О перспективах создания системы автоматического распознавания слитной устной русской речи

Д.Н. Бабин, И.Л. Мазуренко, А.Б. Холоденко

В статье приводится принципиальное описание механизма автоматического распознавания речи. Распознавание речи на акустическом и фонетическом уровнях в настоящее время доведено до совершенства, то есть сравнимо по качеству с надежностью распознавания отдельных звуков человеком, и соответствующие блоки распознавателя уже приобрели канонический вид. Теперь работа по созданию этих блоков распознавателя для русского, равно как и для любого другого естественного языка, может быть выполнена по готовому рецепту: сначала записать представительную базу данных, затем настроить по ней параметры вероятностных автоматов — описателей звуков и их сочетаний. Этот подход и описан в статье.

Сложность задачи распознавания речи перешла в область сокращения числа гипотез распознавания целого предложения. Для английского языка эта задача достаточно эффективно решается на основе статистических методов. Для русского языка такой подход приводит к перебору, непомерному даже для современных компьютеров. Более того, оказывается, что в русском языке просто нет статистически или эвристически представительной выборки текстов для построения статистической языковой модели, и в обозримом будущем невозможно создать такой массив текстов.

Авторы предлагают способ декомпозиции русской языковой модели на формальную и содержательную, при котором удается найти достоверный коэффициент неопределенности следующего слова предложения по двум предыдущим его словам.

1. Введение

Задача распознавания речи состоит в автоматическом восстановлении текста произносимых человеком слов, фраз или предложений на естественном языке. К важным практическим задачам, связанным с распознаванием речи, можно отнести разработку систем диктовки текстов, систем речевого управления различными устройствами, систем речевого диалога (например, по телефону). В настоящее время распознавание речи переживает период бурного роста: десятки крупных коммерческих компаний (IBM, Dragon, Philips, Microsoft,...) создали и активно развивают коммерческие системы распознавания речи [14]. Эксперты в области компьютерных технологий называют распознавание речи одной из важнейших задач XXI века.

Исторически, методы распознавания речи развивались вместе с развитием компьютеров. Задача распознавания речи изначально ставилась как задача восстановления текста раздельно произносимых слов. Только в последние десятилетия компьютерная техника достигла такого уровня, когда стала осмысленной задача распознавания слитной или даже спонтанной устной речи. На этом этапе выяснилось, что для решения задачи распознавания речи недостаточно уметь распознавать отдельные звуки и слова (команды) с надежностью, сравнимой с надежностью распознавания отдельных команд человеком. Как показала практика, человек при распознавании слитной речи для устранения неоднозначности восстановления текста предложения существенно использует свои знания о естественном языке, а также смысл произносимого. Поэтому задачу распознавания речи естественно разделить на две независимые задачи:

- задачу локального распознавания речи (то есть распознавания отдельной команды)
- задачу восстановления текста слитной речи по множеству возможных гипотез распознавания.

Для решения первой задачи существенно знание природы процесса речеобразования [19]. Существуют универсальные модели этого процесса, общие для различных естественных языков [13]. Решение второй задачи, наоборот, сильно зависит от особенностей естественного языка, на котором произносятся слова. Фактически, на этом

уровне построение системы распознавания речи для каждой новой языковой группы требует своих, особых математических и технических подходов и сводится к использованию некоторой последовательной формальной модели этого естественного языка.

На Западе задача распознавания речи входит в число перспективных научно-технических направлений и хорошо финансируется, что обусловило тот факт, что в последние десятилетия созданы и достаточно эффективно работают системы распознавания речи для английского и некоторых других языков. Системы распознавания русской речи с подобными характеристиками не существует. Анализу перспектив создания систем автоматического распознавания русской слитной речи и посвящен настоящий обзор.

2. Устройство системы автоматического распознавания речи

Основными характеристиками современных систем автоматического распознавания речи являются следующие:

- словари размером в десятки и сотни тысяч слов;
- распознавание слитной речи;
- работа в реальном времени;
- возможность работы как с предварительной настройкой на голос диктора, так и без настройки;
- надежность работы 95–98% для грамматически правильных текстов.

Структурная схема работы типичной современной системы распознавания слитной речи изображена на рисунке 1.

3. Функционирование системы происходит следующим образом

Оцифрованный речевой сигнал поступает на вход компьютера. Затем сигнал с некоторым постоянным шагом разбивается на окна,

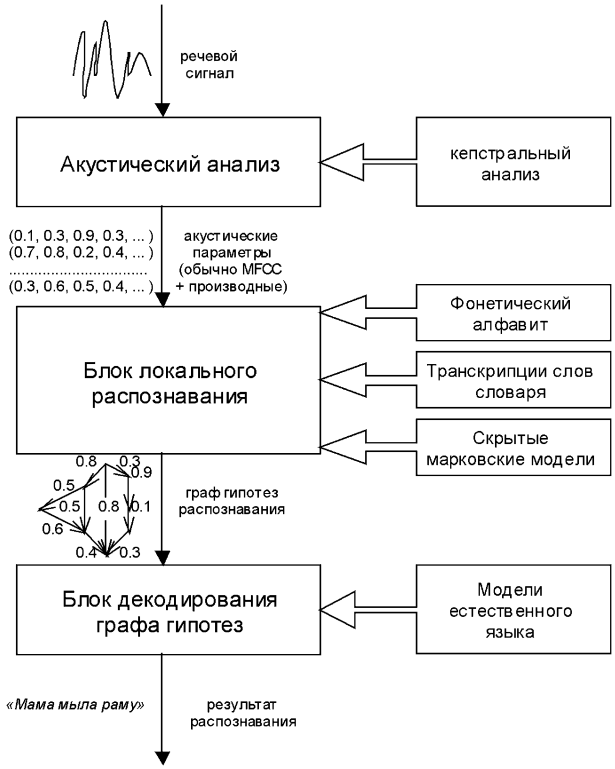

Рис. 1.

и для каждого окна в блоке акустического анализа считается вектор значений некоторых спектральных параметров, чаще всего кепстральных коэффициентов, а также их первой и второй дискретных производных.

Вектора параметров последовательно поступают на вход блока локального распознавания, обычно имеющий в своей основе универсальный монотонный вероятностный автомат [11, 13], объединяющий эталонные вероятностные автоматы всех слов естественного языка, с которыми работает распознающая система. При поступлении на вход этого блока каждого нового окна анализа модифицируется ориентированный нагруженный граф гипотез распознавания — в него добавляются новые гипотезы о произнесенной последовательности слов

языка и удаляются существующие гипотезы, вероятность которых становится меньше некоторого фиксированного порога. Когда поступает последний вектор значений параметров, в графе оставляются только те гипотезы, которые заканчиваются на целое (законченное) слово языка. Для эффективного функционирования блока локального распознавания существенную роль играет выбор фонетического алфавита, что является отдельной исследовательской задачей.

Для выделения из графа гипотез распознавания единственного предложения естественного языка, как результата распознавания, используются наши знания о структуре естественного языка. Модель языка (чаще всего основанная на статистическом подходе) позволяет выбрать среди всего множества путей в графе гипотез один, имеющий максимальную итоговую вероятность. Найденная гипотеза и считается результатом распознавания.

Следует отметить, что функционирование описанной распознающей системы является эффективным только после обучения на основе текстовых и акустических баз данных (корпусов), обладающих достаточно большим объемом и представительностью. Текстовые базы данных необходимы для обучения и проверки эффективности языковых моделей, а речевые — для настройки параметров алгоритмов локального распознавания, чаще всего основанных на применении монотонных вероятностных автоматов. Сбор и обработка таких баз данных является, пожалуй, одним из самых трудоемких этапов в построении систем распознавания речи и требует, помимо всего прочего, наличия достаточного полного словаря естественного языка, систем морфологического анализа, системы автоматического транскрибирования текстов.

4. Акустический уровень

На этапе первичной обработки сигнала основной задачей является извлечение из речи набора признаков, от которых обычно требуется выполнение следующих свойств:

1) Минимизация объема, то есть максимальное сжатие информации и статистическая некоррелированность параметров;

2) Независимость от диктора, то есть максимально возможное удаление информации, связанной с индивидуальными особенностями говорящего;
3) Однородность, то есть параметры должны в среднем иметь одинаковую дисперсию;
4) Возможность применения простых метрик типа Евклидовой для определения близости между наборами признаков, при этом близость участков звука на слух должна быть согласована с близостью в смысле этой метрики.

Наиболее распространенными наборами параметров, используемыми в системах распознавания речи, являются коэффициенты преобразования Фурье [10] (БПФ [2]), коэффициенты линейного предсказания [18] и основанный на них спектр линейного предсказания (сглаженный спектр), а также кепстральные коэффициенты [18].

Коэффициенты кепстра (MFCC — mel-frequency-scaled cepstral coefficients), получающиеся путем последовательного применения к анализируемому отрезку сигнала дискретного преобразования Фурье, спектрального сглаживания, приведения к логарифмической шкале и, наконец, применения действительной части прямого преобразования Фурье, являются наиболее эффективными с точки зрения описанных выше свойств 1–4. Для одной и той же подробности анализа их число (обычно 10–15) значительно меньше числа коэффициентов спектра БПФ, дикторозависимая информация удалена с помощью сглаживания спектра, а информация сжата за счет приведения спектра к логарифмической шкале частот. С целью учета изменения параметров во времени обычно вместе с коэффициентами кепстра рассматривают также их первую и вторую дискретные производные.

Аппарат акустического анализа достаточно развит и переносим с одного языка на другой, что позволяет эффективно применять все наработанные в этой области приемы и подходы при создании распознавателя русской речи. В частности, практически все описанные выше способы расчета акустических параметров речевого сигнала реализованы в известных общедоступных математических компьютерных пакетах обработки сигналов, например, в пакетах SPL и IPPS фирмы Intel [46].

5. Локальное распознавание речи. Метод скрытых марковских моделей

Методы локального распознавания речи [12] можно условно разделить на две большие группы: непараметрические — с использованием непараметрических мер близости к эталонам (к ним можно отнести методы на основе формальных грамматик и методы на основе метрик на множестве речевых сигналов) — и параметрические (вероятностные — на основе метода скрытых марковских процессов, нейросетевые).

Первые устройства автоматического распознавания речи [30, 32, 33] были аналоговыми и использовали пороговую логику, поэтому они не обладали высокой надежностью и были узкоспециализированными. После появления лингвистической теории речи, представляющей речь как производную фонетической транскрипции текста произносимого слова, для распознавания стал использоваться метод фонетической сегментации [39, 45], однако впоследствии выяснилось, что эта задача трудно поддается точному автоматическому решению.

Следующим этапом стало развитие непараметрических подходов, основанных на мерах близости на множестве речевых сигналов. Подход Винцюка [6, 5], основанный на методе динамического программирования (Беллман, [1]), развитый Итакурой [35] и др., позволил сократить время вычисления значений функции близости к эталонным сигналам с экспоненциального (от длины сигнала) до квадратичного. В силу того, что основной спецификой метода являлось нелинейное искажение временной оси одной из сравниваемых функций, метод получил название «динамической деформации времени» (ДДВ). К очевидным достоинствам метода ДДВ относятся простота его реализации и обучения, а основными недостатками метода является сложность вычисления меры близости (пропорционально квадрату длины сигнала) и большой объем памяти, необходимый для хранения эталонов команд (пропорционально длине сигнала и количеству команд в словаре).

Методы, использующиеся в задаче локального распознавания речи в настоящее время, были впервые предложены рядом американских исследователей (Бейкер — CMU — система «Драгон» [23] и,

Джелинек — IBM [7]) в 1970-е годы прошлого века. Они применили теорию скрытых марковских моделей (СММ), созданную Баумом и коллегами [26, 40, 42, 17, 25]. Скрытые марковские модели представляют из себя дважды стохастические процессы — марковские цепи [15] по переходам между состояниями и множества стационарных процессов в каждом состоянии цепи. Для обучения моделей и вычисления вероятности наблюдения слова на выходе СММ был также применен метод динамического программирования (алгоритмы прямого-обратного хода [25], Баума-Уэлча, или EM-алгоритм [31], Виттерби [44, 20]). Достоинствами метода СММ являются достаточно быстрый способ вычисления значений функции расстояния (вероятности) и существенно меньший, по сравнению с методом ДДВ, объем памяти, необходимый для хранения эталонов команд (пропорционально количеству фонем, трифонов и т.п. в языке), а основными недостатками — достаточно большая сложность его реализации, а также необходимость использования больших фонетически сбалансированных речевых корпусов (баз данных) для обучения параметров СММ. По сути, методы ДДВ и СММ имеют очень много общего и могут считаться разными реализациями одного и того же подхода [36].

СММ, возникшие как обобщение цепей Маркова, тесно связаны с понятием вероятностного автомата. Вероятностные автоматы, впервые введенные в общей форме Дж. Карлайлом (1963, [27]) и, независимо от него, Р.Бухараевым (1964, [3]) и П.Штарке (1965, [43]), представляют из себя в практическом плане устройства с конечной памятью, перерабатывающие информацию с входных каналов в выходные, переходы и выходы которых происходят на основе вероятностных законов [4]. Скрытые марковские модели являются частным случаем вероятностных автоматов, а именно, вероятностными автоматами без входа. СММ, используемые в системах распознавания речи, обладают дополнительно тем свойством, что на каждом такте работы автомата переход осуществляется в состояние с тем же или большим номером. Такие модели, предложенные впервые Бакисом [24, 7], называются лево-правыми (left-right), или моделями Бакиса. В [13] предложено называть соответствующие этим моделям вероятностные автоматы монотонными.

Согласно [13], метод скрытых марковских моделей [26] можно из-

ложить на языке вероятностных автоматов. Рассматривается частный случай вероятностных автоматов — инициальные (заданы начальное и финальное состояния) автономные (без входа) монотонные автоматы Мура (выход и переход в следующее состояния осуществляются независимо).

Инициальным автономным монотонным вероятностным автоматом Мура (далее — монотонным вероятностным автоматом) называется шестерка $\mathcal{A} = \langle C, Q, \pi, P, \nu_0, \nu_F \rangle$, где C — конечный выходной алфавит ($|C| = k$), Q — конечный алфавит состояний ($|Q| = m$), π — стохастическая $m \times m$-матрица, такая что π_{ij} задает вероятность перехода из состояния q_i в состояние q_j и $\pi_{ij} = 0$ при $i > j$ и при $i = j = m$, $\pi_{ij} < 1$ при $i = j$, P — стохастическая $m \times k$-матрица, такая что P_{il} задает вероятность выдачи буквы c_l в состоянии q_i, $\nu_0 = (1, 0, \dots 0)$ — вектор длины m, означающий, что q_1 — начальное состояние, $\nu_F = (0, 0, \dots, 0, 1)^T$ — вектор-столбец длины m, означающий, что q_m — финальное состояние автомата.

Считается, что монотонный автомат $\mathcal{A}$ выдал слово $c_1 c_2 \dots c_n$, если автомат начал работу в состоянии q_1, выдал в этом состоянии букву c_1 согласно распределению вероятностей $(P_{11}, P_{12}, \dots, P_{1k})$, далее перешел в следующее состояние согласно распределению вероятностей $(\pi_{11}, \pi_{12}, \dots, \pi_{1m})$ и т.д., наконец, находясь в некотором состоянии q_{i_n}, выдал букву c_n согласно распределению вероятностей в этом состоянии и с вероятностью $\pi_{i_n m}$ перешел в финальное состояние q_m, где и закончил работу. При этом вероятность выдачи автоматом слова $\gamma = c_1 c_2 \dots c_n$ подсчитывается по формуле (*), имеющей смысл для любого автономного вероятностного автомата [4]:

$$p_{\mathcal{A}}(\gamma) = \sum_{\substack{\overline{q}=q_{i_1} q_{i_2} \dots q_{i_{n-1}}: \\ \nu_F(q_{i_{n-1}})=1}} P(\gamma|\overline{q}) =$$

$$= \sum_{\substack{\overline{q}=q_{i_1} q_{i_2} \dots q_{i_{n-1}}: \\ \nu_F(q_{i_{n-1}})=1}} \nu_0(q_{i_1}) P_{i_1}(c_1) \pi_{i_1 i_2} P_{i_2}(c_2) \pi_{i_2 i_3} \cdot \dots \cdot P_{i_n}(c_n) \pi_{i_n i_{n+1}} =$$

$$= \nu_0 M(\gamma) (\nu_F)^T,$$

где $M(\gamma) = M(c_1) M(c_2) \cdot \dots \cdot M(c_n) = \pi_{ij} P_{il}$.

Число операций, необходимых для вычисления $(*)$ можно сократить, используя метод динамического программирования [1]. Алгоритм, реализующий этот эффективный по времени способ вычисления вероятности выдачи слова автоматом, называется алгоритмом прямого-обратного хода [17]. С целью ускорения вычислений часто вместо вычисления вероятности по всем цепочкам состояний, ведущим от начального состояния к финальному, находят цепочку с максимальной вероятностью для данного выходного слова и эту вероятность объявляют искомой вероятностью (алгоритм Витерби [20]). Несмотря на то, что теоретическое обоснование такого подхода неясно, он дает значительный выигрыш во времени, что обуславливает популярность этого метода.

Отдельной и, пожалуй, самой нетривиальной задачей на этапе локального распознавания речи является задача синтеза (обучения параметров) монотонного вероятностного автомата. Для пояснения постановки и методов решения этой задачи требуется описать содержательно, как применяются вероятностные автоматы для решения задачи локального распознавания речи.

Имея транскрипцию $b_1 b_2 \dots b_n$ какого-либо слова словаря, можно построить по ней последовательность трифонов (звуков в контексте двух соседних звуков), из которых состоит произнесение этого слова: $\_b_1b_2, b_1b_2b_3, b_2b_3b_4, \dots, b_{n-2}b_{n-1}b_n, b_{n-1}b_n\_$. Каждому трифону в естественном языке (ясно, что трифонов в языке меньше, чем всех возможных троек звуков) сопоставляется эталон в виде монотонного вероятностного автомата из четырех состояний, так что звуковые сигналы всех возможных произнесений каждого трифона рассматриваются как слова, порожденные этим вероятностным автоматом. Эталонные вероятностные автоматы для слов естественного языка составляются путем последовательного соединения соответствующих эталонных автоматов трифонов, при этом финальные состояния всех таких автоматов, кроме последнего, склеиваются с первым состоянием следующего трифона.

Обучение вероятностного автомата состоит в том, чтобы при заданном числе состояний автомата и, быть может, некотором начальном приближении значений параметров автомата (матриц P и π) по заданному конечному набору выходных слов (обучающей выборке)

таким образом модифицировать матрицы P и π, чтобы итоговая вероятность выдачи этим автоматом каждого из слов обучающей выборки увеличилась (или, по крайней мере, не уменьшилась). Существует ряд алгоритмов решения этой задачи, которые позволяют за конечное число шагов найти некоторый (локальный) максимум функционала вычисления вероятности на обучающей выборке. Самым распространенным из этих методов является EM-алгоритм, или алгоритм Баума-Уэлча [31].

Локальное распознавание речи с помощью монотонных вероятностных автоматов производится методом сравнения с эталонами. Вычисляется вероятность того, что каждый из эталонных вероятностных автоматов выдает распознаваемый отрезок речевого сигнала, и выбирается автомат, максимизирующий эту вероятность. Результатом распознавания считается слово, соответствующее найденному автомату.

Вопросы выполнимости предположений, лежащих в основе применимости метода скрытых марковских моделей, являются открытыми. Тем не менее, практика показывает, что, несмотря на неадекватность модели, этот метод дает хорошие результаты.

6. Выбор фонетического алфавита, транскриптор

Фонетический алфавит является основой работы блока локального распознавания речи. Как уже говорилось, каждый трифон языка моделируется монотонным вероятностным автоматом из четырех состояний. Следовательно, общее число параметров автоматов, которые необходимо настроить в процессе обучения на основе речевого корпуса, линейно зависит от числа звуков, то есть от мощности фонетического алфавита, и уменьшение его размера приводит к ослаблению требований к объему речевой базы данных. С другой стороны, при сокращении алфавита в нем могут быть отождествлены те звуки, различение которых может быть существенным в процессе локального распознавания. Поэтому минимизация размера фонетического алфавита без ущерба для качества распознавания должна быть проведена путем отождествления в алфавите только тех звуков, которые

являются наиболее близкими по звучанию с точки зрения человека.

В работе [13] показано, что на множестве автономных вероятностных автоматов, с помощью которых эффективно моделируются звуки и их сочетания, можно ввести метрику, тесно связанную с вероятностью «спутать» слова при распознавании, то есть с близостью слов естественного языка «на слух». Эта метрика была эффективно использована авторами при решении задачи оптимального выбора фонетического алфавита при разработке системы распознавания русской речи в рамках гранта фирмы Intel Corp., США. С помощью метрики была построена матрица попарных расстояний между фонемами русского языка, представленных в виде автономных вероятностных автоматов, которые были синтезированы на основе русской речевой базы данных. Удалось показать, что алфавит из 150 фонемных символов для русского языка [9] можно сократить без потенциальной потери точности при распознавании до 120 символов.

Автоматический фонетический транскриптор русских текстов по правилам является другим важным элементом при разработке системы распознавания русской речи. При построении размеченной части речевой базы данных (см. ниже) как основы для обучения параметров вероятностных автоматов — эталонов фонем необходимо построить транскрипции текстов, соответствующих этим акустическим данным. Авторами обзора создан такой транскриптор, сравнимый по своим параметрам с существующими транскрибирующими системами для русского языка ([45, 46]).

7. Текстовые и речевые базы данных

В настоящий момент самыми сложными элементами при построении систем распознавания речи являются, как это не покажется странным, не распознающие алгоритмы — их подробные описания можно прочитать в монографиях и патентах, предшествующих появлению той или иной коммерческой системы распознавания, — а построение акустической модели языка и начальное обучение эталонов слов словаря, чаще всего являющихся вероятностными автоматами. Для настройки параметров языковых моделей и эталонов фонетических единиц языка в качестве основы для обучения необходимы тек-

стовые и речевые базы данных достаточно большого объема. Необходимо тщательно учесть все встречающиеся в современном языке слова и языковые обороты, типы голосов и акцентов, имеющихся у носителей языка.

Исследование, проведенное на кафедре Математической теории интеллектуальных систем, позволило произвести оценку параметров существующих русских текстовых и речевых корпусов.

Результаты исследования текстовых баз данных суммированы в сводной таблице.

Речевые базы данных представляют из себя множества записей произнесенных различными дикторами слов, фраз, предложений. Слова могут произноситься как раздельно, так и слитно; каждое предложение в речевом корпусе обычно сопровождается фонетической транскрипцией. Параметры записи могут быть также различными — от узкополосной телефонной записи (моно, частота дискретизации 8 кГц, 8 бит на отсчет) и широкополосной микрофонной (моно, 22 кГц, 16 бит на отсчет) до синхронных многоканальных записей (телефон + микрофон, несколько микрофонов и т.п.) Узкополосные базы данных используются для создания систем распознавания речи по телефону, а широкополосные — для обучения компьютерных систем диктовки текстов.

Объем корпуса характеризуется двумя важными параметрами — числом дикторов и общей длительностью звучания корпуса. Дикторы должны представлять все половозрастные группы, диалекты и т.п. Общая длительность корпуса должна обеспечивать достаточную представительность выборки, позволяющую произвести качественное обучение параметров вероятностных автоматов.

Важно, чтобы текстовая база данных, которая составляет основу речевого корпуса, содержала так называемые фонетически-сбалансированные предложения, то есть такие, в которых в среднем равномерно представлены все звуки и трифоны языка. Кроме того, обычно тексты включают как фрагменты устного диалога, так и письменную речь. Важным элементом любой распознающей системы является распознавание последовательностей чисел и цифр, поэтому часть базы данных, соответствующая набору чисел, должна также присутствовать и быть достаточно большой по объему.

№ п/п	Корпус	Тематика	Общий объем корпуса	Качество корпуса	Комментарии
1	Библиотека Максима Мошкова	Современная художественная литература: фантастика, переводная проза и др.	70 млн. слов	Корпус содержит очень много ошибок	Библиотека расположена в сети Интернет по адресу http://www.lib.ru. Является общедоступной
2	Российская периодика	Электронные версии газет и журналов России за последние 10 лет	более 400 млн. слов	Корпус состоит из текстов, прошедших корректуру	Часть электронных версий газет и журналов общедоступна, большая часть российской прессы доступна через платные Интернет-библиотеки.
3	Стенограммы заседаний депутатов Гос. Думы	Выступления депутатов, законы и т.п. за период с 1994 года	30 млн. слов	Корпус прошел корректуру	Доступен в стенах Гос.Думы
4	Новости	Новостные ленты Интернет-СМИ	100 млн. слов	Корпус проходит корректуру	Общедоступны только архивы за последние годы: www.lenta.ru, www.vesti.ru, www.strana.ru, www.russ.ru, www.utro.ru и т.п.

Поскольку обучение вероятностных автоматов происходит на уровне трифонов, на вход алгоритма обучения должны поступать речевые сигналы, размеченные на отдельные звуки. Фонетическая разметка речевых записей является чрезвычайно трудоемкой работой и требует труда высококвалифицированных специалистов. К счастью, нет необходимости в фонетической разметке всей базы данных, достаточно разметить только ее часть, а затем с помощью алгоритма

Витерби произвести автоматическую разметку всего остального корпуса.

Проведенное нами исследование показало, что в мире существует несколько русских телефонных речевых корпусов, созданных по заказу западных фирм. Общий объем этих баз данных составляет более 100 часов, а общее число дикторов — несколько тысяч человек. Эти базы данных не содержат фонетически размеченную часть.

Что касается широкополосных речевых баз данных, то здесь ситуация несколько хуже. Максимальная по объему база данных из имеющихся на рынке имеет общее время звучания 50 часов, в ней задействовано чуть более 200 дикторов. Тем не менее, в этом частотном диапазоне имеются речевые корпуса, фонетически размеченные составляющие которых имеют значительный объем по сравнению со объемом корпуса в целом (от 3 до 100%).

Известно [34], что хорошие результаты работы систем распознавания достигаются после обучения на речевом корпусе объемом не менее 80–100 часов. Учитывая, что число звуков в русском языке превышает число звуков в английском, можно сделать вывод о том, что в настоящее время имеется недостаток в русских речевых базах данных, записанных в широкополосном диапазоне.

8. Декодирование графа гипотез

Локальное распознавание речи позволяет для каждого отдельного отрезка речевого сигнала проверить гипотезу о том, что этот участок является произнесением какого-либо слова естественного языка. Можно рассматривать всевозможные разбиения сигнала, соответствующего произнесению фразы (предложения, группы последовательных предложений) на отрезки и решать задачу локального распознавания подобных отрезков. Поскольку каждый речевой сигнал можно рассматривать без ущерба для результатов распознавания как конечный вектор действительных чисел [13], такой перебор конечен. Отобрав среди возможных разбиений сигнала и соответствующих им результатов распознавания только те, которые имеют достаточно большую вероятность, мы получим множество возможных гипотез распознавания сигнала, которое традиционно представляется

в виде графа гипотез распознавания.

Граф гипотез обычно является ориентированным ациклическим графом без ориентированных циклов с двумя отмеченными вершинами (полюсами), каждое ребро которого соответствует некоторому слову языка и нагружено вероятностью этого слова. Один полюс этого графа («начальная» вершина) имеет только исходящие ребра, а другой («конечная» вершина), наоборот, только входящие.

Задача декодирования состоит о выборе ориентированного пути, идущего из «начальной» вершины в «конечную» и удовлетворяющего следующему свойству: последовательность слов, соответствующая этому пути, является предложением естественного языка, и вероятность этого пути максимальна. Эта задача требует введения формальной модели языка, позволяющей осуществлять проверку принадлежности языку любой последовательности слов.

Принадлежность языку здесь понимается либо традиционно, как принадлежность множеству, либо как функция вычисления вероятности принадлежности языку. В последнем случае вероятность пути из «начальной» вершины в «конечную» определяется как произведение вероятностей, приписанных всем ребрам, из которых состоит этот путь, и вероятности соответствующей пути последовательности слов в языковой модели.

9. Языковые модели и их применение в распознавании речи

Естественный язык — результат многовековой параллельной работы огромного числа носителей языка. Он принципиально отличается от случайных комбинаций слов и от формально построенных языков. Одной из основных особенностей естественного языка является избыточность, позволяющая понимать искаженную речь. Формализация этого процесса сталкивается с трудностями больших объемов текстов (только в них можно выявить языковые особенности, а не в простом списке фраз). В таких случаях наиболее естественным является исследование различных применений частотных характеристик в текстовых базах данных.

Модели естественного языка интересовали математиков начиная

с конца XIX — начала XX веков. Считается, что Андрей Андреевич Марков (старший) понятие марковской цепи ввел в процессе статистического исследования русского языка [15, 16]. Бурный всплеск теории построения языковых моделей возник после окончания второй мировой войны и был связан с тематикой машинного перевода. Задача машинного перевода была впервые сформулирована в 1946 году (Бут, Уивер). Уже в 1954 году в Джорджтаунском университете был проведен первый эксперимент по машинному переводу [38], а в 1955 году прошел первый эксперимент по машинному переводу в СССР. Существенный вклад в анализ естественных языков внесли работы Н.Хомского, который является основателем нового направления в структурной лингвистике — порождающей лингвистики [28, 29]. В 1966 году Комитет Национальной Академии Наук США делает вывод о нерентабельности применения систем машинного перевода, и работы во всём мире в этом направлении идут на убыль. Тем не менее, к тому времени уже было накоплено достаточно знаний в этой области.

Языковые модели, использовавшиеся в системах машинного перевода, основывались на лингвистических знаниях и не были последовательными в том смысле, что анализ принадлежности цепочки слов языку производился с помощью анализа всей цепочки слов в целом. Распознавание речи, наоборот, требует таких моделей, которые позволяют для каждого нового поступившего на вход слова определять принадлежность (вероятность принадлежности) получившейся последовательности слов языку.

В настоящее время можно выделить следующие подходы к построению формальных моделей естественного языка (здесь $A = \{a_1, \ldots, a_s\}$ — конечное множество слов естественного языка (словарь), а через L обозначена языковая модель):

I. Дискретные модели ($L \subset A^*$).

1) Модели с конечным числом состояний (L — автоматный (регулярный) язык);
2) Модели, основанные на теории формальных языков (различные подходы: применение классических КС-языков, грамматик Вудса, грамматик зависимостей; все перечисленные подходы эквивалентны обычным КС-языкам).

3) Модели, основанные на лингвистических знаниях (экспертные системы, модели с семантикой и т.д.) Обычно характеризуются очень сложным (экспоненциальным) алгоритмом разбора.

II. Статистические модели (L — распределение вероятностей на A^*).

1) n-граммная модель. Основана на формуле Байеса разложения вероятности: $P(a_{i_1}a_{i_2}\dots a_{i_{s-1}}a_{i_s}) = P(a_{i_1})P(a_{i_2}|a_{i_1})\dots P(a_{i_s}|a_{i_1}a_{i_2}\dots a_{i_{s-1}})$ и предположении, что условная вероятность появления очередного слова зависит только от последних $n-1$ слов и не зависит от предыдущих слов.
2) Модели, основанные на деревьях решений (здесь дерево решений — это бинарное дерево, каждой листовой вершине которого приписана вероятность, а остальным вершинам – предикаты $P : A^* \to E_2$). Модель функционирует следующим образом: для слова, поступившего на вход, вычисляется значение предиката, приписанного корневой вершине, и осуществляется переход по левой или правой стрелке в зависимости от значения предиката. Процедура повторяется до тех пор, пока не дойдём до листовой вершины. Приписанная этой вершине вероятность используется вместо вероятности входного слова. Работающих реализаций этой модели для всего языка нет ввиду ручного процесса построения дерева.
3) Статистические обобщения формальных языков (правилам приписываются вероятности, тем самым помимо выводимости в языке появляется ещё вероятность вывода). Эти модели имеют преимущества формальной модели и устойчивость n- граммной.

В настоящее время основным подходом к построению языковых моделей для систем распознавания речи является использование аппарата статистических методов. При этом модель в таком понимании — это просто распределение вероятности на множестве всех предложений языка. Естественно, что хранить модель в таком виде невозможно, поэтому используют более компактные способы задания.

Языковые модели, основанные на n-граммах, используют явное предположение о том, что вероятность появления очередного слова

в предложении зависит только от предыдущих $n-1$ слов. На практике используются модели со значениями $n = 1$, 2, 3 и 4. Наиболее удачной моделью из этого класса для английского языка оказывается триграммная модель. Все новые модели практически всегда оцениваются по отношению к триграммной модели. На сегодняшний день практически все коммерческие системы распознавания речи используют n-граммную модель в той или иной форме. При этом вероятность всего предложения вычисляется как произведение вероятности входящих в него n-грамм.

Основным достоинством данного класса моделей оказывается возможность построения модели по обучающему корпусу достаточно большого размера и высокая скорость работы. Основные недостатки — заведомо неверное предположение о независимости вероятности очередного слова от более длинной истории, что затрудняет работу и не позволяет моделировать более глубокие связи в языке; и колоссальные, но все-таки недостаточные для получения достоверных оценок объемы обучающих данных. В самом деле, если словарь содержит N слов, то число возможных пар слов будет N^2. Даже если только 0,1% от них реально встречаются в языке, то минимально необходимый объем корпуса для получения статистически достоверных оценок будет иметь порядок 125 млрд. слов или около 1 терабайта при специально подобранном корпусе. Для триграмм минимальные корпуса будут достигать размеров в сотни и тысячи терабайт.

Для преодоления этого недостатка используется развитый аппарат техник сглаживания, которые позволяют производить оценку параметров модели в условиях недостаточных или вовсе отсутствующих данных. Другим подходом к решению той же проблемы является кластеризация словаря, позволяющая сократить модель.

Для анализа качества статистических языковых моделей принято использовать так называемый коэффициент неопределенности (perplexity coefficient), введенный в [22], который может быть проинтерпретирован как (геометрическое) среднее ветвление в данной модели [34].

Для n-граммной модели коэффициент неопределенности задается формулой:

$$perplexity = \left(\sqrt[N]{\prod P(\omega_{i_k} | \omega_{i_{k-1}} \dots \omega_{i_{k-n+1}})} \right)^{-1},$$

где $\omega_{i_1}\omega_{i_2}\dots\omega_{i_N}$ — естественный язык, заданный некоторым корпусом текстов.

Нетрудно видеть, что коэффициент неопределенности является функцией от построенной языковой модели и естественно языка (и текстового корпуса). Таком образом, при фиксированном языке он позволяет сравнивать различные языковые модели, а при фиксированном типе модели — оценивать сложность самих естественных языков.

«Чистые» n-граммные модели не применимы для использования в распознавателях русской речи (см., например, [37]), что требует их модификации. Авторами предложен один из подходов к созданию формальных моделей русского языка на основе n-грамм, о чем пойдет речь ниже.

10. Основные отличия русского языка от английского

Основные отличия русского языка от английского приведены в следующей таблице:

	Английский	Русский
Мощность словаря начальных форм слов (типичная и максимальная)	70–90 тыс., до 120 тыс.	120–150 тыс., до 200 тыс.
Мощность словаря словоформ	120–150 тыс.	2.5–3.5 млн.
Число словоформ на 1 слово	1.7	До 100
Информация о связи слов выражается при помощи	порядка слов	Словоформ
Неопределённость (биграммы)	≈ 100	$>> 700$

Нами были проведены исследования, которые состояли в анализе применимости существующих моделей к русскому языку [21]. Были проверены следующие модели: стандартные n-граммы (для $n = 2$ и 3) и n-граммы со свободным порядком слов. Для экспериментов был

разработан пакет программ, который позволяет строить полный список n-грамм по данному текстовому корпусу, использовать сглаживание (реализована техника линейного сбрасывания и торможения [34]), вычислять вероятности предложений, коэффициент неопределенности тестового корпуса, и т.д. Была также создана система морфологического анализа русских текстов, работающая под управлением словаря на 150 тыс. основ.

При проверке простейшей языковой модели (n-грамм с $n = 2$) оказалось, что число пар слов, встретившихся в корпусе из 100 млн. слов по одному разу, составило более 92%, а коэффициент неопределенности превысил 500. Для $n = 3$ ситуация оказалась еще хуже. Тем самым было подтверждено априорное утверждение о неприменимости стандартного статистического подхода для русского языка. После этого были протестированы еще два подхода. Первый подход — с использованием n-грамм со свободным порядком слов — был призван удалить из грамматической информации порядок слов, а второй — преодолеть трудности, связанные с большим количеством словоформ.

n-граммы со свободным порядком слов вводятся следующим образом. Вероятности классических n-грамм $P(\omega_{i_k}|\omega_{i_{k-1}} \dots \omega_{i_{k-n+1}})$ заменяются вероятностями $P(\omega_{i_k}|\{\omega_{i_{k-1}} \dots \omega_{i_{k-n+1}}\})$, где фигурные скобки обозначают множество, то есть языковая модель не учитывает порядок первых $n-1$ слов в n-грамме. Было показано, эта модификация не привела к существенным улучшениям.

Второй подход был основан на разложении общей языковой модели на две составляющие: модель, основанную на морфологии и модель, основанную на начальных формах слов. Для каждой из этих моделей затем был применен подход на основе n-грамм. Модель, использующая только морфологию (названная нами категорной частью языковой модели), была построена для $n = 3$. Итоговый коэффициент неопределенности для этой модели оказался равен 21,93. В качестве второй составляющей грамматики была взята n- граммная языковая модель, построенная на начальных формах слов. В результате экспериментов удалось установить, что коэффициент неопределенности этой модели примерно в 2–2,5 раза выше, чем в случае английского языка (около 230 в нашей модели против примерно 100

в, например, [41]). Ранее получались лишь нижние оценки коэффициента неопределенности. Можно надеяться, что применение этого подхода позволит разработчикам использовать все преимущества n-граммных моделей применительно к русскому языку. Кроме того, выделение морфологической информации в независимую модель позволяет справиться с проблемой акустической похожести различных словоформ одного и того же слова. Предложенное решение проблемы может быть эффективно использовано для многих языков, где число словоформ достаточно велико, например, для языков славянской группы.

11. Заключение

В настоящее время имеются условия для создания полнотекстового распознавателя слитной русской речи. Отличительной особенностью распознавания речи на русском языке является неприменимость языковых n-граммных моделей, успешно используемых в системах распознавания на английском языке. С разработкой языковых моделей русского языка и их проверкой и применением на практике это белое пятно должно быть закрыто. Другим узким местом этого направления долгое время являлось отсутствие текстовых и речевых баз данных, суммарный объем которых был бы достаточен для обучения параметров речевой модели. В последние годы началась работа по созданию подобных баз данных, однако пока их общий объем недостаточен. Кроме того, создание систем распознавания речи невозможно без соответствующего финансирования этих разработок.

Кафедра математической теории интеллектуальных систем имеет коллектив исследователей и разработчиков, занимающихся созданием прикладных систем в области распознавания и синтеза речевых образов в течение последних 10 лет, в том числе в рамках сотрудничества с отечественными и зарубежными партнерами, и открыта для диалога с заказчиками и разработчиками.

Список литературы

[1] Беллман Р. Динамическое программирование. М.: ИЛ, 1960.

[2] Бендат Дж., Пирсол А. Измерение и анализ случайных процессов. М.: Мир, 1974. С. 372, 342, 368.

[3] Бухараев Р.Г. Некоторые эквивалентности в теории вероятностных автоматов // Уч. записки Казан. университета. 1964. 124. №2. С. 45–65.

[4] Бухараев Р.Г. Основы теории вероятностных автоматов. М.: Наука, 1985.

[5] Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наукова думка, 1987.

[6] Винцюк Т.К. Распознавание слов устной речи методами динамического программирования // Кибернетика. 1968. №1. С. 81–88.

[7] Джелинек [Елинек] Ф. Распознавание непрерывной речи с помощью статистических методов // ТИИЭР. 1976. Т. 64. №4. С. 131–160.

[8] Захаров Л.М. Транскрипция текстов при синтезе и анализе русской речи // АРСО-96.

[9] Зиновьева Н.В., Захаров Л.М., Кривнова О.Ф., Фролов А.Ю., Фролова И.Г. Автоматический транскриптор // АРСО-92.

[10] Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. М.: Наука, 1989.

[11] Кудрявцев В.Б., Алешин С.В., Подколзин А.С. Введение в теорию автоматов. М.: Наука, 1985.

[12] Левинсон С.Е. Структурные методы автоматического распознавания речи // ТИИЭР. Т. 73. №11. Ноябрь 1985. С. 100–128.

[13] Мазуренко И.Л. Автоматные методы распознавания речи. Автореферат диссертации на соискание ученой степени кандидата физико-математических наук (на правах рукописи). М., 2001.

[14] Мазуренко И.Л. Компьютерные системы распознавания речи // Интеллектуальные системы. Т. 3. Вып. 1–2. М., 1998.

[15] Марков А.А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Академии наук. СПб. VI. Т. 7. 1913. №3. С. 153–162.

[16] Марков А.А. Об одном применении статистического метода. Доклад в Академии Наук от 17 февраля 1916 года.

[17] Рабинер Л.Р. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: обзор // ТИИЭР. Т. 77. №2. Февраль 1989.

[18] Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов / Пер. с английского. М.: Радио и связь, 1981.

[19] Фант Г. Акустическая теория речеобразования / Пер. под ред. В.С. Григорьева. М.: Наука, 1964.

[20] Форни-мл. Дж.Д. Алгоритм Витерби // ТИИЭР. 1973. Т. 61. №3. С. 12–25.

[21] Холоденко А.Б. О построении статистических языковых моделей для систем распознавания русской речи // Интеллектуальные системы. Т. 6. Вып. 1–4. 2001. С. 381–394.

[22] Bahl L.R., Baker J.K., Jelinek F., Mercer R.L. Perplexity — A measure of the difficulty of speech recognition tasks // J. Acoust. Soc. Amer. Vol. 62. P. S63. 1977. Suppl. no. 1.

[23] Baker J.K. Stochastic modeling for automatic speech understanding // Speech Recognition / ed.: D.R. Reddy. New York: Academic Press, 1975. P. 521–542.

[24] Bakis R. Continuous speech word recognition via senti-second acoustic states // Proc. ASA Meeting (Washington, DC). Apr. 1976.

[25] Baum L.E., Egon J.A. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology // Bull. Amer. Meteorol. Soc. Vol. 73. 1967. P. 360–363.

[26] Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains // Ann. Math. Stat. Vol. 37. 1966. P. 1554–1563.

[27] Carlyle J.W. Reduced forms for stochastic sequencial machines // J. Math. Analysis and Applic. 1963. №7. P. 167–175.

[28] Chomsky N. Transformational analysis. Dissertation. 1955.

[29] Chomsky N. Syntactic Structures. Den Haag: Mouton, 1957. Русский перевод: Хомский Н. Синтаксические структуры // Новое в лингвистике. М., 1962. Вып. 2.

[30] Davis K.H., Biddulph R., Balashek S. Automatic recognition of spoken digits // J. Acoust. Soc. Amer. Vol. 24. 1952. P. 637–642.

[31] Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm // J. Roy. Stat. Soc. Vol. 39. No. 1. 1977. P. 1–38.

[32] Denes P.B., Mathews M.V. Spoken digit recognition using time frequency pattern matching // J. Acoust. Soc. Amer. Vol. 32. 1960. P. 1450–1455.

[33] Dudley H., Balashek S. Automatic recognition of phonetic patterns in speech // J. Acoust. Soc. Amer. Vol. 30. 1958. P. 721–743.

[34] Handbook of Standards and Resources for Spoken Language Systems / Ed. by Gibbon D., Moore R., Winski R. Berlin: Mouton de Gruyter, 1998.

[35] Itakura F. Minimum prediction residual principle applied to speech recognition // IEEE Trans. Acoust., Speech, Signal Processing. Vol. ASSP–23. 1975. P. 67–72.

[36] Juang B.H. On the hidden Markov model and dynamic time warping for speech recognition — A unified view // AT&T Tech. J. Vol. 63. No. 7. Sept. 1984. P. 1213–1243.

[37] Kanevsky D., Monkowsky M., Sedivy J. Large Vocabulary Speaker-Independent Continuous Speech Recognition in Russian Language // Proc. SPECOM'96. St.-Petersburg, October 28–31, 1996.

[38] Kenny H.C. Robot translates nimbly // Christian Science Monitor. 11 January 1954.

[39] Klatt D.H. Review of the ARPA Speech Understanding Project // J. Acoust. Soc. Amer. Vol. 62. No. 6. Dec. 1977. P. 1345–1366.

[40] Levinson S.E., Rabiner L.R., Sondhi M.M. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition // Bell Syst. Tech. J. Vol. 62. No. 4. Apr. 1983. P. 1035–1074.

[41] Manhung S. and others. Integrating a context-dependent phrase grammar in the variable n-gram framework // Proceeding of ICASSP. 2000.

[42] Rabiner L.R., Juang B.H. An introduction to the hidden Markov models // IEEE ASSP Mag. Vol. 3. No. 1. 1986. P. 4–16.

[43] Starke P.H. Theorie Stochastischen Automaten. I, II // Elektron Informationsverarb. und Kybern. 1965. 1. №2.

[44] Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm // IEEE Trans. Informat. Theory. Vol. IT–13. Apr. 1967. P. 260–269.

[45] Zue V.W., Cole R.A. Experiments on spectrogram reading // Proc. ICASSP–79. 1979. P. 116–119.

[46] http://developer.intel.com