ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет компьютерных наук
Образовательная программа «Программная инженерия»

СОГЛАСОВАНО

Научный руководитель, доцент
Факультета Компьютерных Наук

_____ А. А. Харламов

«___» _____ 2025 г.

УТВЕРЖДЕНО

Академический руководитель
образовательной программы
«Программная инженерия», старший
преподаватель департамента
программной инженерии

_____ Н. А. Павлочев

«___» _____ 2025 г.

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ В РАСПОЗНАВАНИИ РЕЧИ.

Исполнители:
Студентка группы БПИ-245
_____ / М. В. Горбачева /
«___» _____ 2025 г.

2025

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ В РАСПОЗНАВАНИИ РЕЧИ.

Реферат

Листов 24

2025

АННОТАЦИЯ

Реферат - это краткое изложение содержания документа или его части, включающее основные фактические сведения и выводы, необходимые для первоначального ознакомления с документом и определения целесообразности обращения к нему.

Сущность реферата – в кратком изложении (с достаточной полнотой) основного содержания источника. Составление рефератов – это процесс аналитико-синтетической переработки первичных документов. Реферируется преимущественно научная и техническая литература, в которой содержится новая информация.

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ

Автоматическое распознавание речи является динамично развивающимся направлением в области искусственного интеллекта. Для обучения нейронных сетей восприятию звучания речи человека и ее анализу, необходимо обратиться к таким наукам, как физика, математический анализ, теория вероятностей, языкознание, лексикология, фонетика.

Первые две главы данного реферата будут посвящены вопросам речеобразования и восприятия. Очевидно, что понимание структуры речевого сигнала и лежащих в его основе движений речеобразующих органов может помочь в решении задачи автоматического распознавания речи. В ещё большей степени это относится к пониманию вопросов, связанных с восприятием звуков вообще и речевых звуков в частности.

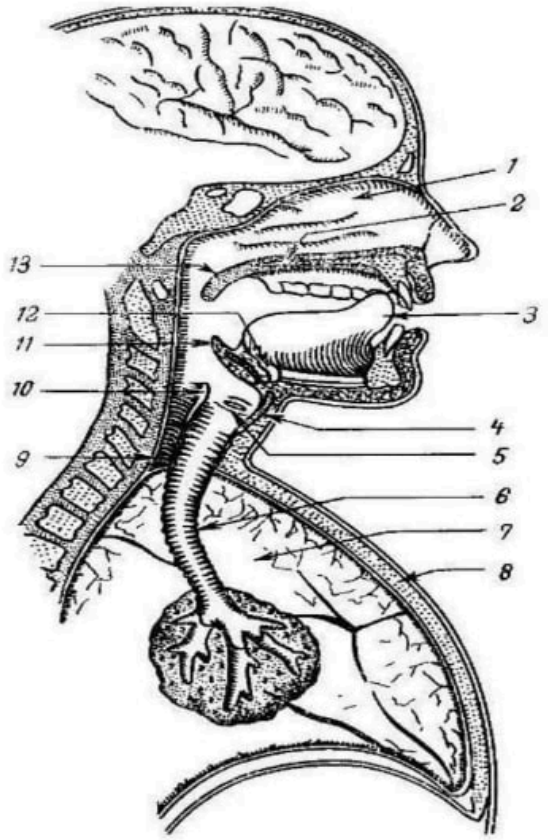
Очень важным вопросом, на который предстоит ответить в ходе изучения речеобразования и восприятия является вопрос о признаках, или параметрах речевого сигнала, которые содержат информацию, распознавания речи. Очевидно, по самому своему смыслу, эти параметры должны являться следствием сознательно контролируемых движений речевых органов. Очевидно также, что выделение этих параметров должно являться главной задачей слуховой системы при распознавании речи.

2. РЕЧЕОБРАЗОВАНИЕ

2.1. Физиология речеобразования

Благодаря создаваемому в лёгких давлению, поток воздуха устремляется в голосовой тракт, проходит через голосовые складки, может устремляться в носовую полость (если нёбная занавеска открыта) и выходит в открытое пространство, минуя возможные зубные и губные сужения.

Процесс речеобразования иллюстрируется на рисунке 1.1



1 — носовая полость, *2* — твердое небо, *3* — язык, *4* — щитовидный хрящ, *5* — голосовые связки, *6* — трахея, *7* — легкое, *8* — грудина, *9* — пищевод, *10* — кольцеобразный хрящ, *11* — надгортанье, *12* — подъязычная кость, *13* — мягкое небо (небная занавеска)

Рис.1.1. Речевой аппарат человека [1].

Речь представляет собой звуковые колебания воздуха в диапазоне частот от 70−100 Гц до нескольких килогерц. Для того чтобы в выходящем воздушном потоке возникли колебания с такими частотами, необходимо наличие источника звука на пути воздушного потока. Источником звука могут являться:

1. Голосовые связки;
2. Турбулентный шум в сужении;
3. Шум внезапно высвободившегося воздуха при смычке (импульсный).

Места сужения или смычки могут быть разными для разных языков (так, в ряде языков существуют необычные для русского языка звуки, источником которых является гортанная смычка, то есть взрыв, образующийся при размыкании голосовых складок). При образовании звука /х/ и шепотной речи шумовым источником являются сведенные, но не колеблющиеся голосовые складки.

В соответствии с типом источника речевые звуки подразделяются на классы:

1. Гласные – источником звука являются только голосовые складки,

проход в носовую полость перекрыт небной занавеской;

2. Щелевые (фрикативные) согласные – источником звука является

турбулентный шум в сужении (глухие согласные /ф/, /с/, /ш/,..), или дополнительно голосовые складки (звонкие /в/, /з/, /ж/,..).

3. Взрывные согласные – источником звука является шум взрыва (глухие /п/, /т/, /к/), или дополнительно импульсы голосовых складок (звонкие /б/, /д/, /г/).

Кроме указанных существуют звуки, которые требуют классификации:

1. Носовые согласные. Характеризуются тем, что излучение полностью

или частично осуществляется через нос. Забегая вперед, отметим, что передаточная функция голосового тракта содержит только полюса, то есть обладает только резонансами; при наличии боковой полости или параллельной ветви передаточная функция содержит также нули.

2. Русское /р/ возбуждается голосовыми складками, однако звук

модулируется дрожанием кончика языка.

3. Звуки, получающиеся сочетанием рассмотренных выше (примеры на

основе общеамериканского диалекта): Полугласные /j/ you, /w/ we; Плавные /r/ read, /l/ let.

4. Звуки, характеризующиеся динамическим характером произнесения:

дифтонги /eI/ say, /Iu/ new, /ɔI/ boy, /aU/ out, /aI/ I, /oU/ go; аффрикаты /tʃ/ chew, /dʒ/ jar.

2.1.1. Процесс образования звуков с голосовым возбуждением

Голосовые складки колеблются при продувании через них потока воздуха под действием эффекта Бернулли. Частота колебаний голосовых складок называется основным тоном (pitch). Частота колебаний голосовых складок при обычной речи находится в пределах 60−180 Гц для мужчин,

160−350 Гц для женщин и 200−650 Гц для детей (указанные границы чисто ориентировочные). При пении частота колебаний голосовых складок может достигать 2 кГц.

2.2. Передаточная функция голосового тракта Передаточная функция голосового тракта рассчитывается, исходя из того, что для значимых для восприятия частот (4000 Гц) акустическая волна с достаточной точностью является плоской. Распространение звука в этом случае описывается одномерным (зависящим от координаты вдоль оси тракта) уравнением Вебстера:

$$\frac{1}{S(x)}\frac{\partial}{\partial x}\left[S(x)\frac{\partial p}{\partial x}\right]=\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2},$$

где S(x) – площадь поперечного сечения как функция расстояния от голосового источника по оси тракта, p – звуковое давление, c – скорость звука, t – время. Даже в одномерном случае для голосового тракта уравнение Вебстера можно решить только численно. При этом не учитывается импеданс стенок тракта и потери энергии на границах и на трение. Таким образом, произнесение конкретного звука человеком преобразуется в строгую математическую формулу, понимаемую машиной.

2.3. Понятие фонемы

Первоначально введенный Фердинандом де Соссюром в 1879 году термин «фонема» (phoneme) практически не отличался от языковедческого термина «звук», как единицы речи, подвергающейся научному анализу. Современное понимание этого термина фонетистами ближе к определению, данному Бодуэном де Куртенэ: «Фонема есть цельное, неделимое во времени представление звука языка». Отметим два момента:

1. Фонема не есть физическая реализация звука, а

является представлением звука в сознании (абстракцией).

2. Фонема воплощает идею атомарности, примененную к субъективному

представлению о речи.

Явная неконструктивность данного определения с точки зрения технической реализации всегда вызывала споры и многочисленные варианты фонетической классификации. Технические специалисты, используя термин «фонема», вкладывали в него свои представления о речеобразовании и восприятии, доводя понятие до вульгарного: «фонема – это то, что я могу определить и выделить на своем приборе». Надо отметить, что и среди самих фонетистов нет единства в представлении о фонеме: московская и ленинградская (петербургская) школы фонетики предлагали алфавиты фонем,

существенно отличающиеся по размеру. Наверное, для того, чтобы избежать этих конфликтов, технические специалисты ввели термин «фон» (phone) – конкретная реализация фонемы. Фоны, принадлежащие к одной фонеме, называются аллофонами.

В процессе речеобразования речевые органы человека – губы, язык, нижняя челюсть, небная занавеска совершают движения, скорость которых зависит от силы мышц и массы самого органа. Скорость эта близка к предельной (иначе скороговорки не были бы популярны). Исследования показывают, что в естественной речи органы практически никогда не занимают положений, характерных для изолированно произнесенных звуков, а лишь обозначают движение в нужном направлении, соответственно, и форманты обозначают движение в нужном направлении. Очевидно, что движение в сторону, характерную для данной фонемы, зависит от предшествовавших и, как показывают эксперименты, даже последующих фонем, то есть, речевой аппарат может готовиться к произнесению некоторых звуков заранее. Этот эффект называется коартикуляцией. Взаимовлияние фонем не ограничивается соседями, а может распространяться на несколько соседних фонем. Можно вычислить, насколько огромное число возможных сочетаний, скажем, хотя бы по три фонемы, существует для языка. Если к этому добавить нерешенную проблему признаков, инвариантных к диктору, то число возможных представлений фонемы становится настолько огромным, что о выуживании признаков вручную не может быть и речи. Таким образом, по-прежнему используя аналогию с атомами, а лучше с квантами, можно заметить, что фонема скорее имеет «волновую» природу, то есть ее признаки «размазаны» по протяженному во времени отрезку, причем признаки различных фонем накладываются друг на друга.

3. КОЛИЧЕСТВЕННАЯ ОЦЕНКА СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Существуют различные по сложности и прикладному значению задачи распознавания: изолированных слов (команд); ключевых слов в потоке речи; связанной речи (тщательное проговаривание текста с паузами между словами); слитной речи (разделяют диктовку в узкой тематической области, и спонтанную речь, например, в диалоге между людьми). Оценка системы, распознающей отдельные команды, не представляет каких-либо трудностей – количество неправильно распознанных команд делится на общее количество испытаний и получается процент ошибки. Для систем, распознающих слитную речь, ситуация не столь проста. Задача оценки систем распознавания речи нетривиальна, так как различные алгоритмы сравниваются на ограниченных базах данных, и каждый из них имеет настраиваемые параметры, да и результаты распознавания можно интерпретировать по-разному. При этом объективное оценивание и сравнение систем распознавания речи важны как для разработчиков, так и для конечных пользователей систем. Существует количественная методика оценки, которая применяется для сравнения и сопоставления различных систем распознавания, в ней различают такие понятия как: критерий, показатель и метод. Критерий – предмет оценки или то, что нам нужно оценить (например, точность распознавания речи, скорость, робастность к шумам и т.д.). Показатель (мера, метрика) определяет конкретное свойство, которое мы оцениваем для выбранного критерия оценки (например, процент правильно распознанных слов, время обработки сигнала, уровень максимально допустимого шума при сохранении работоспособности и т.п.). Метод – способ определения соответствующего значения для данного показателя (сравнение распознанных слов с последовательностью сказанных слов, оценка времени обработки в секундах и т.д.).

Обычно при разработке систем автоматического распознавания речи используются три разных набора данных: обучающий (“train”), отладочный (“dev”) и оценочный/тестовый (“eval”). Обучающий набор данных (обычно это наибольшая часть речевых данных) используется только для создания и обучения моделей системы. Отладочный набор данных используется для настройки и адаптации параметров автоматической системы перед финальной стадией оценки, этот набор данных должен иметь тот же формат, что и тестовые данные. Оценочные данные содержат речевые данные, которые не использовались для обучения и настройки системы, и доступны только при финальной оценке системы. Выделают два основных критерия при оценке работы систем распознавания речи, которые далее рассмотрены детально: качество распознавания и скорость обработки.

3.1. Показатели оценки качества распознавания речи

Для систем автоматического распознавания речи основным показателем оценки по критерию качества является точность распознавания, которая определяется как процент правильно распознан-

ных слов (WRR — Word Recognition Rate) или, наоборот, неправильно распознанных слов (WER — Word Error Rate). Иногда также используется показатель ошибок распознавания фраз/предложений (SER — Sentence Error Rate), который является важным в диалоговых системах, где корректировка гипотезы распознавания невозможна в отличие от задачи диктовки текста. В последнее время в качестве основного показателя точности работы систем распознавания речи используется WER (его абсолютное или относительное значение), если сравниваются различные системы распознавания речи. Поскольку с развитием речевых технологий показатель WER все более приближается к нулю, то значение улучшения WER более наглядно, чем улучшение точности распознавания слов. Метод определения WER состоит в выравнивании двух текстовых строк (первая — это результат распознавания, а вторая — запись того, что было сказано в действительности) путем алгоритма динамического программирования с вычислением расстояния Левенштейна. Расстояние Левенштейна представляет собой “стоимость” редактирования текстовых данных (минимальное количество или взвешенная сумма операций редактирования) для преобразования первой строки во вторую с наименьшим числом операций ручной замены (S), удаления (D) и вставки (I) слов:

$$WER = \frac{S + D + I}{T} \times 100\%,$$

где T — количество слов в распознаваемой фразе. Также для оценки качества распознавания речи используется показатель процента корректно распознанных слов (WCR — Word Correctly Recognized), он не учитывает ошибочные вставки слов, сделанные системой:

$$WCR = \frac{H}{T} \times 100\%, \;\; H = N - D - S \,,$$

где H — количество правильно распознанных слов, а N — количество произнесенных диктором слов. Очевидно, что WER – это интуитивно понятный и адекватный показатель качества распознавания для аналитических естественных языков (класс языков, обладающих достаточно простой морфологией и системой словообразования), в которых грамматические значения однозначно выражаются самим словом (например, английский или французский). Однако другой класс синтетических языков (например, агглютинативные языки: финский, турецкий, венгерский или флективные языки: русский, украинский, казахский и т.д.), напротив, отличается богатой морфологией и развитой

системой словообразования. Такие языки могут синтезировать достаточно длинные словоформы из нескольких составных частей (морфем или слогов), которые определяют грамматические признаки.

При этом в беглой речи конец слова произносится не так четко как начальная часть, что приводит к акустической неопределенности и в среднем к более высоким значениям WER по сравнению с аналитическими языками. Кроме того, многие азиатские языки (например, китайский, корейский и т.п.) используют слоги взамен слов, а тайский и некоторые другие языки не имеют явных разделителей границ слов.

Оценка автоматического распознавания речи по показателю WER предполагает, что все слова во входной фразе одинаково информативны и важны, однако, ясно, что в задачах, отличных от диктовки текста, например, диалоговые системы или понимание (смысла) речи, некоторые значащие слова (ключевые слова) более важны, чем остальные (функциональные слова, предлоги, заполнители и т.п.). Предложено оценивать точность распознавания, используя взвешенный показатель неправильно распознанных слов (WWER — Weighted Word Error Rate), который определяется как

$$WWER = \frac{V_S + V_D + V_I}{V_T} \times 100\%,$$

$$V_T = \sum_{W_i} v_{W_i}, V_I = \sum_{\hat{W}_i \in I} v_{\hat{W}_i}, V_D = \sum_{W_i \in D} v_{W_i}, V_S = \sum_{s_j \in S} v_{s_j},$$

$$v_{s_j} = \max\left(\sum_{\hat{W}_i \in s_j} v_{\hat{W}_i}, \sum_{W_i \in s_j} v_{W_i}\right),$$

, где vW — вес слова Wi , которое является i-м словом во входной фразе, и vi Wˆ —i вес слова Wi , которое является i-м словом в гипотезе распознавания, s j — j-й замененный фрагмент фразы (или одно слово) и vs — вес данного сегмента s j . j Таким образом, в показателе WWER, каждое слово может иметь различный вес.

4. МЕТОД ДИНАМИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ

Для распознавания команд до сих пор иногда используют метод динамического программирования (ДП), впервые использованный в 60-х годах прошлого века. Это объясняется простотой, быстродействием и отсутствием необходимости собирать речевую базу данных. Неизвестная команда в виде последовательности векторов признаков сравнивается с набором эталонов, представленных в таком же виде. Основная проблема – различный темп и нелинейность темпа произнесения. При принятии решения руководствуются критерием минимума расстояния от неизвестного произнесения до эталона. Метод подразумевает, что эталоны, принадлежащие одной команде (одному классу), группируются в кластер, то есть в компактную группу точек в некотором пространстве, в котором существует мера близости.

Идея метода проста и допускает рассмотрение на качественном уровне. Задача состоит в том, чтобы сравнить две совокупности векторов различной длины, причем на пространстве векторов есть метрика или мера близости. Представим, что мы сравниваем эталон сам с собой: отложим векторы признаков эталона по оси X и Y. На плоскости XY на пересечении координат, соответствующих векторам i и j, построим вертикальный отрезок, равный расстоянию (степени близости) между этими векторами. Тогда на квадрате со стороной, равной количеству векторов в эталоне (N), возникнет «гористый ландшафт», симметричный относительно диагонали (0,0) (N,N), однако по диагонали будет пролегать абсолютно прямая «долина» с высотой, равной 0 (поскольку расстояние от вектора до самого себя равно 0). Если мы сравниваем два различных эталона, принадлежащих одному и тому же слову, то «картина местности» исказится, однако, если используемые признаки адекватно отражают процесс восприятия, можно надеяться, что некоторая долина по-прежнему будет пролегать по ломаной, близкой к диагонали, теперь уже прямоугольника.

Метод динамического программирования позволяет сосчитать минимальную сумму высот, набираемую при движении из точки (0,0) в точку (N,M) и, если это требуется, восстановить путь, по которому эта сумма набрана. Полученную сумму обычно нормируют на количество пройденных узлов, либо на сумму длин слов или длину более короткого слова и рассматривают как расстояние между двумя произнесениями. Конечно, используемые в практических системах реализации имеют множество управляемых параметров, оптимизирующих качество распознавания и уменьшающих время счета. Рассмотренный метод позволяет в дикторозависимом варианте распознавать 100–300 слов с вероятностью 90–98%. Для придания системе дикторонезависимых качеств, для каждого слова записывают несколько эталонов от разных дикторов (в процессе обучения добавляют эталон

от нового диктора, если он не распознался). Кроме того, существуют схемы нормализации эталонов относительно дикторов, а также кластеризации дикторов.

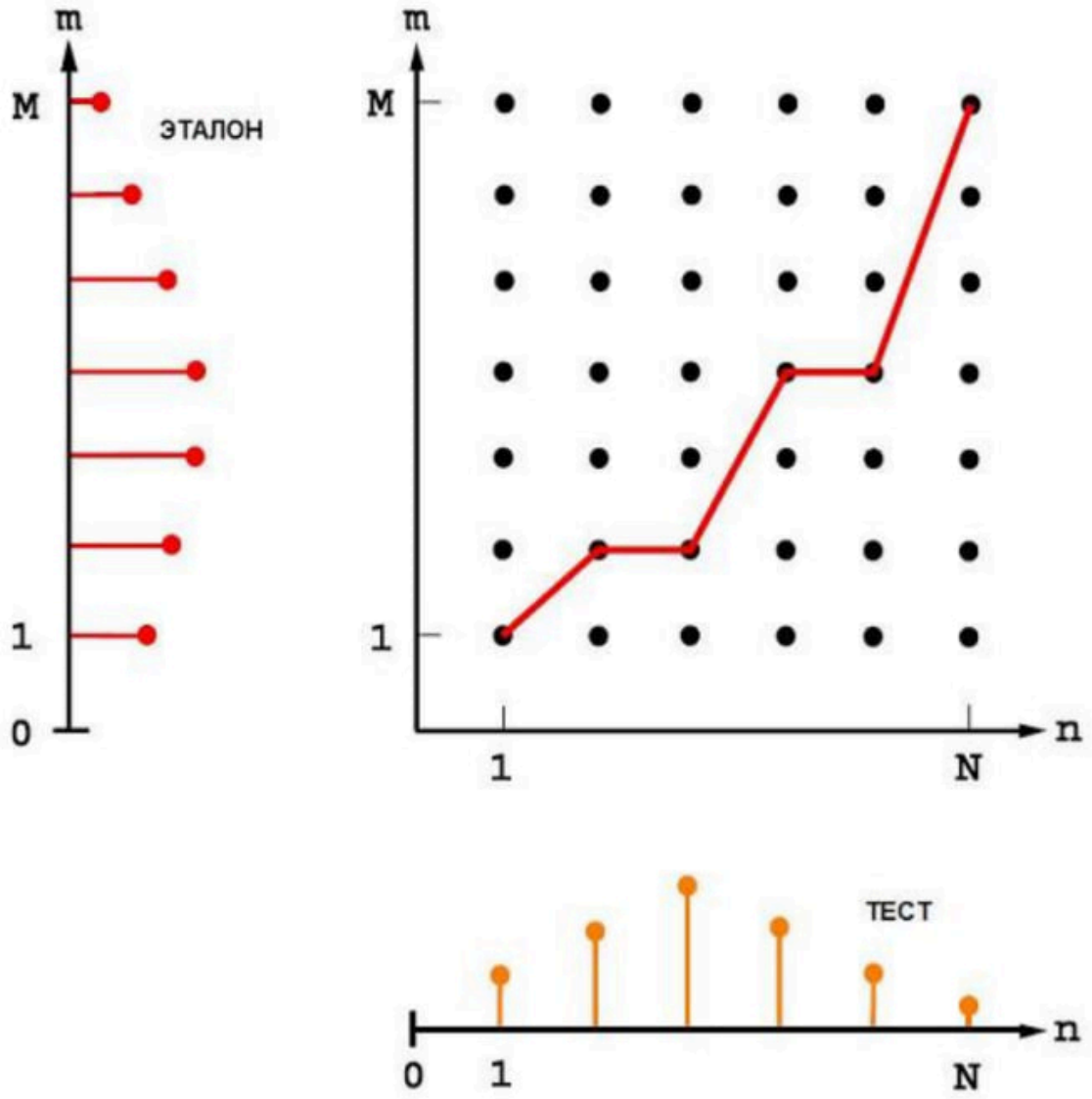


Схема метода динамического программирования

Алгоритм распознавания команд методом ДП прозрачен и не требует подробного рассмотрения. В зависимости от топологии модели (разрешённых переходов) суммарное расстояние очередного узла матрицы [M,N] подсчитывается, исходя из минимума набранного расстояния: Si, j = Dist(i, j) + min(wk ,l Sk ,l) , k ,l∈R где Sn,m – суммарное расстояние в узле (n,m), Dist(i,j) – расстояние между вектором i эталона и вектором j тестового слова, wk,l – вес, который присвоен узлу (k,l) относительно узла (i,j) (например, недиагональные переходы [(i-1,j)- >(i,j) и (i,j-1)->(i,j)] могут «штрафоваться»

бОльшим весом, чем диагональный переход [(i-1,j-1)->(i,j)]), R – множество разрешённых для перехода узлов (обычно это три ближайших узла [(i-1,j),(i-1,j-1) и (i,j-1)]). Кроме суммарного расстояния, каждый узел матрицы [M,N] может содержать информацию об узле, откуда совершён переход – эта информация нужна, если требуется восстановить путь.

5. РАСПОЗНАВАНИЕ РЕЧИ С ПОМОЩЬЮ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

В настоящее время большинство систем распознавания речи опираются на Скрытую Марковскую Модель (СММ). СММ – мощный статистический аппарат, представляющий спектральные свойства речи с помощью параметрического случайного процесса. Каждому моделируемому речевому объекту – фразе, слову, слогу, фонеме или аллофону (фонеме в конкретном окружении) – сопоставляется своя СММ. СММ фразы представляет собой конкатенацию СММ слов, которые представляются конкатенацией СММ более мелких элементов. Рассмотрим математический формализм, определяющий СММ. Процесс называется марковским, если для каждого момента времени вероятность любого состояния системы в следующий момент зависит только от состояния системы в настоящий момент и не зависит от того, каким образом система пришла в это состояние. Марковский процесс называется наблюдаемым, если каждое состояние на выходе взаимно-однозначно соответствует некоторому наблюдаемому явлению.

Пример:

Состояние 1: непогода (дождь, снег, град,…)

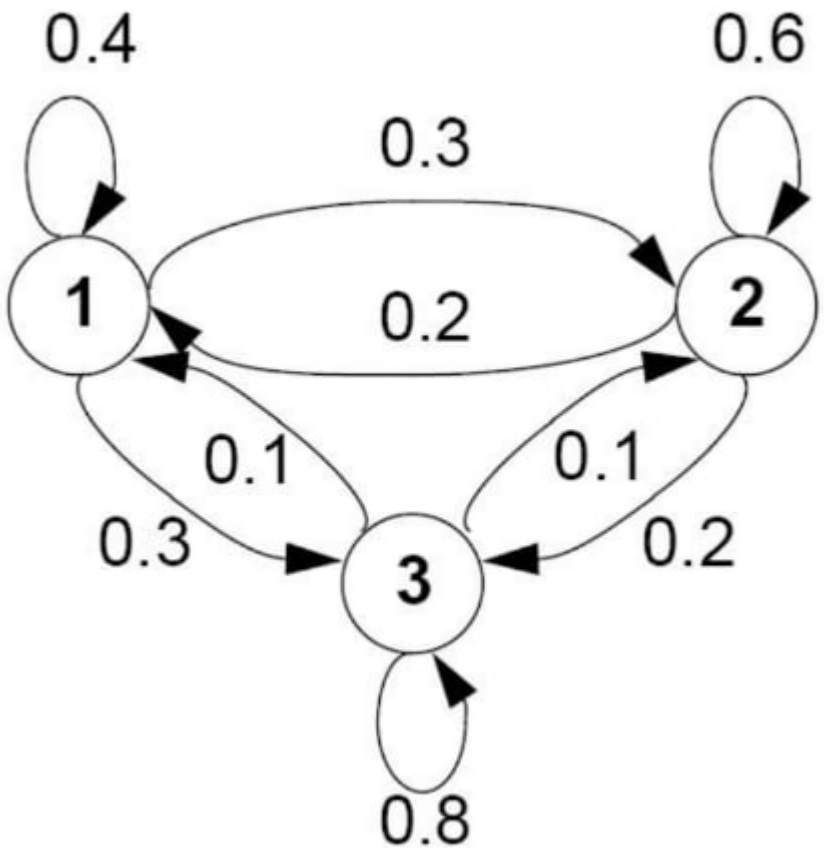
Состояние 2: облачно

Состояние 3: солнечно

Вероятности переходов между состояниями отображены при помощи данной матрицы:

$$A = \{a_{i,j}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}.$$

Или можно посторить схему марковской цепи для наглядного прелставления описания погодной модели.



.

Зная исходные значения вероятностей πj, можно рассчитать вероятность любой последовательности погодных условий в последующие дни, как произведение соответствующих вероятностей. Если состояния связаны с наблюдаемыми явлениями вероятностным образом, то марковская цепь называется скрытой (СММ). В случае распознавания речи наблюдаемыми являются векторы признаков, которые связаны с состояниями вероятностным образом, то есть один и тот же вектор признаков может принадлежать нескольким состояниям. Таким образом, к параметрам модели добавляются распределения вероятностей состояний в пространстве признаков, которые называют вероятностями эмиссии: Bi(x) – функция плотности вероятности состояния si в пространстве признаков или вероятность эмиссии. Если пространство признаков проквантовано, то Bi(x) представляется матрицей Bi(m), где m – номер слова в кодовой книге. Такие модели называются дискретными.

Для непрерывной модели используют аппроксимацию функции плотности вероятности набором стандартных функций – как правило, взвешенной суммой гауссовых функций. Для уменьшения количества оптимизируемых параметров используют гауссовы функции с диагональными матрицами ковариации. Существует разновидность СММ, называемая полунепрерывной – в этой СММ для

аппроксимации функций плотности вероятности всех состояний используются функции из одного пула.

Задачей распознавания является сопоставление набору акустических признаков речевого сигнала или наблюдений X(x1,…,xn) последовательности слов W(w1,…,wk), имеющих наибольшую вероятность правдоподобия среди всех кандидатов:

$$W = \arg\max_{W} P(W|X).$$

Используя теорему Байеса, перепишем это выражение:

$$W = \arg\max_{W} \frac{P(W)P(X|W)}{P(X)}.$$

Поскольку в процессе распознавания вероятность уже полученных акустических признаков P(X) не подлежит оптимизации:

$$W = \arg\max_{W} P(W)P(X|W).$$

Вероятности имеют простую интерпретацию: P(X|W) есть акустическая модель (вероятность порождения данной последовательности наблюдений X данной последовательностью слов W), а P(W) – вероятность существования в рассматриваемом языке данной последовательности слов (модель языка).

Таким образом, марковской моделью l(A,B,π) акустического события (акустической моделью), например, аллофона, называется набор из одного или нескольких состояний si, характеризующийся следующими параметрами: N – количество состояний s: πi – начальное распределение вероятностей

$$\sum_{i=1}^{N} \pi_i = 1.$$

A={ai,j} – вероятность перехода из состояния si в состояние sj

$$\sum_{j=1}^{N} a_{i,j} = 1, \qquad 1 \le i \le N.$$

Bi(x) или Bi(m) – вероятность эмиссии:

$$\int B_i(x)dv = 1, \qquad 1 \le i \le N,$$

где интегрирование проводится по всему объёму пространства признаков,

$$\sum_{m=1}^{M} B_i(m) = 1,$$

где 1≤ i ≤ N , M – размер кодовой книги, то есть количество кластеров.

6. ПРОБЛЕМА ВЫБОРА ЕДИНИЦ ФОНЕТИЧЕСКОГО УРОВНЯ

Понятно, что для реализации распознавания речи состояния должны быть связаны с единицами фонетического уровня. Поскольку речь является процессом, возможно объединение (конкатенация) моделей фонетических фрагментов в непрерывное произнесение. Таким образом, вместо создания моделей для каждого слова, что является непосильной задачей для больших словарей, создаются модели элементов нижнего уровня. В качестве таких элементов исследовались слоги, фонемы и фрагменты фонем. В настоящее время общепринятым является использование контекстно-независимых фонем (монофонов) для средних словарей и контекстно-зависимых фонем (дифонов, трифонов) для больших словарей. Необходимость использовать части фонем и контекстную зависимость объясняется коартикуляцией (взаимным влиянием произносимых звуков друг на разделе. Взаимовлияние фонем не ограничивается соседями, а может распространяется на несколько соседних фонем. Обычно используют информацию об одном (дифоны/бифоны) или левом и правом (трифоны) соседях (по аналогии, фонемы без учёта влияния контекста называют монофонами). При этом количество фонетических единиц настолько возрастает, что даже очень больших баз данных не хватает для оценки их статистики. Приведём данные , относящейся к английскому языку, и широко используемой базе данных Wall Street Journal Pronunciation Lexicon. Для английского языка количество фонем составляет около 50 (количество не является фиксированным – ряд распространённых дифонов (бифонов) или трифонов можно заранее отнести к отдельным фонемам). Полное количество трифонов составляет 503=125000. Часть этих трифонов запрещена фонетическими правилами данного языка и никогда не встречается, остаётся 95221 трифон. В упомянутой базе данных, которая составляет более 57 часов речи и содержит более 36000 предложений, встречается только 22804 трифона, из них только 14545 трифонов встречаются более 10 раз. Понятно, что для обучения СММ требуется значительное количество образцов моделируемого объекта. Число 10 можно признать минимально достаточным для обучения. Таким образом, более 80000 трифонов являются невидимыми (unseen), но могут встретиться при эксплуатации системы распознавания. Количество параметров для одной марковской модели может достигать 1000–2000 (сюда входят матрицы переходов и параметры гауссовых функций, аппроксимирующих функции плотности вероятности). Если умножить это число на количество трифонов (около 100000), общее количество параметров, которое надо оценить в процессе обучения, оказывается порядка 108–109.Таким образом, встаёт нетривиальная задача – оценить миллионы параметров, большинство из которых в обучающей базе данных не проявляются. Для преодоления этой трудности предложено два пути – основанный на фонетических представлениях, то есть в значительной степени субъективный, метод решающего дерева (decision tree) и процесс

образования трифонов, управляемый данными (data driven). Оба метода преследуют одну цель – управлять количеством оцениваемых параметров в зависимости от объёма обучающей выборки, поскольку опыт создания систем распознавания показал, что лучше надёжно обучить систему с небольшим количеством параметров, чем разработать сложную систему с большим количеством параметров, не обеспеченных данными для обучения.

7. МОДЕЛИ ЯЗЫКА

Можно сделать вывод, что задачу распознавания речи можно разбить на три. Первая из них имеет дело с анализом речевого сигнала, выделением и моделированием акустических признаков. Вторая отражает зависимости, существующие между словами в языке и определяющими возможные схемы следования слов друг за другом. Наконец, задачи третьей группы связаны с определением наилучшего кандидата на распознавание среди всех возможных с использованием той информации, которая создается в ходе решения задач первых двух групп. На основании такого разделения образуются три основных модуля любой системы распознавания слитной речи: акустическая модель, модель языка и декодер. До сих пор рассматривались вопросы, относящиеся к первой задаче. Основным понятием лингвистического описания речи является понятие модели языка. Произвольная модель языка позволяет формально описать язык, а точнее, те из его аспектов, которые необходимы для повышения качества автоматического распознавания речи. Определяя возможную последовательность слов, мы поднимаемся на более высокие уровни описания языка по сравнению с фонетическим и, как следствие, должны учитывать системные отношения высших порядков. Используемая модель описания слова в предложении может быть сложной, учитывающей синтаксическую и семантическую структуру высказывания, а может быть очень простой, полагающей, что появление любых слов равновероятно (в таком случае мы, по сути, отказываемся от лингвистического анализа и учета закономерностей и особенностей естественного языка). Языковая модель – обязательная часть систем распознавания слитной речи. Не любая последовательность слов является предложением (в особенности для языков типа немецкого – с жёстким порядком слов), между словами есть грамматические и семантические связи. Языковая модель позволяет узнать, какие последовательности слов в языке более вероятны, а какие менее. В однопроходных декодерах обычно информация от языковой модели учитывается одновременно с информацией от акустической модели (каждая со своим весом), в двухпроходных декодерах языковая модель обычно включается на втором этапе. Использование языковой модели помогает сократить пространство поиска и снять неоднозначность при выборе из нескольких близких по стоимости акустических гипотез (для русского языка, например, помогает правильно распознать слово в нужном падеже). Общепринятой мерой оценки моделей языка в отрыве от акустической модели является перплексия (или коэффициент неопределенности – perplexity), которая соответствует среднему коэффициенту ветвления после каждого слова, согласно модели языка. Перплексия представляет собой меру способности модели предсказывать неизвестные последовательности слов, является функцией кросс-энтропии

8. ДУКОДЕР

В ходе работы системы автоматического распознавания речи задача распознавания сводится к определению наиболее вероятной последовательности слов, соответствующих содержанию речевого сигнала. Наиболее вероятный кандидат должен определяться с учетом как акустической, так и лингвистической информации. Это означает, что необходимо производить эффективный поиск среди возможных кандидатов с учетом различной вероятностной информации. При распознавании слитной речи число таких кандидатов огромно, и даже использование самых простых моделей приводит к серьезным проблемам, связанным с быстродействием и памятью систем. Как результат, эта задача выносится в отдельный модуль системы автоматического распознавания речи, называемый декодером. Декодер должен определять наиболее грамматически вероятную гипотезу для неизвестного высказывания – то есть определять наиболее вероятный путь по сети распознавания, состоящей из моделей слов (которые, в свою очередь, формируются из моделей отдельных фонов). Правдоподобие (likelihood) гипотезы определяется двумя факторами, а именно вероятностями последовательности фонов, приписываемыми акустической моделью, и вероятностями следования слов друг за другом, определяемыми моделью языка. В случае распознавания слитной речи сеть вариантов распознавания оказывается настолько большой, что исследование и оценка всех возможных вариантов представляется невыполнимой с вычислительной точки зрения в режиме, сопоставимым с режимом реального времени. В связи с этим, оказывается необходимой разработка как можно более эффективных алгоритмов быстрого поиска, которые уже не будут гарантировать нахождение оптимального варианта распознавания, однако будут осуществлять поиск за приемлемое время. Как правило, поиск осуществляется в пределах «луча» (Beam Search), то есть все гипотезы, вероятность или правдоподобие которых уступает лучшей гипотезе на величину, превышающую некоторый порог, отбрасываются. Очевидно, что чем чаще производится анализ гипотез и чем раньше лишние гипотезы отбрасываются, тем эффективнее будет работа декодера, поскольку количество гипотез на каждом узле веерообразно увеличивается, напоминая лавину. Ещё одним способом ограничения требований к памяти и быстродействию является сохранение N лучших гипотез.

9. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Фланаган Дж. Анализ, синтез и восприятие речи. «Связь». Москва 1968.
2. MIT Lectures 2003. http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/download-course-materials/
3. Фант. Г. Акустическая теория речеобразования. «Наука». Москва 1964.
4. Picone J. Fundamentals of speech recognition: a short course.1996. http://speech.tifr.res.in/tutorials/fundamentalOfASR_picone96.pdf
5. Алдошина И. Основы психоакустики. http://giga.kadva.ru/files/edu/AldoshinaPsychoacoustics.pdf
6. Слуховая система. серия «Основы современной физиологии». «Наука», Ленинград, 1990.
7. Seneff S. “Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model”, Technical Report 504, January 1985
8. Hermansky H. (1997): “Should recognizers have ears?”, In RSR-1997, 1-10.
9. Маркел Дж.Д., Грей А.Х., Линейное предсказание речи, Москва,«Связь», 1980.
10. Hermansky H., Morgan N., «RASTA Processing of Speech», in IEEE Transaction on Speech and Audio Processing, Vol. 2, No. 4, pp. 587-589, October 1994.
11. Карпов А.А., Кипяткова И.С., Методология оценивания работы систем автоматического распознавания речи
12. Левенштейн В.И., Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965, 163.4:845-848.
13. Khokhlov Y., Tomashenko N., Speech Recognition Performance Evaluation for LVCSR System. In Proc. 14th International Conference “Speech and Computer” SPECOM-2011, Kazan, Russia, 2011, pp. 129-135.
14. Kurimo M., Creutz M., Varjokallio M., Arsoy E., Saraclar M., Unsupervised segmentation of words into morphemes - Morpho challenge 2005 Application to automatic speech recognition. In Proc. INTERSPEECH-2006, Pittsburgh, USA, 2006, pp. 1021-1024.
15. Schlippe T., Ochs S., Schultz T., Grapheme-to-Phoneme Model Generation for Indo-European Languages. In Proc. ICASSP-2012, Kyoto, Japan, 2012.
16. Huang C., Chang E., Zhou J., Lee K. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In Proc. INTERSPEECH-2000, Beijing, China, 2000, pp. 818-821.

17. Ablimit M., Neubig G., Mimura M., Mori S., Kawahara T., Hamdulla A. Uyghur Morpheme-based Language Models and ASR. In Proc. 10th IEEE International Conference on Signal Processing ICSP-2010, Beijing, China, 2010, pp. 581-584.

132

18. Karpov A., Kipyatkova I., Ronzhin A. Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis. In Proc. INTERSPEECH-2011, Florence, Italy, 2011, pp. 3161-3164.
19. Karpov A. A., Tampel I.B. «Automatic speech decoding».