (сравнение следует производить на основании формул (7.6.1) — (7.6.5), полагая для обычного распознавания $N = 1$). За эту «плату» существенно упрощается второй этап обработки информации с целью ее смысловой интерпретации.

Для практического использования представляют интерес и приближенные алгоритмы решения задачи обобщенного распознавания слитной речи с существенно меньшими, чем в случае точного алгоритма, объемами вычислений и памяти. Целый класс приближенных алгоритмов может быть получен на основе обычного алгоритма распознавания слитной речи (гл. 5). Применяя алгоритм распознавания слитной речи (формулы (5.2.1) — (5.2.11)), несколько расширим формулы (5.2.10) — (5.2.11), оставляя остальные неизменными. Вместо одного слова $k_i(0)$ (формула (5.2.10)), которое заканчивается в момент времени i и начинается в момент времени $v_i(0)$ (формула (5.2.11)), найдем на основании $F_i(s_k)$ (см. формулу (5.2.10)) M , $1 < M < K$, наиболее вероят-

ных слов, которые заканчиваются в момент i . Очевидно, что эти слова будут определяться M наибольшими значениями, выбираемыми из массива $F_i(s_k)$, $k = 0 : K$. Пусть $F_i^v(0)$, $v = 1 : M$, являются этими наибольшими значениями и пусть $k_i^v(0)$, $v = 1 : M$, — соответствующие слова, заканчивающиеся в момент i , а $v_i^v(0)$, $v = 1 : M$, — соответствующие моменты начала слов $k_i^v(0)$. Очевидно, что $F_i^1(0) = F_i(0)$, $k_i^1(0) = k_i(0)$, $v_i^1(0) = v_i(0)$.

Пусть тройки $(F_i^v(0), k_i^v(0), v_i^v(0))$, $v = 1 : M$, вычисляются для всех моментов времени $i = 1 : l$. Расширим несколько понятие начала слова. Полагаем, что слово $k_i^v(0)$ из M -ки начинается не только в момент $v_i^v(0)$, а и в соседние моменты времени j

$$v_i^v(0) - \Delta(k_i^v(0)) \leq j \leq v_i^v(0) + \Delta(k_i^v(0)). \quad (7.6.6)$$

Полагаем, что $\Delta(k_i^v(0)) = 0, 1, 2$ или 3 в зависимости от слова $k_i^v(0)$. Будем говорить, что сегмент (участок) речевого сигнала X_{ji} , где j удовлетворяет (7.6.6), принадлежит слову $k_i^v(0)$ с вероятностью, определяемой сходством $G_i^v(0) = F_i^v(0) - F_{v_i^v(0)}(0)$.

Одновременно с вычислениями M -ок $G_i^v(0)$, $k_i^v(0)$, $v_i^v(0)$, $v = 1 : M$, будем находить N -ки последовательностей слов. В приближенном алгоритме N -ки вычисляются не для всех состояний, а только для главного состояния $s = 0$ (именно за счет этого достигается экономия вычислений и памяти).

Пусть N -ки $(\tilde{F}_u^r(0), \tilde{\mathcal{K}}_u^r(0))$, $r = 1 : N$, уже вычислены для всех моментов $u < i$. Тогда N -ку $(\tilde{F}_i^r(0), \tilde{\mathcal{K}}_i^r(0))$, $r = 1 : N$, для момента i составляем на основании сумм

$$\begin{aligned} F_{ii}^{\mu v} &= \tilde{F}_i^{\mu}(0) + G_i^v(0), \quad \mu = 1 : N, \quad v = 1 : M, \\ v_i^v(0) - \Delta(k_i^v(0)) &\leq j \leq v_i^v(0) + \Delta(k_i^v(0)) \end{aligned} \quad (7.6.7)$$

и соответствующих им последовательностей

$$\mathcal{K}_{ii}^{\mu v} = \tilde{\mathcal{K}}_i^{\mu}(0) \cup k_i^v(0). \quad (7.6.8)$$

Далее ищем ранжированную по убыванию F совокупность из N пар $(\tilde{F}_i^r(0), \tilde{\mathcal{K}}_i^r(0))$, $r = 1 : N$, среди пар $(F_{ii}^{\mu v}, \mathcal{K}_{ii}^{\mu v})$, такую, что все $\tilde{F}_i^r(0)$ являются наибольшими, а соответствующие им $\tilde{\mathcal{K}}_i^r(0)$ — разными.

Как и раньше, N -ка $(\tilde{F}_i^r(0), \tilde{\mathcal{K}}_i^r(0))$, $r = 1 : N$, будет содержать обобщенный ответ распознавания.

В данном случае класс приближенных алгоритмов решения обобщенной задачи распознавания слитной речи определяется варьируемыми параметрами M и $\Delta(k)$, $k = 0 : K$.

Сравнивая точный и приближенный алгоритмы решения обобщенной задачи с обычным алгоритмом распознавания слитной речи, убеждаемся, что приближенный алгоритм обобщенного распознавания

слитной речи по трудоемкости и затратам памяти лишь несколько сложнее обычного алгоритма (это обусловлено дополнительными вычислениями по формулам (7.6.7) — (7.6.8) для всех $i = 1 : l$), зато существенно проще точного алгоритма обобщенной задачи.

§ 7.7. АЛГОРИТМ МНОГОЗНАЧНОЙ СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ

Как уже отмечалось в § 7.3, алгоритм многозначной смысловой интерпретации может быть получен на основе алгоритма решения задачи обобщенного распознавания слитной речи, если привлечь априорную информацию о допустимых в языке диалога последовательностях слов.

Пусть для всех моментов времени $v < i$, всех начальных частей $X_v = (x_1, x_2, \dots, x_v)$ распознаваемого сигнала $X_i = (x_1, x_2, \dots, x_i, \dots, x_l)$ и всех состояний s графа слитной речи (рис. 5.1) вычислены N -ки величин:

$$(F'_v(s), \mathcal{K}'_v(s), \mathcal{A}'_v(s)), \quad r = 1 : N, \quad (7.7.1)$$

где $\mathcal{K}'_v(s)$ — наиболее вероятные последовательности слов (точнее, начальные подпоследовательности слов), которые все разные и соответствуют наиболее похожим на X_v эталонным сигналам слитной речи, допустимым в языке диалога; $F'_v(s)$ — соответствующая $\mathcal{K}'_v(s)$ наилучшая интегральная мера сходства; $\mathcal{A}'_v(s)$ — список номеров (мест) подсловарей на сети ОСС, слова из которых могут составить допустимые в языке диалога продолжения начальных подпоследовательностей слов $\mathcal{K}'_v(s)$.

В N -ке (7.7.1) тройки величин ранжируем по убыванию сходства

$$F'_v(s) \geq F^2_v(s) \geq \dots \geq F^r_v(s) \geq \dots \geq F^N_v(s).$$

Далее с приходом в текущий момент i очередного распознаваемого элемента x_i одновременно (параллельно) для всех состояний s графа слитной речи рекуррентно будем вычислять новые N -ки $(F'_i(s), \mathcal{K}'_i(s), \mathcal{A}'_i(s))$, $r = 1 : N$.

С этой целью для всех s , кроме $s = 1$ и главного состояния $s = 0$, будем составлять все возможные суммы $\Phi_u^\omega(s)$, последовательности слов $\mathcal{L}_u^\omega(s)$ и списки $\mathcal{B}_u^\omega(s)$:

$$\left. \begin{array}{l} \Phi_u^\omega(s) = F_{i-u}^\omega(\mu) + G_i(u, s), \\ \mathcal{L}_u^\omega(s) = \mathcal{K}_{i-u}^\omega(\mu), \\ \mathcal{B}_u^\omega(s) = \mathcal{A}_{i-u}^\omega(\mu), \\ u = m(s) : M(s), \quad w = 1 : N, \end{array} \right\} \quad (7.7.2)$$

где

$$G_i(u, s) = \sum_{v=i-u+1}^i g(x_v, e(j(s))) \quad (7.7.3)$$

и полагается $\mu = s - 1$, если s не совпадает с первыми состояниями слов, и $\mu = 0$, если s совпадает с одним из первых состояний слов, причем в этом последнем случае рассматриваются первые состояния s только тех слов $k(s)$, которые принадлежат подсловарям $\mathcal{A}_{i-u}^w(0)$, т. е. $k(s) \in \mathcal{A}_{i-u}^w(0)$. Далее, для фиксированного s выбираем такую N -ку троек $(\Phi_u^w(s), \mathcal{L}_u^w(s), \mathcal{B}_u^w(s))$, чтобы все $\Phi_u^w(s)$ были наибольшими, а соответствующие им $\mathcal{L}_u^w(s)$ — разными. Такие N троек $(\Phi_u^w(s), \mathcal{L}_u^w(s), \mathcal{B}_u^w(s))$, будучи ранжированными по w в порядке убывания величины $\Phi_u^w(s)$, составят новую N -ку троек $(F_i^r(s), \mathcal{K}_i^r(s), \mathcal{A}_i^r(s))$, $r = 1 : N$, для состояния s и момента i .

Особого внимания заслуживают первые состояния слов на графе слитной речи. Как уже говорилось, в случае таких s полагаем в (7.7.2) $\mu = 0$ и рассматриваем соответствующие суммы $\Phi_u^w(s)$, последовательности $\mathcal{L}_u^w(s)$ и списки $\mathcal{B}_u^w(s)$ только для таких пар (u, w) , что s принадлежат словам $k(s) \in \mathcal{A}_{i-u}^w(0)$. И только среди таких допустимых пар (u, w) выбираем наилучшую N -ку для состояния s и момента i . Если в случае таких s окажется, что все соответствующие $\mathcal{A}_{i-u}^w(0)$ равны пустому множеству \emptyset , то принудительно полагаем все $F_i^r(s)$, $r = 1 : N$, равными — ∞ . Аналогично полагаем $F_i^r(s)$, $1 \leq m < r \leq N$, равными — ∞ и в том случае, если среди возможных пар (u, w) в (7.7.2) окажется только всего m , $1 < m < N$, различных $\mathcal{L}_u^w(s)$.

Для состояния паузы $s = 1$ составляем массивы

$$(F_{i-1}^w(\mu), \mathcal{K}_{i-1}^w(\mu), \mathcal{A}_{i-1}^w(\mu)), \quad \mu = 0, 1; \quad w = 1 : N, \quad (7.7.4)$$

и ищем N -ку троек с наибольшими $F_{i-1}^w(\mu)$ и таких, чтобы соответствующие им $\mathcal{K}_{i-1}^w(\mu)$ были все разными. Ранжировав полученную N -ку троек по убыванию сходства и прибавив ко всем сходствам N -ки величину $g(x_i, e(1))$, получим новую N -ку $(F_i^r(1), \mathcal{K}_i^r(1), \mathcal{A}_i^r(1))$, $r = 1 : N$, для состояния $s = 1$ и момента времени i . Таким образом, для $s = 1$ имеет место простая пересылка последовательностей слов и списков номеров (мест) подсловарей. Впрочем, такая пересылка выполнялась и в случае других состояний $s \neq 0$.

И только при обработке главного состояния $s = 0$ происходит дописывание новых слов к начальным подпоследовательностям слов и выполняется формирование новых списков номеров (мест) подсловарей. Рассматриваются все возможные тройки

$$(F_i^w(s_k), \mathcal{K}_i^w(s_k) \xrightarrow{k} k, \mathcal{A}_i^w(s_k)), \quad w = 1 : N, \quad k = 0 : K, \quad (7.7.5)$$

где s_k — последнее состояние слова k на графике слитной речи, \xrightarrow{k} — символ дописывания справа слова k к последовательности слов $\mathcal{K}_i^w(s_k)$, причем это дописывание будет всегда допустимым, поскольку, если только $F_i^w(s_k) \neq -\infty$, выполняется $k \in \mathcal{A}_i^w(s_k)$. Среди рассматриваемых троек выбираем N с наибольшими $F_i^w(s_k)$, для которых соответствующие $\mathcal{K}_i^w(s_k) \xrightarrow{k} k_i^w(s_k)$ все разные, и формируем на основе N отобран-

ных пар $(\mathcal{A}_i^w(s_k), k_i^w(s_k))$ новые списки номеров (мест) подсловарей, слова которых могут составить продолжения отобранных начальных последовательностей $\mathcal{K}_r^w(s_k)$ $\forall k$. Ранжировав вновь сформированные тройки по убыванию сходства, получим N -ку $(F'_i(0), \mathcal{K}'_i(0), \mathcal{A}'_i(0))$, $i = 1 : N$, для главного состояния $s = 0$ и момента времени i .

Очевидно, что N -ка $(F'_i(0), \mathcal{K}'_i(0), \mathcal{A}'_i(0))$, $i = 1 : N$, найденная по рекуррентным формулам для главного состояния $s = 0$ в заключительный момент времени $i = l$ определит ответ смысловой интерпретации. Для записи результата смысловой интерпретации следует проанализировать посредством ОСС или типов смысла и типов предложений последовательность слов $\mathcal{K}^l(0)$.

Как и в случае алгоритма обобщенного распознавания слитной речи, все не определенные в формулах (7.7.1) — (7.7.5) величины сходств полагаются равными $-\infty$.

Алгоритм многозначной смысловой интерпретации начинает свою работу с того, что полагается $(F'_0(0) = 0, \mathcal{K}'_0(0) = \emptyset, \mathcal{A}'_0(0) = \mathcal{A}_{\text{нач}})$, $i = 1 : N$, где $\mathcal{A}_{\text{нач}}$ — список номеров (мест) начальных подсловарей сети ОСС.

Анализируя этот алгоритм, приходим к заключению, что на всех стадиях вычислений он оперирует только с допустимыми в языке диалога начальными подпоследовательностями слов, причем в каждый момент времени, чтобы не оказаться перед фактом локального решения, вводится многозначность решений, т. е. рассматривается не менее $N > > 1$ допустимых начальных подпоследовательностей слов.

Очевидно, что более разумно по ходу обработки речевого сигнала менять количество N многозначных решений. Так, например, исходя из отведенного объема памяти для хранения текущих массивов троек $(F'_i(s), \mathcal{K}'_i(s), \mathcal{A}'_i(s))$, $i = 1 : N$, в первые моменты времени, когда последовательности $\mathcal{K}'_i(s)$ имеют небольшую длину, можно задаваться большими значениями N , постепенно затем уменьшая число N , чтобы оставаться в пределах отведенной памяти. В этом есть смысл еще и потому, что в начале обработки речевого сигнала еще накоплено мало информации о распознаваемом сигнале, еще опасно сужать область поиска решений, и поэтому необходимо предусматривать количество возможных решений как можно большим.

Алгоритм многозначной смысловой интерпретации не гарантирует глобального решения задачи смысловой интерпретации слитной речи, однако в нем предусмотрены такие приемы обработки речевого сигнала, что при надлежащем выборе количества N многозначных решений гарантируется поиск лучшего решения в условиях ограниченных вычислительных ресурсов.

§ 7.8. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

Эксперименты по распознаванию и смысловой интерпретации слитной речи проводились на примере предметной области, относящейся к вопросам о вычислениях. Рассматривалось 17 операций и функций.

Вопросы-задания давались на естественном русском языке и относились к следующим операциям и функциям, в данном случае определяющим типы смыслов: сложение и умножение любого количества целых чисел, вычитание и деление двух целых чисел, логарифм, синус, косинус, обратная величина, корень квадратный и куб числа и др. Целые числа можно было изменять в пределах от —1000 до 1000. Примеры устных вопросов будут даны ниже. Ответ на поставленный устный вопрос выдавался в текстовом виде и путем синтеза речи.

В словаре было около 1000 слов, в том числе основных — 200. Основные слова: ПЯТЬ, СОРОК, ОТНЯТЬ, КУБ, СКОЛЬКО, ПЛЮС и др., вспомогательные: ПЯТЬЮ, ОТ ПЯТИ, ОТНИМИ, ПОЖАЛУЙСТА, СКАЖИ и т. п.

В экспериментах использовались двухэтапный и многозначный алгоритмы распознавания и смысловой интерпретации слитной речи, в точной и приближенной реализациях.

В режиме обучения и дообучения пофонемному распознаванию слов были сначала найдены общая для всех слов совокупность эталонных элементов и акустическая и темпоральная транскрипции каждого слова (гл. 4).

В качестве элементов речи (распознаваемых и эталонных) использовались 48-мерные элементы-коды (§ 2.1, § 4.8, § 5.6), а в качестве элементарной меры сходства — хэммингово расстояние между наблюдаемым и эталонным элементами, взятое со знаком минус. Интервал анализа $\Delta T'$ принимался равным 15 мс, шаг анализа ΔT — 15 мс.

Максимальные значения количества M слов-претендентов и количества N последовательностей слов равнялись 50 и 20 соответственно.

В лингвистическом блоке все возможные предложения языка предметной области задавались с помощью списочных структур на языке ЛИСП. Одному типу предложений соответствует одна списочная структура. Каждый отдельный тип смысла (всего 17 различных типов) задавался небольшим списком структур, представлявшим все возможные варианты предложений естественного языка. Список структур можно было все время пополнять.

Лингвистический анализ заключался в указании одной наиболее вероятной последовательности слов из N -ки, которая удовлетворяет списочной структуре. Направленный перебор всех возможных предложений языка диалога, заданных списочными структурами, и их сравнение с предложениями N -ки велись с помощью процедуры, которая проверяет, есть ли в анализируемом предложении ключевые слова, характерные данному (одному из 17) семантическому заданию (типу смысла). Если да, то дальнейшие анализ и разбор предложения ведутся внутри списочных структур отобранных семантических заданий.

Значительный эффект дает процедура, заключающаяся в том, что для каждого слова словаря заводится набор признаков (двоичный код). Число разрядов в коде равно количеству типов предложений. В i -м разряде кода стоит 1, если слово может входить в состав предложений, относящихся к i -му типу предложений. Любая анализируемая последовательность слов может быть охарактеризована соответствующей

последовательностью двоичных кодов. Логическое поразрядное произведение этих кодов образует некоторый новый двоичный код, единица в i -м разряде которого укажет, что анализируемое предложение принадлежит i -му типу предложений [18, 133].

Завершается семантико-синтаксический разбор указанием, к какому из 17 типов смысла относится устный вопрос и чему равны операнды. Далее формируется ответ на устный вопрос.

Смысловая интерпретация велась в замедленном масштабе времени, что ограничивало объемы экспериментов по интерпретации речевых сигналов.

Всего было проинтерпретировано около 1000 устных заданий, 96 % из них — правильно. По остальным 4 % заданий были получены отказы от интерпретации (в N -ке не было допустимых последовательностей слов).

Представление о работе системы смысловой интерпретации дается следующим диалогом.

Человек (произносит в микрофон): УМНОЖЬ ЧИСЛА СОРОК НА СЕМНАДЦАТЬ И ЧИСЛО ДВА.

ЭВМ (текст на экране дисплея): ОТВЕТ РАВЕН 1360.00.

Ч.: К МИНУС СЕМЬ ПЛЮС МИНУС ДВЕСТИ.

ЭВМ: ПОЛУЧЕН ОТВЕТ — 207.00.

Ч.: НАЙДИ СУММУ ЧИСЕЛ СТО, ДЕВЯТЬСОТ И СОРОК МИНУС ДЕВЯТЬ.

ЭВМ: В РЕЗУЛЬТАТЕ ПОЛУЧЕН ОТВЕТ 1031.00.

Ч.: ОПРЕДЕЛИ, ПОЖАЛУЙСТА, РАЗНОСТЬ ЧИСЕЛ ОТ МИНУС ДВАДЦАТИ ОТНЯТЬ МИНУС СЕМНАДЦАТЬ.

ЭВМ: ОТВЕТ РАВЕН —3.00000.

Ч.: ЧЕМУ РАВЕН ЛОГАРИФМ СЕМИДЕСЯТИ? НАЙДИ, ПОЖАЛУЙСТА.

ЭВМ: ОТВЕТ РАВЕН 4.248.

Ч.: ВЫЧИСЛИ ЛОГАРИФМ НУЛЯ.

ЭВМ: ОШИБКА: АРГУМЕНТ РАВЕН НУЛЬ.

Ч.: ВЫЧИСЛИ КУБ ЧИСЛА СЕМЬ ПЯТЬ.

ЭВМ: ВО ФРАЗЕ ЛИШНИЕ СЛОВА. ВОЗМОЖНЫЙ ОТВЕТ 343.000.

Ч.: СЛОЖИ ДВА.

ЭВМ: НЕ ХВАТАЕТ АРГУМЕНТОВ.

ВЫВОДЫ

1. В рамках КДП-подхода разработаны методы распознавания и смысловой интерпретации слитной речи применительно к устному диалогу человека и ЭВМ на формализованных или усеченных естественных языках предметных областей.

Предлагаемая генеративная модель автоматического понимания (смысловой интерпретации) слитной речи основана на экономном задании всех возможных предложений предметной области с помощью ориентированной семантической сети или структур, названных типами

смыслов и типами предложений. Основным элементом ОСС, типов смыслов и типов предложений является подсловарь и его описание, определяемое его отношением к предметной области. ОСС, типы смыслов и типы предложений позволяют как генерировать все возможные и только допустимые в языке диалога предложения, так и получать передаваемый предложениями смысл в определенной канонической форме.

Используя ОСС или типы смысла и типы предложений, можно построить автоматную грамматику (граф СРОСС), порождающую разнообразные коартикулированные эталонные сигналы слитной речи, отличающиеся темпом и интенсивностью произнесения, длиной пауз между словами и соответствующие, однако, только допустимым в языке диалога предложениям. Задача смысловой интерпретации слитной речи формулируется как отыскание для анализируемого сигнала наиболее правдоподобного эталонного сигнала слитной речи среди множества всех сигналов, порождаемых автоматной грамматикой СРОСС, и как семантико-синтаксический разбор последнего.

Сформулированная задача решается с помощью специальной одноступенчатой схемы ДП, в которой одновременно реализуются три вкладываемых друг в друга ДП-процесса — оптимизация по сходству отрезков речевого сигнала на слова, границам между словами в потоке слитной речи и допустимым последовательностям слов.

Другой возможный путь распознавания и смысловой интерпретации слитной речи заключается в применении двухэтапного метода, согласно которому на первом этапе в предположении свободного порядка следования слов решается так называемая задача обобщенного распознавания, заключающаяся в указании для анализируемого сигнала $N > 1$ наиболее вероятных и ранжированных по убыванию правдоподобия последовательностей слов, и затем на втором этапе из этих N последовательностей слов с помощью ОСС или типов смысла и типов предложения отбирается одна последовательность слов, одновременно и допустимая, и наиболее вероятная. Семантико-синтаксическим разбором этой последней последовательности слов заканчивается формирование канонической формы передаваемого смысла.

Обобщенная задача распознавания слитной речи решается методом, подобным ДП-методу.

Третий путь состоит в том, что в процессе решения обобщенной задачи распознавания слитной речи учитывают синтаксис, семантику и прагматику языка диалога. Для каждого текущего момента времени находят $N > 1$ допустимых в языке диалога начальных подследовательностей слов, а в последующие моменты времени рассматривают только такие слова, которые могут составить допустимые продолжения начальных подследовательностей слов, накопленных в предшествующие моменты времени. Этот путь устраняет недостатки и объединяет достоинства двух других путей. Он реализует многозначную смысловую интерпретацию.

2. При смысловой интерпретации слитной речи существенно (более чем на порядок) возрастают необходимые объемы памяти и вычислений (по сравнению с простым методом распознавания слитной речи, пред-

полагающим свободный порядок следования слов). Это объясняется тем, что дополнительно учитываются ограничения (априорная информация), задаваемые синтаксисом, семантикой и прагматикой предметных областей.

3. Предлагаемые методы обеспечивают высокий процент правильности смысловой интерпретации речевых сигналов. Эти методы имеют универсальный характер и рекомендуются для использования.

ГЛАВА 8

ПРОБЛЕМА ДИКТОРА В РАСПОЗНАВАНИИ РЕЧИ

Индивидуальные особенности голоса человека являются одним из факторов, затрудняющих автоматическое распознавание речевых сигналов [134]. Как это видно из предыдущих глав, высокая надежность распознавания и смысловой интерпретации речи достигается только при условии обучения распознаванию на выбранный словарь и голос диктора. При смене диктора надежность распознавания уменьшается и варьируется в значительных пределах. Так, при распознавании слов «своего» диктора, на которого обучена система распознавания, надежность распознавания словаря из 100 слов составляет 99 % и более. В то же время при распознавании слов, произнесенных новым, «чужим», диктором, эта надежность может упасть даже до 40 %, что зависит как от индивидуальности диктора, так и от используемого способа описания речевых сигналов.

В данной главе в рамках КДП-подхода обсуждаются вопросы адаптации систем распознавания речи к голосу нового диктора, вопросы создания многодикторных систем распознавания. Рассматривается также проблема распознавания дикторов по речевому сигналу, которая имеет и самостоятельное значение. В работах по затронутой проблеме активное участие принимал А. И. Куляс.

§ 8.1. СПОСОБЫ ПРОЯВЛЕНИЯ И УЧЕТА ИНДИВИДУАЛЬНЫХ ОСОБЕННОСТЕЙ ГОЛОСА

При обучении системы распознавания речи на словарь и голос пользователя по обучающей выборке оцениваются исходные эталонные сигналы слов E_k вместе с их громкостной H_k , тональной F_k и темпоральной τ_k транскрипциями (поэлементный метод) либо общая для всех слов совокупность E эталонных элементов и Q_k - и τ_k -транскрипции слов (пофонемный метод). Напомним, что Q -транскрипция — это тройка транскрипций: акустическая R , громкостная H и тональная F . Если же используется глубокое пофонемное распознавание, то оцениванию подлежат общая совокупность E эталонных элементов и Q_d - и τ_d -транскрипции всех дифонов.

Независимо от используемых методов распознавания и обучения распознаванию, все оцениваемые параметры оказываются зависимыми

от индивидуальных особенностей голоса или, проще говоря, от диктора. Эта зависимость имеет место и тогда, когда используется один и тот же словарь.

Прежде всего индивидуально зависимой является совокупность E эталонных элементов. Это следствие того, что наблюдаемые элементы речи x_i зависят от манеры произнесения (способа и места образования) звуков речи данным диктором. Действительно, у каждого индивидуума свои характерные размеры речевого тракта, своя передаточная характеристика речевого тракта для отдельных звуков, свои индивидуальные свойства источника голосового возбуждения (форма импульсов возбуждения).

Но если эталонные элементы из E или E_k выражают индивидуальные особенности произнесения отдельных звуков, точнее, элементов этих звуков, то эталонные сигналы E_k и транскрипции Q_k (Q_d) и τ_k (τ_d) выражают уже индивидуальную манеру произнесения целых слов. Это относится также и к индивидуальным темпу, громкости и тональности произнесения слова, задаваемым транскрипциям τ_k , H_k и F_k соответственно.

Индивидуально также изменяются и оформляются темп, громкость и тональность произнесения целых фраз.

В целом, информация об индивидуальных особенностях голоса распределается по всем параметрам, оцениваемым в режиме обучения.

Сравнение видеоспектрограмм, полученных при анализе речи разных дикторов, показывает, что при всем разнообразии проявления индивидуальных особенностей голоса видеоспектрограммы одних и тех же слов достаточно похожи. Это дает основание искать простые закономерности, позволяющие преобразовывать сигналы одного диктора в сигналы другого.

Предположение о существовании таких закономерностей делает правомочными постановки задачи быстрой перестройки системы распознавания на голос нового диктора и задачи создания многодикторных систем распознавания речи.

Эти постановки задач могут быть сделаны, если ввести дополнительные ограничения на параметры, оцениваемые при обучении.

§ 8.2. ВЗАИМОСВЯЗЬ ЗАДАЧ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ И НАСТРОЙКИ НА ГОЛОС ОПЕРАТОРА

Задачей настройки на голос оператора будем называть такую задачу обучения, когда, располагая результатами обучения на голос одного (опорного) диктора, на основании обучающей (настроечной) выборки нового диктора и дополнительной априорной информации о правилах перехода от сигналов одного диктора к сигналам другого требуется оценить (корректировать) параметры обучения так, чтобы обеспечить распознавание речи нового диктора с приемлемой надежностью распознавания.

Чтобы задача настройки имела смысл, объем настроечной выборки (НВ) и время работы алгоритма настройки должны быть меньше тех

значений, которые обычно требуются или требовались при обучении на словарь и голос опорного диктора. Принципиальная возможность уменьшения значений названных величин вытекает из того, что используется дополнительная априорная информация о закономерностях, связывающих сигналы разных дикторов, а также информация о результатах обучения на словарь и голос опорного диктора.

Задачи настройки и, таким образом, быстрого перевода системы на распознавание речи нового диктора должны удовлетворять еще одному существенному требованию: должна обеспечиваться приблизительно та же надежность распознавания, что и при обучении.

Задачи настройки тесно переплетаются с задачами обучения и, как будет показано в дальнейшем, часто совпадают с частными задачами обучения или дообучения, которые уже были рассмотрены в предыдущих главах.

§ 8.3. НАСТРОЙКА НА ГОЛОС ОПЕРАТОРА ПРИ ПОЭЛЕМЕНТНОМ И ПОФОНЕМНОМ РАСПОЗНАВАНИИ

Изучая, как проявляется «сходство» видеоспектрограмм слов, произнесенных разными дикторами, можно высказать предположение о существовании преобразований, которые позволяют переводить видеоспектрограммы (эталонные сигналы) одного диктора в видеоспектрограммы (эталонные сигналы) другого.

Пусть $E_k = (e_{k1}, e_{k2}, \dots, e_{ks}, \dots, e_{kq_k})$, $k = 1 : K$, — совокупность исходных эталонных сигналов слов вместе с их громкостными и тональными транскрипциями (поэлементный метод) или E — совокупность из небольшого количества эталонных элементов, общих для всех слов, и $O_k = (R_k, H_k, F_k)$, $k = 1 : K$, — совокупность Q_k -транскрипций всех слов (пофонемный метод), полученных для одного опорного диктора. Пусть τ_k , $k = 1 : K$, — темпоральные транскрипции слов для обоих методов.

При пофонемном методе распознавания исходный эталонный сигнал E_k слова k образуется по правилу $E_k = Q_k E$ (гл. 4, 5), причем как в поэлементном методе, так и в пофонемном полагается, что в качестве компонент эталонных элементов, составляющих E_k , входят элементы и громкостной, и тональной транскрипций H_k и F_k .

Одна из возможных постановок задач настройки предполагает существование для каждого диктора некоторого матричного преобразования A , переводящего эталонные сигналы слов опорного диктора в его эталонные сигналы по правилу:

$$E'_k = A E_k \text{ или } E'_k = A Q_k E, \quad k = 1 : K, \quad (8.3.1)$$

где действие оператора A распространяется на все элементы в последовательности

$$A E_k = (A e_{k1}, A e_{k2}, \dots, A e_{ks}, \dots, A e_{kq_k}), \quad (8.3.2)$$

$$A Q_k E = (A e(j_{k1}), A e(j_{k2}), \dots, A e(j_{ks}), \dots, A e(j_{kq_k})). \quad (8.3.3)$$

Таким образом, речь идет об оценивании преобразования A по НВ. Применив преобразование A к эталонным элементам опорного диктора, получим параметры обучения для нового диктора [106, 135—138].

Итак, пусть дана НВ нового диктора

$$(\mathbf{X}'_{l_r}, k(r)), \quad r = 1 : \mathcal{U}, \quad (8.3.4)$$

где $\mathbf{X}'_{l_r} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_{l_r}, \dots, \mathbf{x}'_{l_r})$ — реализация с номером r в НВ; l_r — длина этой реализации; $k(r)$ — слово, к которому относится реализация с номером r ; \mathcal{U} — количество реализаций в НВ.

Требуется на основании НВ и результатов обучения опорного диктора найти оператор A , который вместе с совокупностью операторов $\mathbf{v}' \in \tau_{k(r)}(l_r)$, $r = 1 : \mathcal{U}$, преобразования исходных эталонов слов E_k доставляет максимум критерию качества настройки

$$\Phi(A, \{\mathbf{v}'\}) = \sum_{r=1}^{\mathcal{U}} G(\mathbf{X}'_{l_r}, \mathbf{v}' A E_{k(r)}). \quad (8.3.5)$$

Критерий (8.3.5) определяет нахождение максимально правдоподобных оценок искомого матричного оператора A и контрольных параметров $\{\mathbf{v}'\}$.

Предполагается, что оператор A задается матрицей размером $n \times n$, где n — размерность эталонных элементов e . На оператор A могут накладываться дополнительные ограничения.

Пусть эталонные элементы представляют собой мгновенные амплитудные спектры. Тогда оператор A в общем виде задает амплитудно-частотное преобразование элементов-спектров. Наложим на элементы $a_{\mu s}$ матрицы A следующие ограничения:

- а) $a_{\mu s}$ принимают только значения 0 и 1;
- б) в каждой строке матрицы стоит только одна единица;
- в) если в μ -й строке единица стоит в столбце с номером $s(\mu)$, то в $(\mu + 1)$ -й строке единица может стоять в столбце с номером $s(\mu + 1) \geq s(\mu)$.

В этом частном случае оператор A целиком задается последовательностью $S = (s(1), s(2), \dots, s(j), \dots, s(n))$. Этот S -оператор определяет частотное преобразование спектральных элементов — нелинейное растяжение-сжатие оси частот.

Особого внимания заслуживает случай элементов e с двоичными компонентами, принимающими значения 0 или 1. Если S -преобразование сохраняет двоичный характер элементов, то для общего преобразования A должны быть указаны приемы перехода от Ae к элементам с двоичными компонентами. В [106] обосновывается следующая процедура образования элементов с двоичными компонентами:

$$e'_t = \text{Sign}\left(Ae_t - \frac{1}{2} I\right) \quad (8.3.6)$$

или

$$\dot{e}'_{t\mu} = \text{Sign}\left(\sum_{s=1}^n a_{\mu s} e_{ts} - \frac{1}{2}\right), \quad \mu = 1 : n, \quad (8.3.7)$$

где \mathbf{I} — единичный вектор и

$$\text{Sign } x = \begin{cases} 1, & \text{если } x \geq 0, \\ 0, & \text{если } x < 0. \end{cases}$$

Для этого случая везде, в частности в критерии (8.3.5), вместо $A\mathbf{e}$ следует писать $\text{Sign}\left(A\mathbf{e} - \frac{1}{2} \mathbf{I}\right)$.

Для решения задач настройки предлагается использовать традиционный для КДП-подхода итерационный алгоритм обобщенной по-координатной оптимизации. Обобщенными переменными выступают искомый параметр-оператор A или S (соответственно n^2 или n оцениваемых параметров) и совокупность $\{\mathbf{v}'\}$ операторов $\mathbf{v}' \in \tau_{k(r)}(l_r)$ расстояния исходных эталонов.

Каждая итерация содержит два шага.

На первом шаге $(v + 1)$ -й итерации на основании оператора A^v , найденного на предыдущей итерации, решают задачу нахождения совокупности

$$\{\mathbf{v}^{(v+1)r}\} = \underset{\{\mathbf{v}'\}}{\operatorname{argmax}} \Phi(A^v, \{\mathbf{v}'\}), \quad (8.3.8)$$

которая, в силу сепарабельности критерия (8.3.5) по переменным \mathbf{v}' , распадается на \mathcal{U} независимых задач

$$\mathbf{v}^{(v+1)r} = \underset{\mathbf{v}' \in \tau_{k(r)}(l_r)}{\operatorname{argmax}} G(\mathbf{X}_{l_r}^{-1}, \mathbf{v}' A^v \mathbf{E}_{k(r)}), \quad r = 1 : \mathcal{U}. \quad (8.3.9)$$

Для их решения следует воспользоваться алгоритмом сегментации реализаций, изложенным в § 3.2.

На втором шаге при фиксированной совокупности $\{\mathbf{v}^{(v+1)r}\}$ находим оператор A^{v+1} :

$$A^{v+1} = \underset{A}{\operatorname{argmax}} \Phi(A, \{\mathbf{v}^{(v+1)r}\}). \quad (8.3.10)$$

Способ решения задачи (8.3.10) в значительной мере зависит от используемой элементарной меры сходства $g(\mathbf{x}_i, \mathbf{e}_i)$ и ограничений, накладываемых на A . Так, в случае меры сходства $g(\mathbf{x}_i, \mathbf{e}_i) = -|\mathbf{x}_i - \mathbf{e}_i|^2$ и свободного выбора A задача (8.3.10) сводится к решению системы из n^2 уравнений относительно a_{us} .

Запишем это решение в матричном виде.

Каждую из реализаций НВ запишем в виде матрицы размером $n \times l_r$, $r = 1 : \mathcal{U}$, где n — количество компонент (строк) в элементах; l_r — длина реализации (количество столбцов). Подряд выписывая эти матрицы для всех реализаций НВ, образуем матрицу НВ размером $n \times \sum_{r=1}^{\mathcal{U}} l_r$. Обозначим ее X .

Аналогичным образом составим матрицу \mathbf{E}^{v+1} размером $n \times \sum_{r=1}^{\mathcal{U}} l_r$, из эталонных сигналов $\mathbf{v}^{(v+1)r} \mathbf{E}_{k(r)}$, $r = 1 : \mathcal{U}$, используя результаты первого шага алгоритма настройки. Поскольку для рассматриваемого

случая критерий настройки Φ зависит квадратично от A , то, дифференцируя по A критерий настройки, получим матричное уравнение относительно A :

$$A(\mathcal{E}^{v+1})(\mathcal{E}^{v+1})' = X(\mathcal{E}^{v+1})', \quad (8.3.11)$$

где ' — символ транспонирования.

Из (8.3.11) получаем

$$A^{v+1} = X(\mathcal{E}^{v+1})' ((\mathcal{E}^{v+1})(\mathcal{E}^{v+1})')^{-1}. \quad (8.3.12)$$

Для существования и единственности решения (8.3.12) достаточно, чтобы компоненты эталонных элементов были линейно независимыми, что практически всегда выполняется.

В случае использования S -операторов задача второго шага итерационного алгоритма видоизменяется.

Пусть элементарная мера сходства $g(x_i, e_i)$ может быть представлена в виде суммы покомпонентных слагаемых

$$g(x_i, e_i) = \sum_{\mu=1}^n g_{\mu}(x_{i\mu}, e_{i\mu}). \quad (8.3.13)$$

Тогда критерий настройки Φ при условии фиксированных операторов $v^{(v+1)r}$ можно представить в виде

$$\Phi(S, \{v^{(v+1)r}\}) = \sum_{\mu=1}^n d^{v+1}(\mu, s(\mu)), \quad (8.3.14)$$

где

$$d^{v+1}(\mu, s(\mu)) = \sum_{r=1}^{\mathcal{U}} \sum_{i=1}^{l_r} g_{\mu}(x'_{i\mu}, (v^{(v+1)r} E_{k(r)})_{is(\mu)}). \quad (8.3.15)$$

В выражении (8.3.15) под записью $(v^{(v+1)r} E_{k(r)})_i$ следует понимать i -й эталонный элемент в последовательности из l_r элементов, получаемой в результате применения оператора $v^{(v+1)r}$ к исходному эталону $E_{k(r)}$, а под $(v^{(v+1)r} E_{k(r)})_{is(\mu)}$ — $s(\mu)$ -ю компоненту этого элемента, сопоставляемую μ -й компоненте распознаваемых элементов x'_i .

На втором шаге итерационного алгоритма следует максимизировать критерий (8.3.14) по оператору S или, что то же самое, по монотонно неубывающей последовательности $s(\mu)$, $\mu = 1 : n$.

Сформулированная задача решается с помощью динамического программирования. Граф решения этой задачи для случая $n = 8$ представлен на рис. 8.1. Стрелкам, входящим в вершину (μ, s) , приписывается величина $d(\mu, s)$.

Двигаясь вдоль дуг графа, вычисляем величины $F(\mu, s)$ для всех вершин (μ, s) , пользуясь рекуррентными формулами

$$F(\mu, s) = \max_{1 \leq \xi \leq s} F(\mu - 1, \xi) + d(\mu, s), \quad (8.3.16)$$

и для каждой вершины (μ, s) запоминаем $s(\mu, s)$:

$$s(\mu, s) = \arg \max_{1 \leq \xi \leq s} F(\mu - 1, \xi). \quad (8.3.17)$$

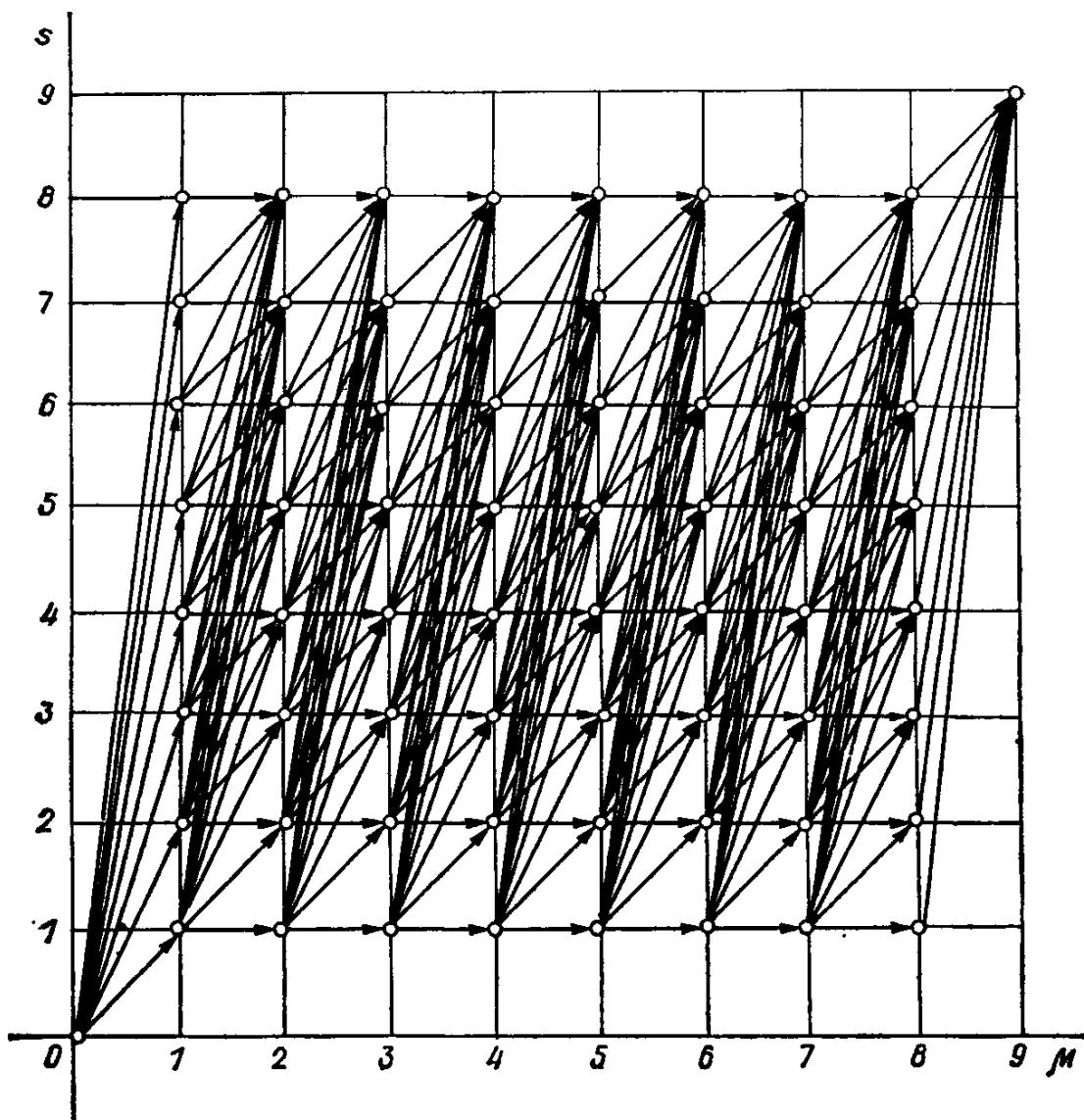


Рис. 8.1. Граф для оценивания оператора настройки S .

Вычисляя по (8.3.16) — (8.3.17) начинаем, полагая, что $F(1, s) = d(1, s)$ и $s(1, s) = s$ для $s = 1 : n$.

Оптимальное значение критерия равно

$$F(n+1, n+1) = \max_{1 \leq s \leq n} F(n, s), \quad (8.3.18)$$

а оптимальное значение $s(n+1, n+1)$:

$$s(n+1, n+1) = \operatorname{argmax}_{1 \leq s \leq n} F(n, s). \quad (8.3.19)$$

Чтобы сформировать оптимальный оператор $S = (s(1), s(2), \dots, s(u), \dots, s(n))$, воспользуемся формулами выписывания:

$$\begin{aligned} s(n) &= s(n+1, n+1), \quad s(u) = s(u, s(u+1)), \quad (8.3.20) \\ u &= n-1, n-2, \dots, 2, 1. \end{aligned}$$

Далее полагаем $S^{v+1} = S$ и приступаем к выполнению очередной $(v+2)$ -й итерации алгоритма настройки.

Также видоизменяется задача второго шага итерационного алгоритма настройки и в случае, когда матрица A произвольна, эталонные

элементы нового и опорного дикторов имеют двоичные компоненты и предоставлена дополнительная информация в виде элементов-векторов, из которых посредством операции Sign по формулам, подобным (8.3.6) — (8.3.7), получают эталонные элементы опорного диктора [106].

Условие останова итерационного алгоритма настройки

$$A^{v+1} = A^v \text{ или } S^{v+1} = S^v \quad (8.3.21)$$

достигается за конечное число итераций.

Для запуска итерационного алгоритма достаточно положить $A^0 = I$, где I — единичная матрица, или $s^0(u) = u$, $u = 1 : n$.

В рассматриваемой задаче настройки предполагалось, что темпоральные и акустические транскрипции слов не меняются при смене дикторов. Остальные же параметры (эталонные элементы, исходные эталоны слов, громкостная и тональная транскрипции) могут изменяться.

Метод настройки на голос оператора исследовался экспериментально.

Поэлементный метод настройки изучался для случая спектрального описания элементов. Размерность векторов-элементов равнялась 20, интервал анализа $\Delta T' = 18$ мс, шаг анализа $\Delta T = 18$ мс. Элементы x_i в реализациях слов нормировались из условия, что максимальный на длине реализации модуль элемента x_i должен равняться единице (см. § 2.6). Используемая мера сходства $g(x_i, e_i) = -|x_i - e_i|^2$.

Настройка велась по НВ, состоящей из 54 реализаций 27 слов (по 2 реализации на слово), а затем осуществлялось распознавание контрольной выборки нового диктора для словаря из 54 слов ($K = 54$).

Эксперименты показали, что A -настройка существенно повышает надежность распознавания (на 10—20 %), в то время как S -настройка в среднем не улучшала результаты распознавания. В целом же эксперименты показали, что настройка при поэлементном распознавании не обеспечивает той практически приемлемой надежности распознавания, которая гарантируется при обучении на голос диктора (99 % при словаре в 100 слов) [106, 135].

При пофонемном распознавании в условиях, подробно описанных в § 4.8, исследовалась эффективность S -настройки. В качестве описания элементов использовались двоичные коды (§ 2.1, § 4.8). В совокупности E было всего $J = 80$ эталонных элементов.

Эксперименты показали эффективность S -настройки: надежность распознавания возрастила от 60—70 до 90—97 %. Объем словаря $K = 50$ слов. Настроенную выборку составляли реализации 15—25 слов [136—137].

Однако наилучшие результаты при пофонемном распознавании обеспечивала такая настройка, которая заключалась в полной замене по НВ старой совокупности эталонных элементов E новой E' (с сохранением неизменными всех транскрипций слов). Задача настройки и алгоритм ее решения в этом случае являются частным случаем задачи и алгоритма обучения пофонемному распознаванию слов речи (гл. 4). В итерационном алгоритме обучения (в данном случае — настройки)

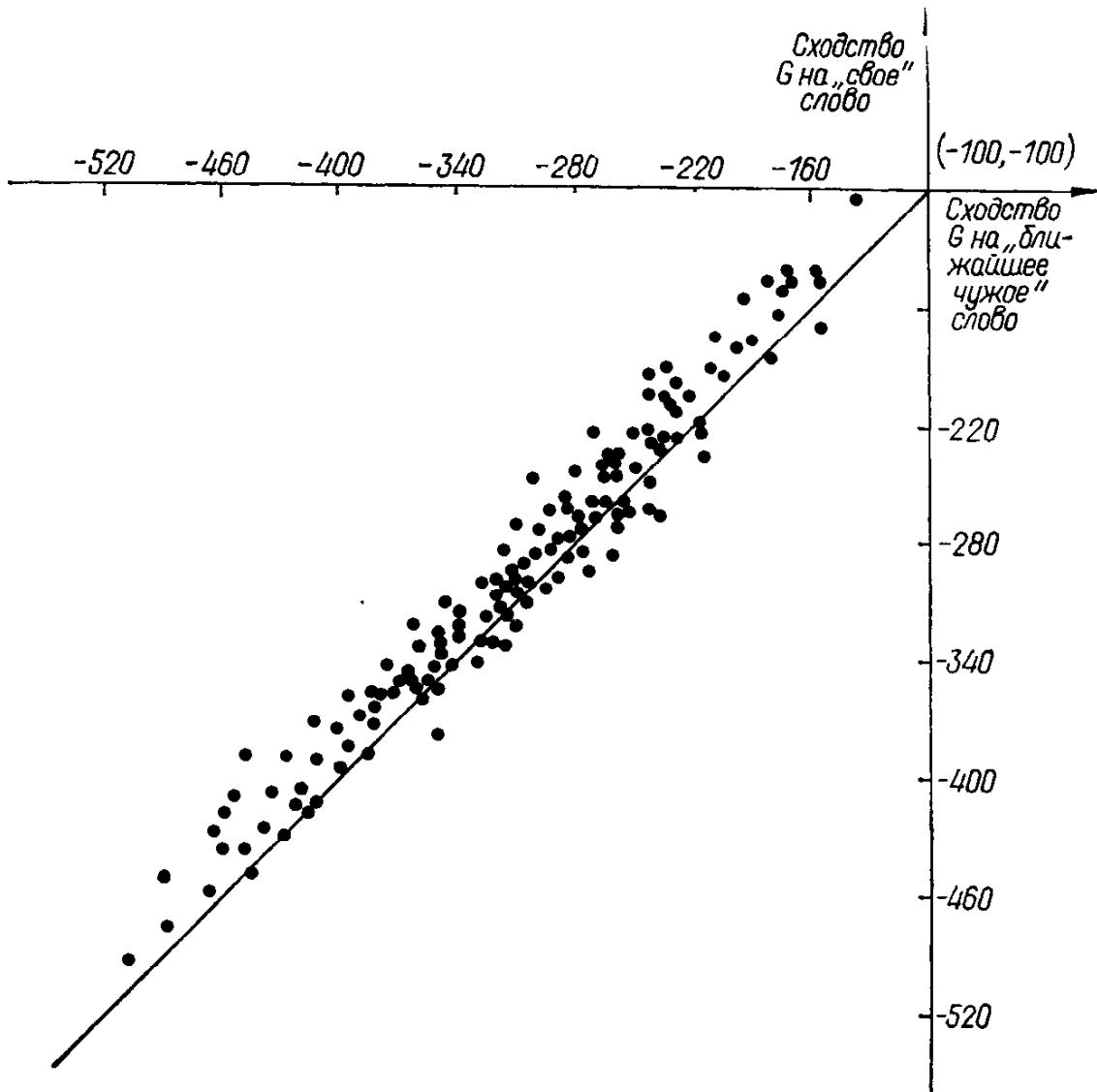


Рис. 8.2. Расположение контрольной выборки в плоскости «ближайший чужой — свой» до настройки.

на каждой итерации следует выполнить только первый и третий шаги. В качестве начальных условий для E' можно брать совокупность эталонных элементов опорного диктора.

Эксперименты для этого случая показали, что если НВ составлять из K реализаций (по одной реализации на слово), то обеспечиваемая надежность распознавания почти такая же, как при полном обучении [138].

Этот метод настройки в случае пофонемного распознавания рекомендуется для практического использования.

Рис. 8.2 и рис. 8.3 иллюстрируют действие настройки в плоскости «ближайший чужой — свой». Одна и та же контрольная выборка нового диктора из 150 реализаций сначала распознавалась по результатам обучения опорного диктора (рис. 8.2), т. е. без настройки, а затем (рис. 8.3) — после настройки. По оси абсцисс отложено сходство реализации на «ближайшее чужое» слово, по оси ординат — сходство на «свое» слово. При безошибочном распознавании вся выборка должна быть расположена выше биссектрисы координатного угла. Сравнивая

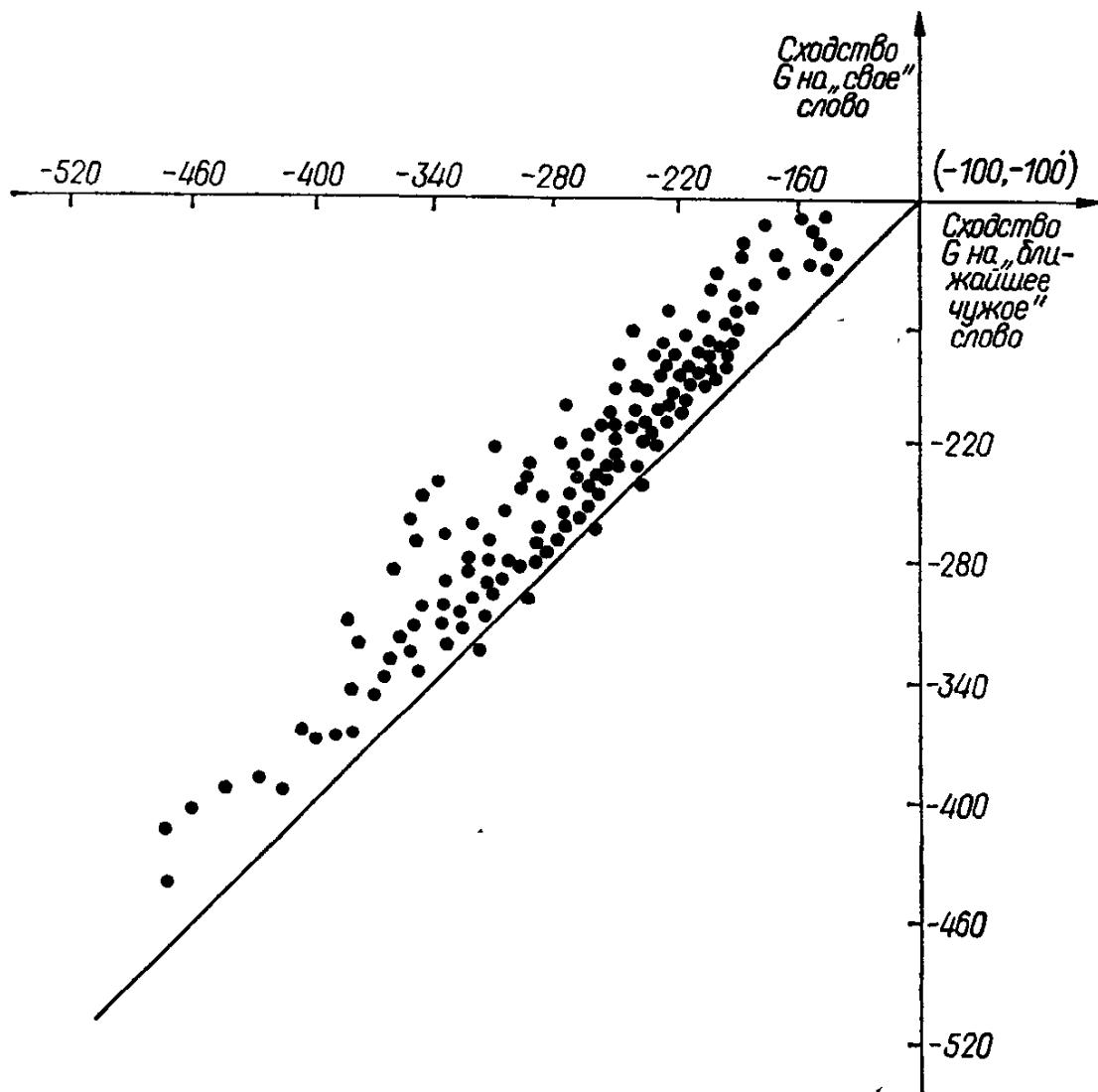


Рис. 8.3. Расположение контрольной выборки в плоскости «ближайший чужой — свой» после настройки

расположение выборок в плоскости «ближайший чужой — свой» до настройки и после настройки, убеждаемся, что после настройки увеличивалась та доля контрольной выборки, которая расположена выше биссектрисы координатного угла.

§ 8.4. ПОСТАНОВКА ЗАДАЧИ СОЗДАНИЯ МНОГОДИКТОРНЫХ СИСТЕМ РАСПОЗНАВАНИЯ

Тот факт, что речь в унисон двух и более дикторов воспринимается как речь некоторого одного диктора, дает основание полагать, что транскрипции слова (акустическая, громкостная и тональная) одинаковы для всех дикторов, а одноименные эталонные элементы $e_p(j)$ из совокупности E для разных дикторов p образуют выпуклые множества Ω_j . Этому предположению не противоречат также результаты экспериментов по настройке.

Сформулированная гипотеза позволяет поставить задачу создания многодикторных систем распознавания речи.

Для ее решения сначала необходимо научиться описывать выпуклые множества Ω_j , $j = 1 : J$, одноименных эталонных элементов $e_p(j)$,

затем разработать средства вычисления акустической, темпоральной, громкостной и тональной транскрипций слова, общих для разных дикторов, по его фонетической транскрипции.

При смене диктора должна осуществляться адаптация системы распознавания на голос нового диктора. В процессе адаптации система распознавания, отправляясь от априорной информации о структуре множеств Ω_j , уточняет свои представления о конкретном дикторе (о значениях параметров $e_p(j)$, $j = 1 : J$). При этом не исключается, что в первое время после смены диктора система распознавания может работать не очень надежно.

Процесс разработки подобной многодикторной системы распознавания может быть автоматизирован, если, задавшись ограничениями на структуру множеств Ω_j , $j = 1 : J$ (например, множества Ω_j не пересекаются), попытаться оценить параметры этих множеств и искомые транскрипции слов речи по ОВ, составленной из реализаций слов и слитной речи в произнесении многих дикторов.

Постановка и решение подобной задачи обучения (самообучения) могут быть сделаны в рамках КДП-подхода. Этому, однако, должно предшествовать изучение структуры множеств Ω_j , $j = 1 : J$, и ее параметризация, выбор адекватных способов описания и представления эталонных элементов.

Простейшим примером применения излагаемого подхода к созданию многодикторных систем распознавания речи является кооперативное распознавание отдельно произносимых слов.

§ 8.5. КООПЕРАТИВНОЕ РАСПОЗНАВАНИЕ

Кооперативное распознавание слов речи — это такое пофонемное распознавание речи группы (кооператива) дикторов, когда на основании ОВ этой группы оцениваются и затем используются при распознавании одна, общая для всей группы, совокупность E из небольшого количества J эталонных элементов $e(j) \in E$, $j = 1 : J$, и одна, общая для всей группы, четверка акустической R_k , темпоральной τ_k , громкостной H_k и тональной F_k транскрипций каждого слова k , $k = 1 : K$. Кооперативное распознавание, таким образом, не отличается от простого пофонемного распознавания слов речи. Различие лишь чисто организационное — ОВ составляют реализации слов не одного, а кооператива дикторов. Например, в кооперативе 10 дикторов, каждый диктор произносит каждое слово по одному разу (10 реализаций на слово при обучении распознаванию) [139, 140].

Кооперативное распознавание — это частный случай излагаемого подхода к созданию многодикторных систем распознавания, когда множества Ω_j , $j = 1 : J$, для кооператива дикторов описываются одним элементом.

Кооперативное распознавание исследовалось экспериментально.

Распознаваемые элементы x_i представлялись 11-мерными векторами автокорреляций, а эталонные элементы $e(j)$, $j = 1 : J$, $J \leq J_o = 80$, — 11-мерными b -параметрами (§ 2.1). Интервал анализа $\Delta T'$ и шаг анализа ΔT были равны 15 мс каждый. Используемая элементарная

Таблица 8.1. Результаты распознавания слов для кооператива мужчин

Дикторы	Номер дикто-ра	Надежность распозна-вания, %	
		I экспери-мент	II экспе-римент
Диктор-мужчи-на, член кооперати-ва	1	99	99
	2	80	97
	3	79	97
	4	79	98
В среднем по ко-оперативу		84	98
Новый диктор-мужчина	5	—	96
	6	—	97
	7	—	95
	8	—	96
Новый диктор-женщина	9	—	88
	10	—	55
	11	—	50
	12	—	50

Таблица 8.2. Результаты распознавания слов для кооператива женщин

Дикторы	Номер дикто-ра	Надеж-ность рас-познава-ния, %
Диктор-женщи-на, член кооперати-ва	9	99
	10	95
	11	96
	12	97
В среднем по ко-оперативу		97
Новый диктор-женщина	13	96
	14	95

мера сходства

$$g(x_i, e(j)) = \\ = - (x_{ii})^{-\frac{3}{4}} (x_i, b(j)),$$

где x_{ii} — энергия элемента x_i (x_{ii} — первая компонента вектора автокорреляции x_i); $(x_i, b(j))$ — скалярное произведение двух векторов x_i и $b(j)$ (см. также § 2.3).

Громкостная и тональная транскрипции слова не использовались.

Кооперативная ОВ была составлена из 400 реализаций 100 слов ($K = 100$) в произнесении четырех дикторов-мужчин (по одной реализации слова на каждого диктора). Контрольная выборка содержала 800 реализаций 100 слов в произнесении этих же дикторов (по две реализации на слово и диктора) и 1600 реализаций этих же слов в произнесении четырех дикторов-мужчин и четырех дикторов-женщин, не вошедших в состав кооператива.

Было проведено два эксперимента. В первом эксперименте распознавались контрольные выборки всех дикторов по ранее полученным результатам обучения первого диктора, во втором — система распознавания сначала была обучена по кооперативной ОВ четырех дикторов, а затем по этим результатам обучения распознавалась контрольная выборка всех дикторов.

Результаты экспериментов по надежности распознавания контрольных выборок сведены в табл. 8.1.

По результатам табл. 8.1 представляла интерес оценка надежности распознавания слов, произнесенных дикторами-женщинами, системой, обученной кооперативом женщин. Результат этого эксперимента представлен в табл. 8.2.

Наибольший интерес представляли оценки надежности распознавания слов, произнесенных как мужчинами, так и женщинами, системой, обученной кооперативом, включающим одинаковое количество дикторов-мужчин и дикторов-женщин (табл. 8.3).

Анализируя табл. 8.1—8.3, заключаем, что при индивидуальном

обучении средняя надежность распознавания по чужим дикторам составляет 84 % и неприемлема для практики. В то же время кооперативное обучение обеспечивает среднюю надежность распознавания для членов кооператива 97—98 %, что приемлемо для практического использования, и существенно повышает надежность распознавания слов для новых дикторов, не являющихся членами кооператива.

Однако высокая надежность распознавания слов в кооперативной системе достигается, если иметь одновременно хотя бы два набора результатов обучения (два кооператива): отдельно для дикторов-мужчин и дикторов-женщин. Выбор нужного набора может осуществляться либо оператором, либо автоматически — путем распознавания, кем произнесена реализация — мужчиной или женщиной.

Эксперименты с кооперативным распознаванием подтверждают перспективность разработки предлагаемого в § 8.4 подхода к созданию многодикторных систем распознавания речи.

§ 8.6. ПРОБЛЕМА РАСПОЗНАВАНИЯ ДИКТОРА ПО РЕЧЕВОМУ СИГНАЛУ

Проблема распознавания диктора по речевому сигналу имеет самостоятельное значение и изучается многими исследователями [134, 141]. В то же время она тесно переплетается с проблемой автоматического распознавания речи.

В системах распознавания речи с ограниченным доступом лиц возникает задача распознавания и верификации (подтверждения) диктора, когда на основании речевого сигнала требуется определить, входит ли данный диктор в число допущенных дикторов, и если входит, то какой это конкретно диктор. Последнее необходимо, чтобы вызвать из внешней памяти результаты обучения на словарь и голос диктора, который представился системе распознавания, и тем самым в дальнейшем гарантировать для этого диктора устный диалог с ЭВМ. Результаты же обучения распознаванию на голос каждого диктора могут быть получены заранее либо путем индивидуального полного обучения на голоса всех допущенных дикторов, либо путем организации многодикторной системы распознавания по способу § 8.4, когда для каждого диктора p заводится своя совокупность E_p , эталонных элементов $e_p(j)$, $j = 1 : J$, а транскрипции слов являются общими для всех дикторов.

Таблица 8.3. Результаты распознавания слов для смешанного кооператива дикторов

Диктор	Номер диктора	Надежность распознавания, %
Диктор-мужчина, член кооператива	1	88
	2	96
Диктор-женщина, член кооператива	9	97
	10	97
Новый диктор-мужчина	3	70
	4	79
	7	79
Новый диктор-женщина	8	80
	11	70
	12	96
	13	70
	14	79

Распознавание и верификация дикторов могут вестись по произвольной речи или парольной фразе.

Использование парольной фразы — наиболее приемлемый для практики способ распознавания и верификации сотрудничающих с системой распознавания дикторов.

Пусть E_p — совокупность эталонных элементов $e_p(j) \in E_p, j = 1 : J$, для p -го диктора, а (Q_ϕ, τ_ϕ) — Q -транскрипция и темпоральная транскрипция τ парольной фразы. Транскрипции Q_ϕ и τ_ϕ являются общими для всех дикторов.

Распознавание диктора на основании сигнала X_l парольной фразы можно вести по формуле

$$p(X_l) = \operatorname{argmax}_p \max_{v \in \tau_\phi(l)} G(X_l, vQ_\phi E_p). \quad (8.6.1)$$

В соответствии с (8.6.1) сигнал X_l считается принадлежащим диктору p (X_l), из совокупности $E_{p(X_l)}$ эталонных элементов которого получается наиболее похожий на X_l эталонный сигнал $vQ_\phi E_p$, $v \in \tau_\phi(l)$.

Задача (8.6.1) подобна той задаче, которая обычно решается при простом фонемном распознавании слов с помощью динамического программирования (см. гл. 4). Разница лишь в том, что делается полный перебор не слов k , а совокупностей E_p эталонных элементов разных дикторов p .

В том случае, когда выполняется условие

$$\max_p \max_{v \in \tau_\phi(l)} G(X_l, vQ_\phi E_p) \leq \gamma l, \quad (8.6.2)$$

где γ — некоторый коэффициент, объявляется отказ от верификации (подтверждения) того, что данный диктор принадлежит к кругу лиц, знакомых системе распознавания. Коэффициент γ подбирается экспериментальным путем.

Распознавание дикторов по парольной фразе можно вести и несколько иначе. Коль скоро при одинаковых транскрипциях слов и фраз у разных дикторов разные совокупности E_p эталонных элементов $e_p(j) \in E_p$, то естественно ожидать, что Q - и τ -транскрипции слов и парольных фраз для разных дикторов, выраженные в именах эталонных элементов $e_{p^*}(j) \in E_{p^*}$ некоторого одного опорного диктора p^* , также будут разными.

Пусть $E = E_{p^*}$ — совокупность эталонных элементов, найденная в режиме полного обучения для некоторого одного опорного диктора p^* . Тогда при фиксированной E_{p^*} в режиме дообучения (§ 4.7) для каждого диктора p могут быть оценены Q - и τ -транскрипции парольного слова или фразы, обозначаемые далее как $Q_{\phi p}$ и $\tau_{\phi p}$. Распознавание диктора по сигналу X_l парольной фразы далее может вестись по формуле

$$p(X_l) = \operatorname{argmax}_p \max_{v \in \tau_{\phi p}(l)} G(X_l, vQ_{\phi p} E), \quad (8.6.3)$$

если только не выполняется условие отказа от верификации того, что диктор знаком системе распознавания:

$$\max_p \max_{v \in \tau_{\phi p}(l)} G(X_l, vQ_{\phi p} E) \leq \gamma l. \quad (8.6.4)$$

Последний прием распознавания дикторов, по существу не отличающийся от пофонемного метода распознавания слов (роль номера слова k играет номер диктора p), исследовался экспериментально. В качестве элементов речи x_i и $e(j)$ использовались двоичные коды (§ 2.1). Интервал $\Delta T'$ и шаг анализа ΔT были равны 18 мс. Используемая элементарная мера сходства — хэммингово расстояние между кодами x_i и $e(j)$, взятое со знаком минус.

В качестве парольных фраз использовались Я — МАШИНА и СЛУШАЙ, ЧЕЛОВЕК.

В экспериментах принял участие 20 человек (мужчин). Каждый диктор в режиме дообучения произнес каждую парольную фразу по 10 раз. Длина транскрипций Q и τ для первой фразы была выбрана равной $q = 19$. В совокупности E_p , опорного диктора было всего $J = 80$ эталонных элементов.

При распознавании контрольных выборок в 95 % случаев было получено правильное распознавание дикторов и в 5 % — отказы от распознавания дикторов [142].

В случае распознавания дикторов по произвольной речи рекомендуется исходить из того, что у каждого диктора p своя индивидуальная совокупность E_p из небольшого количества J эталонных элементов $e_p(j) \in E_p$, $j = 1 : J$, а Q - и τ -транскрипции слов являются общими для всех дикторов (см. § 8.4).

Пусть дано E_p , $p = 1 : P$, где P — количество дикторов, знакомых для системы распознавания. Пусть реализовано глубокое пофонемное распознавание речи (см. гл. 6).

Тогда решая для распознаваемого сигнала X_i ровно P (по числу дикторов) задач глубокого пофонемного распознавания слитной речи (будет равно P общих фонемных графов с одинаковой структурой), сможем указать не только наиболее вероятную последовательность слов или фонем, переданную сигналом X_i , а и номер наиболее вероятного диктора p (X_i), которому принадлежит анализируемый сигнал X_i .

В распознавании дикторов по произвольной речи может принять участие и человек. Например, он может указывать последовательность слов или фонем, которая передается анализируемым сигналом X_i . Тогда номер диктора, которому принадлежит сигнал X_i , определяется номером того фонемного графа, который по данной последовательности фонем генерирует наиболее похожий на X_i эталонный сигнал слитной речи.

Участие человека в процессе распознавания дикторов способствует повышению надежности распознавания.

Отказ от верификации дикторов по произвольной речи формируется по правилу, подобному (8.6.2).

ВЫВОДЫ

1. В рамках КДП-подхода разработаны средства настройки системы распознавания на голос нового диктора, в процессе которой, располагая результатами обучения распознаванию для одного (опорного) диктора, по настроенной выборке нового диктора оцениваются пара-

метры оператора, преобразующего эталонные сигналы опорного диктора в эталонные сигналы нового диктора. При этом считается, что Q - и τ -транскрипции слов не изменяются от диктора к диктору.

Задача настройки принципиально приводит к уменьшению выборок и времени работы алгоритмов по сравнению с полным обучением, поскольку используется дополнительная априорная информация о закономерностях, связывающих сигналы разных дикторов. Эта задача эффективно решается с помощью итерационных алгоритмов обобщенной покоординатной оптимизации.

2. Многодикторные системы пофонемного распознавания речи могут основываться на предположении, что транскрипции слова (акустическая, громкостная и тональная) одинаковы для всех дикторов, а однотипные эталонные элементы $e_p(j)$, $j = 1 : J$, $J \leq 1024$, из совокупности E_p для разных дикторов p образуют выпуклые множества Ω_j , $j = 1 : J$. В этих условиях при смене диктора должна осуществляться адаптация системы распознавания к голосу нового диктора, в процессе которой, отправляясь от структуры множеств Ω_j , уточняются значения параметров $e_p(j)$, $j = 1 : J$, для нового диктора.

В частном случае, множества Ω_j для группы (кооператива) дикторов могут быть приближенно описаны одним элементом $e(j)$. Тогда возможно кооперативное распознавание речи для лиц, вошедших в состав кооператива дикторов. Кооперативное распознавание совпадает с обычным пофонемным распознаванием. Организационное отличие только в том, что в кооперативном обучении параметры обучения оцениваются по смешанной выборке, составленной из реализаций слов разных дикторов — членов кооператива.

Кооперативное распознавание характеризуется высокой надежностью распознавания слов речи, произнесенных как дикторами — членами кооператива, так и многими другими дикторами, не вошедшими в состав кооператива дикторов.

Эксперименты с кооперативным распознаванием показали, что для распознавания речи многих дикторов необходимо иметь, по крайней мере, два кооператива: мужской и женский. Соответственно, для аппроксимации множеств Ω_j необходимо, как минимум, два элемента.

3. Распознавание дикторов по речевому сигналу в рамках КДП-подхода можно вести, используя как произвольную речь, так и парольные фразы. Возникающие постановки задач распознавания дикторов и соответствующие алгоритмы их решения являются аналогичными применяемымся для распознавания слов и слитной речи.

ГЛАВА 9

ИССЛЕДОВАНИЯ ПО ПЕРВИЧНОМУ АНАЛИЗУ РЕЧЕВЫХ СИГНАЛОВ И ВЫБОРУ МЕРЫ СХОДСТВА

В данной главе обсуждаются вопросы первичного анализа речи — получения последовательностей $X_i = (x_1, x_2, \dots, x_i, \dots, x_l)$ элементов x_i , описывающих (представляющих) речевые сигналы на этапе распознавания.

От используемого описания речевых сигналов в значительной мере зависит успех в решении задач распознавания и смысловой интерпретации. Не меньшее значение приобретает и вопрос эффективного использования уже выбранного описания. Применительно к КДП-подходу последняя проблема сводится к выбору адекватной элементарной меры сходства $g(x_i, e(j))$ между наблюдаемыми x_i и эталонными элементами $e(j)$.

Частично вопросы анализа и выбора меры сходства уже рассматривались в предыдущих главах (например, в гл. 1 и 2). Однако в настоящей главе они будут изложены более подробно и в аспектах, не изучавшихся другими исследователями.

§ 9.1. ВЫБОР ИНТЕРВАЛА АНАЛИЗА И МОДЕЛИРОВАНИЕ РАЗЛИЧНЫХ АНАЛИЗАТОРОВ РЕЧЕВЫХ СИГНАЛОВ

При анализе речи обычно пользуются интервалами анализа $\Delta T'$ продолжительностью 10—20 мс. Отсчет же элементов речи делается с шагом ΔT , также выбираемым из диапазона 10—20 мс. Такие интервалы выражают известный компромисс между желанием как можно точнее передать динамику речевого тракта и источником его возбуждения, с одной стороны, и желанием иметь представительную выборку для оценки этой динамики — с другой. Считается само собой разумеющимся, что при сдвиге интервалов анализа, например, в пределах 5 мс, практически не меняются как передаточная характеристика речевого тракта и параметры источников его возбуждения, так и сам результат анализа, например, амплитудный спектр. Если параметры речевого тракта и источников его возбуждения действительно меняются медленно в силу инерционных свойств, то так ли обстоит дело с результатами спектрального анализа? Действительно ли можно считать, что если одну и ту же реализацию слова дважды подвергнуть спектраль-

ному анализу с одними и теми же интервалом и шагом анализа, сдвинув во втором случае начало интервалов на 5 мс, то различия в результатах анализа будут незначительными?

Чтобы ответить на поставленные вопросы и исследовать, как влияет дискретизация речевого сигнала на надежность распознавания, был выполнен ряд экспериментов.

В первом эксперименте изучалась стабильность результатов анализа в зависимости от смещения интервалов анализа относительно первоначального положения.

Дискреты f_n , $n = 1 : M$, речевого сигнала с точностью 9 бит брались с шагом $\Delta t = 80$ мкс (см. обозначения § 2.1). На изучаемые интервалы анализа продолжительностью $\Delta T' = 14,8, 29,6, 44,4, 59,2$ мс приходилось, таким образом, соответственно $M = 185, 370, 555$ и 740 дискрет f_n . Далее по формулам (2.1.1) — (2.1.4) при $m = 10$ вычислялась автокорреляционная функция речевого сигнала $B(s)$, $s = 0 : m$, а по ней тремя способами определялись амплитудные спектры $A(p_r)$, $r = 1 : 48$, на дискретной сетке из 48 частот p_r . По формулам (2.1.3) — (2.1.4) находился прямой спектр (ПС). При этом полагалось $g(s) = 1 - \frac{s}{m+1}$, $s = 0 : m$. По формулам (2.1.5) — (2.1.15) вычислялся так называемый авторегрессионный спектр (АС). Ковариационный спектр (КС) определялся подобно АС-спектру [67]. С этой целью вместо (2.1.1) вычисляли

$$B(u, v) = \sum_{n=m+1}^M f_{n-u} f_{n-v}, \quad u, v = 0 : m, \quad (9.1.1)$$

рассматривая f_n , $n = 1 : m$, как начальные условия для уравнения (2.1.5), и далее находили **a**-параметры, решая систему уравнений

$$\sum_{s=1}^m a_s B(s, u) = -B(0, u), \quad u = 1 : m. \quad (9.1.2)$$

Далее от **a**-параметров перешли к КС-спектру по формулам (2.1.14) — (2.1.15) и (2.1.4), полагая

$$M\sigma^2 = B(0, 0) + \sum_{s=1}^m a_s B(0, s). \quad (9.1.3)$$

Для фиксированных длительностей $\Delta T'$ интервалов анализа и способа вычисления амплитудного спектра изучалось изменение спектра при перемещении интервала анализа вдоль оси времени с шагом $\Delta t = 80$ мкс, т. е. через одну дискрету.

Пусть $A_v(p_r)$, $r = 1 : 48$, $v = 1 : N$, — последовательность спектров, вычисленная для интервалов анализа $\Delta T'$ при его перемещении вдоль оси времени на $v\Delta t$. Выбиралось $N = 185$ или $N = 90$. Это означало, что нестабильность спектра изучалась на протяжении 14,8 мс или 7,2 мс.

В качестве величины, характеризующей нестабильность (разброс) спектра, принималась усредненная нормированная дисперсия норми-

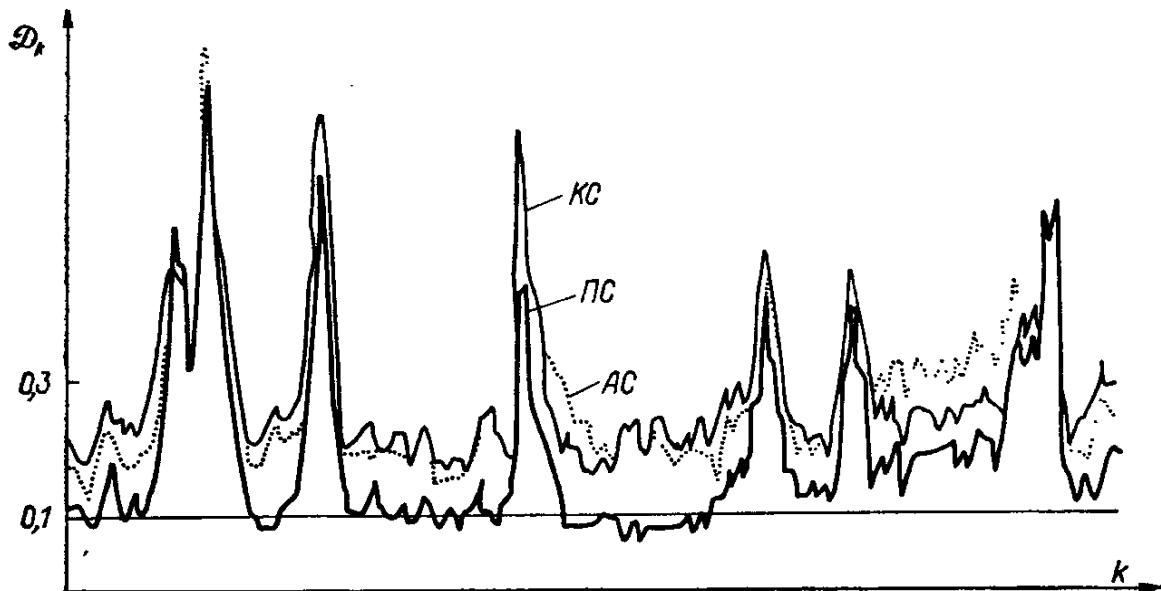


Рис. 9.1. Значения D_k при различных способах вычисления спектра на интервале $\Delta T' = 14,8$ мс.

рованных на единичную сферу векторов амплитудного спектра:

$$\mathcal{D} = \frac{1}{48} \sum_{r=1}^{48} \frac{2\sigma_r}{A_r}, \quad (9.1.4)$$

где

$$A_r = \frac{1}{N} \sum_{v=1}^N \frac{A_v(p_r)}{\sqrt{\sum_{r=1}^{48} A_v^2(p_r)}}, \quad (9.1.5)$$

$$\sigma_r^2 = \frac{1}{N} \sum_{v=1}^N \left(\frac{A_v(p_r)}{\sqrt{\sum_{r=1}^{48} A_v^2(p_r)}} - A_r \right)^2. \quad (9.1.6)$$

Далее изучалась зависимость величины \mathcal{D} от номера k участка речи, начала которого определялось как $t_k = k\Delta\tau$, $\Delta\tau = 5$ мс, т. е. величина D_k измерялась через каждые 5 мс.

На рис. 9.1 — 9.3 приведены графики D_k , полученные для $N = 185$ по одной и той же реализации слова МАШИНА (рассмотрены различные способы вычисления спектра и различные продолжительности интервалов анализа $\Delta T'$). Изучалось также влияние окон на величину \mathcal{D} (прямоугольное, хэммингово и др.) [143].

На основании выполненных исследований были сделаны следующие выводы:

а) относительно высокая стабильность спектра ($\mathcal{D} < 0,1$) получается на стационарных участках звуков, переходные же участки характеризуются значительной нестабильностью ($\mathcal{D} > 0,2$ и часто принимает значения 0,4—0,6);

б) наилучшую стабильность спектра обеспечивает прямой способ вычисления спектра, он лучше ковариационного и авторегрессионного, а сами эти способы отличаются несущественно;

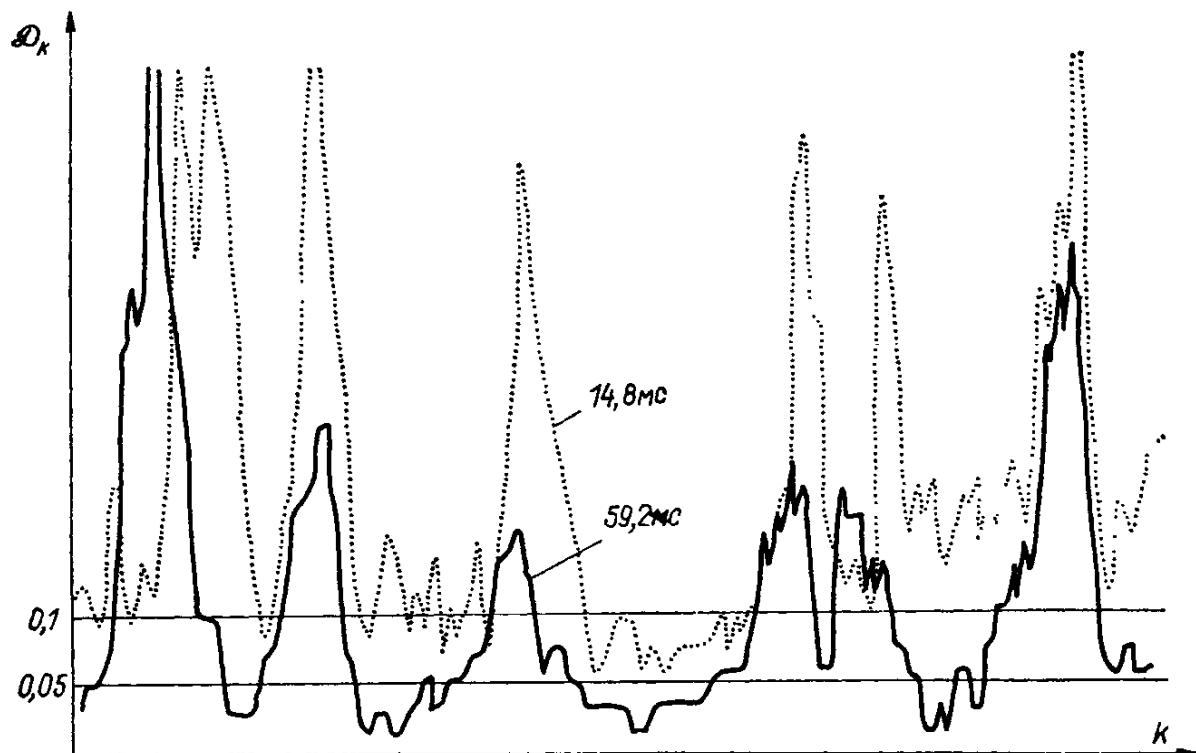


Рис. 9.2. Значения D_k при прямом способе вычисления спектра на интервалах $\Delta T' = 14,8$ мс и $\Delta T' = 59,2$ мс.

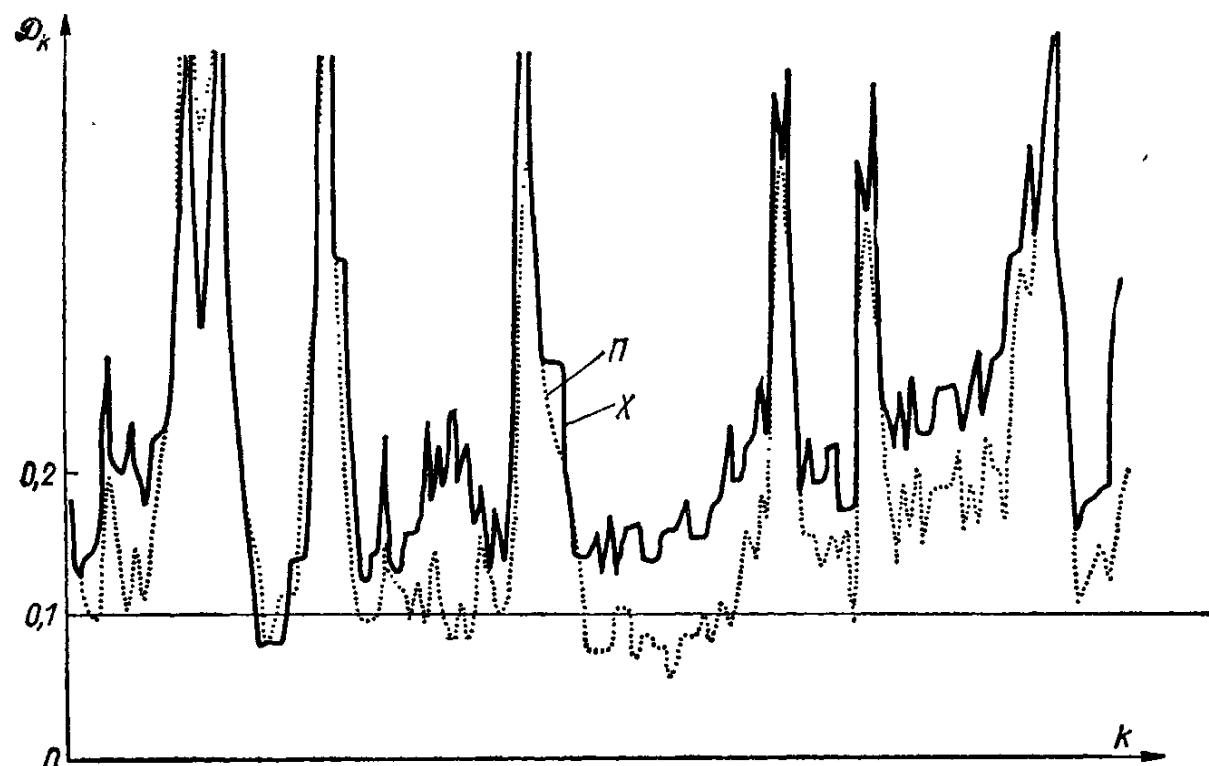


Рис. 9.3. Значения D_k при прямом способе вычисления спектра на интервале 14,8 мс с различными окнами (Π — прямоугольное, X — хэммингово).

- в) наилучшую стабильность спектра обеспечивает прямоугольное окно;
- г) интервалы анализа $\Delta T' = 10 \div 20$ мс характеризуются значительной нестабильностью спектра даже на стационарных участках звуков ($\mathcal{D} > 0,1$);

д) относительно хорошая стабильность спектра достигается на больших интервалах анализа (например, при $\Delta T' = 60$ мс имеем $\mathcal{D} < 0,05$ для стационарных звуков);

е) результат спектрального анализа в значительной мере случаен, что обусловлено как случайностью расположения интервалов анализа, так и малой их продолжительностью; некорректность спектрального анализа речевых сигналов объясняется малостью выборки и большим количеством оцениваемых параметров;

з) на переходных участках звуков стабильность спектрального анализа хотя и улучшается с ростом длины интервалов анализа, однако остается недостаточной;

ж) для устранения эффектов дискретизации и малости выборок необходимо не просто увеличивать длину интервалов анализа, а переходить к зависимому анализу сигналов для соседних интервалов.

К аналогичным выводам приходим, анализируя нестабильность кодового (двоичного) описания речевых сигналов (§ 2.1), получаемого по ПС-, АС- или КС-спектрам.

Во второй серии экспериментов исследовалась зависимость надежности распознавания речи от длины интервала анализа и эффектов случайного расположения интервалов анализа вдоль реализации.

В качестве описания речевых сигналов использовались 48-разрядные двоичные коды, вычисляемые через АС-спектр (см. § 2.1).

Сначала для четырех различных интервалов анализа $\Delta T' = 14,8, 29,6, 44,4, 59,2$ мс с шагом $\Delta T = 14,8$ мс на основании одной и той же обучающей выборки было выполнено четыре процедуры обучения распознаванию ста раздельно произносимых слов (использовался простой пофонемный метод распознавания для случая $K = 100$). Затем одна и та же реализация (исходный речевой сигнал) контрольной выборки подвергалась анализу с различной продолжительностью интервалов анализа $\Delta T' = 14,8, 29,6, 44,4, 59,2$ мс и с различным положением интервалов анализа на оси времени. Шаг анализа выбирался во всех случаях равным $\Delta T = 14,8$ мс. Сдвигая положение интервалов анализа на одну дискрету ($\Delta t = 80$ мкс), для фиксированной продолжительности $\Delta T'$ из одного исходного речевого сигнала получили $N = 185$ различных реализаций одного и того же произнесения слова. С учетом того что $\Delta T'$ еще принимало четыре различных значения, одно и то же произнесение слова давало 4×185 различных реализаций. Поскольку ΔT всегда равнялось 14,8 мс, все 4×185 реализаций имели одинаковую длину.

Всего таким образом было обработано 100 произнесений слов (по одному произнесению на слово) или $100 \times 4 \times 185$ реализаций.

При вычислении двоичных кодов по формуле (2.1.16) пороги Θ_μ , $\mu = 1 : 48$, были выбраны (определенены в режиме обучения) с учетом различий в продолжительности $\Delta T'$ интервалов анализа.

Использование одних и тех же обучающей и контрольной выборок позволяло вести сравнительный анализ в условиях небольших объемов выборок.

Было осуществлено распознавание всех $100 \times 4 \times 185$ реализаций. Для всех четырех интервалов анализа была измерена частота встре-

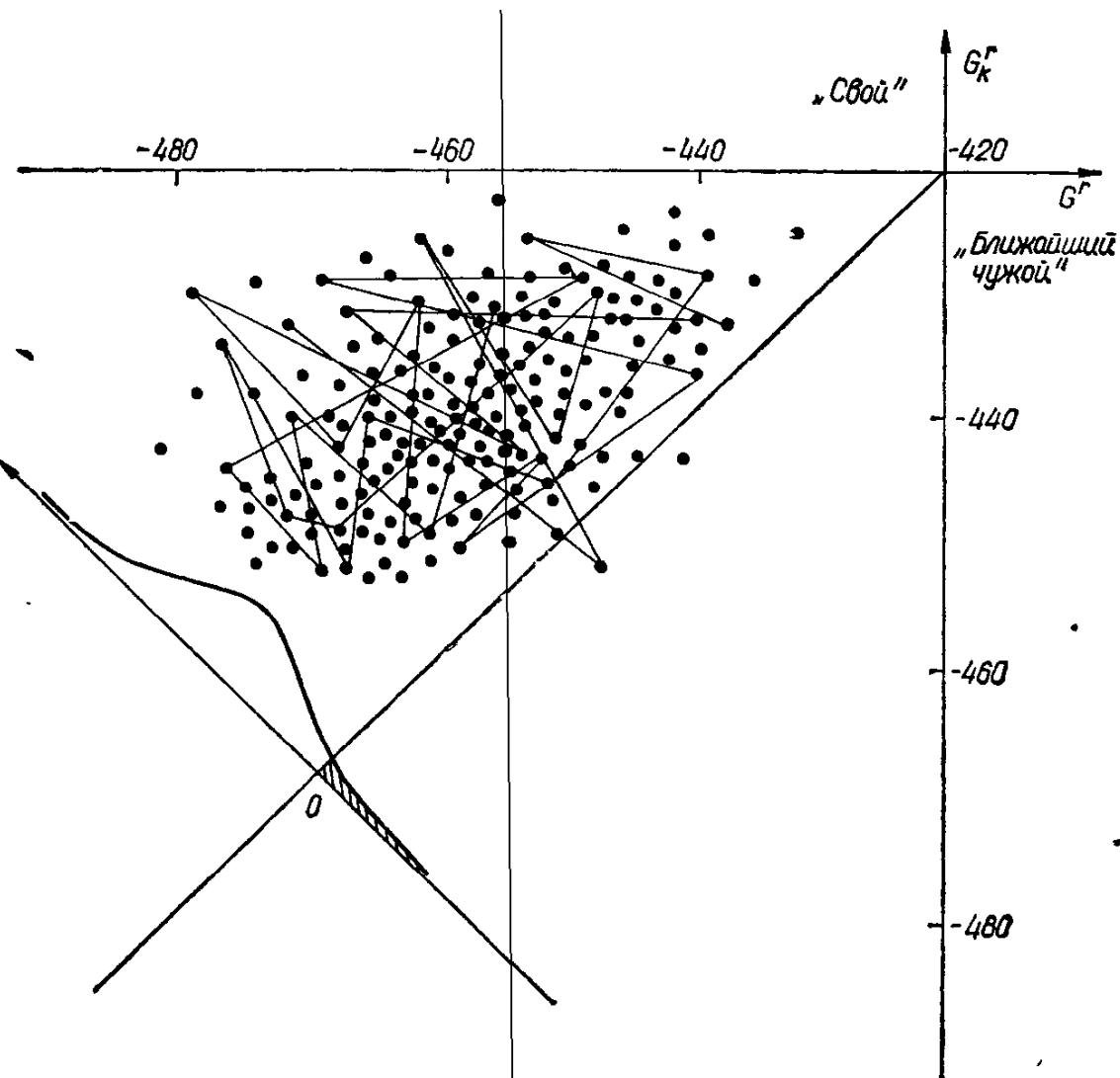


Рис. 9.4. Влияние эффектов дискретизации на надежность распознавания.

чаемости ошибок распознавания. Кроме того, вероятность ошибки распознавания прогнозировалась по распределению случайной величины

$$\xi^r = \frac{1}{\sqrt{2}} (G'_k - G'), \quad r = 1 : N, \quad (9.1.7)$$

где G'_k — максимальное сходство для r -й реализации ($r = 1 : N$, $N = 185$) одного и того же произнесения слова k , полученное при сравнении ее с эталонными сигналами «своего» класса; G' — максимальное сходство для той же r -й реализации, но полученное при ее сравнении с эталонными сигналами всех других, кроме k , слов (в данном случае, при сравнении с эталонными сигналами «ближайшего чужого» слова).

Очевидно, что если $\xi^r < 0$, то r -я реализация произнесения распознается ошибочно.

Найдем среднее a и дисперсию σ случайной величины ξ^r , $r = 1 : N$. Предполагая, что ξ^r распределена по нормальному закону, вычислим величину

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{a}{\sigma}} \exp\left(-\frac{1}{2} x^2\right) dx, \quad (9.1.8)$$

которую можно трактовать как прогноз вероятности ошибочного распознавания, полученный на основании одного произнесения слова k . Очевидно, что это будет прогноз вероятности ошибок распознавания, обусловленных эффектами дискретизации — случайностью расположения интервалов анализа по длине речевого сигнала.

Усредняя оценки P для произнесений разных слов k в условиях фиксированной длины $\Delta T'$ интервала анализа, получим среднюю прогнозируемую вероятность ошибок распознавания, обусловленных эффектами дискретизации, для данного значения длины $\Delta T'$.

Прогнозирование ошибок распознавания позволяет осуществить более тонкое сравнение результатов в условиях ограниченных экспериментов и малой встречаемости ошибок.

Рис. 9.4 иллюстрирует движение точки (G', G_k) на плоскости «ближайший чужой — свой» при смещении положений интервалов анализа $\Delta T' = 14,8$ мс на r , $r = 1 : N$, $N = 185$, дискрет (было произнесено слово ЧЕТЫРЕ).

Результаты сравнительных экспериментов сведены в табл. 9.1. Анализируя данные таблицы, приходим к выводу о необходимости работать с интервалами анализа, продолжительностью до 30 мс. Повидимому, при таких интервалах хотя и относительно сильно сказываются эффекты дискретизации, однако сохраняется достаточно хорошая разрешающая способность анализа во времени, что гарантирует распознавание речи с наибольшей надежностью.

Представляет также интерес и то, как влияют форма весовой функции $g(s)$ и ее длина m в формуле (2.1.3) на надежность распознавания. Можно показать [72, 167], что ответ на эти вопросы равносителен выяснению того, как влияют форма передаточной характеристики фильтрового спектрального анализатора и количество фильтров анализатора на надежность распознавания. Сравнительные эксперименты, описанные подробно в [72, 167], показали, что из четырех исследованных весовых функций $g_1(s) = 1$,

$$g_2(s) = \frac{s\pi}{2(m+1)} \operatorname{ctg} \frac{s\pi}{2(m+1)}, \quad g_3(s) = \cos \frac{s\pi}{2m+1},$$

$g_4(s) = 1 - \frac{s}{m+1}$, $s = 0 : m$, наивысшую надежность распознавания обеспечили вторая и третья весовые функции, хотя вполне приемлемо использование и других весовых функций — соответственно других форм передаточных характеристик фильтрового спектрального анализатора. В частности, весовая функция $g_4(s)$, $s = 0 : m$, взаимнооднозначно определяет $(m+1)$ -канальный спектральный анализатор с равномерной расстановкой спектральных каналов на оси частот и с формой передаточной характеристики каждого канала $K(p) =$

Таблица 9.1. Влияние длины интервала анализа на надежность распознавания

Длина интервала анализа, мс	Частота встречаемости ошибок	Прогноз вероятности ошибок
14,8	0,0123	0,0130
29,6	0,0425	0,0410
44,4	0,0746	0,0754
59,2	0,0856	0,0880

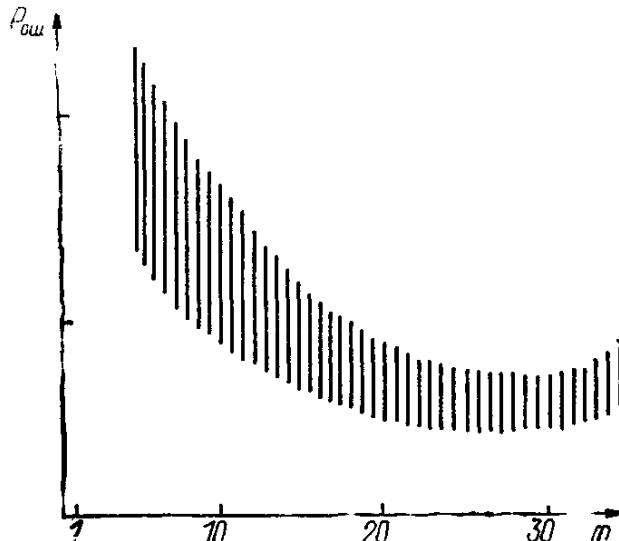


Рис. 9.5. Типовая зависимость частоты встречаемости ошибок от количества фильтров спектрального анализатора.

с ростом n «сужаются» по полосе пропускания таким образом, чтобы равномерно покрыть весь частотный диапазон речевого сигнала. Зависимость снята при частоте дискретизации исходного речевого сигнала, равной 10 кГц ($\Delta t = 100$ мкс) [72, 167]. Из рисунка следует практическая рекомендация по выбору количества n фильтров в анализаторе речевых сигналов:

$$n \leq 30 \frac{\Delta t}{\Delta t'}, \quad (9.1.9)$$

где Δt — период дискретизации сигналов в эксперименте ($\Delta t = 100$ мкс); $\Delta t'$ — предполагаемый период дискретизации сигналов в проектируемом анализаторе.

На рис. 9.5 по оси ординат не указаны конкретные значения частоты ошибок, поскольку значения $P_{\text{ош}}$ зависят от используемого метода распознавания.

§ 9.2. СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАЗЛИЧНЫХ ОПИСАНИЙ РЕЧЕВОГО СИГНАЛА И ЭЛЕМЕНТАРНЫХ МЕР СХОДСТВА

В рамках КДП-подхода успех в решении задач распознавания речи в значительной мере определяется используемыми описанием речевого сигнала и элементарной мерой сходства $g(x_i, e(j))$ между наблюдаемыми x_i и эталонными $e(j)$ элементами.

В рамках задач пофонемного распознавания слов и слитной речи была выполнена серия сравнительных экспериментов с целью оценки пригодности для распознавания тех или иных описаний и мер сходства, основанных на линейном предсказании (§ 2.1).

Используемая частота дискретизации речевого сигнала равнялась 12,5 кГц (шаг дискретизации $\Delta t = 80$ мкс). Длина интервала анализа $\Delta T'$ и шаг анализа ΔT были равны по 15 мс. Количество эталонных элементов в совокупности E равнялось 80 ($e(j) \in E, j = 1 : J, J = 80$).

При использовании линейного предсказания распознаваемые элементы речи x_i , вычисленные на основании дискрет $f_n, n = 1 : M, M = 187$, удобно представлять векторами автокорреляции $B =$

$$= \frac{\sin p - \frac{\pi}{2(m+1)}}{p}, \text{ где } 0 \leq p < 1 — \\ \text{относительная частота (см. также § 2.1).}$$

На рис. 9.5 представлена типичная зависимость частоты встречаемости ошибок распознавания $P_{\text{ош}}$ от длины $n = m + 1$ весовой функции анализатора или, что то же самое, от количества $n = m + 1$ фильтров спектрального анализатора. При этом предполагается, что соседние фильтры «перекрываются» на уровне $\frac{\sqrt{2}}{2}$, а сами фильтры

$\equiv (B_0, B_1, B_2, \dots, B_s, \dots, B_m)$:

$$B_s = \sum_{n=1}^{M-s} f_n f_{n+s}, \quad s = 0 : m, \quad m < M \quad (m = 10), \quad (9.2.1)$$

а эталонные элементы описывать а-параметрами $a = (a_1, a_2, \dots, a_s, \dots, a_m)$ или однозначно связанными с ними б-параметрами $b = (b_0, b_1, b_2, \dots, b_s, \dots, b_m)$ или к-параметрами $k = (k_1, k_2, \dots, k_s, \dots, k_m)$, называемыми коэффициентами отражения (см. § 2.1, формулы (2.1.5) — (2.1.15)).

Вероятность наблюдаемого на интервале анализа сигнала $f_n, n = 1 : M$, при условии его порождения эталонным элементом a может быть вычислена исходя из стохастического уравнения прогноза

$$f_n = - \sum_{s=1}^m a_s f_{n-s} + e_n, \quad (9.2.2)$$

где e_n — ошибка прогноза сигнала $f_n, n = 1 : M$, линейной системой a . Величину

$$g'(\mathbf{x}, \mathbf{e}) = \sum_{n=1}^M e_n^2 = \sum_{n=1}^M \left(f_n + \sum_{s=1}^m a_s f_{n-s} \right)^2 \quad (9.2.3)$$

естественно трактовать как суммарную энергию ошибки прогноза наблюдаемого на интервале анализа сигнала с помощью эталонного элемента, заданного a -параметрами.

Положим в (9.2.3), что $f_{n-s} \equiv 0$ для $n - s < 1$ и $n - s > M$. Тогда $g'(\mathbf{x}, \mathbf{e})$ можно представить в виде

$$g'(\mathbf{x}, \mathbf{e}) = (\mathbf{B}, \mathbf{b}), \quad (9.2.4)$$

где вектор \mathbf{B} определяется посредством (9.2.1), а вектор эталонного элемента \mathbf{b} вычисляется по a с помощью (2.1.14).

Чтобы на основании $g'(\mathbf{x}, \mathbf{e})$ ввести элементарную меру сходства между элементами \mathbf{B} и \mathbf{b} , удовлетворяющую условию (2.3.3)

$$g(\mathbf{x}, \mathbf{e}) = \ln p(\mathbf{x}/\mathbf{e}) \leq 0,$$

воспользуемся уравнением (9.2.2):

$$\ln p(\mathbf{x}/\mathbf{e}) = \ln \left(\frac{1}{(\sqrt{2\pi}\sigma)^M} \exp \left(-\frac{1}{2\sigma^2} g'(\mathbf{x}, \mathbf{e}) \right) \right), \quad (9.2.5)$$

где σ — дисперсия ошибки прогноза для элемента $x: f_n, n = 1 : M$. Естественно полагать σ равной минимально возможному значению для элемента речи x , т. е.

$$M\sigma^2 = \min_a \sum_{n=1}^M \left(f_n + \sum_{s=1}^m a_s f_{n-s} \right)^2. \quad (9.2.6)$$

Если теперь, исходя из формул (2.3.1) — (2.3.4), исключить из (9.2.5) слагаемые, не влияющие на результат распознавания, то получим следующую элементарную меру сходства:

$$g_1(\mathbf{x}, \mathbf{e}) = - \frac{(\mathbf{B}, \mathbf{b})}{M\sigma^2} \leq -1. \quad (9.2.7)$$

Наряду с элементарной мерой сходства $g_1(\mathbf{x}, \mathbf{e})$ изучались и дру-

Таблица 9.2. Сравнение мер сходства при распознавании слов

Мера сходства при обучении	Надежность распознавания при различных мерах сходства, %						
	$-(B, b)$	$\frac{(B, b)}{\sqrt{B_e}}$	$\frac{(B, b)}{M\sigma^2}$	$\frac{(B, b)}{\sqrt[4]{B_0^3}}$	$\frac{(B, b)}{B_e}$	$- k - k_e ^2$	$-\ln \frac{(B, b)}{M\sigma^2}$
$-(B, b)$	95,9	99,3	99,5	99,5	97,3	96,3	99,7
$-(B, b)/\sqrt{B_e}$	95,4	99,0	99,7	99,3	99,5	98,8	99,5
$-(B, b)/M\sigma^2$	94,1	99,7	99,5	99,7	99,7	98,3	99,5
$-(B, b)/\sqrt[4]{B_0^3}$	94,4	99,0	98,8	99,7	97,8	96,6	97,8
$-(B, b)/B_e$	93,4	99,0	98,8	99,5	98,8	98,0	98,8

гие меры:

$$g_2(x, e) = -(B, b); \quad g_3(x, e) = -\frac{(B, b)}{\sqrt{B_e}};$$

$$g_4(x, e) = -\frac{(B, b)}{\sqrt[4]{B_0^3}}; \quad g_5(x, e) = -\ln(B, b),$$

$$g_6(x, e) = -\frac{(B, b)}{B_e},$$

где B_0 — первая компонента (энергия) элемента B .

Все перечисленные меры сходства выражают взаимную энергию ошибки прогноза наблюдаемого элемента B с помощью эталонного элемента b , причем эта энергия берется с некоторым нормировочным множителем. Эта нормировка делается с целью уравнять роли сильных и слабых по интенсивности звуков в интегральной мере сходства между распознаваемой реализацией и эталонными сигналами слов (см. формулу (2.3.4)).

Далее различные меры сходства испытывались при распознавании как отдельно произносимых слов, так и слитной речи.

При распознавании изолированных слов, помимо введенных выше мер сходства, использовалась также мера Итакуры [144]

$$g_7(x, e) = -\ln \frac{(B, b)}{M\sigma^2},$$

а также евклидовы меры в пространстве k коэффициентов отражения, нормированных отсчетов автокорреляции, a -параметров и b -параметров предсказания:

$$g_8(x, e) = -|k - k_e|^2; \quad g_9(x, e) = -|a - a_e|;$$

$$g_{10}(x, e) = -|a - a_e|^2; \quad g_{11}(x, e) = -|a - a_e|^3;$$

$$g_{12}(x, e) = -|b - b_e|^2; \quad g_{13}(x, e) = -\left| \frac{B}{B_0} - \frac{B_e}{B_{el}} \right|^2,$$

где индекс « e » относится к эталонному элементу.

Рассматривался также вариант меры сходства типа g_1 , в которой эталонный элемент e представлялся вектором автокорреляции, норми-

Таблица 9.3. Сравнение мер сходства при распознавании слов

Мера сходства при обучении	Надежность распознавания при различных мерах сходства, %											
	$- a - a_s $	$- a - a_s ^2$	$- a - a_s ^3$	$- b - b_s ^2$	$-\left \frac{B}{B_0} - \frac{B_s}{B_{s0}} \right ^2$	$-\left(b, \frac{B_s}{M\sigma_s^2} \right)$	$g(x, e) = \begin{cases} 0, & \text{если } \left(b, \frac{B_s}{M\sigma_s^2} \right) \leq 1 + \alpha; \\ -1, & \text{в противном случае} \end{cases}$	$\alpha = 0,4$	$\alpha = 0,5$	$\alpha = 0,6$	$\alpha = 0,7$	$\alpha = 0,8$
$-(B, b)/\sqrt[4]{B_0^3}$	92,4	93,2	93,7	71,0	97,8	98,8	94,4	97,6	98,5	98,8	97,3	92,2

Таблица 9.4. Сравнение мер сходства при распознавании слитной речи

Мера сходства при обучении	Проценты правильно распознанных слов и слов-вставок				
	$-(B, b)$	$-\frac{(B, b)}{\sqrt{B_0}}$	$-\frac{(B, b)}{M\sigma^2}$	$-\frac{(B, b)}{\sqrt[4]{B_0^3}}$	$-\frac{(B, b)}{B_0}$
	правильно распознанные слова	правильно распознанные слова-вставки	правильно распознанные слова	правильно распознанные слова	правильно распознанные слова
$-(B, b)$	73	1	—	—	—
$-(B, b)/\sqrt{B_0}$	—	—	89	1	—
$-(B, b)/M\sigma^2$	—	—	—	—	91,7
$-(B, b)/\sqrt[4]{B_0^3}$	—	—	—	—	91,7
$-(B, b)/B_0$	—	—	—	—	—

$-(B, b)$	73	1	—	—	—	—	—	—	—	—	—
$-(B, b)/\sqrt{B_0}$	—	—	89	1	—	—	—	—	—	—	—
$-(B, b)/M\sigma^2$	—	—	—	—	91,7	0,3	89,7	0,3	—	—	—
$-(B, b)/\sqrt[4]{B_0^3}$	—	—	—	—	91,7	0,3	92,7	0,3	—	—	—
$-(B, b)/B_0$	—	—	—	—	—	—	—	—	—	79	3

рованным на собственную ошибку прогноза, а распознаваемый элемент x — b -параметрами:

$$g_{14}(x, e) = -\frac{(b, B_e)}{M\sigma_e^2}.$$

Кроме того, был исследован ряд дискретных мер сходства со значениями ноль и минус единица. Эти меры индуцировались мерой g_{15} :

$$g_{15}(x, e) = \begin{cases} 0, & \text{если } (b, B_e)/M\sigma_e^2 \leq 1 + \alpha; \\ -1 & \text{в противном случае,} \end{cases}$$

для значений $\alpha = 0,4, 0,5, 0,6, 0,7, 0,8, 1,2$.

Эксперименты по распознаванию слов производились в следующих условиях.

Обучающая и контрольная выборки содержали по 410 реализаций каждая (словарь из 82 слов, по пять реализаций на слово). Обучение распознаванию было выполнено для каждой из пяти мер g_1, g_2, g_3, g_4 и g_e .

Затем производилось распознавание реализаций контрольных выборок по всем описанным мерам сходства.

Результаты распознавания (надежность распознавания в процентах) для различных мер сходства в зависимости от меры сходства, использованной при обучении, приведены в табл. 9.2, 9.3.

Была исследована также зависимость надежности распознавания слитной речи от используемой меры сходства и от того, какая мера использовалась при обучении. Произносилось 18 испытательных фраз, по пять раз каждая. Каждая фраза содержала от двух до пяти слов, всего 320 слов. Словарь системы состоял из 50 слов.

В табл. 9.4 приведены результаты экспериментов — процент правильно распознанных слов в слитной речи и процент слов-вставок.

Выполненные исследования позволяют заключить, что для достижения высокой надежности распознавания важно выбрать не только приемлемый метод распознавания и способ представления сигналов, а и правильно задать элементарную меру сходства. Например, евклидова метрика в пространстве a -параметров или, особенно, b -параметров обеспечивает относительно низкую надежность распознавания. Она оказывается приемлемой только для коэффициентов отражения.

В случае использования параметров предсказания можно указать ряд элементарных мер сходства, обеспечивающих распознавание слов с надежностью 99 % и более. К ним относятся, например, меры сходства

$$g(x, e) = -(B, b)/M\sigma^2 \text{ и } g(x, e) = -(B, b)/\sqrt[4]{B_0^3}.$$

§ 9.3. СИНТЕЗ ТАБЛИЧНОЙ ЭЛЕМЕНТАРНОЙ МЕРЫ СХОДСТВА

Элементарную меру сходства $g(x_i, e(j))$ можно попытаться синтезировать (оценить) в режиме обучения по ОВ.

Рассмотрим задачу обучения простому фонемному распознаванию слов и итерационный алгоритм ее решения (см. гл. 4).

Как и ранее, темпоральные транскрипции слов будем вычислять только на заключительной итерации, полагая для всех остальных итераций, что $\tau_k = \tilde{\tau}_k$, $k = 1 : K$ (K — количество слов в словаре). Однако теперь на каждой итерации выполняем еще один, четвертый шаг, заключающийся в оптимизации четвертой обобщенной переменной при фиксированных остальных — в выборе элементарной меры сходства $g(x_i, e(j))$, $j = 1 : J$.

Поскольку

$$g(x, e(j)) = \ln p(x/e(j)) \leq 0, \quad (9.3.1)$$

то всегда должно выполняться

$$\sum_x \exp(g(x, e(j))) = 1. \quad (9.3.2)$$

Пусть элементарная мера сходства $g^n(x, e(j))$, $j = 1 : J$, была найдена в результате n -й итерации обучения и пусть выполнены первые три шага $(n + 1)$ -й итерации обучения, т. е. при условии фиксированной $g^n(x, e(j))$, $j = 1 : J$, уже найдены E^{n+1} , Q_k^{n+1} , $k = 1 : K$, $v^{(n+1)r}$, $r = 1 : \mathcal{U}$. Тогда на четвертом шаге $(n + 1)$ -й итерации будем вычислять новую элементарную меру сходства $g^{n+1}(x, e(j))$, $j = 1 : J$, удовлетворяющую (9.3.1), (9.3.2).

В общем случае элементарная мера сходства должна задаваться в виде прямоугольной таблицы, один размер которой равен J , а дру-

гой — количеству N возможных значений элемента \mathbf{x} в дискретном пространстве сигналов, например, $N = 2^{48}$, если \mathbf{x} — 48-мерный элемент-код с двоичными компонентами.

Чтобы уменьшить размеры этой таблицы, введем вспомогательную меру сходства $g^*(\mathbf{x}, \mathbf{e}(j))$, которую менять не будем. На основании этой меры для каждого наблюдаемого элемента \mathbf{x} укажем наиболее похожий на него эталонный элемент $\mathbf{e}(j^*)$:

$$j^* = \operatorname{argmax}_{j=1:J} g^*(\mathbf{x}, \mathbf{e}(j)). \quad (9.3.3)$$

Эта операция позволит задавать искомую элементарную меру сходства в виде квадратной таблицы размером $J \times J$.

Применение вспомогательной меры $g^*(\mathbf{x}, \mathbf{e}(j))$ позволяет разбить все пространство сигналов \mathbf{x} на J областей и считать распределения $p(\mathbf{x}/\mathbf{e}(j))$ принимающими постоянное значение на отдельных областях.

По результатам первых трех шагов при фиксированных границах $\mathbf{v}^{(n+1)r}$, $r = 1 : \mathcal{U}$, сегментов всех реализаций ОВ, фиксированной совокупности E^{n+1} эталонных элементов $\mathbf{e}^{n+1}(j) \in E^{n+1}$ и фиксированных транскрипциях \mathbf{Q}_k^{n+1} , $k = 1 : K$, в частности, акустических транскрипциях \mathbf{R}_k^{n+1} , $k = 1 : K$, на четвертом шаге сначала выписываем (собираем) одноименные по j сегменты всех реализаций.

Заменим каждый элемент \mathbf{x} одноименных по j сегментов всех реализаций ОВ соответствующим эталонным элементом $\mathbf{e}(j^*)$ согласно формуле (9.3.3), используя для этого вспомогательную меру сходства $g^*(\mathbf{x}, \mathbf{e}(j))$, и вычислим частоты встречаемости элемента j^* , $j^* = 1 : J$, по всем элементам сегментов с именем j . Натуральные логарифмы этих частот встречаемости составят j -й столбец $g^{n+1}(\mathbf{x}, \mathbf{e}(j))$ искомой квадратной табличной меры сходства. Выполнив аналогичную процедуру для всех $j = 1 : J$, получим искомую табличную меру сходства $g^{n+1}(\mathbf{x}, \mathbf{e}(j))$, $j = 1 : J$.

Чтобы пользоваться этой табличной мерой сходства, необходимо при сравнении элемента \mathbf{x} с эталонным элементом $\mathbf{e}(j)$ сначала найти по формуле (9.3.3) для \mathbf{x} его ближайший элемент j^* , а затем выбрать элемент g_{j^*} табличной меры сходства. Это и будет значение элементарной меры сходства при сравнении \mathbf{x} с $\mathbf{e}(j)$.

При итерационном вычислении табличной меры сходства необходимо осуществлять консервативный выбор новых значений табличной меры (о консервативном выборе см. гл. 4), а в условие останова добавить еще одно равенство

$$g_{\omega j}^{n+1} = g_{\omega j}^n, \quad \omega, j = 1 : J. \quad (9.3.4)$$

Для запуска итерационного алгоритма необходимо задаться начальным значением табличной меры сходства $g_{\omega j}^0$, $\omega, j = 1 : J$.

Так, если использовать двоичные 48-мерные элементы-коды \mathbf{x}_t и $\mathbf{e}(j)$, можно положить, что $g_{\omega j}^0 = \ln 0.5$, если $\omega = j$, считая, что в остальных случаях $g_{\omega j}^0 = \ln \frac{0.5}{J-1}$. В качестве вспомогательной меры сходства $g^*(\mathbf{x}, \mathbf{e}(j))$ можно взять все ту же меру, связанную с хэмминговым расстоянием: $g^*(\mathbf{x}, \mathbf{e}(j)) = -H(\mathbf{x}, \mathbf{e}(j))$. Очевидно, что в результате обучения фонемному распознаванию будет синтезирована

мера сходства $g_{\omega j}$, $\omega, j = 1 : J$, отличающаяся от первоначально выбранной $g^0_{\omega j}$, $\omega, j = 1 : J$.

Включение табличной меры сходства в число искомых при обучении параметров несколько изменяет характер вычислений на первом, втором и третьем шагах итерационного алгоритма обучения. Так, на третьем шаге новую совокупность E^{n+1} эталонных элементов $e^{n+1}(j) \in E^{n+1}$ необходимо будет находить на основании вспомогательной меры $g^*(x, e(j))$: сначала собрать все элементы x ОВ, которые аппроксимируются по сходству $\max_{v=1:J} g^*(x, e^n(v))$ одним и тем же эталонным элементом с номером j , а затем найти новый элемент

$$e^{n+1}(j) = \max_e \sum_{(i,r) \in I(j)} g^*(x'_i, e), \quad (9.3.5)$$

где

$$I(j) = \{(i, r) : \operatorname{argmax}_{v=1:J} g^*(x'_i, e^n(v)) = j\}. \quad (9.3.6)$$

Однако такой способ вычисления E^{n+1} показывает, что совокупность E из эталонных элементов $e(j) \in E, j = 1 : J$, можно находить и отдельно от других искомых параметров обучения, не обязательно пользуясь итерационным алгоритмом обучения. Так, совокупность E можно заранее оценить по ОВ, используя один из методов § 4.5. Разумеется, вспомогательная мера сходства $g^*(x_i, e(j))$ при этом должна быть задана.

Если совокупность E находится отдельно, то итерационный алгоритм обучения будет заключаться в оценивании транскрипций слов и табличной меры сходства. Каждая итерация будет содержать три шага: на первом — осуществляется оптимальная сегментация реализаций, на втором — оцениваются Q-транскрипции слов и на третьем — вычисляется табличная мера сходства.

Очевидно, что если пользоваться заранее вычисленной совокупностью E эталонных элементов $e(j) \in E, j = 1 : J$, и вспомогательной мерой сходства $g^*(x_i, e(j))$, то и ОВ, и распознаваемые сигналы будут достаточно далее представлять не последовательностями элементов $X_i = (x_1, x_2, \dots, x_i, \dots, x_l)$, а только последовательностями $J_i = (j_1, j_2, \dots, j_i, \dots, j_l)$ из номеров j_i элементов $e(j) \in E$, таких, что

$$j_i = \operatorname{argmax}_i g^*(x_i, e(j)). \quad (9.3.7)$$

В целом, убеждаемся, что в рамках КДП-подхода к распознаванию речи элементарная мера сходства наблюдаемого и эталонного элементов может быть синтезирована в процессе обучения распознаванию слов речи. Такая возможность является одним из достоинств КДП-подхода.

§ 9.4. МОДЕЛИ АНАЛИЗА РЕЧЕВОГО СИГНАЛА. ВЫЧИСЛЕНИЕ ПРИЗНАКОВ ТОНАЛЬНОСТИ (ПРИЗНАКА ТОН—ШУМ И ПЕРИОДА ОСНОВНОГО ТОНА)

Отдельную проблему при описании речевых сигналов составляют признаки тональности, вычисляемые на основе значений признака тон — шум и периода основного тона (ОТ). Хотя этому вопросу уде-

ляется значительное внимание (см., например, библиографию в [64—68]), в целом состояние дел нельзя считать удовлетворительным. Поиски надежных и относительно простых способов вычисления признаков тональности продолжаются.

В наших исследованиях в этой области, в частности, была показана целесообразность синхронного с ОТ анализа речевого сигнала, например, по три периода основного тона в интервале анализа [75, 145—147]. Изучались также различные модели совместного оценивания параметров речевого тракта и характеристик источников его возбуждения по речевому сигналу [75, 145—150].

Некоторой обобщающей все эти исследования моделью речевого сигнала на интервале анализа является следующая модель, задаваемая разностным уравнением:

$$f_n = - \sum_{i=1}^m a_i f_{n-i} + B \sum_{k=0}^q \varphi_k \delta_{n-k} + C \sum_{u=0}^r c_u \eta_{n-u} + D \varepsilon_n, \quad (9.4.1)$$

где f_n , $n = 1 : M$ — отсчеты речевого сигнала на интервале анализа (наблюдения); $a = (a_1, a_2, \dots, a_t, \dots, a_m)$ — параметры речевого тракта как линейной системы (эти параметры считаются постоянными для данного интервала анализа); $\varphi = (\varphi_0, \varphi_1, \dots, \varphi_k, \dots, \varphi_q)$ — форма сигнала возбуждения для голосового источника (обычно считается заданной или постоянной на интервале анализа); δ_n — квазипериодическая последовательность импульсов ($\delta_n = 0$ для всех n , кроме точек n_s ; $n_{s+1} - n_s = T_s$, T_s — текущий период ОТ, причем на изменение периода T_s накладывается ограничение $|T_s - T_{s-1}| \leq \Delta$, δ_{n_s} — амплитуда сигнала возбуждения в момент n_s); η_n и ε_n — независимые источники дискретного белого шума с дисперсией σ_η и σ_ε соответственно; B, C и D — параметры, принимающие значения от 0 до 1, например, $B = 1$, если звук звонкий, и $B = 0$, если он шумный; m, q и r — параметры, определяющие порядок системы (обычно $r \leq m < 20$).

В рамках приведенной модели возможно совместное оценивание по речевому сигналу как передаточной характеристики речевого тракта (его полюсов и нулей), так и характеристик источника его возбуждения (мгновенного периода основного тона, амплитуды сигнала возбуждения, признаков типа тон — шум и т. п.). Все эти характеристики передаются искомыми при оценивании параметрами a , φ , $c = (c_0, c_1, \dots, c_u, \dots, c_r)$, $\{n_s\}$, $\{\delta_{n_s}\}$, B , C , D .

Максимально правдоподобные оценки искомых параметров могут быть получены с помощью традиционного для КДГ-подхода метода обобщенной покоординатной оптимизации с использованием динамического программирования для оценки мгновенного периода ОТ [75, 145—150].

Все же возникающие при этом алгоритмы оказываются весьма громоздкими и неудобными для непосредственного использования в технических системах.

Практически оказывается вполне приемлемой процедура раздельного оценивания передаточной характеристики речевого тракта и периода ОТ по речевому сигналу.

Передаточную характеристику и форму импульсов возбуждения речевого тракта в неразделенном виде обычно оценивают по а-параметрам исходя из уравнения

$$f_n = - \sum_{t=1}^m a_t f_{n-t} + \varepsilon_n, \quad (9.4.2)$$

что приводит к эквивалентному описанию элементов речи автокорреляционной функцией или амплитудным спектром (см. § 2.1). Что же касается признаков тональности, то их значения вычисляем исходя из так называемой нулевой модели речевого сигнала [146—147]:

$$f_{n_s+j} = \beta_{n_s} f_{n_s-1+j} + \varepsilon_{n_s+j}, \quad j = 0, 1, 2, \dots \quad (9.4.3)$$

Согласно этой модели сигнал на текущем периоде ОТ является результатом случайного искажения сигнала предыдущего периода ОТ, взятого с определенными амплитудой β_{n_s} и периодом, удовлетворяющим ранее приведенным ограничениям.

Пусть T_{\min} и T_{\max} — соответственно минимальное и максимальное значения периода ОТ, выраженные количеством дискрет f_n на периоде.

Исходя из модели (9.4.3) и полагая форму периода речевого сигнала неизменной на интервале анализа и равной $f_j, j = j^* : (j^* + T_{\min} - 1)$,

где $j^* = \underset{n=1:(M-T_{\min}+1)}{\operatorname{argmax}} \sum_{k=0}^{T_{\min}-1} f_{n+k}^2$, выражение правдоподобия для оценивания текущего периода ОТ $T_s = n_s - n_{s-1}$ можно записать в виде

$$\Phi(\{n_s\}) = \sum_{s=0}^P \left(\sum_{j=0}^{T_{\min}-1} f_{n_s+j} f_{j^*+j} \right)^2, \quad (9.4.4)$$

где $(P + 1)$ — количество периодов, в том числе неполных, на интервале анализа.

В критерии (9.4.4) полагаем, что $f_{n_s+j} = 0$ для $n_s + j < 0$ и $n_s + j > M$.

Оптимизация критерия (9.4.4) по местоположениям n_s периодов ОТ решается с помощью динамического программирования [145—151].

Введем дополнительное ограничение, согласно которому местоположение n_s одного из периодов совпадает с точкой j^* .

На рис. 9.6 представлен график решения некоторой упрощенной задачи для нахождения текущего периода ОТ.

По оси абсцисс отложены возможные значения периода ОТ $T = T_{\min}, T_{\max}$, а по оси ординат — порядковый номер периода $s = 0, 1, 2, 3, \dots$. Рассмотрен случай $|T_s - T_{s-1}| \leq \Delta, \Delta = 2$.

Каждой стрелке, входящей в вершину (T, s) , припишем длину

$$d(T, s) = \left(\sum_{j=0}^{T_{\min}-1} f_{sT+j-n^*(T)} f_{j^*+j} \right)^2, \quad (9.4.5)$$

где полагается локально (это и есть упрощение задачи), что $n_s = sT - n^*(T)$. Величина $n^*(T_{\min})$ равна такому минимальному целому

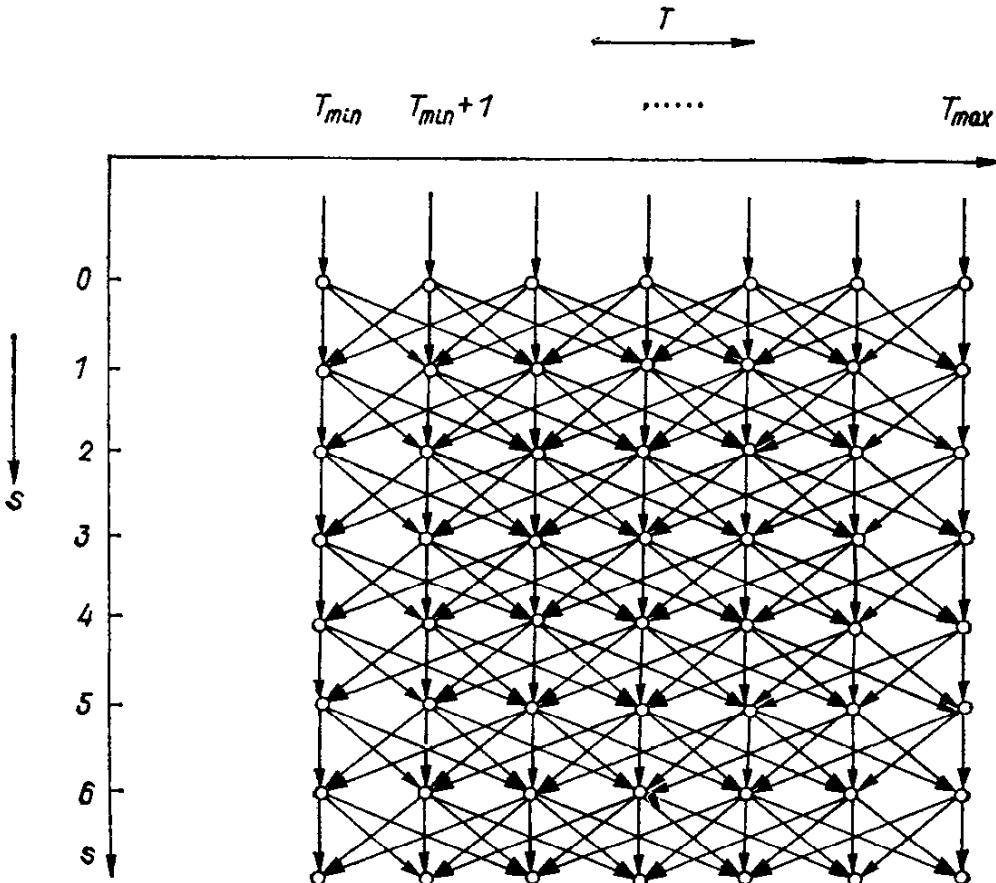


Рис. 9.6. Граф решения задачи о мгновенном периоде основного тона.

положительному числу, что

$$\left[\frac{n^*(T_{\min}) + j^*}{T_{\min}} \right] T_{\min} = n^*(T_{\min}) + j^*,$$

а для остальных T величина $n^*(T)$ выбирается из условия

$$n^*(T) = \left[\frac{n^*(T_{\min}) + j^*}{T_{\min}} \right] T - j^*.$$

Такой выбор смещений $n^*(T)$ гарантирует, что начало n_s одного из периодов лежит в точке j^* .

Максимально возможное значение номера периода $s = 0, 1, \dots, P$ равно

$$P = \frac{\mathcal{M} + n^*(T_{\min}) + 1}{T_{\min}}$$

либо

$$P = \left[\frac{\mathcal{M} + n^*(T_{\min}) + 1}{T_{\min}} \right] + 1$$

в зависимости от того, делится ли $\mathcal{M} + n^*(T_{\min}) + 1$ нацело на T_{\min} или нет.

Для максимизации критерия (9.4.4) на графике рис. 9.6 воспользуемся рекуррентными формулами ДП:

$$F(T, s) = \max_{T'; |T-T'| \leq \Delta} F(T', s-1) + d(T, s), \quad (9.4.6)$$

$$I(T, s) = \underset{T': |T-T'| \leq \Delta}{\operatorname{argmax}} F(T', s-1), \quad (9.4.7)$$

$$T = T_{\min} : T_{\max}, \quad s = 1 : P.$$

При этом для $s = 0$ полагаем

$$F(T, 0) = d(T, 0) \quad I(T, 0) = T.$$

Найдем

$$\Phi^* = \max_{T=T_{\min} : T_{\max}} F(T, P) \quad (9.4.8)$$

и

$$I_{P+1}^* = \underset{T=T_{\min} : T_{\max}}{\operatorname{argmax}} F(T, P). \quad (9.4.9)$$

Оптимальная последовательность периодов T_s^* основного тона определяется следующим образом:

$$T_{s-1}^* = I(T_s^*, s), \quad s = P, P-1, \dots, 2, 1, \quad (9.4.10)$$

причем $T_P^* = I_{P+1}^*$.

Из полученной последовательности (9.4.10) выбросим все первые и все последние члены, которым соответствует

$$d(T_s^*, s) = 0,$$

и для оставшихся членов T_s^* найдем среднее значение.

Это и будет среднее значение периода ОТ на интервале анализа.

Для принятия решения о значении признака тон — шум вычисляется отношение

$$\eta = \frac{\Phi^*}{\sum_{n=1}^M f_n^2 \cdot \sum_{i=0}^{T_{\min}-1} f_{i+1}^2}, \quad (9.4.11)$$

и звук считается тональным, если значение η оказывается больше некоторого порогового значения Θ , либо шумным — в противном случае

Параметр Θ подбирался экспериментально.

Хотя изложенный метод вычисления признаков тональности хорошо зарекомендовал себя в действующих программных моделях вокодеров [146, 147, 151], были предприняты дополнительные усилия по упрощению метода с целью его последующего использования в инженерных разработках.

Представляют интерес, прежде всего, методы, определяющие средний период ОТ без предварительного оценивания мгновенных значений периода.

Приемлемым для использования оказался ВОТ-метод вычисления признаков тональности, использующий отдельные достоинства автокорреляционного метода выделения ОТ [66] и метода SIFT [67].

Заданный на интервале анализа речевой сигнал f_n , $n = 1 : M$, сначала подвергается низкочастотной фильтрации с частотой среза 1500 Гц и далее прореживается с коэффициентом \mathcal{L} . Пусть y_s , $s =$

$= 1 : \mathcal{N} \left(\mathcal{N} = \frac{\mathcal{M}}{\mathcal{L}} \right)$ — прореженный сигнал (расстояние между дискретами y_s равно $\mathcal{L}\Delta t$).

Сигнал y_s , $s = 1 : \mathcal{N}$, подвергается операции выравнивания по амплитудному спектру, в результате которой огибающая амплитудного спектра преобразованного сигнала становится линией, параллельной оси частот.

С этой целью сначала вычисляется автокорреляционная функция сигнала y_s , $s = 1 : \mathcal{N}$:

$$B(v) = \sum_{s=1}^{\mathcal{N}-v} y_s y_{s+v}, \quad v = 0 : (\mathcal{N} - 1), \quad (9.4.12)$$

а затем определяется фильтр $a = (a_1, a_2, \dots, a_r, \dots, a_m)$ путем решения системы уравнений

$$\sum_{r=1}^m a_r B(|v-r|) = -B(v), \quad v = 1 : m, \quad m = 4. \quad (9.4.13)$$

Сигнал \tilde{y}_s с выравненным спектром вычисляется как

$$\tilde{y}_s = \sum_{r=0}^m a_r y_{s-r}, \quad a_0 \equiv 1, \quad s = 1 : (\mathcal{N} + m), \quad (9.4.14)$$

где полагается $y_{s-r} \equiv 0$ для $s-r < 1$ и $s-r > \mathcal{N}$.

Автокорреляционную функцию $\tilde{B}(v)$ сигнала \tilde{y}_s можно рассчитать, пользуясь только $B(v)$, $v = 0 : (\mathcal{N} - 1)$, и b -параметрами фильтра $a = (a_1, a_2, \dots, a_r, \dots, a_m)$:

$$\begin{aligned} b &= (b_0, b_1, b_2, \dots, b_\mu, \dots, b_m), \\ b_0 &= \sum_{\mu=0}^m a_r^2, \quad b_\mu = \sum_{r=0}^{m-\mu} a_r a_{r+\mu}, \quad a_0 \equiv 1, \quad \mu = 1 : m; \end{aligned} \quad (9.4.15)$$

$$\tilde{B}(v) = \sum_{\mu=0}^m b_\mu (B(|v-\mu|) + B(|v+\mu|)), \quad v = 0 : (\mathcal{N} + m - 1), \quad (9.4.16)$$

где полагается $B(|v|) \equiv 0$ для $|v| \geq \mathcal{N}$.

Поскольку сигнал \tilde{y}_s является результатом «выбеливания» исходного сигнала y_s , то \tilde{y}_s — это сигнал со свойствами дискретного «белого» шума. Это квазипериодический «белый» шум, если сигнал y_s квазипериодический, или просто «белый» шум, если сигнал y_s шумный.

Соответственно, автокорреляционная функция $\tilde{B}(v)$ квазипериодического «белого» шума имеет существенные (соизмеримые с $\tilde{B}(0)$) значения в точках, близких к

$$v = v_k, \quad v_k = kT', \quad k = 1, 2, \dots, Q, \quad (9.4.17)$$

где T' — среднее значение периода ОТ, выраженное в количестве дискрет прореженного сигнала y_s . В случае же просто «белого» шума

значения $\tilde{B}(v)$ в упомянутых точках будут малы по сравнению с $\tilde{B}(0)$. Очевидно, что $Q = \left\lceil \frac{\mathcal{N} + m - 1}{T'} \right\rceil$.

С целью выражения среднего периода ОТ в количестве дискрет исходного непрореженного сигнала и для повышения точности измерения периода ОТ далее будем действовать следующим образом.

Пусть T'_{\min} и T'_{\max} — соответственно минимальное и максимальное значения периода ОТ в количестве прореженных дискрет (например, $T'_{\min} = 8$, $T'_{\max} = 40$).

Найдем среди массива $\tilde{B}(v)$, $v = T'_{\min} : T'_{\max}$, несколько, например $\mathcal{W} = 7$, наибольших значений $\tilde{B}(v)$. Пусть v_r^1 , $r = 1 : \mathcal{W}$, — соответствующие значения v .

Обозначим $\Phi_r^1 = \tilde{B}(v_r^1)$, $r = 1 : \mathcal{W}$.

Составим далее суммы

$$\Phi_r^2 = \Phi_r^1 + \max_{u=-1, 0, 1} \tilde{B}(v_r^1 + v_r^1 + u), \quad (9.4.18)$$

$$v_r^2 = v_r^1 + v_r^1 + \operatorname{argmax}_{u=-1, 0, 1} \tilde{B}(v_r^1 + v_r^1 + u), \quad r = 1 : \mathcal{W}. \quad (9.4.19)$$

Вычисления, подобные (9.4.18)–(9.4.19), повторим для каждого r несколько раз

$$\Phi_r^s = \Phi_r^{s-1} + \max_{u=-1, 0, 1} \tilde{B}(v_r^{s-1} + v_r^1 + u), \quad (9.4.20)$$

$$v_r^s = v_r^{s-1} + v_r^1 + \operatorname{argmax}_{u=-1, 0, 1} \tilde{B}(v_r^{s-1} + v_r^1 + u) \quad (9.4.21)$$

и остановимся для данного r при таком $s = Q_r$, что $v_r^s \leq \mathcal{N} + m - 1$, а $v_r^s + v_r^1 \geq \mathcal{N} + m$.

Далее найдем

$$r^* = \operatorname{argmax}_{r=1:\mathcal{W}} \frac{\Phi_r^{Q_r}}{Q_r} \quad (9.4.22)$$

и средний период

$$T = \left[\frac{v_{r^*}^{Q_{r^*}} \mathcal{L}}{Q_{r^*}} \right], \quad (9.4.23)$$

выраженный уже в количестве непрореженных дискрет.

Наконец, величина $\eta_{r^*} = \Phi_{r^*}^{Q_{r^*}} / Q_{r^*}$ сравнивается с $\tilde{B}(0)$ и, если выполняется

$$\kappa = \frac{\eta_{r^*}}{\tilde{B}(0)} < \Theta_1, \quad (9.4.24)$$

где Θ_1 — некоторый экспериментально подбираемый порог, то звук на интервале анализа объявляется как шумный (признак тон — шум принимает значение 0). В противном случае он объявляется звонким (признак тон — шум принимает значение 1), а период ОТ полагается равным T (см. формулу (9.4.23)).

Приведенный ВОТ-метод вычисления признаков тональности требует существенно меньше вычислений, чем метод, основанный на динамическом программировании.

Экспериментальные исследования показали, что ВОТ-метод не уступает по надежности и точности ДП-методу и обеспечивает восстановление речи с лучшим качеством, чем при использовании SIFT-метода.

Важным резервом повышения надежности выделения признака тон—шум и точности вычисления периода ОТ явилась регуляризация этих параметров для соседних интервалов анализа.

Во-первых, если тональный интервал анализа оказывался среди трех соседних шумных (двух предшествующих и одного последующего), то он объявлялся шумным. Аналогично поступаем с шумным интервалом среди трех соседних тональных — объявляем его тональным.

Во-вторых, движение периода ОТ на 3—5 соседних звонких интервалах можно регуляризовать линейными или квадратичными функциями, пользуясь данными измерений для отдельных интервалов анализа.

Исследованные методы вычисления признаков тональности нашли применение в распознавании речи и особенно в системах компрессированной передачи речи [152, 153].

§ 9.5. ОБЩАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РЕЧЕВОГО СИГНАЛА. НЕОБХОДИМОСТЬ В КУСОЧНО-ЛИНЕЙНЫХ МОДЕЛЯХ

Способ анализа речевых сигналов, заключающийся в независимом анализе сигналов на перемещаемом с шагом ΔT интервале анализа продолжительностью $\Delta T'$, выглядит слабо обоснованным. Во-первых, потому, что сами-то речевые сигналы на соседних интервалах анализа являются зависимыми. Во-вторых, случайное равномерное расположение интервалов анализа на оси времени плохо согласуется со структурой речевых сигналов и приводит к нестабильности результатов анализа. В-третьих, этот способ анализа не учитывает свойства моделей речевых сигналов (в частности, кусочно-постоянной модели), применяемых на уровне распознавания (см. § 2.2 гл. 1). Напомним, что кусочно-постоянная модель предполагает, что на стационарных участках звуков элементы речи x_i повторяются, а на переходных — эта повторяемость сведена к минимуму — образуются цепочки из элементов, в которых каждый элемент повторяется не более одного раза. Из-за эффектов дискретизации элементов, даже при абсолютной адекватности кусочно-постоянной модели, будет плохая воспроизведение результатов анализа — вычисления элементов (цепочек из

элементов) и на стационарных, и особенно на переходных участках.

Стабильность и воспроизводимость результатов анализа можно обеспечить, если осуществлять зависимый для соседних интервалов анализ, привлекая различную априорную информацию о структуре и свойствах речевых сигналов. Так, на стационарных участках результаты анализа должны быть близкими, а на переходных — последовательности элементов должны подчиняться определенным закономерностям, отражающим коартикуляцию звуков. Это в конечном счете приводит к заключению о необходимости совмещения предварительной обработки и распознавания в некотором едином процессе (гл. I).

Однако, оставаясь в рамках раздельного решения задач предварительной обработки и распознавания, в принципе можно осуществить зависимый анализ сигналов и на уровне предварительной обработки с сохранением кусочно-постоянной модели на уровне распознавания.

Внимания заслуживает кусочно-линейная модель речевых сигналов, которая в простейшем случае предполагает, что если стационарные части двух звуков передаются соответственно элементами e_1 и e_2 , то переходный участок одного звука в другой (например, e_1 в e_2) задается уравнением

$$e_t = \frac{t}{T_{12}} e_1 + \left(1 - \frac{t}{T_{12}}\right) e_2, \quad (9.5.1)$$

где $t = 0 : T_{12}$ — дискретное время; T_{12} — длительность переходного участка для перехода звука e_1 в e_2 , выраженная в дискретном времени.

Чтобы задать стационарный звук, характеризующийся элементом e , достаточно в (9.5.1) положить $e_1 = e_2 = e$, а T_{12} принять равным возможному значению T длительности звука.

Кусочно-линейная модель более явно, чем кусочно-постоянная, указывает на необходимость зависимого анализа для соседних интервалов. Более того, она определяет характер этой зависимости: получаемые в результате анализа цепочки из элементов должны удовлетворять уравнениям, подобным (9.5.1).

В кусочно-линейной модели переходные участки одного звука в другой передаются всего двумя элементами e_1 и e_2 и одним параметром длительности T , а промежуточные элементы должны рассчитываться по формуле линейной интерполяции. В то же время в кусочно-постоянной модели эти переходные участки задаются последовательностями элементов. Таким образом, кусочно-линейная модель более компактна, но требует дополнительных вычислений (интерполяции).

Зависимый анализ речевого сигнала в условиях кусочно-линейной модели предполагает оценивание сравнительно небольшого количества элементов и параметров, существенно меньшего, чем в случае независимого анализа. Это дает основание надеяться на хорошие стабильность и воспроизводимость результатов при зависимом кусочно-линейном анализе. Значительное улучшение стабильности ожидается прежде всего для переходных участков.

Апеллирование к зависимому анализу сигналов приводит к необходимости совмещения математических моделей речевого сигнала,

используемых как на уровне предварительной обработки, так и на уровне распознавания. В самом деле, если модели на уровне предварительной обработки (см., например, § 9.4) указывают на связь параметров анализа только внутри одного интервала, то связи возможных значений этих параметров в последовательностях элементов уже выражаются в моделях на уровне распознавания.

Проиллюстрируем, как можно осуществить совместный процесс предварительной обработки и распознавания на примере распознавания слов, сравнивая одновременно кусочно-постоянную и кусочно-линейную модели.

Поскольку и распознавание, и анализ объединены в одном процессе, значит, теперь будет обрабатываться сигнал $f_n, n = 1 : \mathcal{L}$, заданный отсчетами (\mathcal{L} — количество отсчетов f_n) на выходе микрофона, без какого-либо предварительного разбиения на интервалы анализа.

Кусочно-постоянная модель, независимо от того, какой метод распознавания используется (поэлементный, пофонемный, глубокий пофонемный), предполагает, что каждое слово k в конечном счете задается своим исходным эталонным сигналом E_k и темпоральной транскрипцией τ_k (подчеркнем, что теперь темпоральная транскрипция выражена в дискретном времени с шагом Δt между дискретами f_n). Как эталон E_k , так и транскрипция τ_k при кусочно-постоянной модели только набираются (составляются) с помощью графов или транскрипций.

В случае кусочно-линейной модели исходный эталон слова E_k и темпоральная транскрипция слова τ_k частично набираются, а частично вычисляются, причем одинаково во всех методах: поэлементном, пофонемном, глубоком пофонемном. В самом деле, в случае кусочно-линейной модели исходный эталон слова E_k (аналогично граф слова и транскрипция τ_k) может быть рассчитан путем линейной интерполяции (см. формулу (9.5.1)) по некоторым первоначальным заготовкам, задающим информацию о словах более экономно, чем в случае кусочно-постоянной модели.

Таким образом, использование кусочно-постоянной и кусочно-линейной моделей принципиально не отличается. Это всего лишь несколько разные средства задания одного и того же. Различие сводится к вопросу о том, что важно в данный момент экономить память или вычисления.

Более глубокое отличие имеет место лишь при постановке и решении соответствующих задач обучения.

Зададим исходные эталонные элементы с помощью a -параметров (§ 2.1) и, помня о том, что мы отказались от первичного анализа, сформулируем задачу распознавания отдельно произносимых слов в следующем виде (по аналогии с § 2.3 и § 3.2);

$$k(f_n, n = 1 : \mathcal{L}) = \operatorname{argmax}_k \sum_{\{w_{ks}\}} \sum_{s=1}^{q_k} \sum_{v=w_{k(s-1)}+1}^{w_{ks}} \left(f_v + \sum_{r=1}^m (e_{ks})_r f_{v-r} \right)^2, \quad (9.5.2)$$

где $f_n, n = 1 : \mathcal{L}$ — распознаваемый сигнал; e_{ks} — s -й эталонный элемент k -го слова; $(e_{ks})_r$ — его r -я компонента ($r = 1 : m$); w_{ks} — верх-

ная граница s -го сегмента реализации, аппроксимируемого s -м эталонным элементом k -го слова ($w_{k0} = 0$, $w_{kq_k} = \mathcal{L}$), ограничения на w_{ks} задаются темпоральной транскрипцией слова $v_{ks} = w_{ks} - w_{k(s-1)}$, $m_{ks} \leq v_{ks} \leq M_{ks}$, $s = 1 : q_k$; q_k — длина исходного эталона k -го слова. В (9.5.2) полагается также, что $f_{v-r} \equiv 0$ для $v - r < 1$.

В соответствии с критерием (9.5.2) границы сегментов w_{ks} находятся с точностью до одной дискреты f_n речевого сигнала.

Из (9.5.2) следует, что если осуществлять распознавание речевого сигнала на уровне дискрет f_n , т. е. совмещать предварительную обработку с распознаванием, то автоматически снимаются проблемы выбора интервалов анализа и их размещения на оси времени. Однако за это необходимо платить значительным увеличением объема вычислений. Так, формулы (2.3.7) — (2.3.10) надо будет применить \mathcal{L} раз вместо $[\mathcal{L}/M]$, где M — шаг ΔT интервала анализа, выраженный в количестве дискрет.

Задачи, подобные (9.5.2), будем называть сформулированными на уровне общих математических моделей речевого сигнала. Это все те же математические модели, обычно используемые на уровне распознавания, однако теперь уже выраженные на уровне дискрет.

Использование общих математических моделей снимает многие проблемы предварительной обработки речевого сигнала.

На основе общих моделей можно сформулировать более точные, чем в (9.5.2), критерии для сравнения сегмента речевого сигнала f_n , $n = M_1 : M_2$ (M_1 и M_2 — границы сегмента) с эталонными элементами. Так, отправляясь от уравнения (9.4.1), нетрудно записать критерий, аналогичный используемому в (9.5.2) и учитывающий дополнительную информацию об интенсивности и тональности эталонных элементов, способе их образования. В любом случае, однако, необходимо будет производить оптимизацию по границам сегментов с точностью до одной дискреты.

Основной целью данного параграфа было показать, что выбор и дискретизация интервалов анализа должны быть увязаны с моделью речевых сигналов, используемой на уровне распознавания, и что эта увязка неизбежно приводит к совмещению предварительной обработки и распознавания в одном процессе, т. е. к использованию общих математических моделей речевого сигнала [150].

Одновременно мы убедились, что нет принципиальной разницы в использовании кусочно-постоянной и кусочно-линейной моделей. Есть различие лишь в технике вычисления исходных эталонов и транскрипций слов. Можно обратить внимание на то, что в отличие от кусочно-постоянной модели кусочно-линейная пригодна не для всякого описания речевого сигнала. Так, спектральное и автокорреляционное описание элементов e_1 и e_2 в модели (9.5.1) не обеспечивают адекватных плавных переходов формант. Не удовлетворяет этому требованию описание с помощью параметров предсказания. Наиболее адекватными в обсуждаемом смысле являются описания с помощью коэффициентов отражения (§ 2.1) или формант [125].

ВЫВОДЫ

1. Показано, что анализ речевых сигналов интервалами продолжительностью $\Delta T' = 10\text{--}20$ мс и дискретизация результатов анализа с шагом $\Delta T = 10\text{--}20$ мс характеризуются значительной нестабильностью. Наилучшую стабильность анализа обеспечивают интервалы анализа продолжительностью $\Delta T' = 30\text{--}50$ мс с шагом дискретизации $\Delta T = 10\text{--}20$ мс. Однако компромиссом между стабильностью анализа и разрешающей способностью анализа во времени, обеспечивающим наилучшую надежность распознавания, являются интервалы анализа продолжительностью менее 30 мс с шагом дискретизации $\Delta T = 15$ мс.

2. Надежность распознавания речи в значительной степени зависит от используемой элементарной меры сходства. Экспериментально показано, что меры сходства, основанные на использовании евклидового расстояния между a - или b -параметрами предсказания, обеспечивают относительно низкую надежность распознавания. Среди мер, рекомендуемых для использования, можно выделить такие, которые основаны на хэмминговом расстоянии между кодами и скалярном произведении вектора автокорреляции речевого сигнала и вектора b -параметров предсказания.

Предложен итерационный алгоритм синтеза табличной элементарной меры сходства в процессе обучения распознаванию речи.

3. Анализ речевых сигналов сводится к совместному или раздельному оцениванию по речевому сигналу как передаточной характеристики речевого тракта (его полюсов и нулей), так и характеристик источников его возбуждения (мгновенного периода ОТ, амплитуды сигнала возбуждения, признаков типа тон — шум и т. п.).

Вычисление признаков тональности составляет одну из трудных задач первичного анализа речевых сигналов. Разработаны нулевой метод и ВОТ-методы выделения признаков тональности, основанные на использовании глобальных или локальных процедур динамического программирования и характеризующиеся высокими надежностью и точностью выделения признака тон — шум и периода ОТ и приемлемой для практики трудоемкостью.

4. Показано, что проблемы первичного анализа, в том числе такие, как выбор и дискретизация интервалов анализа, автоматически разрешаются, если распознавание речи вести не на уровне результатов первичного анализа, а на уровне дискрет сигнала микрофона.

Надежный первичный анализ можно проводить, если его согласовывать с математическими моделями речевых сигналов, используемыми на уровне распознавания. Последовательная реализация принципов согласования приводит к необходимости совмещения предварительной обработки и распознавания в едином процессе, при котором распознавание следует вести на уровне дискрет сигнала микрофона.

Показана перспективность кусочно-линейной модели речевого сигнала в распознавании речи.

ГЛАВА 10

НИЗКОСКОРОСТНЫЕ СИСТЕМЫ КОМПРЕССИРОВАННОЙ ПЕРЕДАЧИ РЕЧИ

Системы компрессированной передачи речи (вокодерные системы) занимают промежуточное положение между системами распознавания и системами синтеза речи. В последние годы, особенно в связи с использованием низкоскоростных интегрированных цифровых систем связи, интерес к разработке компрессированных систем передачи речи возник с новой силой.

В прошлое десятилетие уже появились коммерческие вокодерные системы на информационные скорости от 9600 до 2400 бит/с [64—68]. Среди отечественных разработок можно выделить гармонический вокодер на 4800—2400 бит/с [65]. При всех успехах по-прежнему остро стоят проблемы снижения информационной скорости передачи информации (до 600 бит/с и ниже), увеличения разборчивости, качества и натуральности восстанавливаемой речи, проблема создания малогабаритной цифровой аппаратуры.

Оказывается, что чем ниже информационная скорость вокодера, тем теснее связь вокодерной передачи речи с распознаванием речи. Общими становятся не только анализаторы речи, все в большей степени анализирующая часть вокодера напоминает систему распознавания.

В настоящей главе показано, что в рамках КДП-подхода могут быть предложены низкоскоростные системы компрессированной передачи речи на информационную скорость до 600 бит/с и ниже. В работах по созданию компрессированных систем передачи речи активное участие принимал Е. К. Людовик.

§ 10.1. ПОСТАНОВКА ЗАДАЧИ

Речевой сигнал на выходе микрофона характеризуется информационной скоростью около 250 000—300 000 бит/с. В самом деле, при верхней граничной частоте речевого сигнала около 12 кГц и разрядности преобразователя аналог-код 10—12 бит получим как раз эти величины. Для телефонного речевого сигнала информационная скорость уменьшается приблизительно в два раза.

В связи с вводом в эксплуатацию цифровых интегрированных сетей связи со стандартной скоростью 2400 бит/с существует необходимость

передачи речевой информации по этим каналам в реальном масштабе времени, с сохранением разборчивости, качества и натуральности звучания, с сохранением индивидуальных особенностей речи. Более того, желательно по каналу связи передавать одновременно разговоры нескольких лиц. Речь идет, таким образом, о сжатии объема исходного речевого сигнала до 1 200, 600, 300 и 150 бит/с.

Представляется, что 150 бит/с составляет некоторое предельное сжатие объема речевого сигнала, при котором еще можно надеяться на передачу индивидуальных особенностей голоса.

На приемном конце линии связи, очевидно, должны быть приняты меры по восстановлению первоначального объема передаваемого речевого сигнала. Эту часть системы компрессированной передачи речи принято называть синтезатором речи (синтезирующей частью вокодера) в отличие от анализатора речи (анализирующей части вокодера), располагаемого на передающем конце линии связи.

Принципиальная возможность компрессированной передачи речи вытекает из того, что, хотя речевой сигнал на выходе микрофона и описывается быстроосциллирующими функциями, сама же динамика передаточной характеристики речевого тракта человека и параметры источников его возбуждения описываются медленно изменяющимися функциями времени, которые как раз и следует передавать при компрессии речи.

Практически это означает, что речевой сигнал разбивается, например, на примыкающие друг к другу интервалы анализа продолжительностью $\Delta T' = 25$ мс, что составляет 40 интервалов (кадров) на секунду речи. Внутри одного кадра (интервала анализа), полагая, что передаточная характеристика речевого тракта и параметры источников его возбуждения не меняются, как раз и оценивают и передают в линию связи некоторые величины, представляющие параметры речевого тракта и источников его возбуждения. Традиционно это приводило к 240—60 битам информации на один кадр [64—68].

Далее будет показано [152, 153], как, применяя процедуры распознавания в анализирующей части вокодера, можно снизить информационную скорость вокодера до 600 бит/с (до 15 бит информации на один кадр).

Дальнейшее снижение информационной скорости возможно за счет использования математических моделей речевых сигналов, применяемых обычно при распознавании речи в рамках КДП-подхода.

Так, используя кусочно-постоянную модель (§ 2.2), согласно которой на квазистационарных участках звуков элементы речи, представляющие отдельные кадры, не меняются, можно сократить информационную скорость в среднем не менее чем в два раза, если применить предварительную сегментацию речевого сигнала на квазистационарные участки (сегменты) (§ 3.2) и в линию связи передавать один кадр, представляющий сегмент, и количество его повторений в сегменте.

Еще более значительное сокращение информационной скорости (до 150 бит/с) может быть получено, если использовать кусочно-линейную модель (§ 9.5). Согласно этой модели следует применить предварительную кусочно-линейную сегментацию речевого сигнала, а

затем в линию связи передавать по два элемента (кадра) на сегмент и количество элементов в сегменте. Тем самым будет достигнута значительная экономия не только на стационарных, но и на переходных участках звуков друг в друга.

Что касается синтезирующей части, то недостающие кадры сегментов должны быть восстановлены путем линейного интерполяирования по двум переданным элементам сегментов.

Наконец, можно ограничиться передачей не самих элементов (кадров), а их номеров из заданного множества эталонных элементов, каждый раз передавая номер того эталонного элемента (кадра), который в том или ином смысле наиболее похож на наблюдаемый элемент.

Таким образом, показано, что применение процедур и моделей распознавания приводит к предельному сжатию объемов передаваемой информации до 150 бит/с.

§ 10.2. НУЛЬ-ПОЛЮСНЫЕ ВОКОДЕРЫ НА 2 400 И 1 200 БИТ/С

Созданию квазифонемного вокодера на 600 бит/с предшествовали работы по нуль-полюсной модели анализа и восстановления речи [75, 146, 147]. Эти работы, прежде всего, проводились с целью изучения различных описаний (представлений) речевого сигнала.

Исходный речевой сигнал представлялся последовательностью дискрет f_n , $n = 1 : M$, где M — количество дискрет на интервале анализа продолжительностью $\Delta T'$. Пусть Δt — шаг отсчета дискрет f_n .

Полюсное представление речевого сигнала основывалось на z -передаточной характеристике речевого тракта в виде

$$H(z) = \frac{1}{1 + \sum_{i=1}^m a_i z^{-i}} = \frac{1}{\prod_{i=1}^{m/2} (1 + \mathcal{A}_i z^{-1} + \mathcal{B}_i z^{-2})}, \quad (10.2.1)$$

где $a = (a_1, a_2, \dots, a_t, \dots, a_m)$ — обычные параметры предсказания; $(\mathcal{A}_i, \mathcal{B}_i)$ — полюсные параметры, $i = 1 : m/2$, m выбрано четным. Параметры $(\mathcal{A}_i, \mathcal{B}_i)$ названы полюсными, так как непосредственно определяют резонансные свойства речевого тракта. Так, если корни уравнения $(1 + \mathcal{A}_i z^{-1} + \mathcal{B}_i z^{-2}) = 0$ являются комплексно-сопряженными, то круговая резонансная частота ω_i и круговая полуполоса пропускания σ_i связаны с \mathcal{A}_i и \mathcal{B}_i соотношениями:

$$\mathcal{A}_i = -2e^{-\sigma_i \Delta t} \cos \omega_i \Delta t, \quad (10.2.2)$$

$$\mathcal{B}_i = e^{-2\sigma_i \Delta t}. \quad (10.2.3)$$

При оценке полюсных параметров исходим из того, что сигнал f_n , $n = 1 : M$, должен удовлетворять стохастическому разностному уравнению

$$f_n = - \sum_{i=1}^m a_i f_{n-i} + e_n, \quad n = (m+1) : M, \quad (10.2.4)$$

причем f_n , $n = 1 : m$, рассматриваются как начальные условия.

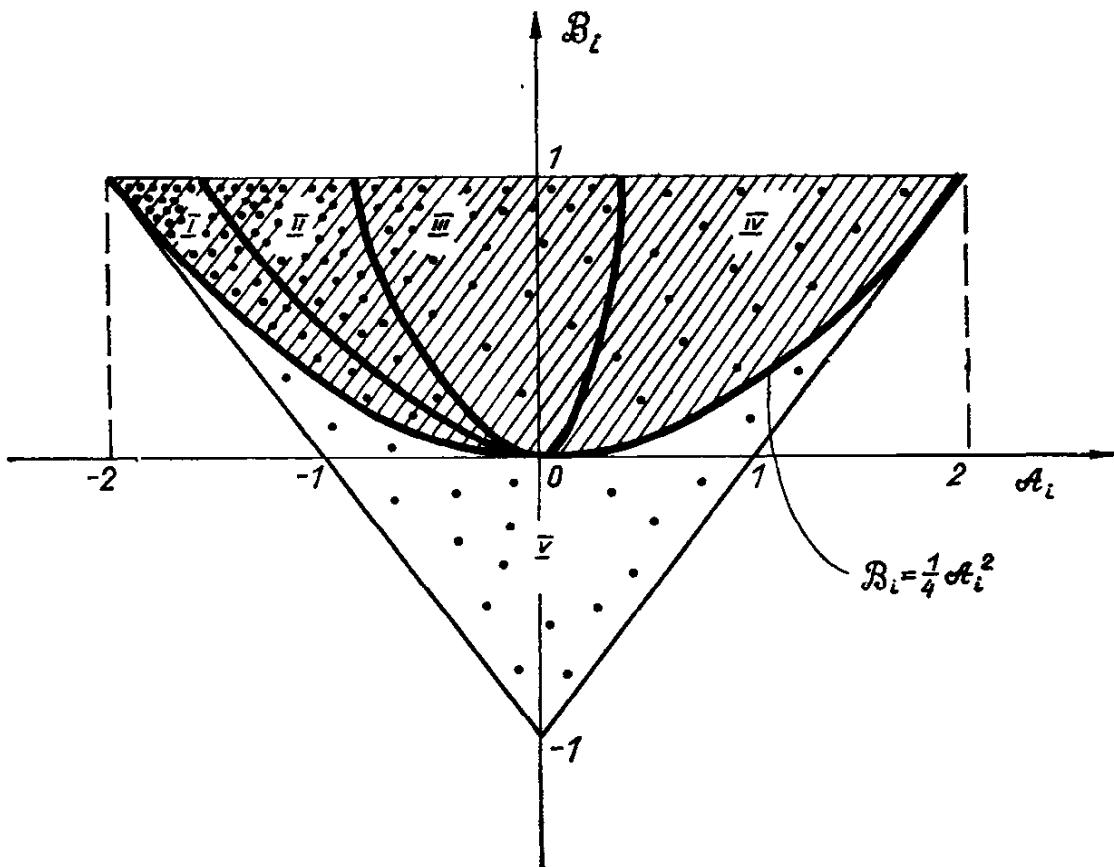


Рис. 10.1. Треугольник устойчивости для полюсных параметров.

В [75] предложен итерационный алгоритм максимально правдоподобного оценивания полюсных параметров (A_i, B_i) , $i = 1 : m/2$, по речевому сигналу f_n , $n = (m + 1) : M$, в условиях связей (10.2.1), (10.2.4), начальных условий f_n , $n = 1 : m$, и ограничений устойчивости [154]:

$$\begin{aligned} 1 + A_i + B_i &> 0, \\ 1 - B_i &> 0, \\ 1 - A_i + B_i &> 0, \quad i = 1 : m/2. \end{aligned} \quad (10.2.5)$$

Соблюдение последнего условия необходимо для гарантирования восстановления (синтеза) речи по полюсным параметрам.

Условия (10.2.5) определяют треугольник устойчивости полюсных параметров (рис. 10.1). Заштрихованная часть треугольника соответствует системам второго порядка с резонансными свойствами (σ_i, ω_i) , определяемым посредством (10.2.2), (10.2.3).

Каждая итерация алгоритма состоит из $m/2$ шагов, а каждый шаг заключается в нахождении максимально правдоподобных оценок одной пары полюсных параметров (A_i, B_i) при условии фиксированных остальных пар и ограничений (10.2.5) на параметры (A_i, B_i) [75].

Использование полюсных параметров удобно при квантовании и кодировании. Так, квантование A_i (соответственно квантование B_i) для различных $i = 1 : m/2$ может быть выполнено одним и тем же способом и на любое требуемое количество бит.

В частности, треугольник устойчивости можно разбить на пять пересекающихся или непересекающихся областей, причем первые

четыре из них могут быть интерпретированы как области первых четырех формант (см. рис. 10.1). Каждая из областей представлялась дискретным набором пар $(\mathcal{A}_i, \mathcal{B}_i)$. Области и значения пар внутри областей нумеруются.

При анализе сигнала f_n , $n = 1 : M$, необходимо из каждой области выбрать по одной паре полюсных параметров, таких, чтобы они составляли максимально правдоподобную (среди возможного выбора пар) оценку полюсных параметров. В такой постановке мы фактически определили некоторую процедуру совместного оценивания и квантования полюсных параметров.

Количество информации на кодирование пары $(\mathcal{A}_i, \mathcal{B}_i)$ внутри каждой области определяется количеством дискретных наборов пар в области и треугольнике в целом.

Сами же дискретные наборы пар внутри областей выбирались так, чтобы они в определенном смысле наилучшим образом аппроксимировали реально встретившиеся значения полюсных параметров. Это своеобразная задача самообучения (таксономии, развода на кучи).

Информация о дискретных наборах полюсных параметров внутри всех пяти областей закладывалась также и в синтезирующую часть вокодера.

Для определения значений признака тон-шум и вычисления периода ОТ использовался нулевой метод (§ 9.4), основанный на динамическом программировании. В этом методе полагается, что сигнал на текущем периоде ОТ является случайным искажением сигнала предыдущего периода, взятого с некоторой амплитудой и другим периодом. Нулевой метод позволяет находить текущее значение периода ОТ. Это свойство метода позволяет, наряду с постоянным интервалом анализа, реализовать синхронный с периодом ОТ анализ речевого сигнала, например, по три периода в интервале анализа.

При кодировании информации кадра 1 бит информации использовался для представления признака тон-шум и 4 бита — для представления среднего периода ОТ (точнее, приращения среднего периода ОТ). При этом полное значение периода ОТ в синтезаторе речи определялось 8-ю битами. Восьмибитовый код периода ОТ для звонкого элемента (кадра) после шумного элемента (кадра) передавался 4-мя разрядами периода ОТ на предшествующем шумном элементе (4 старших разряда) и 4-мя разрядами периода ОТ текущего звонкого элемента (4 младших разряда). На звонком же элементе после звонкого элемента полный период ОТ определялся добавлением текущего приращения периода к значению полного периода на предыдущем интервале синтеза.

На основе нуль-полюсной модели анализа речевого сигнала были созданы машинные модели компрессированной передачи речи на скорость 9600, 4800, 2400 и 1200 бит/с. Частота дискретизации сигналов равнялась 16 кГц ($\Delta t = 65$ мкс). Длина интервала анализа $\Delta T'$ и шаг анализа ΔT были равны и выбирались из диапазона 20—30 мс. Применились синхронный (по три периода в интервале анализа) с ОТ и несинхронный (с постоянной длительностью $\Delta T'$) анализы. Для скоростей 9600 и 4800 бит/с применялось равномерное квантование по-

полюсных параметров \mathcal{A}_i и \mathcal{B}_i с учетом того, что $-2 < \mathcal{A}_i < 2$ и $-1 < \mathcal{B}_i < 1$.

Для скоростей 2400 и 1200 бит/с линейное квантование параметров оказалось неприемлемым. Поэтому применялись дискретные пронумерованные наборы $(\mathcal{A}_i, \mathcal{B}_i)$ для пяти пересекающихся областей.

Признак тон-шум и период ОТ кодировались 5-ю битами. Амплитуда сигнала возбуждения задавалась 5-ю битами для скоростей 9600 и 4800 бит/с и 2-я битами для скоростей 2400 и 1200 бит/с. В обоих случаях использовалась равномерная логарифмическая шкала, так что код амплитуды возбуждения определялся выражением

$$W = \text{INT}(\ln(\mathcal{D}/\mathcal{D}_0)/\Delta\mathcal{D}), \quad (10.2.6)$$

где \mathcal{D} — неквантованное значение амплитуды возбуждения, равное среднеквадратичному значению сигнала ошибки предсказания ϵ_n в формуле (10.2.4):

$$\mathcal{D} = \sqrt{\frac{1}{m-m} \sum_{n=m+1}^M \left(f_n + \sum_{i=1}^m a_i f_{n-i} \right)^2}; \quad (10.2.7)$$

параметры $\mathbf{a} = (a_1, a_2, \dots, a_i, \dots, a_m)$ вычислялись на основании (10.2.1) по полюсным параметрам $(\mathcal{A}_i, \mathcal{B}_i)$, $i = 1 \cdot m/2$; m изменялось от 6 до 16, в качестве основного значения бралось $m = 10$; \mathcal{D}_0 — некоторое пороговое значение (выбиралось отдельно для звонких и шумных кадров); $\Delta\mathcal{D}$ — шаг квантования (также выбирался отдельно для звонких и шумных кадров); INT — выделение целой части числа

Значения параметров \mathcal{D}_0 и $\Delta\mathcal{D}$ для разных скоростей и типов кадров приведены в табл. 10.1

Экспериментальные исследования в условиях внешних акустических шумов и помех с уровнем 75 дБ показали, что качество восстановленной речи на скоростях 9600 и 4800 бит/с незначительно отличается от качества речи, восстановленной по неквантованным полюсным параметрам. При этом разборчивость и натуральность вокодерной речи были достаточно высоки. Натуральность и качество восстановленной речи были лучше при синхронном с ОТ анализе.

При моделировании вокодеров на скорости 2400 и 1200 бит/с применялись совместное оценивание и квантование полюсных параметров, что сводилось к выбору одного наиболее правдоподобного стандартного набора $(\mathcal{A}_i, \mathcal{B}_i)$ в каждой из пяти областей (рис. 10.1). Так, для скорости 1200 бит/с в областях первой, второй, третьей и четвертой формант было соответственно 16, 8, 16 и 8 типовых стандартных значений $(\mathcal{A}_i, \mathcal{B}_i)$, в пятой области было 16 стандартных пар $(\mathcal{A}_i, \mathcal{B}_i)$.

Речь, восстановленная вокодерами со скоростями 2400 и 1200 бит/с, была вполне разборчива, хотя по качеству и натуральности заметно уступала речи для вокодера в 9600 бит/с.

Таблица 10.1. Пороги квантования сигналов возбуждения

Информационная скорость, бит/с	Звонкие кадры		Шумные кадры	
	\mathcal{D}_0	$\Delta\mathcal{D}$	\mathcal{D}_0	$\Delta\mathcal{D}$
9600—4800	0,010	1,07	0,025	1,08
2400—1200	0,015	1,50	0,030	1,60

При восстановлении речи в синтезирующей части всех вокодеров осуществлялся переход от полюсных параметров к параметрам предсказания и непосредственный синтез речи по этим параметрам согласно уравнению (10.2.4).

Эксперименты с нуль-полюсными вокодерами показали перспективность использования полюсных параметров, приемлемость нулевого метода выделения признаков тональности, целесообразность совмещения процедур оценивания и квантования параметров в одном процессе.

Выполненные работы позволили приступить к разработке квазифонемного вокодера на скорость 600 бит/с и меньше.

§ 10.3. КВАЗИФОНЕМНЫЙ ВОКОДЕР НА 600 БИТ/С

Существенное уменьшение информационной скорости вокодера с сохранением разборчивости и качества восстановленной речи может быть достигнуто, если от независимого квантования параметров речевого сигнала перейти к зависимому квантованию, т. е. многомерному квантованию параметров, передающих текущие значения передаточной характеристики речевого тракта и характеристик источников возбуждения. Реализация этой идеи приводит к необходимости распознавания элементов речи, представляющих речевой сигнал на интервале анализа [152, 153].

При многомерном квантовании пространство параметров речевого сигнала разбивается на непересекающиеся области. Каждая область представляется одним набором параметров — одним вектором, интерпретируемым как эталон фонемы или, что точнее, части фонемы. Этот эталон в равной мере можно называть эталонным элементом.

Перенумеруем все области, соответственно все эталонные элементы.

Каждый текущий наблюдаемый в пространстве параметров элемент речи теперь будем рассматривать как принадлежащий определенной области в пространстве параметров и в линию связи будем передавать не сам наблюдаемый элемент, а только номер области, которой этот элемент принадлежит, или номер эталонного элемента, который эту область представляет.

В этом будет заключаться многомерное квантование (распознавание) текущего элемента речи на множестве эталонных элементов.

Приняв переданный по каналу связи номер эталонного элемента, синтезатор речи извлекает из памяти по номеру элемента сам эталонный элемент и использует его для синтеза речи.

Поскольку эталонные элементы интерпретируются как представляющие фонемы или их части, то предлагаемый вокодер назван квазифонемным или поэлементным.

Далее более детально описывается квазифонемный (поэлементный) вокодер на 600 бит/с.

В анализирующей части вокодера содержится J эталонных элементов b_j , $j = 1 : J$, например, $J = 512$ или $J = 1024$. Эти J эталонных элементов представляются b -параметрами (b -векторами) пред-

сказания $\mathbf{b}_j = (b_{j0}, b_{j1}, \dots, b_{js}, \dots, b_{jm})$, m — порядок системы предсказания ($m = 10$).

Каждый вектор \mathbf{b}_j взаимооднозначно определяется а-параметрами предсказания $\mathbf{a} = (a_0, a_1, a_2, \dots, a_s, \dots, a_m)$, $a_0 \equiv 1$:

$$b_{j0} = \sum_{i=0}^m a_{ji}^2; \quad b_{js} = 2 \sum_{i=0}^{m-s} a_{ji} a_{j(i+s)}, \quad s = 1 : m. \quad (10.3.1)$$

Поскольку б-параметры (или а-параметры) можно трактовать как параметры некоторой линейной системы, моделирующей речеобразование, то естественно поставить вопрос, какой именно одной из J линейных систем \mathbf{b}_j может быть с наибольшей вероятностью синтезирован наблюдаемый на интервале анализа речевой сигнал f_n , $n = 1 : M$.

Пусть для определенности речевой сигнал разбивается на неперекрывающиеся участки продолжительностью $\Delta T' = 26$ мс. Пусть шаг дискретизации исходного речевого сигнала Δt равен 65 мкс. Тогда на интервале анализа будет $M = 400$ дискрет речевого сигнала f_n , $n = 1 : M$.

Рассматривая линейные системы \mathbf{a}_j или \mathbf{b}_j как системы с авторегрессией, естественно величину

$$\tilde{f}_n = - \sum_{v=1}^m a_{jv} f_{n-v} \quad (10.3.2)$$

интерпретировать как прогноз с помощью системы \mathbf{a}_j значения сигнала для момента времени n по предыдущим m наблюденным значениям речевого сигнала f_{n-v} , $v = 1 : m$.

Величину же $\epsilon_{jn} = f_n - \tilde{f}_n$ в таком случае следует рассматривать как ошибку прогноза либо как сигнал возбуждения для линейной системы \mathbf{a}_j в момент времени n . Применяя этот сигнал возбуждения, можно точно находить значение сигнала f_n с помощью системы \mathbf{a}_j по предыдущим отсчетам f_{n-v} , $v = 1 : m$:

$$f_n = - \sum_{v=1}^m a_{jv} f_{n-v} + \epsilon_{jn}, \quad j = 1 : J. \quad (10.3.3)$$

Придадим уравнениям (10.3.3) некий приближенный смысл, полагая, что отсчеты ϵ_{jn} не зависят от j и представляют собой отсчеты дискретного белого шума ϵ_n с постоянной дисперсией σ и нулевым средним (§ 9.4). Таким образом, точные уравнения (10.3.3) заменим приближенными:

$$f_n = - \sum_{v=1}^m a_{jv} f_{n-v} + \epsilon_n, \quad j = 1 : J, \quad (10.3.4)$$

где ϵ_n — дискретный белый шум или периодический белый шум с дисперсией σ_j .

Наиболее подходящей линейной системой \mathbf{a}_j или \mathbf{b}_j , позволяющей интерпретировать (и синтезировать) наблюдаемый речевой сигнал с помощью уравнения (10.3.4), естественно назвать ту, которая обеспечивает минимальную (суммарную для интервала анализа) энергию

ошибки прогноза или энергию сигнала возбуждения e_n , $n = 1 : M$:

$$\begin{aligned} j^*(f_n, n = 1 : M) &= \operatorname{argmin}_{i=1:J} \sum_{n=1}^M \left(\sum_{v=0}^m a_{iv} f_{n-v} \right)^2 = \\ &= \operatorname{argmin}_{i=1:J} \sum_{u,v=0}^m a_{iu} a_{iv} \sum_{n=1}^M f_{n-u} f_{n-v} = \operatorname{argmin}_{i=1:J} (\mathbf{B}, \mathbf{b}_i), \end{aligned} \quad (10.3.5)$$

где обозначено $\mathbf{B} = (B_0, B_1, \dots, B_s, \dots, B_m)$:

$$B_s = \sum_{n=1}^{M-s} f_n f_{n+s}, \quad s = 0 : m, \quad (10.3.6)$$

и полагается, что $f_n \equiv 0$ для $n \leq 0$ и $n > M$.

Конечно, минимизация энергии ошибки прогноза вовсе не означает, что для оптимальной линейной системы j^* сигнал

$$e_n = \sum_{v=0}^m a_{iv} f_{n-v}, \quad n = 1 : M, \quad (10.3.7)$$

строго будет отрезком белого или периодического белого шума. Однако практически получается, что минимизация энергии ошибки прогноза и фильтрация исходного речевого сигнала согласно уравнению (10.3.7) приводят к выбелыванию исходного речевого сигнала. В результате последовательность e_n , $n = 1 : M$, вычисленная по (10.3.7), обладает многими свойствами белого (периодического белого) шума.

Дисперсия этого шума определяется выражением

$$\sigma_{j^*} = \sqrt{\frac{1}{M} (\mathbf{B}, \mathbf{b}_{j^*})}. \quad (10.3.8)$$

Из всего сказанного следует, что в анализирующей части вокодера необходимо на основании наблюдаемого речевого сигнала f_n , $n = 1 : M$, вычислить элемент автокорреляции \mathbf{B} , воспользовавшись формулой (10.3.6), а затем распознать этот элемент

$$j^*(\mathbf{B}) = \operatorname{argmin}_{i=1:J} (\mathbf{B}, \mathbf{b}_i) \quad (10.3.9)$$

на множестве эталонных элементов \mathbf{b}_j , $j = 1 : J$.

При этом в линию связи будем передавать девяты- или десятиразрядный код (9 или 10 бит) о номере эталонного элемента $j^*(\mathbf{B})$, а также двухразрядный код (2 бита) об амплитуде сигнала возбуждения, получаемый нелинейным квантованием величины σ_{j^*} (см. формулу (10.3.8)):

$$W = \operatorname{INT} \left(\frac{1}{\Delta \mathcal{D}} \ln \frac{\sigma_{j^*}}{\mathcal{D}_0} \right), \quad (10.3.10)$$

где INT — оператор выделения целой части; $\mathcal{D}_0 = 0,015$ и $\Delta \mathcal{D} = 1,5$ для шумных звуков (признак тон-шум равен нулю); $\mathcal{D}_0 = 0,030$ и $\Delta \mathcal{D} = 1,6$ для звонких звуков (признак тон-шум равен единице).

Для кодирования признака тон-шум и среднего для интервала анализа периода основного тона используются 4 бита, причем в линию

связи передается приращение периода ОТ для двух соседних интервалов анализа. Фактически же информация об ОТ представляется 8-ю битами, однако при передаче приращений вполне достаточно для кодировки 4 бит. Подробности кодирования описаны в § 10.2.

Для вычисления признака тон-шум и среднего периода ОТ используется нулевой метод или ВОТ-метод. Последний обеспечивает хорошую надежность и точность и прост в технической реализации. Поэтому он рекомендуется для практического использования.

В результате убеждаемся, что в квазифонемном (поэлементном) вокодере для кодирования одного кадра длительностью $\Delta T' = 26$ мс требуется всего 15 бит, что составляет информационную скорость не более 600 бит/с.

В синтезирующей части вокодера содержится также J эталонных элементов, однако представленных не b -параметрами, а a -параметрами. Для восстановления речи используется уравнение (10.3.4), в котором следует положить $j = j^*(B)$. Сигнал возбуждения e_n для уравнения (10.3.4) является просто белым шумом, если значение признака тон-шум равно нулю, либо периодическим белым шумом, если переданный кадр был звонким. Амплитуда сигнала возбуждения определяется двухбитовым кодом амплитуды (см. формулу (10.3.10)). Информация о периоде ОТ используется для формирования периодического (с требуемым периодом) белого шума в случае синтеза звонких звуков.

Синтезированный сигнал подается на цифро-анalogовый преобразователь, а с него — на громкоговоритель.

Для функционирования квазифонемного вокодера необходимо задать эталонные элементы a_i или b_j , $j = 1 : J$. Они вычисляются один раз в процессе конструирования (разработки) вокодера. Оценивание эталонных элементов выполняется в лабораторных условиях по ОВ, представляющей речь многих дикторов.

Пусть B_i , $i = 1 : \mathcal{L}$ — ОВ, состоящая из \mathcal{L} элементов автокорреляций.

Математическая задача обучения может быть сформулирована так:

$$\{b_j^*, j = 1 : J\} = \underset{\{b_j, j = 1 : J\}}{\operatorname{argmin}} \sum_{i=1}^{\mathcal{L}} \min_{j=1 : J} \frac{(B_i, b_j)}{\sigma^2(B_i)}, \quad (10.3.11)$$

или

$$\{b_j^*, j = 1 : J\} = \underset{\{b_j, j = 1 : J\}}{\operatorname{argmin}} \max_{i=1 : \mathcal{L}} \min_{j=1 : J} \frac{(B_i, b_j)}{\sigma^2(B_i)}, \quad (10.3.12)$$

где

$$\sigma^2(B_i) = \min_b (B_i, b), \quad (10.3.13)$$

а на b -параметры накладываются ограничения (10.3.1).

Задачи (10.3.11) — (10.3.12) являются обычными задачами самообучения (таксономии, кластер-анализа, развода на кучи), уже обсуждавшимися подробно в гл. 3 и 4. Решаются эти задачи в основном итерационными алгоритмами (см., в частности, [152—153]).

Рассмотрим итерационный алгоритм решения задачи (10.3.12). Очередная итерация заключается в последовательном переборе

элементов ОВ, при котором некоторые из них порождают эталонные элементы.

Пусть к моменту обработки элемента B_i уже имеется $J' < J$ эталонных элементов. Если оказывается, что

$$\min_{j=1:J'} \frac{(B_i, b_j)}{\sigma^2(B_i)} \leq \Theta, \quad (10.3.14)$$

где Θ — фиксированный для данной итерации порог качества аппроксимации, то осуществляется переход к анализу очередного элемента.

Если неравенство (10.3.14) не выполнилось, т. е. элемент B_i не может быть аппроксимирован с точностью Θ ни одним из уже имеющихся эталонных элементов, осуществляется проверка, не достигло ли количество эталонных элементов J' заданной границы J . Если еще нет, то полагаем $J' := J' + 1$ и порождаем новый эталонный элемент

$$b_{J'} = \operatorname{argmin}_b (B_i, b), \quad (10.3.15)$$

после чего обрабатывается очередной элемент ОВ.

Если же число эталонов J' уже равно J и неравенство (10.3.14) не выполняется, то делается вывод, что с помощью выбранной совокупности эталонных элементов вся ОВ не может быть аппроксимирована с точностью Θ . Поэтому Θ увеличивается и начинается новая итерация — новый просмотр всех элементов ОВ с целью найти новую систему эталонных элементов, аппроксимирующих ОВ с точностью Θ .

Если в результате очередной итерации оказалось, что для аппроксимации ОВ с точностью Θ достаточно меньшего, чем J , количества эталонных элементов, то Θ уменьшается с последующим переходом к очередной итерации.

Увеличивая или уменьшая Θ методом деления шага пополам, будем гарантировать сходимость процесса обучения.

Квазифонемный (поэлементный) вокодер многосторонне исследовался экспериментально. Он обеспечил приемлемые разборчивость, качество и натуральность восстановленной речи с сохранением распознаваемости индивидуальности голосов.

§ 10.4. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ. ПЕРСПЕКТИВЫ СОЗДАНИЯ ФОНЕМНОГО ВОКОДЕРА

Нуль-полюсный и квазифонемный вокодеры моделировались на ЭВМ как для отработки принципов компрессированной передачи речи, так и с целью проверки, разработки и доводки конкретных систем компрессированной передачи речи.

В частности, на основе нуль-полюсной системы компрессированной передачи речи был смоделирован формантный вероятностный вокодер параллельного типа и для него были найдены оптимальные режимы работы, что способствовало сокращению сроков разработки [155]. В этом вокодере каждая форманта задавалась четырьмя параметрами: частотой, добротностью, амплитудой и фазой. Эти параметры вычислялись по полюсным параметрам $(\mathcal{A}_i, \mathcal{B}_i)$ (см. § 10.2). Полюса параллель-

ной форманты определялись так:

$$z_i = -\mathcal{A}_i/2 \pm j \sqrt{\mathcal{B}_i - \mathcal{A}_i^2/4}, \quad j^2 = -1. \quad (10.4.1)$$

Частота f_i и ширина Δf_i форманты были равны

$$f_i = \frac{1}{2\pi\Delta t} \arg z_i, \quad (10.4.2)$$

$$\Delta f_i = \frac{1}{2\pi\Delta t} \ln \mathcal{B}_i. \quad (10.4.3)$$

Соответственно добротность $Q_i = f_i/\Delta f_i$.

Амплитуда d_i и фаза ϕ_i параллельной форманты определяются выражением h_{in} для составной части импульсного отклика, порожденной i -й параллельной формантой:

$$h_{in} = d_i |z_i|^n \cos(2\pi f_i n \Delta t + \phi_i), \quad n = 0 : \infty \quad (10.4.4)$$

Для их вычисления передаточную функцию (10.2.1) линейной системы представляем в виде суммы дробей типа $D_i/(1 - z_i/z)$. По правилам разложения дробно-рациональной функции на элементарные дроби получаем

$$D_i = \frac{z_i^{m-1}}{\frac{dH(z_i)}{dz_i}} \quad (10.4.5)$$

и затем вычисляем амплитуду и фазу форманты:

$$d_i = 2 |D_i|, \quad (10.4.6)$$

$$\phi_i = \arg D_i. \quad (10.4.7)$$

Особенно тщательно исследовался квазифонемный вокодер на 600 бит/с в связи с острой необходимостью его разработки для применения в интегрированных цифровых системах связи с коммутацией пакетов. Моделировались различные варианты технической реализации вокодера на быстродействующих микропроцессорах (гл. 12). Особое внимание уделялось моделированию вычислений с учетом разрядности представления используемых величин. Это делалось с целью выяснения минимально допустимых разрядностей, что способствовало уменьшению веса и стоимости вокодера.

Такие исследования по разрядности представления проводились для всех используемых величин (см. § 9.4 и § 10.3): отсчетов исходного речевого сигнала, отсчетов автокорреляций, компонент эталонных элементов, представленных а- и б-параметрами, отсчетов прореженного сигнала, коэффициентов выбеляющего фильтра, отсчетов сигналов обратной фильтрации, отсчетов автокорреляции сигналов обратной фильтрации, отсчетов сигнала возбуждения, отсчетов синтезированного сигнала. Таким образом, разработка и изготовление квазифонемного вокодера сопровождались экспериментами и исследованиями на адекватной информационной модели.

Последующие эксперименты были направлены на уменьшение информационных скоростей квазифонемного вокодера до 300 и 150 бит/с.

Это достигалось применением временной компрессии и линейной интерполяции элементов речи, т. е. использованием кусочно-постоянных и кусочно-линейных моделей речевого сигнала, а также применением более изощренных приемов многомерного квантования и распознавания элементов речи.

ВЫВОДЫ

1. Предложена и исследована нуль-полюсная система компрессированной передачи речи на информационные скорости 9600, 4800, 2400 и 1200 бит/с. Показана целесообразность использования полюсных параметров для представления речевого сигнала, продемонстрирована простота их квантования. Обращено внимание на необходимость совмещения процедур оценивания и квантования параметров в одном процессе, на избыточность кодирования информации при независимом квантовании параметров.

2. Показано, что устранение избыточности в кодировании параметров речевого сигнала приводит к зависимому квантованию параметров, т. е. многомерному квантованию параметров, которое заключается в распознавании текущего элемента речи на множестве небольшого количества эталонных элементов и в передаче в линию связи номера наиболее похожего эталонного элемента.

Многомерное квантование приводит к существенному уменьшению информационной скорости вокодерных систем.

Разработан и исследован квазифонемный вокодер на 600 бит/с, содержащий 512 (1024) эталонных элементов и основанный на реализации идеи многомерного квантования путем распознавания элементов речи. Восстанавливаемая (синтезируемая) речь характеризуется приемлемыми разборчивостью, качеством и натуральностью звучания, сохраняет существенные индивидуальные особенности голоса.

4. Показано, что применение кусочно-постоянных и кусочно-линейных моделей распознавания речи для компрессированной передачи речи делает конструктивным создание квазифонемного вокодера на 300 и 150 бит/с с обеспечением приемлемых разборчивости, качества и натуральности звучания и сохранением индивидуальных особенностей голоса.

ГЛАВА 11

ЭКСПЕРИМЕНТАЛЬНЫЕ СИСТЕМЫ РАСПОЗНАВАНИЯ, СМЫСЛОВОЙ ИНТЕРПРЕТАЦИИ И КОМПРЕССИРОВАННОЙ ПЕРЕДАЧИ РЕЧИ

В настоящей главе кратко описываются техника экспериментирования, универсальный моделирующий стенд, экспериментальные системы распознавания, смысловой интерпретации и компрессированной передачи речевых сигналов, структуры технического и программного обеспечения, результаты использования моделирующего стенда и экспериментальных систем.

§ 11.1. УНИВЕРСАЛЬНЫЙ МОДЕЛИРУЮЩИЙ СТЕНД

Поскольку распознавание, смысловая интерпретация, компрессированная передача и синтез речи являются естественно-научными проблемами, то наряду с теоретическими разработками с самого начала велись и экспериментальные исследования на моделирующих стенах, создаваемых на основе мощных универсальных ЭВМ, оснащенных средствами ввода и вывода речевого сигнала. Моделирование на таких стенах позволяло полностью отработать информационный процесс обработки речевого сигнала с целью его распознавания, смысловой интерпретации, компрессированной передачи и синтеза и создать соответствующие экспериментальные системы.

Первые моделирующие стены в наших исследованиях были сначала разработаны для ЭВМ «Киев» и М-50. «Киев» была оснащена устройствами ввода и вывода клипированного речевого сигнала (1962 г.), а М-50 — пятибитовыми преобразователями аналог-код и код-аналог для ввода и вывода сигналов (1965 г.). Эти стены позволили получить первые результаты, которые показали, что в целях моделирования следует использовать более мощные вычислительные машины.

Основной моделирующий стенд был разработан в 1970 г. и используется до сих пор. Его основу составляют БЭСМ-6 и устройство ввода-вывода речевого сигнала [62]. Устройство ввода-вывода содержит девятивитовые преобразователи аналог-код и код-аналог. Набор стандартных частот преобразований: 40, 33, 25, 20, 16, 10, 8 и 6 кГц. Предусмотрено подключение внешнего генератора, с тем чтобы можно было вести преобразование с любой желаемой частотой.

Сигнал в ЭВМ вводится либо с микрофона непосредственно в машинном зале, либо с магнитофона. Предусмотрены предварительное

усиление и частотное корректирование сигналов. Вывод сигналов из ЭВМ осуществляется либо на громкоговоритель, либо на магнитофон.

Устройство ввода-вывода речевого сигнала подключено к БЭСМ-6 через коммутатор внешних устройств (КВУ) с использованием восьмого разряда периферийного регистра прерываний (ПРП). По сигналу прерывания ЭВМ считывает текущий код-измерение с помощью команды EXT 4120B, где EXT — команда обмена с периферийными устройствами, а 4120B — адрес устройства в режиме ввода речевого сигнала в ЭВМ.

При выводе сигнала используются сигнал прерывания по восьмому разряду и команда вывода дискреты EXT 0060B, где 0060B — адрес устройства при выводе речевого сигнала.

Особая роль отводится кнопке связи: сигнал поступает в ЭВМ только тогда, когда она нажата. Поэтому во всех случаях ввод сигналов в ЭВМ осуществляется следующим образом: диктор-оператор нажимает кнопку, затем произносит слово, словосочетание или слитную фразу, после чего отпускает кнопку. Благодаря кнопке связи в ЭВМ вводятся только интересующие человека-оператора сигналы.

Прием и выдача речевого сигнала в (из) ЭВМ осуществляются с помощью программы ввода-вывода. Управление работой устройства — программное. Инициатива принадлежит ЭВМ. По индикации, вмонтированной в микрофон, диктор-оператор видит, вышла ли ЭВМ на связь с устройством и если вышла, то какой при этом режим — ввод или вывод сигналов. Разумеется, вводить речевые сигналы в ЭВМ можно только после того, как ЭВМ выставила на устройстве режим ввода.

Ввод (вывод) речевого сигнала в (из) ЭВМ осуществляется в режиме произвольных физических действий, экстракода реального времени и совместим с операционными системами ДУБНА и ДИСПАК.

Максимальная продолжительность вводимых за один акт ввода сигналов составляет 10 с.

Для индикации результатов распознавания в виде бегущего текста слова или последовательности слов используется световое табло, в качестве которого служит индикаторное поле регистров БРЗ пульта ЭВМ. Это же табло используется для высвечивания служебной информации при накоплении ОВ, в процессах обучения, записи и считывания на магнитную ленту и т. п.

Регистровый пульт БЭСМ-6 используется для выбора режимов работы, параметров программ, ветвления процессов.

Если устройство ввода-вывода речевого сигнала является неотъемлемой частью моделирующего стенда, то суть и возможности стенда определяются комплексом программных средств — совокупностью программ для анализа, распознавания, смысловой интерпретации, компрессированной передачи и синтеза речевых сигналов.

Информация о структуре и возможностях разработанного программного обеспечения содержится в [62, 106, 108, 147, 156, 157]. Комплекс программ для решения той или иной задачи набирается как из отработанных базовых подпрограмм, так и из вновь разрабатываемых подпрограмм.

В числе базовых подпрограмм находятся следующие: ввода-вывода речевого сигнала, распаковки-упаковки речевого сигнала (5 дискрет в одной ячейке), автокорреляционного анализа, цифрового спектрального анализа, предиктивного анализа (вычисления а-, б- и к-параметров предсказания), вычисления признаков тональности (признака тон-шум и периода ОТ), сегментации и самосегментации реализаций, сравнения реализаций с исходными эталонами слов, обучения и самообучения распознаванию, распознавания слов и слитной речи, смысловой интерпретации слитной речи, высвечивания символной информации на световом табло, восстановления (синтеза) речи, синтеза речи и многие другие.

Программное обеспечение составлено в основном на языке МАДЛЕН. Управляющие программы составляются на ФОРТРАНе. Используются также языки БЭМШ и ЛИСП. Программирование на языке МАДЛЕН обусловлено стремлением создать системы реального времени для обработки речевых сигналов.

Базовые подпрограммы объединяются в программные модули, имеющие самостоятельное назначение. Эти модули могут загружаться в память ЭВМ в том или ином порядке, организуя моделирование различных процессов обработки речевого сигнала с целью его распознавания, смысловой интерпретации или компрессированной передачи. Примеры программных модулей приведены в следующем параграфе.

§ 11.2. ЭКСПЕРИМЕНТАЛЬНЫЕ СИСТЕМЫ 1966—1983 гг.

На основе моделирующих стендов были созданы экспериментальные системы распознавания, смысловой интерпретации, компрессированной передачи и синтеза речи. Вводу в действие этих систем предшествовали предварительные исследования по выбору параметров.

Далее в хронологическом порядке дается краткая характеристика разработанных и используемых экспериментальных систем.

1966—1967 гг. Распознавание отдельно произносимых слов. Объем словаря — 10 слов. Метод распознавания основывался на евклидовом расстоянии и предполагал временную нормализацию описаний слов до распознавания. Надежность распознавания 93—97 % [93—94]. Были вскрыты основные недостатки методов, основанных на временной нормализации, что привело к разработке поэлементного метода распознавания. Использовалась ЭВМ М-50.

1966—1967 гг. Распознавание отдельно произносимых слов. Объем словаря — 12 слов. Метод распознавания — поэлементный. На контрольной выборке ошибок распознавания зарегистрировано не было. Использовалось спектральное описание. Элементы речи нормировались из условия равенства единице длины наблюдаемых элементов-векторов. Элементарная мера сходства — скалярное произведение векторов-элементов со знаком минус. Использовалась ЭВМ М-50. Эксперименты с этой системой распознавания показали перспективность разработки поэлементного метода (§ 2.6, [1—4]).

1969—1973 гг. Экспериментальная система поэлементного распознавания слов и слитной речи. Объем словаря — до 200 слов. Применяется обучение поэлементному распознаванию. Метод распознавания — поэлементный. Процент ошибок и процент отказов от распознавания при распознавании 200 слов соответственно 0,5 и 2,5 %. Аналогичные показатели для слов в случае распознавания слитной речи равны соответственно 8 и 5 % (объем словаря 66 слов). Распознавание велось в замедленном масштабе времени — время запаздывания ответа распознавания составляло в среднем 15 с после окончания произнесения слова (распознавание слов) и более 1 мин — в случае распознавания слитной речи. Использовалось спектральное описание. Элементы речи нормировались из условия, что максимальная длина элемента-вектора, возможная на всей длине реализации, равна единице. Элементарная мера сходства основывалась на евклидовом расстоянии. Эксперименты показали целесообразность использования темпоральной транскрипции слова и информации об интенсивности элементов. Значительные затраты памяти и времени приходятся на хранение исходных эталонов слов и вычисление значений элементарной меры сходства (§ 2.6, § 5.6, [16, 61, 62, 89, 90]).

1973—1976 гг. Экспериментальная система пофонемного распознавания слов и слитной речи. Объем словаря — 300 слов. Применяются обучение и дообучение распознаванию слов. Надежность распознавания трехсот слов — 98 %, надежность распознавания слов в слитной речи — 90 % (3 % ошибок и 7 % отказов от распознавания). Распознавание ведется в замедленном масштабе времени. Запаздывание ответа распознавания обусловлено главным образом вычислениями на предварительную обработку речевого сигнала. Использовалось описание элементов речи 48-разрядными двоичными кодами, вычисляемыми по текущему спектру речевого сигнала. Элементарная мера сходства основывалась на хэмминговом расстоянии между кодами. Затраты памяти на хранение совокупности из 80 эталонных элементов и акустических и темпоральных транскрипций слов, а также затраты времени на вычисление значений элементарной меры сходства были во много раз меньше в сравнении с системой поэлементного распознавания (§ 4.8, § 5.6, [78, 80, 85, 105]).

1977—1983 гг. Экспериментальная система квазиреального времени для пофонемного распознавания слов и слитной речи. Эта система аналогична предыдущей и отличается от нее только квазиреальным временем распознавания. Время обучения на распознавание двухсот слов составляет 2 ч машинного времени БЭСМ-6, включая накопление обучающей выборки. Время дообучения на одно слово составляет в среднем 15 с. Объем словаря — 1000 слов. Надежность распознавания слов для словарей из 200, 500 и 1000 слов составляет соответственно 99,5, 98 и 96 %. При этом распознавание есть выбор одной гипотезы из 200, 500 и 1000 соответственно. Запаздывание ответа распознавания после окончания произнесения слова составляет 1, 4 и 8 с для словаря из 200, 500 и 1000 слов соответственно. Надежность распознавания слитной речи в пересчете на надежность распознавания слов составляет 93 % при словаре в 200 слов. На хранение информации об одном

слове (акустической и темпоральной транскрипций) требуется всего 36 байт. Квазиреальный масштаб времени распознавания достигнут за счет вычисления элементов автокорреляций в процессе ввода речевого сигнала. Переход к двоичным кодам и распознавание осуществляются после накопления (ввода) элементов автокорреляций в памяти ЭВМ (§ 4.8, § 5.6, [17, 77, 104, 108, 116]).

1977—1983 гг. Экспериментальная система квазиреального времени для пофонемного распознавания слов и слитной речи по параметрам предсказания. Эта система является вариантом предыдущей системы для случая, когда распознаваемые элементы представляются 11-мерными векторами автокорреляций, а эталонные — описываются 11-мерными b -параметрами предсказания. Используемая элементарная мера сходства основывается на скалярном произведении элемента автокорреляции и эталонного элемента. Остальные данные такие же, как у предыдущей системы. На хранение общей совокупности из 80 эталонных элементов здесь требуется не 80, а 11×80 ячеек памяти (§ 4.8, § 5.6, [81, 108, 115]).

1979—1983 гг. Кооперативная система распознавания слов устной речи. Система предназначена для распознавания речи определенной группы (кооператива) дикторов и является вариантом многодикторной системы распознавания. Подобна предыдущей системе. Отличается от нее только способом формирования ОВ при обучении пофонемному распознаванию и тем, что распознавание слитной речи не реализуется. Объем словаря — 200 слов. Надежность распознавания слов для членов кооператива составляет 97—98 %. При этом гарантируется относительно высокая надежность распознавания для дикторов, не являющихся членами кооператива. Рекомендуется иметь, по крайней мере, два кооператива — один для мужских голосов и один для женских (§ 8.5, [139—140]).

1981—1983 гг. Система распознавания и смысловой интерпретации слитной речи, отвечающая на вопросы, заданные на естественном языке относительно операций над целыми числами. Объем словаря — до 1000 слов. Надежность смысловой интерпретации — 95 %, отказов от смысловой интерпретации — 5 %. Система является дальнейшим развитием системы квазиреального времени для распознавания слов и слитной речи. Использует априорную информацию о синтаксисе, семантике и прагматике предметной области. Может быть настроена на любую другую предметную область, что обеспечивается применением списочных структур, позволяющих задавать типы предложений и типы смыслов для различных предметных областей (§ 7.8, [18, 133, 158]).

1975—1980 гг. Нуль-полосные программные модели анализа, компрессированной передачи и восстановления (синтеза) речи на информационные скорости 9600, 4800, 2400 и 1200 бит/с. В этих моделях применяется нулевой метод для вычисления признаков тональности (§ 10.2, [75, 145—147]).

1977—1983 гг. Квазифонемный вокодер на информационную скорость 600 бит/с. Использование ВОТ-метода для вычисления признаков тональности. Полная информационная модель всех блоков и вокодера в целом (§ 10.3, § 10.4, [152—153]).

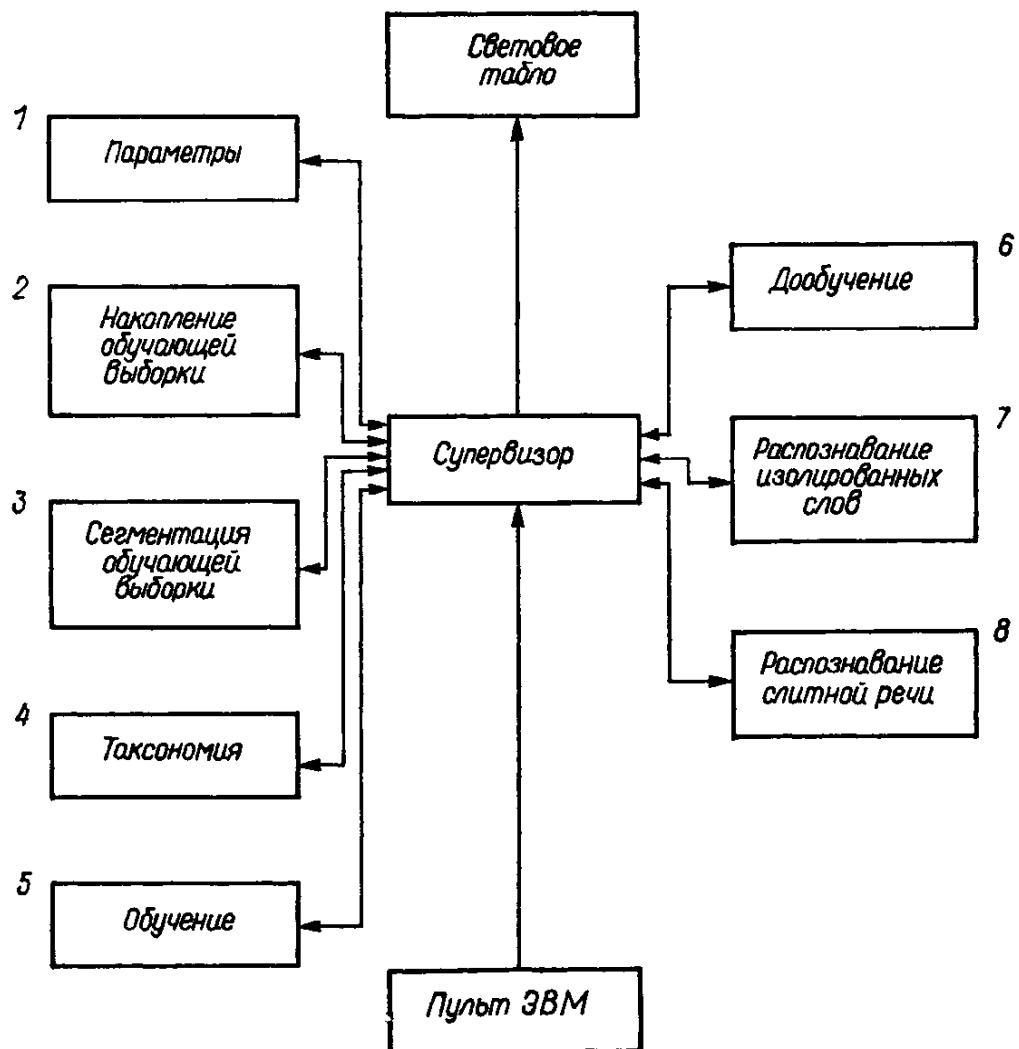


Рис. 11.1. Структура программного обеспечения системы квазиреального времени для пофонемного распознавания слов и слитной речи.

Техническое обеспечение всех систем описано в предыдущем параграфе, основное программное обеспечение приведено в [62, 106, 108, 147, 156, 157].

На рис. 11.1 в качестве примера представлена структура программного обеспечения системы квазиреального времени для пофонемного распознавания слов и слитной речи. Программное обеспечение содержит восемь независимых модулей, сменяющих друг друга в памяти ЭВМ. Вызов необходимого модуля или блока модулей осуществляется оператором через пульт ЭВМ. Взаимодействие модулей обеспечивается управляющей программой СУПЕРВИЗОР.

Составной частью модулей 1, 2, 6, 7, 8 является программа предварительной обработки речевого сигнала, параметрами которой являются количество отсчетов автокорреляционной функции, длина интервала анализа, частота дискретизации речевого сигнала и др. Выбор порогов, необходимых для вычисления элементов-кодов, осуществляется с помощью модуля 1.

Обучение распознаванию реализуется с помощью первых пяти модулей. Модули 3 и 4 являются вспомогательными по отношению к модулю 5. Они находят начальное приближение для итерационного

алгоритма обучения. Назначение остальных модулей следует из их названия.

Весь диалог пользователя с системой распознавания речи ведется через пульт ЭВМ, с помощью которого в процессе работы осуществляется выбор того или иного режима работы, задаются количество распознаваемых слов, номер корректируемого слова и т. п. Для индикации режимов работы и высвечивания результатов распознавания используется световое табло [108, 156—157].

Экспериментальные системы служили не только для отработки принципов анализа, распознавания, смысловой интерпретации и компрессированной передачи речевых сигналов, а и для демонстрационных целей и, главное, разработок и проектирования конкретных систем.

§ 11.3. ИСПОЛЬЗОВАНИЕ МОДЕЛИРУЮЩЕГО СТЕНДА И ЭКСПЕРИМЕНТАЛЬНЫХ СИСТЕМ

Приведем примеры использования разработанных экспериментальных систем и моделирующего стендса.

В 1976—1978 гг. по заказу РИВЦ Минсвязи УССР и Киевского института автоматики выполнялась работа по расчету управляющих параметров для синтезаторов речи применительно к справочной службе с речевым ответом [157]. Применялся пословный синтез речи. Для составления речи из небольшого количества слов слова предварительно записывались в оперативную память ЭВМ М-6000 в виде управляющих параметров для синтезаторов речи гармонического типа [65], а затем считывались в требуемом порядке. Управляющие сигналы далее подавались на гармонический синтезатор речи, а с него через АТС — в линию связи. Информационная скорость управляющих параметров составляла 2400 бит/с.

Для расчета управляющих параметров синтезаторов гармонического типа были использованы отдельные программные модули систем распознавания и вокодерных систем, в том числе модуль анализа речи с помощью гребенки цифровых резонансных фильтров, модули вычисления признаков тональности (признака тон-шум и периода основного тона).

Было закодировано, в частности, 56 слов для справочной службы «Автоответ» РИВЦ Минсвязи УССР. Эта служба предназначена для ведения расчетов с квартирными абонентами за междугородние телефонные разговоры. Система выдавала речевую справку о задолженности за услуги электросвязи вместо высылки почтовых карточек-счетов. Абонент набирал по телефону код 083, а затем номер телефона, о котором желал получить справку.

Система «Автоответ» прошла опытную эксплуатацию на АТС 53 и 58 г. Киева.

Примеры синтезируемых фраз: АБОНЕНТ ПЯТЬДЕСЯТ ТРИ СОРОК ТРИ ДВАДЦАТЬ ШЕСТЬ ЗАДОЛЖЕННОСТЬ ЗА ЯНВАРЬ ПЯТЬ РУБЛЕЙ ТРИ КОПЕЙКИ, АБОНЕНТ ПЯТЬДЕСЯТ ВОСЕМЬ ТРИДЦАТЬ НОЛЬ ПЯТЬ ЗАДОЛЖЕННОСТИ НЕ ЧИСЛИТСЯ.

Нуль-полюсные модели и модель квазифонемного вокодера использовались в разработках систем компрессированной передачи речи в интересах предприятий машиностроительных ведомств (см., например, [155]).

Экспериментальная система квазиреального времени для распознавания слов и слитной речи была использована в составе комплекса автоматизированной обработки графической информации [159].

Комплекс предназначен для обработки графических знаков, нанесенных вручную на типографские карты. Геологическая карта в целом или по частям высвечивается на экране графического дисплея, человек находит на карте интересующие его геологические знаки (уголь, нефть, газ, руда, линия излома, район исследований, место бурения, местоположение геологической партии, цифры, буквы и т. п.), касается световым пером характерной точки знака и голосом называет (именует) его. При этом в ЭВМ поступают координаты характерной точки знака (считывание координат светового пера) и его наименование (результат распознавания речи).

Кроме именования объектов голосом человек также произносит устные команды редактирования и управления ДАЛЬШЕ, СМЕСТИТЬ, ОШИБКА, КОНЕЦ. По команде ДАЛЬШЕ осуществляется переход к именованию следующего знака, по команде СМЕСТИТЬ смещается характерная точка на экране, если она была поставлена неверно. По команде КОНЕЦ заканчивается решение задачи.

При разрешении ввода команды (по индикатору ввода) оператор, держа в левой руке кнопку связи (расположенную на конце гибкого шланга), нажимает ее, затем произносит в микрофон команду, после чего отпускает кнопку. Правая рука используется оператором для действий со световым пером.

Система распознавания речи заранее настраивалась на голоса многих дикторов — на каждого в отдельности. Результаты обучения (объемом 2 кбайта для словаря из 50 слов) записывались на магнитную ленту или диски ЭВМ. При работе с экраном оператор набирает свой шифр, и по смене шифра в оперативную память ЭВМ считывается информация о результатах обучения на оператора с новым шифром.

Для обработки графической информации оказалось достаточно словаря из 50 слов. Надежность распознавания — выше 99 %. Результат распознавания устной команды высвечивается в форме текста на экране. В случае неправильного распознавания произносится команда ОШИБКА, предыдущее имя знака исключается на один сеанс ввода из числа возможных гипотез и знак именуется заново.

Количество дикторов-операторов практически неограничено и определяется емкостью магнитной ленты или диска.

Время распознавания — 0,5 с после окончания произнесения слова.

Управление работой дисплея голосом оказалось весьма перспективным применением автоматического распознавания речи.

Самое же главное использование экспериментальных систем распознавания речи и моделирующего стенда обусловлено предоставленными возможностями опробования различных вариантов структуры и принципов функционирования разрабатываемых устройств и систем

распознавания речи, выбора оптимальной архитектуры, исследования упрощенных процедур распознавания, приемов ускорения принятия решений при распознавании и т. п. Именно с использованием возможностей экспериментальных систем и моделирующего стенда были спроектированы и отработаны устройство распознавания 200 устных команд, а затем система речевого диалога (СРД) «Речь-1» на основе микро-ЭВМ «Электроника-60 М» (глава 12, [128]) и другие модели СРД типа «Речь» [168]. Разработки, проектирование, изготовление и внедрение систем речевого диалога и систем распознавания речи велись в интересах предприятий машиностроительных ведомств.

ВЫВОДЫ

1. Разработан универсальный стенд для моделирования и создания систем распознавания, смысловой интерпретации, компрессированной передачи и синтеза речевых сигналов. Основу стенда составляют БЭСМ-6, устройство ввода и вывода речевого сигнала, использующее аналого-цифровой и цифро-аналоговый преобразователи, соответствующее системное и проблемно-ориентированное программное обеспечение.

2. Разработаны экспериментальные системы распознавания, смысловой интерпретации и компрессированной передачи речевых сигналов, основанные на методах и алгоритмах обработки речевого сигнала, сформулированных в рамках КДП-подхода. Эти системы предназначены для использования при разработке, проектировании и испытании средств распознавания, смысловой интерпретации, компрессированной передачи и синтеза речи.

3. На примерах конкретного применения моделирующего стенда и экспериментальных систем показаны эффективность и целесообразность использования предлагаемых методов и средств распознавания, смысловой интерпретации и компрессированной передачи речевых сигналов.

ГЛАВА 12

ПРОЕКТИРОВАНИЕ СИСТЕМ РАСПОЗНАВАНИЯ. СИСТЕМЫ РЕЧЕВОГО ДИАЛОГА

В настоящей главе кратко описываются разрабатываемые или реализованные системы распознавания и компрессированной передачи речи. Показана целесообразность разработки систем речевого диалога. Дано описание систем речевого диалога (СРД) типа «Речь», которые распознают и синтезируют речь. Приведены примеры использования систем речевого диалога.

§ 12.1. АРХИТЕКТУРА УСТРОЙСТВ И СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ. ТРЕБОВАНИЯ, ПРЕДЪЯВЛЯЕМЫЕ К УСТРОЙСТВАМ И СИСТЕМАМ

Чтобы предлагаемые в рамках КДП-подхода методы распознавания слов и слитной речи могли быть реализованы в устройствах и системах, последние должны обладать определенными производительностью (скоростью вычислений) и объемом памяти. Так, для пофонемного распознавания слов и слитной речи (гл. 4 и 5) в реальном масштабе времени (объем словаря — 500 слов, порядок следования слов свободный) требуется быстродействие (в пересчете на однопроцессорную ЭВМ) около 5—7 млн. операций типа сложения в секунду и оперативная память на 128 кбайт. При этом, для большей определенности, в продолжительность выполнения операции сложения включены времена выбора двух операндов и засылки результата. При смысловой интерпретации слитной речи быстродействие должно быть увеличено более чем на порядок, объемы памяти также возрастают, однако в меньшее число раз (гл. 7).

Однако как это было показано ранее в предыдущих главах, процессы распознавания, смысловой интерпретации и компрессированной передачи речи очевидным образом распараллеливаются, поскольку однотипные вычислительные процессы выполняются над однотипными массивами данных. По этой причине автономные системы распознавания и смысловой интерпретации речи могут создаваться, прежде всего, как мульти микропроцессорные системы с использованием серийно выпускаемых микропроцессорных наборов или как системы на основе микро-ЭВМ с быстродействующими специализированными вычислителями типа процессора динамического программирования, спек-

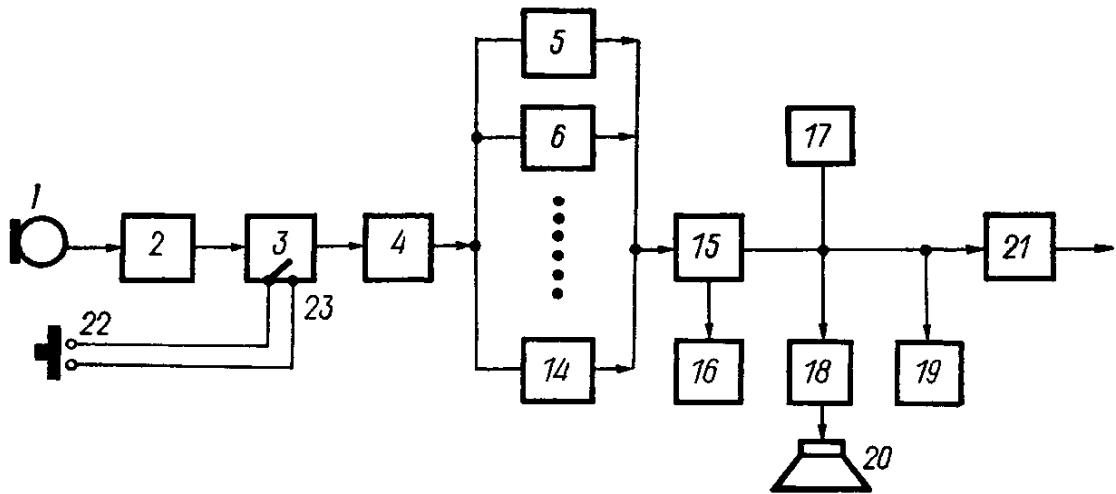


Рис. 12.1. Типовая структура системы распознавания и смысловой интерпретации речи.

трального анализатора, вычислителя элементарных мер сходства и т. п. [160—162]. Возможный набор специализированных вычислителей упоминается в главах 2, 5—7, 9—10. Сами же специализированные вычислители могут разрабатываться в виде универсальных или специализированных больших или сверхбольших интегральных схем для цифровой обработки сигналов. Перспективно также использование универсальных быстродействующих микропроцессоров, объединенных в конвейер для обработки информации.

Некоторая типовая архитектура системы распознавания, смысловой интерпретации и синтеза речи приведена на рис. 12.1.

Сигнал микрофона 1 подается в цепь усиления и частотной коррекции 2 и далее поступает в анализатор речи 3, который осуществляет предварительную обработку речевого сигнала.

Анализатор речи, являясь параллельной вычислительной системой, каждые, например, $\Delta T = 15$ мс передает в микропроцессор 4 результаты предварительной обработки — элементы речи x_i (§ 2.1, гл. 9 и 10).

В микропроцессоре 4, который содержит совокупность эталонных элементов $e(j)$, $j = 1 : J$ ($J = 128, 256, 512$ или 1024), вычисляются элементарные сходства $g(x_i, e(j))$ текущего наблюдаемого элемента x_i на эталонные элементы $e(j)$, $j = 1 : J$. Вычисления эти в случае необходимости (например, если $g(x_i, e(j))$ выражается через скалярное произведение) распараллеливаются. Важно лишь, чтобы все вычисления элементарных сходств были закончены за время, меньшее $\Delta T = 15$ мс, с тем, чтобы быть готовым к приему и обработке очередного элемента измерений x_{i+1} .

Массив сходств $g(x_i, e(j))$, $j = 1 : J$, далее передается в микропроцессоры 5—14 (специализированные процессоры), которые, работая параллельно, по рекуррентным формулам динамического программирования осуществляют процесс распознавания и смысловой интерпретации, причем так, чтобы все вычисления опять же закончились за время, меньшее ΔT , что необходимо для подготовки к приему и обработке очередного массива сходств $g(x_{i+1}, e(j))$, $j = 1 : J$.

Таким образом, вся информация об алфавите эталонных элементов, их громкости и тональности заносится в процессор 4, а вся информация о словаре, синтаксисе и семантике закладывается в процессоры 5—14. Процесс вычислений в микропроцессорах 5—14 распараллеливается таким образом, что каждый из вычислителей 5—14 обслуживает группу подсловарей или один подсловарь, причем, в свою очередь, каждый из этих вычислителей состоит из однотипных процессоров, по которым процесс вычислений также распараллеливается, но уже по словам или состояниям внутри слов (гл. 2, 4—11). Основная информация, априори вносимая в эти вычисления,— акустические, темпоральные, громкостные и тональные транскрипции слов, подсловари и связи между ними. На рис. 12.1 связи между процессорами 5—14 не показаны.

Обратим внимание на то, что процессоры 5—14 должны выполнять только самые простые операции: складывать, сравнивать, пересылать числа и выполнять условные переходы.

Количество слов, которое может обслуживать один из процессоров 5—14, определяется производительностью процессора. Так, центральный процессор микро-ЭВМ «Электроника-60» за время $\Delta T = 15$ мс успевает обработать только 2—3 слова, а центральный процессор БЭСМ-6 — около 100 слов.

Процессор 15 является заключительным. Он предназначен для формирования ответа распознавания и смысловой интерпретации по информации, поступающей от вычислителей 5—14, и для управления работой синтезатора речи. Так, в случае распознавания отдельно произносимых слов, процессор 15 принимает от вычислителей 5—14 величины $G(k)$, характеризующие принадлежность распознаваемой реализации к слову с номером k , находит номер k^* слова, для которого $G(k^*)$ является наибольшим, и выдает его в качестве ответа распознавания.

В случае распознавания слитной речи со свободным порядком следования слов процессор 15 формирует для каждого текущего момента времени i потенциально-оптимальные слова и индексы распознавания, по которым, после окончания произнесения, составляет ответ распознавания в виде последовательности слов, содержащейся в распознаваемом сигнале. Аналогично, но несколько сложнее, действует процессор 15 при смысловой интерпретации слитной речи. Завершается смысловая интерпретация тем, что в процессоре 15 формируется каноническая форма переданного речевым сигналом смысла (гл. 7).

Для контроля правильности распознавания и смысловой интерпретации служат дисплей 16 и синтезатор речи 18 с громкоговорителем 20. Использование синтезатора речи предпочтительнее дисплея, поскольку при этом полностью высвобождается зрение человека.

Внешняя память 17 предназначена для хранения программ и результатов обучения на словарь и голос дикторов. В ней же хранятся данные о синтаксисе и семантике используемых языков диалога.

Клавиатура управления 19 используется для задания указаний учителя при обучении, выбора результатов обучения из внешней па-

мяти на требуемый словарь и заданного диктора, загрузки программ и т. п.

Результат распознавания и смысловой интерпретации речи через интерфейс 21 передается во внешнюю ЭВМ. Через него также поступает от внешней ЭВМ информация, которая подлежит озвучиванию с помощью синтезатора речи 18.

Кнопка связи 22 и тумблер связи 23 предназначены для управления вводом речевого сигнала. Если тумблер выключен, человек-оператор нажимает кнопку, затем произносит устную команду — слово, фразу или устный текст, после чего отпускает кнопку. Подвергаются анализу и обработке только те сигналы, которые поступают в систему распознавания речи во время нажатия кнопки. Благодаря кнопке связи осуществляется управляемый человеком отбор сигналов для последующего распознавания.

Режим работы с кнопкой связи рекомендуется при сильных внешних акустических помехах, имеющих явно выраженный нестационарный характер, например, при разговоре рядом стоящих посторонних лиц. В остальных случаях работа осуществляется без кнопки связи. Тогда сигнал микрофона все время поступает в систему распознавания. Для работы в этом режиме необходимо включить тумблер связи 23.

Обучение и дообучение распознаванию речи, замена слов в словаре, пополнение словаря — все это реализуется программными средствами с использованием всех вычислительных ресурсов системы распознавания. Обучение может осуществляться и в замедленном масштабе времени.

Конфигурацию системы распознавания речи изменяют в зависимости от решаемых диалоговых задач. Так, внешняя память 17 выносится в состав внешней ЭВМ, а загрузка программ и результатов обучения на словарь и голоса дикторов выполняется из внешней ЭВМ через интерфейс 21. В ряде случаев можно отказаться от дисплея 16 и клавиатуры 19.

Помимо автономных устройств и систем распознавания, смысловой интерпретации и синтеза речи, возможны и другие варианты создания этих устройств и систем. Во-первых, устройство (или система) распознавания и смысловой интерпретации речи (УРИСИР) может быть задумано как составная часть большой высокопроизводительной ЭВМ. В этом случае к большой ЭВМ подключаются анализатор и синтезатор речи, которые через каждые, например, $\Delta T = 15$ мс передают в ЭВМ результаты предварительной обработки речевого сигнала или принимают от ЭВМ сигналы управления синтезатором речи. Большая ЭВМ, обладая высокой производительностью, принимает сигналы анализатора и на фоне основных задач решает задачу распознавания и смысловой интерпретации речевого сигнала и задачу генерации управляющих параметров синтезатора речи как вспомогательные. Таким образом, УРИСИР вполне может быть составной частью таких ЭВМ, как «Эльбрус», ПС-2000 или макроконвейер. В рассматриваемом случае вся трудность создания УРИСИР переносится на разработку проблемно-ориентированного математического обеспечения и привязку его к операционной системе ЭВМ.

Возможен вариант полуавтономного УРИСИР, который предлагается с целью упрощения УРИСИР как полностью автономного устройства. Так, можно упростить УРИСИР, отказавшись от обучения на словари и голос пользоваеля, перенеся эти функции на универсальные ЭВМ. Для этого необходимо, чтобы полуавтономное УРИСИР имело средства связи с универсальными ЭВМ. При наличии каналов связи и сетей ЭВМ использование полуавтономного УРИСИР может быть вполне экономически оправданным.

В целом, УРИСИР должны создаваться как составная часть вычислительной техники и систем управления с учетом всех тенденций и закономерностей их развития.

Архитектура УРИСИР в значительной мере определяется теми требованиями, которые к нему предъявляются, а именно:

а) должно распознавать отдельно произносимые слова из словаря объемом 20—1000 слов;

б) должно распознавать слитную речь, составляемую из выбранного словаря объемом 20—1000 слов;

в) должно осуществлять смысловую интерпретацию слитной речи и, таким образом, обеспечивать устный диалог человека и ЭВМ на формализованных или усеченных естественных языках предметных областей;

г) должны обеспечиваться показатели надежности распознавания и смысловой интерпретации (типичными показателями могут быть: надежность распознавания слов при раздельном произнесении — в интервале 90—99 %; надежность смысловой интерпретации — 95 %; допускаются отказы от распознавания и смысловой интерпретации в пределах 0—5 %);

д) должен быть предусмотрен контроль правильности распознавания и смысловой интерпретации, например, зрительный — с помощью дисплея или звуковой — с помощью синтезатора речи. Синтезатор речи должен обеспечивать словесную разборчивость синтезированной речи не хуже 98 %. Должны быть предусмотрены эргономические и диалоговые средства, гарантирующие 100 %-ную надежность восприятия устных заданий;

е) требования по надежности распознавания и смысловой интерпретации должны обеспечиваться в условиях внешних акустических шумов и помех (типичное значение допустимого уровня помех — 85 дБ);

з) требование фиксирования микрофона относительно головы говорящего может быть не обязательным;

ж) для обеспечения высокой надежности распознавания и смысловой интерпретации УРИСИР должно быть обучаемым, т. е. способным настраиваться на словари и голос пользоваеля;

и) УРИСИР должно быть удобным в использовании, должны быть предусмотрены возможности оперативной замены, пополнения и исключения отдельных слов в словаре;

й) распознавание и смысловая интерпретация должны осуществляться в реальном масштабе времени, т. е. ответ распознавания и смысловой интерпретации должен выдаваться не позже, чем через 0,3 с после окончания произнесения.

Сформулированные требования относятся к УРИСИР, ориентированным как на распознавание слов и слитной речи, так и на смысловую интерпретацию речи. Часть требований может быть снята или уточнена в зависимости от конкретного назначения УРИСИР.

Требования к надежности распознавания, требование настройки на голос оператора отражают компромисс между тем, что может современная наука по распознаванию речи, и тем, что приемлемо для практики.

Особого внимания заслуживает проблема достижения приемлемого компромисса между стоимостью, объемом и весом УРИСИР, с одной стороны, и гарантированной надежностью распознавания и смысловой интерпретации, с другой. В связи с противоречивостью этих требований большое значение придается такому упрощению методов и алгоритмов распознавания, которое приводит к существенным уменьшениям объемов, веса и стоимости аппаратуры за счет незначительного снижения надежности распознавания.

§ 12.2. ПАРАЛЛЕЛЬНАЯ МАШИНА ДЛЯ РАСПОЗНАВАНИЯ СЛОВ И СЛИТНОЙ РЕЧИ

В данном параграфе кратко описывается разработка системы для распознавания слов и слитной речи. С целью обеспечения реального времени распознавания в ней реализуется распараллеливание вычислений (см. гл. 2, 4 и 5, [118, 160—163]).

Структура параллельной машины приведена на рис. 12.2. Ее основу составляют микро-ЭВМ типа «Электроника-60 М» и специализированные вычислители динамического программирования. Вся адресуемая память микро-ЭВМ З разбита на части: резидентная часть

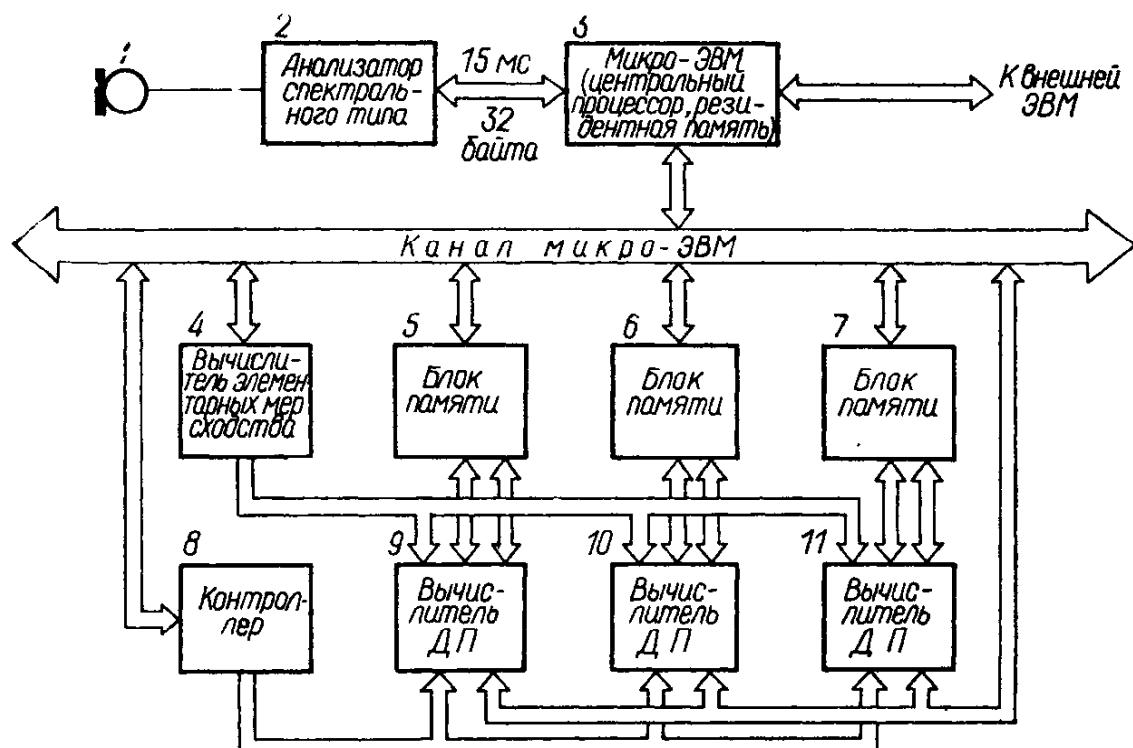


Рис. 12.2. Структура параллельной машины для распознавания слов и слитной речи.

и блоки памяти. Блоки памяти 5—7 организованы таким образом, что они доступны как микро-ЭВМ 3 через канал 12, так и специализированным вычислителем 9—11, причем каждый из них параллельно обращается к своему блоку памяти 5—7.

Сигнал микрофона 1 поступает на анализатор спектрального типа 2, основу которого составляют 32 цифровых резонансных фильтра. В каждом из 32 фильтровых каналов накапливается энергия на интервале анализа $\Delta T' = 15$ мс, и через $\Delta T = 15$ мс отсчеты этой энергии, определяемые одним байтом, считываются и передаются в микро-ЭВМ 3. Последняя, приняв результаты анализа, вырабатывает 32-разрядный двоичный код, имеющий смысл знака производной спектра по частоте на дискретной сетке частот (§ 2.1). В числе компонент этого кода находятся также двоичные признаки тональности и громкости произнесения.

Затем микро-ЭВМ 3 передает два 16-разрядных слова, представляющие распознаваемый 32-разрядный двоичный код-элемент, на вычислитель элементарных мер сходства 4.

В памяти вычислителя 4 находятся 128 эталонных элементов, представленных 32-разрядными двоичными кодами.

Одновременно с запуском вычислителя 4 микро-ЭВМ выставляет в вычислителях динамического программирования 9—11 некоторые начальные условия, необходимые для обработки текущего наблюдаемого элемента речи.

После получения сигнала готовности от вычислителя 4 о том, что элементарные меры сходства $g(x_i, e(j))$ текущего распознаваемого элемента x_i с эталонными $e(j)$, $j = 1 : 128$, уже вычислены, микро-ЭВМ 3 запускает контроллер 8, управляющий параллельной работой вычислителей динамического программирования 9—11. Эти вычислители извлекают из блоков 5—7 информацию о транскрипциях слов, ранее вычисленных интегральных мерах сходства и потенциально-оптимальных индексах, уточняют и записывают в память текущие интегральные меры сходства и потенциально-оптимальные индексы для всех состояний графа слитной речи (см. гл. 5). При этом в соответствии с транскрипциями слов используются значения элементарных мер сходства, хранящиеся в вычислителе 4.

Каждый из процессоров 9—11 обслуживает целую группу слов и вырабатывает номер слова, которое могло закончиться в данный текущий момент, момент его начала и соответствующую интегральную меру сходства.

После истечения отведенного времени, например 14 мс, контроллер 8 и вычислители 9—11 заканчивают свою работу и контроллер передает сигнал микро-ЭВМ о готовности результатов. Микро-ЭВМ считывает результаты вычислений процессоров и окончательно формирует тройку $F_i(0), k_i(0), v_i(0)$ о текущей максимальной интегральной мере сходства $F_i(0)$, потенциально-закончившемся слове $k_i(0)$ и моменте его начала $v_i(0)$ (см. гл. 5). После чего микро-ЭВМ переходит в режим ожидания сигналов готовности от анализатора 2 с тем, чтобы начать очередной 15-миллисекундный цикл обработки речевой информации.

Если сигналы готовности не поступают от анализатора 2 в течение более чем 15 мс, микро-ЭВМ переходит на формирование ответа распознавания в виде последовательности слов, переданных речевым сигналом. Одновременно осуществляется первоначальная установка данных в блоках памяти 5—7 с целью подготовки к распознаванию очередной реализации речевого сигнала.

Используя микропроцессорные элементы К580—К589 и К1801—К1810, к «Электронике-60 М» можно подключить три блока памяти 5—7 и три процессора динамического программирования 9—11, что достаточно для распознавания слов и слитной речи при словарях объемом в 500—1000 слов.

Аналогичную структуру имеет параллельная машина, осуществляющая смысловую интерпретацию слитной речи по алгоритму многозначной интерпретации (§ 7.7).

Использование универсальной микро-ЭВМ, специализированных вычислителей, многосекционной памяти, доступной как микро-ЭВМ, так и специализированным вычислителям, общей шины позволяет создать гибкую и экономичную структуру, которая простым комплексированием может быть превращена в одну из следующих систем: автономную мульти микропроцессорную систему распознавания речи с возможностями наращивания объема словаря; универсальную микро-ЭВМ с речевым вводом информации; собственно универсальную микро-ЭВМ.

§ 12.3. СТРУКТУРА КВАЗИФОНЕМНОГО ВОКОДЕРА НА 600 БИТ/С

Структура квазифонемного (поэлементного) вокодера на информационную скорость 600 бит/с представлена на рис. 12.3. Эта структура реализует информационные процессы обработки речевого сигнала, описанные в § 10.3 и § 9.4. Квазифонемный вокодер выполняется на элементах цифровой вычислительной техники.

Анализирующая часть вокодера содержит два канала обработки информации: 1) вычисления передаточной характеристики речевого тракта; 2) вычисления характеристик источников возбуждения речевого тракта.

Вокодер состоит из нескольких специализированных вычислителей, работа которых синхронизируется микро-ЭВМ 6. Она же выполняет функции кодировщика при передаче речи и декодировщика при приеме речи. Связь микро-ЭВМ 6 с цифровым каналом связи 8 осуществляется через согласователь 7.

В случае передачи речи речевой сигнал проходит через микрофон 1, фильтр нижних частот 2, аналого-цифровой преобразователь 3 и далее поступает на автокоррелятор 4, где по рекуррентным формулам по мере поступления отсчетов f_n речевого сигнала вычисляется 11-мерный вектор автокорреляции \mathbf{B} на интервале анализа 26 мс. В конце текущего интервала анализа вектор автокорреляции \mathbf{B} передается в скалятор 5, а автокоррелятор 4 обнуляется с тем, чтобы начать вычисление следующего наблюдаемого вектора автокорреляции \mathbf{B} .

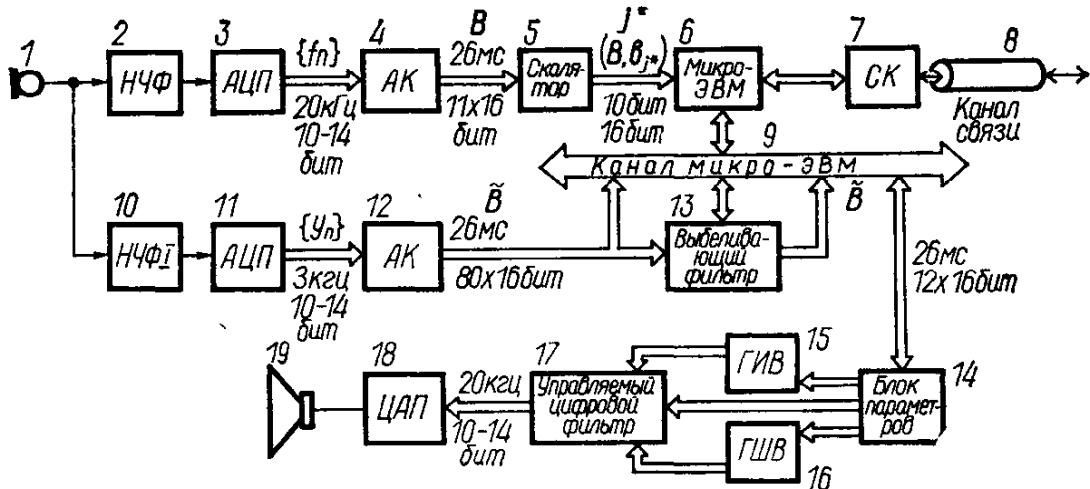


Рис. 12.3. Структура квазифонемного вокодера на 600 бит/с.

В скаляторе 5 наблюдаемый вектор автокорреляции \mathbf{B} сравнивается с $J = 512$ или $J = 1024$ эталонными элементами \mathbf{b}_j , $j = 1 : J$, и в микро-ЭВМ 6 поступает номер j^* наиболее похожего на \mathbf{B} эталонного элемента и соответствующая величина скалярного произведения $(\mathbf{B}, \mathbf{b}_{j^*})$.

Скалятор 5 является распознавателем элементов речи. Он указывает номер j^* линейной системы, которая является самой подходящей для синтеза сигнала, наиболее похожего на анализируемый речевой сигнал. Величина $(\mathbf{B}, \mathbf{b}_{j^*})$ определяет амплитуду возбуждения этой линейной системы. Квантование величины $(\mathbf{B}, \mathbf{b}_{j^*})$ и ее представление двумя битами выполняется в кодировщике, роль которого выполняет микро-ЭВМ 6.

Одновременно с определением передаточной характеристики речевого тракта и амплитуды сигнала возбуждения по параллельному каналу осуществляется вычисление значений признаков тональности. С этой целью исходный речевой сигнал подвергается низкочастотной фильтрации в блоке 10, затем аналого-цифровому преобразованию в блоке 11 и далее поступает на автокоррелятор 12. Последний вычисляет автокорреляционную функцию прореженного сигнала на интервале анализа той же длины 26 мс. Отличие этого автокоррелятора от автокоррелятора 4 только в том, что он считает все возможные отсчеты автокорреляции (в данном конкретном случае 80 отсчетов), а не первые 11, как в автокорреляторе 4. Далее первые семь отсчетов вектора автокорреляции на выходе блока 12 через канал 9 микро-ЭВМ поступают в микро-ЭВМ 6, где путем решения системы из шести уравнений (9.4.13) и применения формул (9.4.15) находятся \mathbf{b} -параметры выбеливающего фильтра, которые через канал 9 выставляются в процессоре выбеливания 13. Последний затем на основании автокорреляции блока 12 по формуле (9.4.16) вычисляет автокорреляционную функцию \mathbf{B} выбеленного речевого сигнала. По этой функции микро-ЭВМ 6 в соответствии с алгоритмом § 9.4 вычисляет признак тон — шум и период ОТ.

Данные анализа, полученные из обоих каналов анализа, кодируются и с информационной скоростью 600 бит/с через согласователь 7 передаются в канал связи 8.

При приеме сигналов из канала связи микро-ЭВМ 6 обслуживает синтезирующую часть вокодера, которую составляют блоки 14—19. Микро-ЭВМ декодирует принятые сигналы и через каждые 26 мс выставляет в блоке параметров 14 истинные значения а-параметров предсказания для линейной системы j^* , амплитуды возбуждения, признака тон — шум и периода ОТ. Для генерации речевых сигналов используются управляемый цифровой фильтр 17, управляемый генератор дискретного периодического белого шума 15 и просто генератор дискретного белого шума 16. Синтезируемый сигнал через цифро-аналоговый преобразователь 18 поступает на громкоговоритель 19.

Очевидно, что для обеспечения двухсторонних переговоров на другом конце канала связи должна быть расположена аналогичная аппаратура, содержащая микро-ЭВМ, анализирующую и синтезирующую части вокодера.

Нетрудно убедиться, что вокодер использует некоторые унифицированные блоки: автокорреляторы 4, 12, скалятор 5 и цифровые фильтры 13 и 17.

Использование микро-ЭВМ и специализированных вычислителей обеспечивает гибкость в изменении кодировки сигналов и архитектуры всего вокодера, гарантируя работу в реальном масштабе времени и с низкой информационной скоростью 600 бит/с.

§ 12.4. НЕОБХОДИМОСТЬ СИСТЕМ РЕЧЕВОГО ДИАЛОГА, ОБЪЕДИНЯЮЩИХ ФУНКЦИИ РАСПОЗНАВАНИЯ И СИНТЕЗА РЕЧИ

В наибольшей мере преимущества речевого ввода и вывода информации проявляются в системах речевого диалога, обеспечивающих двустороннюю связь человека с машиной. В общем случае системы речевого диалога (СРД) осуществляют распознавание, смысловую интерпретацию и синтез речевых сигналов и обеспечивают устный диалог с ЭВМ на формализованных или усеченных естественных языках.

Представляется, что именно системы речевого диалога, а вовсе не отдельные устройства распознавания и синтеза речи, найдут наибольшее распространение. Даже простейшие устройства распознавания отдельно произносимых слов, оснащенные синтезатором речи для озвучивания результатов распознавания, в ряде применений намного выигрывают по сравнению с устройствами с визуализацией результатов распознавания. Другой пример — информационно-справочные системы с применением синтезаторов речи существенно выиграли бы по эффективности, если бы запросы в эти системы формировались автоматически по речевым сигналам.

Опыт работы с СРД «Речь-1» [128] и другими моделями СРД серии «Речь» показал, что объединение функций распознавания и синтеза речи в одной системе создает эффект подлинного диалога (беседы, присутствия собеседника). Благодаря этому объединению достигается исключительно акустический диалог человека с машиной.

Представляется также, что СРД должна быть принципиально проще, чем два устройства — распознающее и синтезирующее — вместе взятые. Во всяком случае в рамках КДП-подхода очевидна возможность создания общей информационной базы, пригодной как для

распознавания, так и для синтеза речи (§ 6.6, [158]). В самом деле, используемые при распознавании генераторы эталонных сигналов, обычно задаваемые с помощью автоматных порождающих грамматик или графов, в равной мере могут быть использованы и для синтеза речи, поскольку генерируемые эталонные сигналы имеют смысл управляющих сигналов для собственно синтезаторов речи.

СРД могут создаваться на принципах распознавания как отдельно произносимых слов, так и слитной речи и могут использовать и пословный, и пофонемный принципы синтеза речи.

§ 12.5. РАЗРАБОТКА И ПРИМЕНЕНИЕ СИСТЕМ РЕЧЕВОГО ДИАЛОГА СЕРИИ «РЕЧЬ»

Работы по созданию СРД были начаты в 1978 г. Сначала ставилась цель создать устройство распознавания устных команд на основе серийно выпускаемой микро-ЭВМ «Электроника-60». В 1979 г. к этой работе подключились сотрудники Минского научного отдела ЦНИИС. Заказчиком работы выступило ПО им. С. П. Королева (г. Киев). В конце 1980 г. был изготовлен и сдан заказчику действующий макет устройства для распознавания 120 устных команд. В 1980 г. были начаты также работы по расширению объема словаря, оснащению устройства синтезатором речи (здесь использованы в основном результаты минских коллег) и преобразованию его, таким образом, в СРД «Речь». Одновременно в Институте кибернетики им. В. М. Глушкова АН УССР и в его СКБ ММС была выполнена соответствующая ОКР, изготовлена конструкторская документация и в 1981 г. выпущена опытная партия из 10 устройств под названием СРД «Речь-1».

Первые СРД «Речь-1» стали применяться начиная с 1982 г. Работы по развитию СРД типа «Речь» продолжались. Было выполнено ряд ОКР по созданию полностью цифровых СРД «Речь», осуществляющих распознавание, смысловую интерпретацию и синтез слов и слитной речи. В числе этих моделей варианты СРД «Речь» на специализированных СБИС или на основе типовых модулей анализа, распознавания, смысловой интерпретации и синтеза речи.

С самого начала предполагалось, что СРД должны создаваться на основе серийно выпускаемой микро-ЭВМ, например типа «Электроника-60», и содержать дополнительную к микро-ЭВМ аппаратуру, по объему, весу и стоимости не превышающую соответствующие показатели микро-ЭВМ в два раза.

Созданию СРД «Речь-1» предшествовал значительный объем исследований по распознаванию речи и синтезу речи. Достаточно сказать, что для обеспечения реализуемости алгоритмов распознавания на микро-ЭВМ в реальном времени предстояло существенно упростить (приблизительно в 50 раз) объемы вычислений и уменьшить приблизительно в 10 раз объемы памяти. На моделирующем стенде Института кибернетики им. В. М. Глушкова АН УССР было опробовано и на больших выборках испытано около 25 вариантов алгоритмов упрощенного поэлементного распознавания. В результате был найден вариант, обеспечивающий 95 %-ную надежность распознавания 200 устных команд. Одновременно велись исследования в Минском науч-

ном отделе ЦНИИ связи по совершенствованию фонемного синтезатора речи и его реализации с помощью «Электроники-60».

СРД «Речь-1», как и последующие модели СРД типа «Речь», предназначена для обеспечения двустороннего взаимодействия человека и ЭВМ посредством голоса. Она ориентирована на применение в АСУ, АСУТП, САПР, ГАП, робототехнике, ИПС и в других человеко-машинных системах сбора, обработки информации и управления [128].

Основные технические характеристики системы:

а) базовая микро-ЭВМ — «Электроника-60 М» с блоком питания БПС6-1;

б) максимальное количество распознаваемых слов — 200;

в) надежность распознавания словарей из 50, 100 и 200 слов соответственно — 98, 96, 95 %;

г) словесная разборчивость синтезируемой речи, в общем случае произвольной — 98 %;

д) задержка ответа распознавания после ввода сигнала — не более 0,3 с;

е) тип словаря — сменный, по выбору пользователя;

ж) способ адаптации к словарю и голосу пользователя — настройка на словарь и голос пользователя в режиме обучения;

з) скорость обучения распознаванию — 20 слов в минуту;

и) допустимый уровень внешних акустических шумов и помех — 85 дБ;

й) объем используемой памяти — 32 кбайт;

к) объем математического обеспечения на языке АССЕМБЛЕР — свыше 5000 операторов;

л) вес системы (без периферии микро-ЭВМ) — около 16 кг;

м) общая стоимость, включая микро-ЭВМ и периферийное оборудование, — 9 тыс. р.

Связь с внешней ЭВМ осуществляется через стандартный или модифицированный интерфейсный модуль И2 микро-ЭВМ с помощью драйвера обмена.

Отметим, что СРД «Речь-1» способна распознавать до 200 устных команд на любом индоевропейском языке и озвучивать произвольный текст на русском или украинском языках.

Технически СРД «Речь-1» состоит из двух блоков:

1) микро-ЭВМ «Электроника-60 М» вместе с периферийными устройствами (в основном варианте используется полупостоянное запоминающее устройство на 16 тысяч 16-разрядных слов для хранения программ и поэтому можно ограничиться одним периферийным устройством — алфавитно-цифровым дисплеем; в числе возможных периферийных устройств, если в этом есть необходимость, также могут быть электропищащая машинка (ЭПМ) КОНСУЛ 260, фотосчитыватель ФС 1501 и выходной перфоратор ПЛ-150);

2) блок анализа, синтеза и визуализации (блок АСВ) (в его состав входят 16-битовый спектральный анализатор, 8-битовый формантный синтезатор речи, прибор ПИУ-2 на 16 знакомест для высвечивания результатов распознавания и служебной информации, микрофон, звуковая колонка, кнопка связи на гибком шланге, тумблер связи).

В состав блока АСВ входят восемь модулей (точнее, полумодулей в конструктивах «Электроники-60 М»).

Питание системы осуществляется от стандартного блока питания БПС6-1, входящего в комплект «Электроники-60 М».

СРД «Речь-1» структурирована и организована таким образом, что может быть использована и как универсальная микро-ЭВМ, и как диалоговая вычислительная система с речевыми вводом и выводом информации.

Основные режимы работы СРД «Речь-1» — распознавание устных команд и синтез речи.

Человек-оператор произносит в микрофон устные команды. Результат распознавания в форме текста высвечивается на индикаторной панели ПИУ-2 и (или) произносится (озвучивается) синтезатором речи.

Синтезатор речи также озвучивает различного рода служебную информацию, которая одновременно выдается на ПИУ-2.

СРД «Речь-1» как терминальное устройство для внешней ЭВМ передает результаты распознавания в эту ЭВМ и озвучивает произвольный текст, который поступает в СРД из этой ЭВМ.

Возможны два режима распознавания: с кнопкой связи и без нее. В режиме с кнопкой связи тумблер связи должен быть выключен. Для обеспечения ввода и распознавания устной команды в этом случае диктор должен сначала нажать кнопку, затем произнести команду, после чего отпустить кнопку. Режим с кнопкой связи рекомендуется при работе в условиях значительных нестационарных акустических помех, в том числе при разговоре посторонних лиц.

Более удобен режим без кнопки связи. При этом должен быть включен тумблер связи. Для ввода и распознавания команды достаточно произнести отдельную команду в микрофон, не прибегая к нажатию кнопки связи. В этом режиме рекомендуется работать в условиях стационарных акустических помех и шумов до 85 дБ.

Если анализатор и синтезатор есть являются неотъемлемой частью СРД «Речь-1», то главной компонентой этой системы, определяющей решение задачи распознавания устных команд и синтеза речи, является программное обеспечение (ПО).

ПО СРД «Речь-1» организовано по блочно-модульному принципу.

После включения питания происходит автоматическая загрузка ПО СРД из модуля УЗПП в оперативную память СРД. По окончании загрузки система выходит на начальный диалог: на индикаторе ПИУ-2 высвечивается, а синтезатором озвучивается вопрос **КАКОЙ РЕЖИМ?** Ответ оператора, набираемый на клавиатуре и оканчивающийся нажатием клавиши ВК, позволяет перейти к одному из режимов, которые можно условно разделить на три группы: рабочие, служебные и тестовые.

Первую группу составляют три рабочих режима.

По команде М производится переход в рабочий режим связи с внешней ЭВМ. СРД становится речевым терминалом внешней ЭВМ. Основные функции терминала в этом режиме будут описаны далее.

Команды Р, РН, РК переводят СРД в рабочий режим распозна-

вания изолированно произносимых слов из выбранного словаря. Предполагается, что обучение на словарь и голос диктора уже произведено. На индикаторе ПИУ-2 высвечивается объем (количество слов) используемого словаря, и СРД приглашает оператора к работе, синтезируя ГОВОРИТЕ. Оператор произносит в микрофон слово-команду, далее производится распознавание, и в результате на индикаторе ПИУ-2 высвечивается, а синтезатором озвучивается ответ распознавания. Команда РН отличается от команды Р тем, что отменяет синтез ответа распознавания. Команда РК задает комментированный режим распознавания устных команд. В этом случае предполагается использование двух служебных комментирующих слов ВЕРНО и ОШИБКА. Когда оператор произносит какое-либо слово из основного словаря, то результат распознавания в этом случае высвечивается и синтезируется в вопросительной форме, как бы запрашивая подтверждение правильности ответа распознавания. Если ответ правильный, оператор произносит комментирующее слово ВЕРНО, СРД синтезирует ПРИЯТО, высвечивая на панели ПИУ-2 символы «+++», и передает результат распознавания основной команды на исполнение. Если основная команда была распознана неправильно, оператор комментирует это служебной командой ОШИБКА, после чего СРД исключает ошибочное слово из словаря на один акт распознавания и предлагает оператору, синтезируя ЕЩЕ РАЗ, повторно произнести основную команду. При этом на индикаторе ПИУ-2 высвечиваются символы «— — —». При распознавании комментирующих слов ошибки распознавания практически исключена, поскольку в этом случае распознаются два непохожих между собой слова ВЕРНО и ОШИБКА. Благодаря режиму РК гарантируется 100 %-ная надежность и правильность интерпретации вводимой информации, центральной, правда, удлинения процедуры ввода.

Команды А и АК переводят СРД «Речь-1» в рабочие режимы демонстрации игровой задачи АРИФМЕТИКА. СРД сообщает голосом правила игры и переходит в режим ожидания, синтезируя СПРАШИВАЙТЕ. Далее предоставляется возможность вводить в ЭВМ голосом простейшие задания на арифметические вычисления. Оператор произносит по цифрам первый operand (не более трех цифр), затем операцию (ПЛЮС, СЛОЖИТЬ, МИНУС, ВЫЧЕСТЬ, УМНОЖИТЬ, РАЗДЕЛИТЬ), наконец, второй operand (также не более трех цифр). Если второй operand содержит одну или две цифры, подается команда завершения СКОЛЬКО, СЧИТАЙ или ОТВЕЧАЙ. СРД вычисляет введенное выражение, сообщает голосом результат вычисления и снова переходит в состояние ожидания задания, синтезируя СПРАШИВАЙТЕ. Команда АК позволяет демонстрировать в этом же режиме комментированный ввод данных. Предполагается, что в режимах А и АК система «Речь-1» обучена не менее чем на 21 слово демонстрационного словаря.

Группу служебных составляют три режима. В служебный режим можно перейти из начального (после запроса КАКОЙ РЕЖИМ?) и из любого рабочего режима. После выполнения служебной функции происходит автоматический возврат в исходный режим.

В служебный режим обучения переходим по командам О, ОС или ОН. Синтезируется и высвечивается на индикаторе ПИУ-2 вопрос СКОЛЬКО СЛОВ? Оператор заказывает посредством клавиатуры дисплея или ЭПМ восьмеричное число — желаемое количество слов в словаре. СРД высвечивает текст (если вошли в режим по команде О) или восьмеричный номер слова (если была команда ОН), синтезирует это слово (если поступила команда ОС) и ожидает ввода реализации этого слова. Оператор произносит слово. СРД предлагает произнести следующее, и так далее до тех пор, пока не будет исчерпан заданный оператором объем словаря. Команда ОН сигнализирует, что оператор собирается работать с нестандартным словарем.

В служебный режим «коррекция» переходим по командам К и КС. Назначение этого режима — коррекция (замена) эталона слова, которое по каким-либо причинам распознается с недостаточной надежностью (например, при обучении оператор плохо произнес слово или была помеха). СРД высвечивает КАКОЕ СЛОВО? и синтезирует КАКОЕ СЛОВО ЗАМЕНИТЬ? Оператор набирает на клавиатуре восьмеричный номер слова в словаре или его текст. СРД высвечивает текст слова и синтезирует его (если была команда КС). Оператор произносит слово, старый эталон заменяется новым, и СРД возвращается в исходный режим.

Тестовые режимы предназначены для контроля анализатора, синтезатора и результатов обучения. Подобно служебным, тестовые режимы завершаются автоматической передачей управления на исходный (начальный или рабочий) режим.

В режим тестирования анализатора переходим по команде Т. СРД высвечивает на индикаторе ПИУ-2 текст ПРОВЕРКА и синтезирует СКАЖИТЕ ЧТО-НИБУДЬ. Оператор произносит в микрофон слово или короткую фразу (время экспозиции не более 2 с). На дисплей или ЭПМ выводятся видеоспектrogramma первичного описания произнесенного речевого сигнала, озаглавленная словом РЕАЛИЗАЦИЯ, а затем видеоспектrogramma сокращенного описания этого произнесения, озаглавленная словом ОПИСАНИЕ.

Для перехода в режим тестирования синтезатора необходимо подать команду С. На индикаторе ПИУ-2 высветится, а синтезатором будетзвучен текст ВВЕДИТЕ ТЕКСТ. Оператор набирает на клавиатуре фонемный текст фразы на русском или украинском языке и нажимает клавишу ВК. СРД синтезирует эту фразу.

Подав команду У, оператор имеет возможность услышать повторно набранный фонемный текст или последнее синтезированное СРД сообщение.

Для контроля результатов обучения служит тест эталонов, на который передается управление по команде Е. Исходным в этом случае является режим распознавания Р. Если какое-либо слово было распознано неправильно и оператор хочет выяснить причину ошибки, он подает команду Е. СРД высвечивает и синтезирует вопрос ЧТО СКАЗАНО? Оператор вводит текст или номер произнесенного слова. СРД выводит на дисплей или на ЭПМ видеоспектrogramму сокращенного описания распознаваемой реализации под заголовком РЕАЛИ-

ЗАЦИЯ, эталон произнесенного слова, озаглавленный его текстом или номером, и эталон ошибочного слова — ответа распознавания, озаглавленный также его текстом или номером. Сравнивая эталоны и распознаваемую реализацию, оператор или исследователь имеет возможность определить причину ошибки: плохой эталон, помеха при произнесении реализации, отказ анализатора,— а затем принять соответствующие меры.

В каждом конкретном случае использования СРД «Речь-1» принимаются эргономические меры по обеспечению достоверного восприятия устных команд. Помимо уже упоминавшегося режима комментированного ввода устных команд возможен и другой способ, ориентированный на практическое использование. Вводится служебное слово **ОТМЕНИТЬ** и включается в основной словарь. В случае ошибки распознавания произносится команда **ОТМЕНИТЬ**. По этой команде ошибочный результат распознавания стирается, слово, вызвавшее ошибку распознавания, исключается из словаря на один очередной акт распознавания, а ошибочно распознанная команда произносится повторно. В отличие от комментированного ввода этот способ ввода данных голосом требует незначительного удлинения сеанса ввода, однако предполагает дополнительные меры по повышению надежности распознавания служебного слова **ОТМЕНИТЬ**, включенного в основной словарь.

Распознавание речи в СРД «Речь-1» основано на описании 15-миллисекундных отрезков речевого сигнала с помощью 16-разрядного двоичного кода, задающего знаки разности энергий в соседних спектральных каналах (всего 17 каналов) и на вводе этих кодов в память микро-ЭВМ через каждые 15 мс (см. § 2.1).

Каждый спектральный канал представляет собой цепочку фильтр — квадратор — интегратор — нуль-орган. Нуль-орган сравнивает значения энергий, накопленных в соседних каналах. Показания нуль-органов считаются в конце интервала анализа продолжительностью 15 мс. После отсчета 16-разрядного двоичного кода интеграторы за короткое время «сбрасываются в ноль» и начинают накопление энергии от нулевого значения. Об информативности и приемлемости получаемого описания речевого сигнала говорилось в главах 2, 4, 5, 9, 11, а также в [17, 77, 78].

В процессе ввода элементов-кодов в память микро-ЭВМ осуществляется адаптивная дискретизация (сегментация) речевого сигнала, приводящая к укорочению (сокращению) распознаваемой последовательности элементов не менее чем в три раза. Соседние элементы-коды объединяются в сегменты. В сегмент входит не менее трех и не более пяти элементов.

Пусть по мере поступления элементов-кодов был уже образован тем или иным способом очередной сегмент. Верхняя граница этого сегмента одновременно будет нижней границей следующего сегмента. Образуем новый сегмент, включив в него три очередных элемента-кода. Попытаемся включить в этот сегмент и следующий, четвертый, элемент. Измерим «диаметр» сегмента из четверки элементов — максимальное хэммингово расстояние между элементами сегмента. Если

«диаметр» меньше порогового значения Θ (например, $\Theta = 3$), то пробуем включить в текущий сегмент следующий, пятый, элемент. Если «диаметр» вновь образованного сегмента из пяти элементов оказывается меньше порогового значения Θ , то включаем пятый элемент в сегмент и завершаем формирование нового сегмента. Если попытка включить в сегмент четвертый или пятый элемент приводит к значению «диаметра» сегмента больше порога Θ , то соответственно верхняя граница нового сегмента ставится после третьего или четвертого элемента; соответственно четвертый или пятый элемент уже будет считаться первым элементом последующего сегмента.

Последовательность из «средних» элементов-кодов полученных сегментов образует укороченное не менее чем в три раза (сокращенное) представление реализации речевого сигнала, которое и используется непосредственно в режимах обучения и распознавания [158, 165].

Описанная процедура сегментации исходных реализаций гарантирует некоторое локально-оптимальное разбиение на сегменты. В процессе ввода кодов в микро-ЭВМ может осуществляться и оптимальная сегментация (самосегментация или m -самосегментация, $m = 3$) речевого сигнала с помощью динамического программирования, заключающаяся в таком разбиении последовательности элементов-кодов на подпоследовательности (сегменты) не менее чем из трех элементов, что достигается минимальное суммарное отклонение элементов последовательности от «средних» элементов сегментов, которые как раз и составляют укороченное (сокращенное) описание (§ 3.2, [98, 165—167]).

Оптимальная сегментация реализаций хотя и реализуема на микро-ЭВМ одновременно с вводом элементов-кодов, однако требует в 5—7 раз больше вычислений, чем локально-оптимальная процедура. Так, из времени 15 мс между двумя последовательными поступлениями элементов-кодов тратится около 7 мс на рекуррентные вычисления оптимальной сегментации. Остающихся 8 мс недостаточно, чтобы организовать последовательное (рекуррентное по мере поступления элементов) распознавание 200 устных команд.

Локально-оптимальная же сегментация занимает около 1 мс из 15 мс, и остающегося времени 14 мс оказывается вполне достаточно для организации последовательных алгоритмов распознавания и обеспечения, таким образом, практически мгновенного ответа распознавания после окончания произнесения.

Заметим только, что может быть предложено несколько (целый ряд) алгоритмов адаптивной локально-оптимальной сегментации и сокращения реализации речевого сигнала, с приблизительно одинаковыми возможностями для практического использования.

Исходным эталоном слова в СРД «Речь-1» объявляется сокращенное описание какого-либо произнесения этого слова.

В процессе распознавания сокращенное описание предъявленного для распознавания сигнала сравнивается со всеми исходными эталонами всех слов с помощью локального динамического программирования (ЛДП) — приближенно реализуется один из вариантов поэлементного метода распознавания слов (см. гл. 2, [165, 167]). При этом используются процедуры ускорения принятия решений [169]. Ответом рас-

познавания объявляется то слово, эталон которого дал наибольшее сходство на распознаваемый сигнал.

Дадим более определенное описание процедур распознавания, реализованных в СРД «Речь-1».

Способ вычисления величины F_k сходства распознаваемой реализации $\mathbf{X}_k = (x_1, x_2, \dots, x_l, \dots, x_l)$ с эталоном $\mathbf{E}_k = (e_{k1}, e_{k2}, \dots, e_{kj}, \dots, e_{kq_k})$ слова k учитывает нелинейные изменения темпа произнесения. Ясно, что не должны допускаться чрезмерные вариации длительности как слова в целом, так и его частей. Для описания алгоритма вычисления F_k используем элементы алгольной символики.

Начальные условия: $i := j := 1; g := r := F_k := 0$.

Основной шаг (выполняется одновременно для всех k с появлением очередного распознаваемого элемента x_i):

$$\begin{aligned} (\mu^*, v^*) &:= \underset{(\mu, v) \in \Psi(i, j, r)}{\operatorname{argmin}} H(x_{i+\mu}, e_{k(j+v)}); \\ i &:= i + \mu^*; \quad j := j + v^*; \quad \rho := r; \quad r := \mu^* - v^*; \\ F_k &:= F_k + H(x_i, e_{kj}) + r \rho; \\ g &:= H(x_i, e_{kj}); \quad r := r | r + \rho|. \end{aligned}$$

Ограничения $\Psi(i, j, r)$ на параметры μ и v очередного шага задаются следующим образом:

$$\Psi(i, j, r) = \begin{cases} (\mu, v) : \mu, v = 0, 1; \quad \mu v \neq 0, \quad i + \mu \leq l, \quad j + v \leq q_k; \\ \quad |\mu - v + r| \leq 1, \quad 1 < i < l, \quad 1 < j < q_k; \\ \quad \min(0, l - q_k) - 2 \leq (i + \mu) - (j + v) \leq \\ \quad \leq \max(0, l - q_k) + 2. \end{cases}$$

Неравенство во второй строке ограничивает локальный темп произнесения, а неравенство в последней задает ограничения на изменения длительности слова в целом.

Вычисления основного шага выполняются до тех пор, пока не окажется $i = l$ и $j = q_k$. Тогда величина F_k , вычисленная в этот момент, будет характеризовать сходство распознаваемой реализации на k -е слово.

Приведенный алгоритм ЛДП для решения задачи распознавания иллюстрирует, насколько упрощаются вычисления по сравнению с поэлементным методом, и одновременно обращает внимание на возможность порождения большого семейства ЛДП-алгоритмов распознавания речи [167].

Важное значение в СРД «Речь-1» играют процедуры ускорения принятия решений в процессе распознавания. Сущность их сводится к тому, в частности, чтобы организовать вычисления таким образом, что перебираются (сравниваются с эталонами) не все слова, а только часть из них, и при этом гарантируется та же надежность распознавания, что и при полном переборе. Помимо точных представляют интерес и приближенные процедуры ускорения принятия решений при распознавании речи. Оставляя в стороне эти важные вопросы,

заметим, что может быть указан целый класс процедур ускорения принятия решений при распознавании речи, гарантирующий получение оптимального ответа распознавания (см., например, [169]).

Применение процедур сокращения реализаций, ЛДП, ускорения принятия решений позволило обеспечить в СРД «Речь-1» реальный масштаб времени распознавания 200 слов с помощью серийно выпускаемой микро-ЭВМ.

Синтез речи в СРД «Речь-1» ведется по артикуляционно-фонемному методу, разработанному в Минском научном отделе ЦНИИ связи [125—127]. Предъявленный для синтеза речи фонетический текст должен быть предварительно размечен — необходимо указать ударения в словах, синтагматические и фразовые ударения, интонационные знаки. Размеченная последовательность фонем далее трансформируется в последовательность открытых слогов СГСГСГСГ... Алгоритм образования этой последовательности прост: вставляем нейтральную гласную Γ_0 или нейтральную согласную C_0 там, где это необходимо. Каждому слогу СГ соответствует своя определенная динамика артикуляторов, которая рассчитывается исходя из статистических (характерных) значений положения артикуляторов для каждой отдельной фонемы — согласной и гласной. Далее динамика артикуляторов пересчитывается в движение формантных параметров, которые, управляя работой формантного синтезатора речи, обуславливают генерацию (синтез) речевого сигнала.

Всего используется 10 управляющих формантных параметров: F_0 — частота основного тона голосового источника возбуждения; F_1 , F_2 и F_3 — первая, вторая и соответственно третья формантные частоты; $F_{\text{фр}}$ — частота (точнее, две частоты) фрикативных формант; A_p , A_n , $A_{\text{фр}}$, $A_{\text{ас}}$ — амплитуды соответственно ротовых, носовых, фрикативных формант и аспиративных звуков; $B_{\text{фр}}$ — добротность фрикативных формант.

Выдача динамически изменяющихся формантных параметров на синтезатор речи осуществляется через каждые 10 мс. Генерация (расчет) управляющих формантных параметров производится в процессе синтеза речи.

Формантный синтезатор речи содержит три параллельных канала: а) канал ротовых формант — три последовательно включенных резонансных фильтра с нерегулируемыми полосами пропускания, регулировать можно только частоты фильтров F_1 , F_2 и F_3 ; б) канал назальных формант — два последовательно включенных резонансных фильтра с фиксированными частотами и полосами пропускания; в) шумовой канал с двумя последовательно включенными резонансными фильтрами с регулируемыми частотами $F_{\text{фр}}$ и полосами пропускания $B_{\text{фр}}$. Каналы ротовых и назальных формант возбуждаются сигналами голосового источника возбуждения с амплитудами A_p и A_n соответственно, причем импульсы ОТ следуют с частотой F_0 . Канал ротовых формант может дополнительно возбуждаться аспиративным шумом с амплитудой $A_{\text{ас}}$. Канал шумовых формант возбуждается шумом с амплитудой $A_{\text{фр}}$.

Как уже отмечалось, связь СРД «Речь-1» с внешней ЭВМ осуществ-

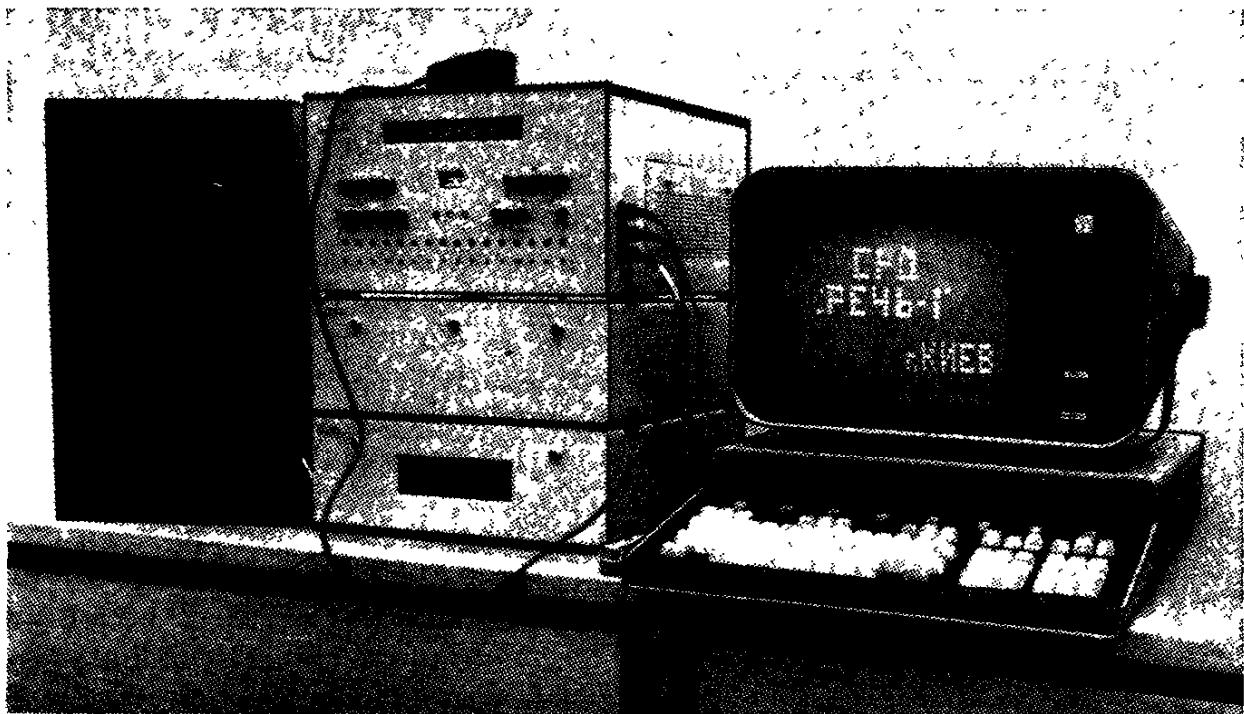


Рис. 12.4. Внешний вид системы речевого диалога СРД «Речь-1».

ляется через интерфейсный модуль И2 или через модифицированный интерфейсный модуль И2 с использованием драйвера обмена.

Рассмотрим средства сопряжения СРД «Речь-1» с СМ-4. В состав технических средств сопряжения со стороны СМ-4 входит стандартный согласователь интерфейсов ИРПР/ОШ, обычно используемый для подключения видеотерминалов типа ВТА 2000-10 или ВТА 2000-12. Он состоит из блоков БЭ 002, БЭ 810М, ПІ и кабеля интерфейсного ИРПР/ОШ.

Для обмена информацией между СРД «Речь-1» и СМ-4 используется четырехуровневая организация межмашинного обмена, при котором обеспечивается загрузка математического обеспечения СРД из СМ-4, обмен словарями распознаваемых слов и текстами синтезируемых фраз, результатами распознавания устных команд, осуществляется управление режимами работы СРД [170—171].

Внешний вид СРД «Речь-1» представлен на рис. 12.4.

С 1983 г. разрабатываются другие модели СРД серии «Речь». Все они создаются на основе «Электроники-60 М» и отличаются как техническим исполнением, так и функциональными возможностями. Рассмотрим некоторые из разрабатываемых моделей СРД серии «Речь» [168].

СРД «Речь-2» содержит, как и СРД «Речь-1», 8 нестандартных аналоговых полумодулей. Однако в микро-ЭВМ через каждые 15 мс передается не 16-разрядный двоичный код, а значения текущей энергии сигнала в 16 спектральных полосах. Реализует распознавание команд в процессе поступления данных от анализатора. Использует более совершенный, чем в СРД «Речь-1», ЛДП-метод распознавания.

СРД «Речь-12» является полным цифровым аналогом СРД «Речь-2». Восемь нестандартных модулей выполнены на быстродействующих микропроцессорах серий 1802 и 1804.



Рис. 12.5. Внешний вид СРД «Речь-1001».

Самой старшей моделью типа, которая задумана для смысловой интерпретации слитной речи и синтеза осмысленной слитной речи. Эта модель реализует метод многозначной смысловой интерпретации (см. § 7.7).

В части синтеза речи модели отличаются наличием программных модулей автоматического фонетического транскрибирования орфографических текстов, автоматического интонирования текстов, задания ритмики. Наряду с формантными синтезаторами применяются и предиктивные синтезаторы речи.

Подчеркнем, что в случае смысловой интерпретации квазислитной речи применяется все тот же алгоритм многозначной смысловой интерпретации речи с той существенно упрощающей особенностью, что сам способ произнесения фраз определяет автоматическую сегментацию речевого сигнала на сегменты — слова (см. § 7.7, а также [172]).

Пусть N -ка ($F_r^v, \mathcal{K}_r^v, \mathcal{A}_r^v$), $v = 1 : N$, — результат смысловой интерпретации начальной подпоследовательности из r слов распознаваемой фразы и пусть $\mathcal{K}_r^v, v = 1 : N$, — наиболее вероятные и разные начальные подпоследовательности из r слов, ранжированные по убыванию сходства F_r^v , а \mathcal{A}_r^v — номера мест подсловарей в ориентированной семантической сети, слова из которых могут составить допустимое продолжение \mathcal{K}_r^v (см. § 7.7).

Тогда после произнесения и распознавания $(r + 1)$ -го слова фразы на основании сходств на отдельные слова $G_{r+1}^k, k = 1 : K$, K — объем словаря, составляем все возможные подпоследовательности из $(r + 1)$ -го слова \mathcal{K}_r^v $b_k, k \in \mathcal{A}_r^v$, путем дописывания к \mathcal{K}_r^v справа нового слова $k \in \mathcal{A}_r^v$ и вычисляем соответствующие им сходства $F_r^v + G_{r+1}^k, k \in \mathcal{A}_r^v$.

СРД «Речь-112» и «Речь-113» являются полностью цифровыми моделями СРД серии «Речь», создаваемыми на базе параллельной машины для распознавания 500 слов и слитной речи (см. § 12.2). Эти модели содержат не более 24 полумодулей.

СРД «Речь-1001» (см. рис. 12.5) и «Речь-1012» осуществляют смысловую интерпретацию квазислитной (с паузами между словами) речи применительно к устному диалогу человека и ЭВМ на формализованных и усеченных естественных языках предметных областей. Отличаются от своих младших моделей более совершенным математическим обеспечением, в основном наличием лингвистического блока семантико-синтаксического анализа.

«Речь» является СРД «Речь-1112»,

интерпретации слитной речи и синтеза осмысленной слитной речи.

далее отбираем такую новую N -ку $(F_{r+1}^v, \mathcal{K}_{r+1}^v)$, $v = 1 : N$, чтобы все \mathcal{K}_{r+1}^v были разными, имели наибольшие сходства и были ранжированы по убыванию сходства. Затем формируем для каждого \mathcal{K}_{r+1}^v соответствующее продолжение \mathcal{A}_{r+1}^v .

В результате получаем N -ку $(F_{r+1}^v, \mathcal{K}_{r+1}^v, \mathcal{A}_{r+1}^v)$, $v = 1 : N$, для начальной подпоследовательности из $(r + 1)$ -го слова.

При этом возможно, что в новой N -ке все \mathcal{A}_{r+1}^v , $v = 1 : N$, одновременно пустые. Это значит, что семантически нет необходимости ждать следующего произнесения слова. Тогда проверяем, является ли \mathcal{K}_{r+1}^1 законченным предложением в языке диалога или нет. Если да, то используем \mathcal{K}_{r+1}^1 для формирования ответа распознавания и смысловой интерпретации. Если нет, то проверяем \mathcal{K}_{r+1}^2 на законченность и используем ее, если эта последовательность слов закончена, для формирования ответа распознавания. Анализируя подряд все \mathcal{K}_{r+1}^v , $v = 1 : N$, либо найдем законченное предложение, либо обнаружим, что среди \mathcal{K}_{r+1}^v , $v = 1 : N$, нет законченных предложений. В этом последнем случае выдаем отказ от интерпретации (СРД сообщает ВАС НЕ ПОНЯЛА. ЗАДАЙТЕ СВОЙ ВОПРОС КАК-НИБУДЬ ИНАЧЕ.).

В тех же случаях, когда продолжение возможно, т. е. среди \mathcal{A}_{r+1}^v , $v = 1 : N$, хотя бы для одного v имеет место $\mathcal{A}_{r+1}^v \neq \emptyset$, СРД «Речь» ждет следующего слова и, если новое, $(r + 2)$ -е, слово не появилось в течение более чем 2с после произнесения $(r + 1)$ -го слова, на основании \mathcal{K}_{r+1}^v , $v = 1 : N$, аналогично предыдущему пункту, формирует результат интерпретации, в том числе, быть может, и отказ от интерпретации.

В случае, если результат смысловой интерпретации сформирован, подготовка к обработке следующей фразы начинается с того, что полагается $F_0^v = 0$, $\mathcal{K}_0^v = \emptyset$, $\mathcal{A}_0^v = \mathcal{A}_{\text{нач}}$, $v = 1 : N$, где $\mathcal{A}_{\text{нач}}$ — номера мест всех начальных подсловарей в ориентированной семантической предметной области.

Таким образом, членение квазислитной речи на осмыслиенные фразы происходит автоматически.

Состоянием на конец 1984 г. было выпущено около 40 образцов СРД «Речь-1» и несколько образцов СРД «Речь-1001». Перечислим только некоторые примеры использования СРД типа «Речь» в составе человеко-машинных комплексов: АРМ проектировщика печатных плат, управление роботизированным комплексом для сборки микросхем, управление манипуляционным роботом с системой технического зрения, автоматизированная система обработки чертежно-графической информации, САПР машиностроения, ситуационное управление в АСУ, информационно-справочная система, конвейерное производство, автоматизированный контроль изделий, АСУ «Склад».

Использование систем речевого диалога увеличивает производительность труда, повышает эффективность использования техники, создает благоприятные условия труда и, таким образом, обеспечивает как экономический, так и социальный эффекты.

ВЫВОДЫ•

1. Показано, что для распознавания слов и слитной речи в реальном масштабе времени (объем словаря — 500 слов, порядок следования слов — свободный) требуются быстродействие, в пересчете на однопроцессорную ЭВМ, около 10 млн. операций типа сложения в секунду и оперативная память на 128 кбайт. При смысловой интерпретации слитной речи требования к скорости вычислений и объемам памяти увеличиваются не менее чем на порядок.

Процессы распознавания, смысловой интерпретации и компрессированной передачи речи легко поддаются распараллеливанию на структуре с общим потоком команд.

Автономные системы распознавания, смысловой интерпретации и компрессированной передачи речи следует создавать как мультимикропроцессорные системы с использованием выпускаемых микропроцессорных наборов или как системы на основе микро-ЭВМ с быстродействующими специализированными вычислителями типа спектрального анализатора речи, вычислителя элементарных мер сходства, процессора динамического программирования и т. п. Роль этих вычислителей могут выполнять сверхбольшие интегральные схемы для цифровой обработки сигналов.

2. Предложены конкретные структуры параллельной машины для распознавания слов и слитной речи и квазифонемного (поэлементного) вокодера для компрессированной передачи речи по цифровым каналам связи со скоростью 600 бит/с.

3. Показана необходимость и целесообразность создания систем речевого диалога, объединяющих функции распознавания и синтеза речевых сигналов. Разработано ряд моделей речевого диалога серии «Речь», обеспечено их производство и применение. Показана перспективность развития систем речевого диалога серии «Речь».

ЗАКЛЮЧЕНИЕ

В настоящей монографии показано, как одна и та же идея оказалась плодотворной в решении различных задач из области анализа, распознавания, смысловой интерпретации, компрессированной передачи и синтеза речевых сигналов. Это идея экономного задания (описания, структурирования) разнообразных и изменяющихся речевых сигналов с помощью автоматных порождающих грамматик и идея направленного поиска и последующего анализа (разбора) с помощью динамического программирования наиболее правдоподобного сигнала на множестве сигналов, генерируемых автоматными порождающими грамматиками. Положенные в основу КДП-метода анализа, распознавания, смысловой интерпретации, компрессии и синтеза речевых сигналов, эти идеи привели к конструктивному воплощению общей идеи анализа сигналов посредством их синтеза. В КДП-методе характерно то, что синтез (генерация) сигналов включен в обратную связь по отношению к процессу распознавания. При этом в едином процессе обработки речевого сигнала используются различные источники знаний об акустике и фонетике, о закономерностях речеобразования и преобразования речевых сигналов, о лексике, синтаксисе и семантике языков устного диалога человека и ЭВМ. Показано, что иерархическая организация и структурирование этой априорной информации приводят к взаимосвязанным процедурам обработки речевых сигналов на всех уровнях. При этом многие процедуры, такие, например, как сегментация речевого сигнала, хотя и выполняются, однако самостоятельного значения не имеют, более того, их окончательные результаты, если таковы интересны, устанавливаются одновременно с завершением процесса распознавания и смысловой интерпретации.

Монография иллюстрирует, как разработка упомянутых идей приводит к решению таких задач обработки речевой информации, как распознавание отдельно произносимых слов и слитной речи, смысловая интерпретация слитной речи, компрессированная передача речевых сигналов, синтез речи. В равной мере это относится и к вспомогательным задачам: обучение и самообучение распознаванию речи, оптимальная сегментация речевых сигналов, подстройка под диктора и настройка на голос оператора, реализация принципов пофонемного распознавания, выбор параметров анализатора, синтез меры сходства и др.

Конструктивный характер развивающегося КДП-метода анализа и распознавания речи подчеркнут многочисленными экспериментальными системами распознавания и смысловой интерпретации речи, разработками систем речевого диалога серии «Речь» и их применением.

Хотя КДП-метод как научное направление в области анализа, распознавания, смысловой интерпретации, компрессии и синтеза речи и позволяет трактовать разнообразные задачи обработки речевой информации с некоторой единой позиции, это все не отвергает другие возможные подходы к решению проблемы. В частности, ряд новых результатов может быть получен и в рамках подхода, основанного на иерархическом принципе обработки информации с применением многозначных решений на всех уровнях этой обработки.

Как бы то ни было, монография, раскрывающая КДП-метод, была задумана прежде всего таким образом, чтобы проиллюстрировать возможности КДП-метода, показать не только возникающие постановки задач и методы их решения, а и подвести к другим возможным постановкам задач, которые еще могут быть сделаны в рамках КДП-метода. Последнее особенно относится к глубокому фонемному распознаванию, распознаванию речи многих дикторов, распознаванию дикторов по речевым сигналам, смысловой интерпретации слитной речи.

Предстоит большая работа по упрощению процедур распознавания и ускорению принятия решений в процессе распознавания речи. Необходимо решить ряд инженерных проблем по проектированию систем. Важны также эргономические и системные вопросы. Немаловажное значение приобретают разработки языков устного диалога, создание надлежащей элементной базы для обработки речевой информации. Словом, можно надеяться, что КДП-метод сыграет свою роль и в дальнейшем решении проблем устного диалога человека и ЭВМ.

Можно также надеяться, что КДП-метод будет полезен при разработке как автоматической пишущей машинки, печатающей и редактирующей тексты под диктовку, так и машин, осуществляющих устный перевод с одного языка на другой. Иначе говоря, развитие КДП-метода представляется перспективным и полезным в создании средств устного диалога человека и ЭВМ на формализованных и естественных языках.

В заключение, перечислим ряд проблем, требующих своего безотлагательного решения для успешного продвижения в области анализа, распознавания, смысловой интерпретации и синтеза речевых сигналов.

Первая проблема — это проблема описания речевых сигналов (вычисления значений признаков). До сих пор нет четкой ясности, как, располагая речевым сигналом, определять динамически изменяющуюся передаточную характеристику речевого тракта и параметры источников его возбуждения, которые можно было бы взять за основу для экономного описания речевых сигналов. В равной мере это относится к вычислению артикуляционных признаков, качественно характеризующих движение артикуляционных органов (положение кончика языка, тела языка, губ, мягкого неба, место-

положение и размеры сужений речевого тракта, наличие-отсутствие колебаний голосовых связок и частота этих колебаний и т. п.). Существующее положение дел с вычислением собственных частот речевого тракта, декрементов затуханий, формы импульсов возбуждения речевого тракта, параметров шумового возбуждения и других аналогичных параметров также нельзя считать удовлетворительным. Решаемые задачи вычисления признаков часто базируются на необоснованном и фиксированном задании количества полюсов и нулей речевого тракта, количества параметров источников возбуждения; не учитываются энергетические потери в самом тракте, анатомические особенности речевого тракта человека. Словом, оцениваемые в процессе первичного анализа величины хотя и позволяют экономно описывать речевые сигналы, в целом имеют недостаточную физическую интерпретацию и только опосредованно передают информацию о движении артикуляционных органов и изменении параметров источников возбуждения речевого тракта. Приходится констатировать тот факт, что в большинстве случаев решаемые задачи предварительной обработки речевых сигналов являются некорректными, и, к сожалению, здесь недостаточно используются приемы регуляризации решений. Представляется, что описание речевых сигналов должно быть наглядным, экономным, легко интерпретируемым, таким, чтобы по нему можно было восстановить исходный речевой сигнал, с передачей характерных особенностей первоначального сигнала (разборчивости, качества, натуральности, индивидуальности голоса и пр.). Артикуляционные признаки, качественно характеризующие динамику артикуляционных органов, могут быть примером подобного описания.

Какой длины должен быть интервал анализа, должен ли он быть постоянным или переменным, чему должно равняться количество фильтров спектрального анализатора, каковы форма и размещение этих фильтров по оси частот, какие размерности параметров в моделях идентификации речевого сигнала и каковы должны быть сами эти модели, как вычислять значения признака тон-шум и значение мгновенной (или текущей средней) частоты основного тона речевого сигнала, как выполнить зависимый анализ сигнала для соседних интервалов анализа — эти и другие аналогичные вопросы все еще ждут своего решения.

Вторая проблема — экономное описание (задание) множеств разнообразных и изменяющихся сигналов речевых образов (классов). Если в первой проблеме речь идет об экономном представлении каждого отдельного сигнала, то по второй проблеме имеется в виду, как, используя эти экономные представления отдельных сигналов классов, экономно задавать все множество речевых сигналов, во всем их разнообразии и изменчивости. На сегодняшний день более-менее удовлетворительно удается учесть вариативность речи за счет нелинейного изменения темпа произнесения. Однако, хотя этот фактор изменчивости речи и является основным, он все же не отражает реального разнообразия речи. Даже при условии одного диктора более глубокого изучения заслуживают эффекты нелиней-

нного изменения громкости произнесения, явления коартикуляции звуков, редукции звуков, влияние просодии речи, контекста, особенности громкой и тихой речи, шепотной речи и др. Речь идет, таким образом, о разработке таких математических моделей речевого сигнала, которые более точно аппроксимируют реальные множества сигналов. Другими словами, изучению подлежат допустимые преобразования речевых сигналов. Необходимо расширить и уточнить эти преобразования. Наряду с автоматными порождающими грамматиками, для экономного задания множеств речевых сигналов надлежит найти другие приемы структурирования данных о речевых сигналах, удобные для отображения не отдельных, а целого комплекса факторов изменчивости речи. Кроме экономности, к средствам задания множеств сигналов классов предъявляются требования наглядности для человека и, самое главное, возможности направленного поиска сигнала из множества, который в том или ином смысле наиболее похож на распознаваемый сигнал.

Составной частью второй проблемы является третья, которая выделена особо в силу ее чрезвычайной важности. Это проблема вариативности речи, обусловленной индивидуальными особенностями голоса. Вопрос здесь в том, какими закономерностями преобразования связаны речевые сигналы разных дикторов или как речи одного диктора придать индивидуальные особенности речи другого. Приходится констатировать, что попытки создания многодикторных систем распознавания речи рассчитаны скорее на удачу, а не основаны на мало-мальских научных результатах. Речь идет о дикторской вариативности речи, разработке научных основ создания многодикторных систем распознавания, моделях речевого сигнала, отражающих индивидуальные особенности голоса и позволяющих придавать речи те или иные индивидуальные свойства.

Четвертая проблема — проблема сравнения речевых сигналов. Необходимо научиться отвечать на вопрос, насколько похожи или различны два анализируемых речевых сигнала. Речь идет, таким образом, об определении как элементарной, так и интегральной мер сходства речевых сигналов и эффективном их вычислении. Здесь особое внимание должно быть уделено соотношению вкладов отдельных звуков (громких и слабых, гласных и согласных, коротких и продолжительных) в интегральную меру сходства. Должны быть также рассмотрены вопросы учета громкости, ритмики и просодии речи и их роли при вычислении мер сходства.

Пятая проблема — взаимоотношение акустики и фонетики, установление связей между понятиями фонемы, дифона, слога, слова и их акустическими реализациями, формулировка фонологических правил, увязка правил коартикуляции и редукции звуков с фонетическим транскрибированием.

Шестая проблема заключается в выяснении роли априорной информации в распознавании речи и порядка ее использования, анализе и изучении задач обучения и самообучения распознаванию речи, выяснении количества оцениваемых параметров при обучении, формулировке требований к объему обучающей выборки, формиро-

вании самой обучающей выборки. К этой проблеме примыкают вопросы перенастройки систем распознавания речи на голос нового диктора, изменения словаря, разработки адаптивных систем распознавания речи.

Седьмая проблема — лингвистическое обеспечение систем распознавания речи, множественное фонетическое транскрибирование слов, формулировка правил трансформации фонетической транскрипции слова под влиянием контекста. В рамках седьмой проблемы видится изучение фонетической вариативности речевых сигналов слов и разработка эффективных приемов ее структурирования на лексико-грамматическом уровне.

Восьмая проблема касается семантики предметных областей. В рамках этой проблемы должны быть установлены эффективные способы задания множеств предложений, выражающих один и тот же смысл, и должны быть структурированы все возможные смысловые высказывания языка предметной области. Эти множества, будучи структурированы некоторым экономным образом, должны допускать направленный поиск предложений и сигналов, наиболее похожих на предъявленный для распознавания сигнал. Таким образом, здесь речь идет о конструировании языка устного диалога, о таком его экономном задании и структурировании, чтобы легко реализовалась семантическая интерпретация анализируемых речевых сигналов. С этой целью средства задания языка диалога посредством соответствующей иерархии образов должны быть детализированы вплоть до акустического уровня. Представляется, что в рамках восьмой проблемы должны найти отражение такие категории, как синтаксис и прагматика языков диалога. К этой проблеме также относятся вопросы перенастройки систем речевого диалога на ту или иную предметную область. Синтез осмыслинной речи, по-видимому, следует рассматривать как составную часть восьмой проблемы. Таким образом, смысловая интерпретация речевых сигналов с целью обеспечения устного диалога человека и ЭВМ на формализованных и усеченных естественных языках составляет основное содержание восьмой проблемы.

В девятую проблему вынесены все вопросы, связанные с обеспечением работоспособности систем распознавания речи в условиях внешних акустических шумов и помех как стационарного, так и нестационарного характера. К этой проблеме следует отнести задачи очистки речевых сигналов от помех, восстановления речи, улучшения разборчивости, качества и натуральности звучания естественной и синтезированной речи.

Десятую проблему составляет круг вопросов, связанных со спецификой анализа, распознавания, смысловой интерпретации, восстановления и синтеза речи в условиях передачи речи по коммутируемым каналам связи, когда в явной форме приходится иметь дело с изменяющимися свойствами каналов связи и окружающей среды. Учет этой «вариативности» речевых сигналов имеет важное теоретическое и практическое значение для повышения устойчивости ра-

боты систем распознавания речи и расширения областей их применения.

Перечисленные проблемы являются фундаментальными естественно-научными проблемами, имеющими принципиальное научное значение для анализа, распознавания, смысловой интерпретации и синтеза речевых сигналов. Эти проблемы составляют научную основу создания систем устного диалога человека и ЭВМ на формализованных и естественных языках.

Создание систем речевого диалога требует решения, кроме естественно-научных, целого ряда научно-технических проблем.

Сюда относятся вопросы разработки архитектур систем обработки речевой информации, распараллеливания вычислений, выбора элементной базы, в частности, синтеза типовых процессоров и реализации их на специализированных СБИС. Например, уже сейчас можно предложить в качестве типовых процессоров такие процессы как скаляр и ДП-процессор. Так, на основе скалатора можно создавать цифровые спектральные анализаторы, автокорреляторы, вычислители элементарных мер сходства. На ДП-процессорах можно решать задачи вычисления признаков тональности (основного тона и признака тон-шум), вычисления интегральных мер сходства наблюдаемых и эталонных сигналов, сегментации и компрессии речи, синтаксического и семантического анализа.

К числу научно-технических проблем следует отнести конструирование базы знаний о речи и языке, поиск структур представления данных, разработка процедур направленного поиска и анализа сигналов в иерархической структуре знаний, поиск адекватного математического языка для описания и преобразования речевых сигналов и языковых структур, разработка соответствующих средств моделирования (технических и программных), а также научной аппаратуры для визуализации, озвучивания и другого наглядного представления речевой информации.

Отдельную группу научно-технических проблем, составляют вопросы конструирования диалоговых систем обработки информации, использующих в той или иной форме речевое общение человека и ЭВМ. Это не только вопросы эргономики, разработки языков устного диалога, а и многие другие вопросы использования средств речевого ввода-вывода информации в локальных или интегрированных сетях ЭВМ, в системах «коллективного разума», в различных человеко-машинных системах сбора, обработки информации и управления, таких, как САПР, ГАП, робототехника, информационно-справочные системы и т. п. Решению подлежат и социально-экономические вопросы применения средств речевого диалога.

Сегодня трудно предсказать, какое окончательное место займут системы речевого диалога в жизни и деятельности человека. Но ясно одно, что эти системы в перспективе найдут такое же широкое применение и использование как дисплеи, клавиатура, пишущая машинка, без которых не мыслима деятельность человека, оснащенного ЭВМ. Тем актуальнее становится проблема создания систем речевого диалога, тем смелее она должна решаться.

ПРИЛОЖЕНИЯ

Приложение 1. Словарь из 200 слов

№ п/п	Текст слова	Длина исход- ного эталона	№ п/п	Текст слова	Длина исходного эталона
1	один	9	40	индекс	16
2	два	12	41	оператор	18
3	три	12	42	нет	10
4	четыре	19	43	или	12
5	пять	13	44	также	12
6	шесть	17	45	метка	17
7	семь	13	46	больше	13
8	восемь	20	47	меньше	16
9	девять	23	48	программа	23
10	ноль	11	49	подпрограмма	24
11	делить	17	50	функция	16
12	плюс	15	51	реальный	21
13	минус	13	52	целый	14
14	пробел	13	53	точный	15
15	точка	16	54	комплекс	23
16	открыть	19	55	логический	18
17	закрыть	19	56	вход	9
18	умножить	20	57	внешний	12
19	запятая	19	58	вызвать	19
20	равно	17	59	возврат	16
21	неравно	17	60	размер	18
22	кавычка	20	61	общий	18
23	двоеточие	23	62	эквивалентно	27
24	данные	13	63	кодировать	21
25	перейти	14	64	экспонента	21
26	если	9	65	логарифм	21
27	цикл	10	66	синус	13
28	стоп	11	67	косинус	19
29	читать	13	68	корень	17
30	печатать	23	69	абсолют	17
31	пробить	10	70	арктангенс	21
32	писать	11	71	тангенс	14
33	конец	17	72	остаток	20
34	поле	15	73	минимум	17
35	перемотать	21	74	цифра	14
36	назад	16	75	латинский	19
37	истинно	15	76	русский	15
38	ложно	13	77	ввод	15
39	описание	14	78	степень	18

Продолжение прилож. 1

№ п/п	Текст слова	Длина исходного эталона	№ п/п	Текст слова	Длина исходного эталона
79	градус	17	134	сумматор	18
80	минута	18	135	библиотека	24
81	секунда	17	136	ресурс	14
82	влево	17	137	заказ	16
83	вправо	17	138	время	11
84	вверх	17	139	личный	16
85	вниз	13	140	рабочий	16
86	слушай	15	141	отказ	14
87	интеграл	16	142	ячейка	17
88	радиан	19	143	ошибка	18
89	квадрат	14	144	управление	17
90	куб	9	145	резидент	22
91	натуральный	19	146	монитор	16
92	арксинус	21	147	прерывание	22
93	массив	17	148	регистр	15
94	начало	18	149	телетайп	14
95	примечание	20	150	включить	16
96	иначе	14	151	выключить	18
97	собственный	23	152	система	20
98	шаг	9	153	пакет	15
99	значение	18	154	транслятор	19
100	пока	12	155	задача	18
101	скобка	15	156	список	20
102	идентификатор	24	157	редактор	16
103	результат	16	158	сервис	16
104	переключить	20	159	автокод	20
105	присвоить	17	160	сегмент	19
106	символ	16	161	стандартный	24
107	магнитофон	20	162	головная	20
108	барабан	14	163	фактический	21
109	лента	12	164	параметр	17
110	зона	11	165	заголовок	23
111	тракт	15	166	имя	10
112	адрес	17	167	обмен	12
113	команда	18	168	временный	18
114	константа	23	169	постоянный	17
115	число	14	170	формат	13
116	счет	15	171	память	16
117	сдвиг	15	172	текст	10
118	запись	14	173	информация	18
119	пуск	14	174	карта	11
120	сумма	16	175	перфолента	17
121	стереть	18	176	двоичный	13
122	повторить	21	177	восьмеричный	20
123	вектор	14	178	десятичный	21
124	матрица	14	179	разряд	13
125	форTRAN	15	180	решать	16
126	алгол	13	181	лист	12
127	диспетчер	19	182	каталог	19
128	пульт	13	183	номер	15
129	арккосинус	20	184	часть	15
130	котангенс	21	185	останов	18
131	максимум	16	186	сравнить	16
132	производная	20	187	уравнение	17
133	машина	16	188	тождество	23

Продолжение прилож. 1

№ п/п	Текст слова	Длина исходного эталона	№ п/п	Текст слова	Длина исходного эталона
189	интервал	18	195	длина	15
190	случайный	21	196	переменная	22
191	вперед	14	197	режим	11
192	продолжать	18	198	выполнить	20
193	канал	15	199	исправить	18
194	процессор	17	200	формула	16

Приложение 2. Словарь из 1000 слов

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
1	один	43	или	85	вниз
2	два	44	также	86	слушай
3	три	45	пометить	87	интеграл
4	четыре	46	больше	88	радиан
5	пять	47	меньше	89	квадрат
6	шесть	48	программа	90	куб
7	семь	49	подпрограмма	91	натуральный
8	восемь	50	функция	92	арксинус
9	девять	51	реальный	93	массив
10	ноль	52	целый	94	начало
11	разделить	53	точный	95	примечание
12	сложить	54	комплексный	96	иначе
13	вычесть	55	логический	97	собственный
14	пробел	56	вхождение	98	шаг
15	точка	57	внешний	99	значение
16	открыть	58	вызвать	100	пока
17	закрыть	59	возвратить	101	скобка
18	умножить	60	размер	102	идентификатор
19	запятая	61	общий	103	результат
20	равно	62	эквивалентно	104	переключить
21	неравно	63	кодировать	105	присвоить
22	кавычка	64	экспонента	106	символ
23	двоеточие	65	логарифм	107	магнитофон
24	данные	66	синус	108	барабан
25	перейти	67	косинус	109	лента
26	если	68	корень	110	зона
27	цикл	69	абсолют	111	носитель
28	стоп	70	арктангенс	112	адрес
29	читать	71	тангенс	113	команда
30	печатать	72	остаток	114	константа
31	пробить	73	минимум	115	число
32	писать	74	цифра	116	счет
33	конец	75	латинский	117	сдвиг
34	пространство	76	русский	118	запись
35	перемотать	77	ввести	119	пуск
36	назад	78	степень	120	сумма
37	истинно	79	градус	121	затереть
38	ложно	80	минута	122	повторить
39	описание	81	секунда	123	вектор
40	индекс	82	влево	124	матрица
41	оператор	83	вправо	125	форTRAN
42	нет	84	вверх	126	алгол

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
127	диспетчер	182	каталог	237	Павел
128	пульт	183	номер	238	Руслан
129	арккосинус	184	часть	239	Семен
130	котаигенс	185	останов	240	Тарас
131	максимум	186	сравнить	241	Ульяна
132	производная	187	уравнение	242	Феликс
133	машина	188	тождество	243	Харитон
134	сумматор	189	интервал	244	Цезарь
135	библиотека	190	случайный	245	Чеслав
136	ресурс	191	вперед	246	Шура
137	заказ	192	продолжать	247	щека
138	время	193	связь	248	Юлия
139	личный	194	процессор	249	Яша
140	рабочий	195	длина	250	Элеонора
141	отказ	196	переменная	251	твердый
142	ячейка	197	режим	252	мягкий
143	ошибка	198	выполнить	253	арифметическое
144	управление	199	исправить	254	выражение
145	резидент	200	формула	255	устройство
146	монитор	201	конус	256	операция
147	прерывание	202	треугольник	257	положительный
148	регистр	203	ромбик	258	отрицательный
149	тететайп	204	алмаз	259	младший
150	включить	205	бульдозер	260	старший
151	выключить	206	вездеход	261	статистический
152	система	207	кирпич	262	динамический
153	пакет	208	горизонталь	263	прямой
154	транслятор	209	вертикаль	264	обратный
155	задача	210	чистый	265	необходимо
156	список	211	косой	266	достаточно
157	редактор	212	двойной	267	изменить
158	сервис	213	острый	268	использовать
159	автокод	214	кривая	269	следующий
160	сегмент	215	забор	270	масштаб
161	стандартный	216	область	271	шаблон
162	головная	217	стрелка	272	единица
163	фактический	218	отрезок	273	отчет
164	параметр	219	сместить	274	баланс
165	заголовок	220	далъше	275	комментарий
166	наименование	221	не знаю	276	цена
167	обмен	222	Александр	277	час
168	временный	223	Борис	278	каждый
169	постоянный	224	Василий	279	скаляр
170	формат	225	Геннадий	280	знак
171	память	226	Дмитрий	281	порядок
172	текст	227	Елена	282	мантийса
173	информация	228	Жанна	283	индикатор
174	перфокарта	229	Зинаида	284	буква
175	перфолеита	230	Иван	285	слово
176	двоичный	231	и краткое	286	предложение
177	восьмеричный	232	Ксения	287	строка
178	дестичный	233	Людмила	288	язык
179	разряд	234	Мария	289	состояние
180	решать	235	Надежда	290	ситуация
181	обучение	236	Оксана	291	событие

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
292	структура	347	плоскость	402	интерполяция
293	раздел	348	плотность	403	источник
294	доступ	349	распределение	404	сообщение
295	загрузка	350	поправка	405	схема
296	диагностика	351	совместное	406	передача
297	четность	352	условное	407	компонента
298	экстракод	353	усеченное	408	контур
299	ключ	354	нормальное	409	метрика
300	процедура	355	равномерное	410	модель
301	ожидать	356	рассеяние	411	мощность
302	магазин	357	эксперимент	412	энергия
303	резерв	358	смещение	413	энтропия
304	буфер	359	сокращение	414	нагрузка
305	инструкция	360	среднее	415	обнаружение
306	шифр	361	таблица	416	огибающая
307	пользователь	362	частота	417	фаза
308	признак	363	широта	418	амплитуда
309	дисплей	364	долгота	419	ограничитель
310	диск	365	элемент	420	пассивный
311	счетчик	366	эллипс	421	период
312	сигнал	367	эффективность	422	приемник
313	график	368	диагональ	423	передатчик
314	эталон	369	высота	424	продукт
315	чужой	370	детерминант	425	скорость
316	класс	371	окрестность	426	производство
317	вероятность	372	полином	427	потребление
318	достоверность	373	подмножество	428	ограничение
319	анализ	374	теорема	429	преобразование
320	синтез	375	автокорреляция	430	разложение
321	множество	376	фильтр	431	индукция
322	днalog	377	форманта	432	свертка
323	позиция	378	алфавит	433	распознавание
324	выборка	379	движение	434	совокупность
325	гистограмма	380	инерция	435	функционал
326	спектр	381	масса	436	способ
327	дисперсия	382	ускорение	437	статистика
328	отношение	383	процесс	438	интенсивность
329	предел	384	обработка	439	свойство
330	уровень	385	теория	440	напряжение
331	значимость	386	импульс	441	усилитель
332	композиция	387	выход	442	устойчивость
333	корреляция	388	выпрямитель	443	условие
334	коэффициент	389	детектор	444	понимание
335	расхождение	390	гетеродин	445	фокус
336	регрессия	391	гистерезис	446	реакция
337	критерий	392	отклик	447	частное
338	момент	393	вокодер	448	шум
339	медиана	394	непрерывный	449	температура
340	метод	395	дискретный	450	ядро
341	группа	396	количество	451	техника
342	фактор	397	качество	452	математика
343	величина	398	вещество	453	физика
344	нормировка	399	емкость	454	химия
345	гипотеза	400	резистор	455	биология
346	оценка	401	индуктивность	456	генетика

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
457	история	512	двадцать	567	явление
458	кибернетика	513	тридцать	568	природа
459	философия	514	сорок	569	охрана
460	астрономия	515	пятьдесят	570	понедельник
461	механика	516	шестьдесят	571	вторник
462	симметрия	517	семьдесят	572	среда
463	аргумент	518	восемьдесят	573	четверг
464	абсцисса	519	девяносто	574	пятница
465	ордината	520	сотня	575	суббота
466	алгебра	521	двести	576	воскресенье
467	аппроксимация	522	триста	577	неделя
468	вариация	523	четыреста	578	месяц
469	базис	524	пятьсот	579	эпоха
470	надстройка	525	шеестьсот	580	январь
471	вершина	526	семьсот	581	февраль
472	основание	527	восемьсот	582	март
473	волна	528	девятьсот	583	апрель
474	ветвь	529	дюжина	584	май
475	правило	530	тысяча	585	июнь
476	гармоника	531	миллион	586	июль
477	состав	532	миллиард	587	август
478	диаграмма	533	алгоритм	588	сентябрь
479	диаметр	534	автомат	589	октябрь
480	радиус	535	выпуклый	590	ноябрь
481	одночлен	536	вогнутый	591	декабрь
482	многочлен	537	конечный	592	перевод
483	принцип	538	поток	593	институт
484	суперпозиция	539	экран	594	выпуск
485	приближение	540	стратегия	595	абстракция
486	решение	541	материал	596	проект
487	дополнение	542	гласный	597	серия
488	итерация	543	согласный	598	триггер
489	числитель	544	носовой	599	плюс
490	знаменатель	545	фрикативный	600	минус
491	прогресс	546	щелевой	601	цилиндр
492	изображение	547	известие	602	нападение
493	инвариант	548	взрывной	603	дисковая
494	экономика	549	передний	604	упаковка
495	экстраполяция	550	задний	605	компилятор
496	испытание	551	квантовать	606	пересылка
497	колебание	552	восприятие	607	проверить
498	затухание	553	компрессор	608	указатель
499	конъюнкция	554	оборона	609	лето
500	дизъюнкция	555	фонема	610	зима
501	отрицание	556	артикуляция	611	весна
502	одиннадцать	557	защита	612	осень
503	двенадцать	558	автор	613	абонент
504	тринадцать	559	проблема	614	аванс
505	четыриадцать	560	поведение	615	авиация
506	пятнадцать	561	контроль	616	почта
507	шестнадцать	562	сознание	617	активный
508	семнадцать	563	организм	618	анатомия
509	восемнадцать	564	понятие	619	анкета
510	девятнадцать	565	механизм	620	ансамбль
511	десять	566	адаптация	621	аппарат

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
622	армия	677	граница	732	Дания
623	атомный	678	громко	733	Днепр
624	аэропорт	679	тихо	734	Донбасс
625	багаж	680	давно	735	Дунай
626	бассейн	681	далеко	736	Европа
627	башня	682	действие	737	Енисей
628	бедный	683	декада	738	Женева
629	безусловно	684	делать	739	Индия
630	беречь	685	деньги	740	Индонезия
631	билет	686	дерево	741	Ирландия
632	ближе	687	деталь	742	Испания
633	близкий	688	дешево	743	Италия
634	близость	689	диагноз	744	Кавказ
635	богатство	690	динамик	745	Казахстан
636	потенциал	691	диктор	746	Камчатка
637	большинство	692	директор	747	Канада
638	бороться	693	добавить	748	Карпаты
639	бригада	694	договор	749	Киев
640	бронза	695	документ	750	Кишинев
641	будущее	696	допускать	751	Колумбия
642	бумага	697	достичь	752	Корейская
643	быстрый	698	доход	753	народная
644	вылет	699	дружба	754	Крым
645	введение	700	единство	755	Куба
646	везде	701	Азия	756	Латвия
647	великий	702	Америка	757	Лейпциг
648	верный	703	Англия	758	Китай
649	вечер	704	океан	759	Ленинград
650	утро	705	Атлантический	760	Литва
651	вклад	706	Африка	761	Лондон
652	владеть	707	Байкал	762	Люксембург
653	влиять	708	море	763	Мадагаскар
654	вложить	709	Белоруссия	764	Мадрид
655	внимание	710	штаты	765	Мексика
656	власть	711	Бельгия	766	Минск
657	вначале	712	Болгария	767	Милан
658	ничья	713	Варшава	768	Молдавия
659	вода	714	Вашингтон	769	Монголия
660	возможно	715	Шотландия	770	Москва
661	вопрос	716	Вена	771	река
662	воспитать	717	Венгрия	772	Неаполь
663	восток	718	Вильнюс	773	близкий
664	запад	719	Владивосток	774	дальний
665	север	720	Волга	775	Конго
666	встреча	721	Волгоград	776	Новосибирск
667	вчера	722	Вьетнам	777	Амазонка
668	сегодня	723	Гавана	778	Норвегия
669	завтра	724	Германская	779	Нью-Йорк
670	выразить	725	демократическая	780	Одесса
671	выяснить	726	республика	781	Оттава
672	гарантия	727	федеративная	782	Пакистан
673	глава	728	Гималаи	783	Памир
674	голос	729	Голландия	784	Париж
675	готовить	730	Греция	785	Польша
676	грамматика	731	Грузия	786	Португалия

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
787	Прага	842	Баренцево	897	офицер
788	Россия	843	Батуми	898	лейтенант
789	Алжир	844	Белград	899	капитан
790	Алтай	845	Белое	900	майор
791	Альпы	846	Берлин	901	полковник
792	Румыния	847	Бразилия	902	космос
793	Севастополь	848	Будапешт	903	космонавт
794	Северный	849	Бухарест	904	станция
795	Ледовитый	850	Венеция	905	стыковка
796	Сибирь	851	Ереван	906	невесомость
797	Советский	852	Баку	907	вакуум
798	Союз	853	Душанбе	908	скафандр
799	София	854	Каспийское	909	шлюз
800	Сочи	855	Красное	910	отсек
801	Средиземное	856	Мраморное	911	топливо
802	Средняя	857	Египет	912	окислитель
803	Таллин	858	Охотское	913	спутник
804	Ташкент	859	Южная	914	посадка
805	Тбилиси	860	меридиан	915	шлем
806	Тюмень	861	Антарктида	916	орбита
807	Тихий	862	континент	917	Солнце
808	Токио	863	материк	918	Меркурий
809	Туркмения	864	остров	919	Венера
810	Турция	865	полуостров	920	Земля
811	Узбекистан	866	Аравия	921	Марс
812	Украина	867	Аргентина	922	астероид
813	Урал	868	топология	923	Юпитер
814	Финляндия	869	панель	924	Сатурн
815	Франция	870	проводник	925	Нептун
816	Фрунзе	871	соединение	926	Плутон
817	Харьков	872	слой	927	кольцо
818	Хельсинки	873	микросхема	928	пояс
819	Хиросима	874	модуль	929	горизонт
820	Цейлон	875	узел	930	атмосфера
821	Черное	876	компоновка	931	герметичность
822	Чехословакия	877	трассировка	932	двигатель
823	Швейцария	878	отверстие	933	агрегат
824	Швеция	879	дорожка	934	комплекс
825	Эстония	880	площадка	935	экипаж
826	Югославия	881	Уэльс	936	торможение
827	Ялта	882	зачем	937	вахта
828	Япония	883	почему	938	тренажер
829	Армения	884	артиллерия	939	акватория
830	Азербайджан	885	пехота	940	территория
831	Таджикистан	886	ракета	941	самочувствие
832	Киргизия	887	самолет	942	чувство
833	Индийский	888	солдат	943	страх
834	Австралия	889	граната	944	радость
835	Австрия	890	вертолет	945	горе
836	Азовское	891	корабль	946	уверенность
837	Амур	892	пушка	947	торжество
838	Анды	893	командир	948	гнев
839	Арктика	894	генерал	949	отвращение
840	Афины	895	адмирал	950	удовольствие
841	Балтийское	896	маршал	951	эмоция

Продолжение прилож. 2

№ п/п	Текст слова	№ п/п	Текст слова	№ п/п	Текст слова
952	разум	968	женский	984	обстоятельство
953	ощущение	969	избыточный	985	объект
954	речь	970	инверсия	986	окончание
955	абзац	971	интонация	987	отчество
956	агент	972	коммутация	988	перенос
957	акцент	973	контекст	989	прошедшее
958	аффриката	974	контраст	990	пустой
959	бинарный	975	легкий	991	семантика
960	будущее	976	лингвистика	992	сильный
961	вставка	977	многоточие	993	синтаксис
962	восклицание	978	монолог	994	слабый
963	гипербола	979	мужской	995	специальный
964	готический	980	мышление	996	субъект
965	губной	981	настоящее	997	тяжелый
966	диалект	982	обозначение	998	фамилия
967	диапазон	983	обращение	999	школа
				1000	выставка

Приложение 3. Примеры акустических транскрипций слов

№ п/п	Текст слова	Акустическая транскрипция слова
1	один	1, 2, 3, 1, 30, 22, 47, 24, 1
2	два	1, 5, 73, 5, 73, 2, 1
3	три	1, 67, 68, 38, 25, 1
4	четыре	1, 51, 67, 1, 47, 27, 56, 36, 1
5	пять	1, 46, 65, 4, 1, 29, 1
6	шесть	1, 43, 61, 46, 72, 77, 32, 1, 29, 1
7	семь	1, 10, 63, 7, 9, 72, 5, 1
8	восемь	1, 6, 71, 6, 50, 7, 77, 68, 8, 1
9	девять	1, 64, 59, 78, 49, 1, 29, 1
10	ноль	1, 24, 45, 6, 61, 48, 1
11	разделить	1, 8, 65, 28, 76, 1, 30, 9, 27, 22, 49, 1, 29, 1
12	сложить	1, 10, 11, 12, 61, 52, 22, 1, 29, 1
13	вычесть	1, 5, 13, 1, 70, 9, 54, 7, 1, 29, 1
14	пробел	1, 8, 60, 8, 1, 67, 72, 69, 6, 1
15	точки	1, 14, 50, 1, 70, 51, 1, 18, 8, 1
16	открыть	1, 69, 1, 50, 40, 52, 47, 1, 29, 1
17	закрыть	1, 10, 7, 69, 1, 50, 40, 13, 1, 29, 1
18	умножить	1, 14, 11, 19, 15, 12, 39, 61, 13, 1, 29, 1
19	запятая	1, 10, 77, 65, 1, 68, 1, 16, 69, 17, 8, 1
20	равно	1, 18, 60, 20, 45, 19, 15, 20, 5, 1
21	неравно	1, 14, 35, 72, 50, 45, 19, 24, 20, 21, 1
22	кавычка	1, 18, 57, 5, 52, 22, 1, 51, 1, 50, 18, 1
23	двоеточие	1, 5, 57, 46, 67, 1, 5, 8, 1, 75, 27, 23, 1
24	данные	1, 16, 69, 53, 24, 23, 35, 17, 8, 1
25	перейти	1, 67, 59, 38, 1, 10, 29, 49, 25, 1
26	если	1, 26, 64, 59, 46, 54, 1, 27, 1
27	цикл	1, 34, 28, 65, 52, 1, 14, 1
28	стоп	1, 55, 63, 1, 11, 6, 1
29	читать	1, 75, 51, 67, 1, 65, 3, 1, 29, 1
30	печатать	1, 9, 1, 75, 67, 65, 69, 1, 17, 1, 29, 1
31	пробить	1, 8, 17, 1, 22, 38, 1, 62, 29, 1, 29, 1

Продолжение прилож. 3

№ п/п	Текст слова	Акустическая транскрипция слова
32	писать	1, 30, 31, 77, 16, 3, 1, 29, 1
33	конец	1, 3, 23, 42, 72, 9, 1, 32, 10, 1
34	пространство	1, 5, 69, 77, 34, 1, 16, 79, 5, 1, 32, 1, 5, 33, 1
35	перемотать	1, 72, 9, 72, 40, 45, 2, 1, 16, 18, 4, 1, 29, 1
36	назад	1, 79, 69, 7, 54, 7, 58, 69, 79, 58, 1
37	истинно	1, 42, 38, 74, 7, 1, 77, 27, 23, 24, 2, 1
38	ложно	1, 14, 15, 19, 77, 39, 1, 2, 8, 1
39	описание	1, 17, 1, 67, 76, 34, 16, 4, 14, 35, 4, 36, 1
40	индекс	1, 38, 27, 1, 13, 1, 37, 10, 1
41	оператор	1, 17, 1, 56, 65, 50, 18, 58, 1, 58, 33, 1
42	нет	1, 8, 45, 38, 72, 17, 58, 1
43	или	1, 49, 38, 78, 59, 78, 1
44	также	1, 16, 18, 1, 39, 40, 8, 1
45	пометить	1, 21, 69, 41, 42, 38, 1, 29, 49, 1, 29, 1
46	больше	1, 71, 14, 40, 39, 43, 40, 8, 1
47	меньше	1, 78, 59, 42, 35, 1, 43, 40, 8, 1
48	программа	1, 33, 57, 8, 1, 50, 40, 8, 18, 60, 5, 8, 1
49	подпрограмма	1, 73, 44, 1, 6, 44, 50, 1, 8, 57, 33, 5, 45, 8, 1
50	функция	1, 55, 14, 17, 71, 1, 66, 28, 46, 17, 1

СПИСОК ЛИТЕРАТУРЫ

1. Винцюк Т. К. Сравнительные характеристики анализаторов в системе распознавания слов речи // II симпозиум по кибернетике.— Тбилиси : Ин-т киберн. АН ГССР, 1965.— С. 125.
2. Винцюк Т. К. Распознавание некоторых классов речевых сигналов : Автореф. дис. ... канд. техн. наук.— Киев, 1967.— 26 с.
3. Винцюк Т. К. Распознавание слов устной речи методами динамического программирования // Кибернетика.— 1968.— № 1.— С. 81—88.
4. Винцюк Т. К. Распознавание слов устной речи методами направленного синтеза последовательностей, составленных из эталонных элементов : VI Всесоюз. акуст. конф.— М. : Акуст. ин-т АН СССР, 1968.— 4 с.
5. Величко В. М., Загоруйко Н. Г. Автоматическое распознавание ограниченного набора устных команд // Вычисл. системы.— Ин-т мат. СО АН СССР, 1969.— Вып. 36.— С. 101—110.
6. Петров Г. М. и др. Система ввода речевых сигналов для ЭВМ/Г. М. Петров, С. Б. Аврин, А. Б. Копейкин, Т. М. Малыгина // Автоматическое распознавание слуховых образов.— Ереван : Ерев. политехн. ин-т, 1980.— С. 280—282.
7. Sakoe H., Chiba S. Оценка подобия речевых образов методом динамического программирования // Dig. 1970 Nat. Meeting Inst. Electron. Comm. Eng. Japan.— Tokyo, 1970.— Р. 136.— На яп. яз.
8. Bars J. F., Gresser J. Y., Querre M. Application de la programmation dynamique à la reconnaissance des mots // Journées d'Etudes sur la Parole.— Bruxelles, 1973.— Р. 327—334.
9. Coker M. J. An improved isolation word recognition system based upon the linear prediction residual // Record 1976 IEEE Intern. conf. on ASSP.— Philadelphia : IEEE, 1976.— Р. 206—209.
10. Woods W. et al. Speech understanding systems / W. Woods, M. Bates, G. Brown et al. In 5 vol. // BBN Inc. Report No. 3438.— Boston, 1976.
11. Westendorf C. M. Ein Erkennungssystem für fließende Lautsprache: Probleme und Strategie des Systementwurfs Modellierung fließender Sprache.— Dresden, 1979.— 16 S.— (Techn. Univ. Dresden, Sektion Informationstechnik, 1979; Inform. 09/09/79).
12. Винцюк Т. К. Распознавание непрерывной речи, составленной из слов заданного словаря // VII Всесоюз. акуст. конф. : Тез. докл.— Л., 1971.— С. 19.
13. Винцюк Т. К. Поэлементное распознавание непрерывной речи, составленной из слов заданного словаря // Кибернетика.— 1971.— № 2.— С. 133—143.
14. Система распознавания связной речи фирмы NEC // Зарубеж. радиоэлектрон.— 1980.— № 4.— С. 108—120.
15. Величко В. М. Алгоритм распознавания слитной речи с использованием семантико-синтаксических ограничений // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 342—345.
16. Винцюк Т. К., Гаврилюк О. Н., Пучкова Н. Г. Экспериментальная система ввода данных в ЦВМ посредством голоса // VIII Всесоюз. акуст. конф. Распознавание звуковых образов.— М. : Акуст. ин-т АН СССР, 1973.— С. 65—68.
17. Винцюк Т. К. и др. Система реального времени для распознавания слов и слитной речи / Т. К. Винцюк, О. Н. Гаврилюк, А. И. Куляс, А. Г. Шинкаж // Автома-

- тическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 176—178.
18. Биатов К. М., Винюк Т. К. Система смысловой интерпретации слитной речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 365—368.
19. *Speech Understanding systems // Summary of results of the five-year effort at Carnegie-Mellon University, Department of Computer Science.*— Pittsburg, 1977.— 175 р.
20. Mercier G., Quinton P., Vives R. KEAL: un système pour le dialogue oral avec une machine // Actes du Congrès de l'AFCET.— Gif-sur-Yvette, 1978.— Р. 304—314.
21. Ли У. А. Распознавание речи: прошлое, настоящее и будущее // Методы автоматического распознавания речи : Пер. с англ.— М. : Мир, 1983.— Вып. 1.— С. 65—141.
22. Клятт Д. Х. Основные результаты работ по проекту ARPA // Там же.— Вып. 2.— С. 334—360.
23. Вакита Х., Макино М. Исследования в области распознавания речи в Японии // Там же.— С. 613—629.
24. Устройства речевого ввода—вывода информации // Радиоэлектроника в 1978 году: Обзор по материалам иностр. печати.— М. : НИИ эконом. и информ. по радиоэлектрон., 1979.— Вып. 1.— С. 74—85.
25. *Proceedings of the ICASSP: In 3 vol.*— Paris, 1982.
26. Бондарко Л. В., Загоруйко Н. Г., Кожевников В. А. и др. Модель восприятия речи человеком. Новосибирск : Наука, 1968.— 60 с.
27. Галунов В. И. Бионическая модель системы распознавания речи // Исследование моделей речеобразования и речевосприятия.— Л., Науч. совет по комплекс. проблем. физиологии человека и животных АН СССР, 1981.— С. 36—51.
28. Деркач М. Ф. Автоматическое распознавание речи, основанное на дедуктивном принципе // Symposium Franco-Soviétique sur la parole.— Grenoble: ENSERG, 1981.— Р. 274—285.
29. Высоцкий Г. Я., Рудный Б. Н., Трунин-Донской В. Н., Цемель Г. И. Опыт речевого управления вычислительной машиной // Изв. АН СССР. Техн. кибернетика.— 1970.— № 2.— С. 134—143.
30. Величко В. М., Загоруйко Н. Г. Система распознавания речевых команд // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 173—175.
31. Винюк Т. К. Состояние и перспективы создания и внедрения систем распознавания и смысловой интерпретации речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 13—16.
32. Система распознавания речевого сообщения для произвольного диктора / Ю. Г. Василевский, В. И. Галунов, В. И. Золотарев и др. // Там же.— С. 510—512.
33. Голубцов С. В. Задачи и перспективы распознавания речи // Автоматическое распознавание слуховых образов (APCO VI).— Таллин : АН ЭССР, 1972.— С. 64—73.
34. Гура Б. М., Деркач М. Ф., Чабан М. Е. Распознавание слов как восстановление динамики фонемообразующих усилий по оценкам последовательностей предсказываемых фонемных тембров // Автоматическое распознавание слуховых образов.— Ереван : Ерев. политехн. ин-т, 1980.— С. 298—301.
35. Какауридзе А. Г. Экспериментальное устройство для автоматического различения ограниченного набора речевых команд // Элементы вычислительной техники и машинный перевод.— Тбилиси : Мецниереба, 1964.— С. 143—163.
36. Петров Г. М., Аврин С. Б., Копейкин А. Б. Аппаратурные и программные решения задачи ввода устных команд в диалоговых периферийных устройствах ЭВМ // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 213—215.
37. Рамишвили Г. С., Сердюков В. Д., Чикоидзе Г. Б. Система распознавания фраз-команд с предварительной верификацией оператора // Автоматическое распознавание слуховых образов.— Ереван : Ерев. политехн. ин-т., 1980.— С. 287—290.
38. Слуцкер Г. С., Беляев В. К., Кукушкин В. П. Реализация метода динамического программирования в устройстве распознавания речевых команд // Там же.— С. 309—310.
39. Пажитнов А. Л., Трунин-Донской В. Н. Обучаемая система распознавания

- изолированных слов, работающая с множеством дикторов // Вопросы кибернетики. Анализ и синтез речи в системах управления.— М., 1981.— С. 18—32.
40. Фролов Г. Д. Система распознавания речевых образов с обучением и редактированием // Программный и аппаратный контроль ЭЦВМ.— М. : Сов. радио, 1973.— С. 89—93.
41. Цемель Г. И. Опознавание речевых сигналов.— М. : Наука, 1971.— 148 с.
42. Martin T. B. Practical Applications of Voice Input to Machines // Proc. IEEE.— 1976.— 64, No. 4.— Р. 487—501.
43. Sakoe H., Chiba S. Dynamic programming algorithm optimisation for spoken word recognition // IEEE Trans. on ASSP.— 1978.— 26, No. 1.— Р. 43—49.
44. Le système de compréhension de la parole KEAL/C. Gagnoulet, F. Gillet, D. Gillet et al. // Symp. Fr.-Sov. sur la Parole.— Grenoble : ENSERG, 1981.— Р. 11 —129.
45. Liénard J. S., Mariani J. J. La reconnaissance de la parole au LIMSI: réalisation actuelle et projets // Ibid.— Р. 263—270.
46. Vintsik T. K. Generative grammars and dynamic programming in phoneme-by-phoneme recognition and semantic interpretation of connected speech // Ibid.— Р. 106—109.
47. Галунов В. И., Орлова М. И. Архитектура диалоговой системы, построенной на основе модели восприятия речи человеком // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР.— С. 16—18.
48. Восприятие речи в распознающих моделях / М. Деркач, Р. Гумецкий, Л. Мшин и др.— Львов : Изд-во Льв. ун-та, 1971.— 186 с.
49. Бондарко Л. В., Величко В. М., Загоруйко Н. Г. Словообразовательный словарь и его использование для автоматического распознавания речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 442—445.
50. Петров А. Н., Туркин В. Н., Дашико Е. Ю. Организация проблемно-ориентированной системы понимания речи и принципы распознавания // Там же.— С. 28—30.
51. Вадова З. А., Высоцкий Г. Я., Пятков В. С., Трунин-Донской В. Н. Аппаратурио-программная система автоматического понимания речи // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 108—110.
52. Liénard J. S. Les processus de la communication parlée.— Paris etc.: Masson, 1977.— 189 р.
53. Лауэр Б., Редди Д. Р. Система понимания речи Нарпу // Методы автоматического распознавания речи : Пер. с англ.— М. : Мир, 1983.— С. 459—481.
54. Sakoe H. Two Level DP-Matching — A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition // IEEE Trans. on ASSP.— 1979.— 27, No. 6.— Р. 588—595.
55. Андерсен Т. У. Введение в многомерный статистический анализ : Пер. с англ.— М. : Физматгиз, 1963.— 500 с.
56. Андерсен Т. Статистический анализ временных рядов : Пер. с англ.— М. : Мир, 1976.— 755 с.
57. Блекьюэлл Д., Гиришк М. Теория игр и статистических решений : Пер. с англ.— М. : Изд-во иностр. лит., 1958.— 374 с.
58. Миддлтон Д. Введение в статистическую теорию связи : Пер. с англ.— М. : Сов. радио, 1961.— Т. 1.— 782 с.
59. Острем К. Введение в стохастическую теорию управления : Пер. с англ.— М. : Мир, 1973.— 322 с.
60. Линник Ю. В. Статистические задачи с мешающими параметрами.— М. : Наука, 1966.— 252 с.
61. Винюк Т. К., Пучкова Н. Г. Распознавание слов речи с помощью обучаемого алгоритма // Тр. VII Всесоюз. шк.-семинара «Автоматическое распознавание слуховых образов» (1972).— Алма-Ата : Наука КазССР, 1973.— С. 16—20.
62. Винюк Т. К. Распознавание слов и фраз в экспериментальной системе ввода данных голосом // Проблемы развития средств ввода и вывода информации для ЭВМ : Тез. докл. науч.-техн. конф.— М. : Минрадиопром, 1973.— С. 50—51.
63. Фант Г. Акустическая теория речеобразования : Пер. с англ.— М. : Наука, 1964.— 284 с.
64. Фланаган Дж. Анализ, синтез и восприятие речи : Пер. с англ.— М. : Связь, 1968.— 396 с.
65. Вокодерная телефония / Под ред. А. А. Пирогова.— М. : Связь, 1974.— 535 с.

66. Сапожков М. А. Речевой сигнал в кибернетике и связи.— М. : Связьиздат, 1963.— 452 с.
67. Маркел Дж. Д., Грей А. Х. Линейное предсказание речи : Пер. с англ.— М. : Связь, 1980.— 308 с.
68. Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов : Пер. с англ.— М. : Радио и связь, 1981.— 495 с.
69. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов : Пер. с англ.— М. : Связь, 1979.— 416 с.
70. Винюк Т. К. Особенности описания речевых сигналов с помощью автокорреляционных функций // Распознавание образов и конструирование читающих автоматов.— 1967.— Вып. 1.— С. 42—57.
71. Тиман А. Ф. Теория приближения функций действительного переменного.— М. : Физматгиз, 1960.— 624 с.
72. Винюк Т. К. Сравнительные характеристики анализаторов, используемых при распознавании речи // Распознавание образов и конструирование читающих автоматов.— 1968.— Вып. 1.— С. 46—67.
73. Гуттер Р. С., Кудрявцев Л. Д., Левитан Б. М. Элементы теории функций : Справ. мат. б-ка.— М. : Физматгиз, 1963.— 244 с.
74. Atal B. S., Hanauer S. L. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave // J. Acoust. Soc. Amer. 1971.— No. 50.— P. 637—655.
75. Винюк Т. К., Людовик Е. К. Оценивание параметров речевого тракта как устойчивой линейной стохастической системы // Распознавание образов.— Киев : ИК АН УССР, 1975.— С. 3—30.
76. El Malavany I. Détermination de la fonction d'aire du conduit vocale par codage prédictif // Compte Rendu des Journées d'Études sur la parole organisées Les 31 mai, 1-er et 2 juin 1972 au CNET à Lannion.— Lannion : CNET, 1972.— P. 279—306.
77. Винюк Т. К., Куляс А. И. Универсальная программа анализа речи в реальном масштабе времени // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 182—184.
78. Винюк Т. К., Шинкаж А. Г. Пофонемное распознавание слов устной речи: алгоритмы обучения, распознавания и экспериментальные результаты // Автоматическое распознавание слуховых образов: VIII Всесоюз. семинар.— Львов : Изд-во Льв. ун-та, 1974.— Ч. 3.— С. 19—24.
79. Винюк Т. К. Конструктивный подход к распознаванию речи. Применение математического программирования в распознавании слуховых образов // Тр. IV Всесоюз. шк.-семинара «Автоматическое распознавание слуховых образов» : (APCO-IV, 1968).— Киев : ИК АН УССР, 1969.— С. 21—41.
80. Винюк Т. К., Шинкаж А. Г. Автоматическое транскрибирование образов по обучающей выборке // Обработка и распознавание сигналов.— Киев : ИК АН УССР, 1975.— С. 102—120.
81. Винюк Т. К., Гаврилюк О. Н., Людовик Е. К. Пофонемное распознавание слов и слитной речи по параметрам предсказания // Математические и технические средства робототехники и распознавания образов.— Киев : ИК АН УССР, 1981.— С. 71—79.
82. Беллман Р. Динамическое программирование : Пер. с англ.— М. : Изд-во иностр. лит., 1960.— 400 с.
83. Беллман Р., Дрейфус С. Прикладные задачи динамического программирования : Пер. с англ.— М. : Наука, 1964.— 457 с.
84. Михалевич В. С. Последовательные алгоритмы оптимизации и их применение // Кибернетика.— 1965.— № 1.— С. 45—56; № 2.— С. 85—89.
85. Винюк Т. К., Гаврилюк О. Н., Шинкаж А. Г. Пофонемное распознавание слитной речи, составляемой из слов выбранного словаря // Распознавание образов.— Киев : ИК АН УССР, 1977.— С. 34—43.
86. Слуцкер Г. С. Нелинейный метод анализа речевых сигналов // Тр. НИИРадио.— 1968.— Вып. 2.— С. 76—81.
87. Itakura F. Minimum Prediction Residual Principle Applied to Speech Recognition // IEEE Symposium on Speech Recognition.— Pittsburgh : IEEE, 1974.— P. 181—185.
88. Sakoe H., Chiba S. Dynamic programming algorithm optimisation for spoken

- word recognition // IEEE Trans. on Acoust., Speech and Signal Processing.— 1978.— 26, No. 1.— Р. 43—49.
1. 89. Винцюк Т. К. Методы обучения, самообучения и распознавания речи, основанные на составлении эталонных сигналов из элементарных частей // VIII Всесоюз. акуст. конф.: Пленар. докл.— М. : Акустич. ин-т АН ССР, 1973.— С. 75—90.
 90. Винцюк Т. К., Гаврилюк О. Н., Пучкова Н. Г. Алгоритмы распознавания слов и слитной речи и результаты их моделирования // Автоматическое распознавание слуховых образов: VIII Всесоюз. семинар.— Львов : Изд-во Льв. ун-та, 1974.— Ч. 3.— С. 33—37.
 91. Загоруйко Н. Г. Методы распознавания и их применение.— М. : Сов. радио, 1972.— 207 с.
 92. Зайцев В. Г., Тимофеев Б. Б. Распознавание клипированной речи с помощью вычислительной машины // Автоматика и приборостроение.— 1965.— № 2.— С. 19—22.
 93. Винцюк Т. К. Распознавание ограниченного набора речевых сигналов // Распознавание образов и конструирование читающих автоматов.— 1966.— Вып. 1.— С. 135—149.
 94. Винцюк Т. К. Опыты по линейному разделению слов, произнесенных многими дикторами // Там же.— 1967.— Вып. 1.— С. 58—70.
 95. Винцюк Т. К. Обучение поэлементному распознаванию речи // Там же.— 1969.— Вып. 2.— С. 23—35.
 96. Винцюк Т. К. Алгоритм определения эталонных элементов слова по совокупности его реализаций // Тр. Акуст. ин-та, 1970.— Вып. 12.— С. 163—168.
 97. Винцюк Т. К. Алгоритм оптимального членения речевого сигнала на части (сегменты) // Тр. IV Всесоюз. шк.-семинара «Автоматическое распознавание слуховых образов» (APCO-IV, 1968).— Киев : ИК АН УССР, 1969.— С. 135—142.
 98. Винцюк Т. К. Оптимальное разбиение последовательности элементов на подпоследовательности // Кибернетика.— 1970.— № 4.— С. 128—133.
 99. Винцюк Т. К. Постановка и метод решения задачи пофонемного распознавания речевых сигналов // VI Всесоюз. семинар «Автоматическое распознавание слуховых образов» (APCO-VI).— Таллин : АН ЭССР, 1972.— С. 41—48.
 100. Вінцюк Т. К. Пофонемне розпізнавання зв'язної мови. Вихідні передумови і постановка задачі // Автоматика.— 1972.— № 6.— С. 40—49.
 101. Вінцюк Т. К. Пофонемне розпізнавання зв'язної мови. Алгоритми розпізнавання, навчання і самонавчання // Там же.— 1973.— № 1.— С. 63—72.
 102. Винцюк Т. К. Работы Института кибернетики АН УССР по автоматическому распознаванию речевых сигналов // Распознавание образов : Тр. Междунар. симп. 1971 года по практическим применениям методов распознавания образов.— М. : ВЦ АН ССР, 1973.— С. 106—117.
 103. Шинкаж А. Г. Разработка и исследование алгоритмов пофонемного распознавания слов речи : Автореф дис. ... канд. техн. наук.— Киев, 1980.— 24 с.
 104. Винцюк Т. К. Распознавание и смысловая интерпретация речи // Кибернетика.— 1982.— № 5.— С. 101—111.
 105. Винцюк Т. К., Шинкаж А. Г. Автоматическое транскрибирование слов речи по обучающей выборке // Распознавание образов.— Киев : ИК АН УССР, 1975.— С. 30—47.
 106. Vintsuk T. K., Gavriluk O. N., Shinkazh A. G. Phoneme-by-phoneme Recognition of Speech Composed of the Words of Given Vocabulary // Record 1976 IEEE Intern. conf. ASSP.— Philadelphia : IEEE, 1976.— Р. 450—452.
 107. Винцюк Т. К., Гаврилюк О. Н., Шинкаж А. Г. Система пофонемного распознавания речи с обучением // IX Всесоюз. акуст. конф. Распознавание звуковых образов.— М. : Акустич. ин-т АН ССР, 1977.— С. 13—16.
 108. Винцюк Т. К., Гаврилюк О. Н., Куляс А. И. и др. Экспериментальная система пофонемного распознавания речи // Упр. системы и машины.— 1982.— № 5.— С. 17—22.
 109. Шлезингер М. И. О самопроизвольном различении образов // Читающие автоматы.— Киев : Наук. думка, 1965.— С. 38—45.
 110. Ивахненко А. Г. Самообучающиеся системы распознавания и автоматического управления.— Киев : Техніка, 1969.— 392 с.
 111. Васильев В. И. Распознающие системы : Справочник.— Киев : Наук. думка, 1969.— 292 с.

112. Цыпкин Я. З. Адаптация и обучение в автоматических системах.— М. : Наука, 1968.— 399 с.
113. Классификация и кластер : Пер. с англ. / Под ред. Дж. Вэн Райзин.— М. : Мир, 1980.— 389 с.
114. Шинкаж А. Г. Начальное приближение алгоритма обучения пофонемному распознаванию слов речи // Распознавание образов.— Киев : ИК АН УССР, 1977.— С. 43—53.
115. Винцюк Т. К., Людовик Е. К., Шинкаж А. Г. Распознавание речи на основе параметров предсказания // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 181—182.
116. Винцюк Т. К., Шинкаж А. Г. Распознавание 1 000 слов // Там же.— С. 180—181.
117. Pat. 4059725 USA, IC³ G 10 L 1/00. Automatic continuous speech recognition system employing dynamic programming / Н. Sakoe.— Publ. 22.11.77.
118. А. с. № 1159059 ССР, МКН³ G 10 L 1/00. Т. К. Винцюк, А. Б. Лысенко. Способ распознавания слитной речи и устройство для его осуществления // Открытия. Изобретения.— 1985.— № 20.
119. Винцюк Т. К. Учет синтаксиса языка при распознавании слитной речи // Обработка и распознавание сигналов.— Киев : ИК АН УССР, 1975.— С. 86—102.
120. Параллельная обработка информации. Том 2. Параллельные методы и средства распознавания образов / Под ред. А. Н. Свенсона.— Киев : Наук. думка, 1985.— 280 с.
121. Бондарко Л. В. Звуковой строй современного русского языка.— М. : Пропаганда, 1977.— 175 с.
122. Панов М. В. Современный русский язык. Фонетика.— М. : Высш. шк., 1979.— 255 с.
123. Алгоритмы преобразования русских орфографических текстов в фонетическую запись / Л. В. Златоустова, С. В. Кодзасов, О. Ф. Кривнова, И. Г. Фролова.— М. : Изд-во Моск. ун-та, 1970.— 130 с.
124. Величко В. Г., Зимовина Г. В. Автоматическое фонетическое транскрибирование печатного текста // Обработка и распознавание сигналов.— Киев : ИК АН УССР, 1975.— С. 150—179.
125. Лобанов Б. М., Марченков М. А. Алгоритм синтеза формантных параметров по тексту // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 414—417.
126. Лобанов Б. М., Марченко М. А. Алгоритм синтеза по тексту мелодического и ритмического контуров фразы // Там же.— С. 412—414.
127. Карневская Е. Б., Лобанов Б. М. Модели синтеза мелодического контура русских и английских фраз // Там же.— С. 399—402.
128. Винцюк Т. К., Лобанов Б. М., Шинкаж А. Г. Система распознавания речи и система устного диалога СРД «Речь-1» на основе микро-ЭВМ // Там же.— С. 516—521.
129. Винцюк Т. К. Проблема автоматического понимания речи // Распознавание образов.— Киев : ИК АН УССР, 1977.— С. 28—34.
130. Винцюк Т. К. Автоматическое понимание речи при устном диалоге с ЭВМ // Распознавание графических и звуковых сигналов.— Киев : ИК АН УССР, 1979.— С. 3—20.
131. Винцюк Т. К. Два основных пути создания систем распознавания и смысловой интерпретации слитной речи // Автоматическое распознавание слуховых образов.— Ереван : Ерев. политехн. ин-т, 1980.— С. 221—225.
132. Винцюк Т. К. Обобщенная задача распознавания слитной речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 345—348.
133. Биатов К. М. Подсистема анализа смысла в системе смысловой интерпретации слитной речи // Там же.— С. 369—371.
134. Рамишвили Г. С. Автоматическое опознавание говорящего по голосу.— М. : Радио и связь, 1981.— 224 с.
135. Винцюк Т. К., Куляс А. И. Задача подстройки под диктора при распознавании речи // Обработка и распознавание сигналов.— Киев : ИК АН УССР / 1975.— С. 134—149.
136. Винцюк Т. К., Куляс А. И. Подстройка под диктора при пофонемном рас-

- познавании речи // IX Всесоюз. акуст. конф. : Программа.— М. : Акустич. ин-т АН СССР, 1977.— С. 20—2.
137. Куляс А. И. Подстройка под диктора при пофонемном распознавании речи // Распознавание образов.— Киев : ИК АН УССР, 1977.— С. 54—59.
138. Винценок Т. К., Куляс А. И. Сравнительный анализ одного класса методов подстройки под диктора // Автоматическое распознавание слуховых образов.— Ереван : Ерев. политехн. ин-т, 1980.— С. 76—79.
139. Винценок Т. К., Куляс А. И., Людовик Е. К., Шинкаж А. Г. Кооперативная система распознавания речи // Там же.— С. 316—319.
140. Винценок Т. К., Куляс А. И., Людовик Е. К., Шинкаж А. Г. Эксперименты с кооперативной системой распознавания речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 298—300.
141. Кузнецов П. Г. Статистические методы верификации дикторов // Там же.— С. 203—205.
142. Винценок Т. К., Куляс А. И., Шинкаж А. Г. Система распознавания речи распознает дикторов по голосу // Речь, эмоции и личность: Материалы и сообщения Всесоюз. симп. 1978 г.— Л., 1978.— С. 65—70.
143. Винценок Т. К., Дьяченко Л. И., Куляс А. И. Некорректность анализа речевых сигналов. Выбор интервала анализа // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 76—80.
144. Itakura F. Minimum prediction residual principle applied to speech recognition // IEEE Trans. on Acoust. Speech and Signal Proc. 1975.— 23.— Р. 67—72.
145. Винценок Т. К., Людовик Е. К. Идентификация речевого тракта в классе устойчивых линейных систем, возбуждаемых белым шумом и почти периодическим сигналом // Всесоюз. науч. шк.-семинар «Автоматическое распознавание и синтез речи» : Тез. докл.— Минск : БелНИИТИ, 1976.— С. 51.
146. Винценок Т. К., Людовик Е. К. Нуль-полюсная модель анализа и восстановления речи // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 16—17.
147. Винценок Т. К. Альтернативные пути решения проблемы распознавания и смысловой интерпретации слитной речи для устного диалога человека и ЭВМ // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 8—12.
148. Людовик Е. К. Определение периода основного тона с помощью динамического программирования // Всесоюз. науч. шк.-семинар «Автоматическое распознавание и синтез речи» : Тез. докл.— Минск : БелНИИТИ, 1978.— С. 48.
149. Людовик Е. К. Алгоритм совместного определения параметров предсказания, формы, амплитуды и местоположений импульсов основного тона // Распознавание образов (изображений и речи).— Киев : ИК АН УССР, 1980.— С. 28—36.
150. Винценок Т. К. О математических моделях речевого сигнала, используемых в распознавании речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 34—37.
151. Людовик Е. К. Анализ, синтез и распознавание речи на основе параметров предсказания : Дис. на соиск. канд. техн. наук.— Киев, 1981.— 154 с.
152. Винценок Т. К., Людовик Е. К. Модель фонемного вокодера на 600 бит/с // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 184—187.
153. Винценок Т. К., Людовик Е. К. Низкоскоростной фонемный вокодер на 600 бит/с // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 466—469.
154. Техническая кибернетика. Теория автоматического регулирования / Под ред. В. В. Соловникова.— М. : Машиностроение, 1967.— Кн. 2.— 679 с.
155. Кац Б. А., Петрова Н. К., Родионова Т. М. и др. Исследование программной модели формантного синтезатора речи // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 148—149.
156. Винценок Т. К., Гаврилюк О. Н., Пучкова Н. Г., Шинкаж А. Г. Комплекс алгоритмов и программ для распознавания речи // Структура и организация пакетов программ : Междунар. конф. : Тез. докл.— Тбилиси : Мецниереба, 1976.— С. 92—95.
157. Винценок Т. К., Гаврилюк О. Н., Куляс А. И., Шинкаж А. Г. Комплекс программ для обработки и распознавания речевых сигналов // Распознавание графических и звуковых сигналов.— Киев : ИК АН УССР, 1979.— С. 71—81.

158. Винцюк Т..К. Системы речевого диалога // Материалы пятой шк.-семинара «Интерактивные системы».— Тбилиси : Мецниереба, 1983.— С. 16—22.
159. Винцюк Т. К., Шинкаш А. Г. Автоматизированная система обработки графической информации с помощью системы распознавания речи и дисплея // Автоматическое распознавание слуховых образов.— Тбилиси : Мецниереба, 1978.— С. 178—179.
160. Винцюк Т. К. Архитектура устройств распознавания речи // Распознавание образов (изображений и речи).— Киев : ИК АН УССР, 1980.— С. 3—19.
161. Винцюк Т. К. Распараллеливание вычислений в процессе анализа и распознавания речевых сигналов // Распараллеливание алгоритмов при поиске и распознавании образов в реальном времени: Докл. и сообщ. Всесоюз. шк.-семинара «Распараллеливание обработки информации».— Львов : ФМИ АН УССР, 1981.— С. 7—9.
162. Винцюк Т. К. Распараллеливание обработки информации при распознавании речевых сигналов // Распараллеливание обработки информации : IV Всесоюз. шк.-семинар : Тез. докл. и сообщ.— Львов : ФМИ АН УССР, 1983.— Ч. 2.— С. 13—14.
163. Винцюк Т. К., Лысенко А. Б. Структура параллельной машины для распознавания речи // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 449—452.
164. Винцюк Т. К., Гринчук В. М., Калмыков В. Г. Автоматизированное рабочее место (АРМ) конструктора, использующее систему речевого диалога СРД «Речь-1» // Там же.— С. 512—515.
165. Винцюк Т. К. Перспективы практического использования систем речевого диалога // Проблемы практического использования систем автоматического распознавания и синтеза речи : Тез. докл. симпоз.— Л. : НТО радиоэлектрон. и связи им. Попова, 1983.— С. 40—43.
166. Людовик Е. К. Алгоритм оптимального квазилинейного сокращения речевых сигналов // Автоматическое распознавание слуховых образов 1982.— Киев : ИК АН УССР, 1982.— С. 114—116.
167. Винцюк Т. К. Распознавание некоторых классов речевых сигналов: Дис. ... канд. техн. наук / Ин-т кибернетики АН УССР.— Киев, 1967.— 322 с.
168. Винцюк Т. К. Новые модели систем речевого диалога типа «Речь» // Автоматическое распознавание слуховых образов (АРСО-13).— Новосибирск : Ин-т мат. СО АН СССР, 1984.— Ч. 1.— С. 33—34.
169. Винцюк Т. К., Людовик Е. К. Способы ускорения принятия решений при распознавании слов речи // Автоматическое распознавание слуховых образов: VIII Всесоюз. семинар.— Львов : Изд-во Льв. ун-та, 1974.— Ч. 3.— С. 14—18.
170. Биднюк С. А. и др. Речевой диалог в САПР на базе СРД «Речь-1» / С. А. Биднюк, И. М. Блощаневич, Т. К. Винцюк и др. Автоматическое распознавание слуховых образов (АРСО-13).— Новосибирск : Ин-т мат. СО АН СССР, 1984.— Ч. 2.— С. 163.
171. Протоколы речевого обмена в интерактивных системах / С. А. Биднюк, И. М. Блощаневич, Т. К. Винцюк и др. // Там же.— С. 174.
172. Винцюк Т. К. Системы речевого диалога типа «Речь-1000» для смысловой интерпретации квазислитной речи // Там же.— С. 133—134.

Монография

Тарас Климович Винценок

**АНАЛИЗ, РАСПОЗНАВАНИЕ И ИНТЕРПРЕТАЦИЯ
РЕЧЕВЫХ СИГНАЛОВ**

*Утверждено к печати ученым советом Института кибернетики
имени В. М. Глушкова АН УССР*

Редактор Т. С. Мельник

Художественный редактор И. П. Антонюк

Технические редакторы Б. М. Кричевская, А. М. Капустина

Корректоры С. А. Снегур, Т. В. Пантелеимонова,

Л. М. Тищенко