Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
Национальный Исследовательский Университет
“Высшая Школа Экономики”

На правах рукописи

Ермилов Алексей Валерьевич

Методы, алгоритмы и программы решения задач идентификации языка и диктора

Специальность 05.13.11 —
«Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор технических наук
Гостев И.М.

Москва – 2014

Содержание

Введение

В современном мире все большее значение уделяется интерфейсам, использующим речевой ввод и вывод для взаимодействия между пользователем и компьютером. Поэтому разработчику систем распознавания речи, реализующих акустический интерфейс, приходиться принимать во внимание всё большую вариативность в голосовых сообщениях.

Задача распознавания речи (во многих своих проявлениях: от транскрибирования слитной речи до верификации и идентификации диктора) в настоящее время является крайне актуальной. Свидетельством этому служит растущее число публикаций и конференций по данной тематике (таких как ICASSP, INTERSPEECH), а также то, что в крупнейших транснациональных корпорациях (таких как Microsoft, Google, IBM) открываются департаменты, ориентированные на исследования по данной тематике.

Исследовательские усилия в сфере речевых технологий привели к появлению большого числа коммерческих систем распознавания речи. Такие компании как Nuance, IBM, ScanSoft предлагают большой набор программных решений как для серверных, так и для десктопных приложений.

Улучшение существующих систем распознавания речи позволит существенно упростить взаимодействие человека с компьютером в том случае, когда использование классических интерфейсов невозможно (например, при управлении автомобилем или в сложных условиях, таких как ликвидация последствий чрезвычайных ситуаций) или затруднено

(например, людям, обладающим слабым зрением, или с ограниченными физическими возможностями), а также сделать работу с компьютером или иной техникой более комфортной. Также следует отметить, что применение систем распознавания речи весьма велико в работе правоохранительных служб (например, при идентификации говорящего или в системе защиты свидетелей).

Необходимость исследований по этой тематике объясняется малоудовлетворительными результатами существующих систем при низком соотношении сигнал/шум, зависимостями результата от диктора и, в ряде задач, невысокой скоростью работы систем.

Существующие системы распознавания речи в основном построены на Скрытых Марковских Моделях (HMM), которые задают динамику перехода от одной фонемы в речи к другой и обеспечивают вариативность наблюдаемого сигнала посредством моделирования вероятностного распределения признаков посредством Гауссовой Смеси (GMM) [1]. Такой подход был предложен в 1989 Лоуренсом Рабинером и долгое время являлся основным для моделирования речевого сигнала.

Быстро развивающейся альтернативой HMM становятся Deep Belief Networks [2], которые обеспечивают высокую точность распознавания. Работы, посвященные байесовским сетям, были начаты в середине 80-х годов, но особую распространённость получили после публикаций серии работ Д. Хинтона, в которых приводились эффективные алгоритмы подобных сетей, а также примеры их использования.

Для описания речевого сигнала в системах автоматического распознавания речи со времен работы Л. Рабинера используются так называемаы мел-кепстральные коэффициенты (MFCC Mel Frequency Cepstral Coefficients), начало развитию которых положил Пол Мермельстайн в 1976 [3].

Также следует отметить, что в последнее время альтернативой используемым сейчас MFCC становятся признаки, устойчивые к вариабельности речевого тракта у диктора (например, bottleneck features [4]), что позволяет строить более робастные системы. В данном исследовании предлагается новая вероятностная модель (расширение Грам — Шарлье для функции распределения) для дикторонезависимых признаков и использование Фишеровских ядер в алгоритме опорных векторов, а также используется новые вычислительные методы для оценки этих модели (алгоритм симуляции отжига), использующие преимущества параллельных вычислений. Следует отметить, что указанные методы пока не получили широкого распространения при решении задачи распознавания речи и их применение является новаторским и может послужить базой для дальнейшего развития этого направления. При применении этих моделей имеется прирост в точности распознавания языка и диктора, а также ускорение работы всей системы распознавания.

Следует также отметить, что более широкому распространению компьютерных систем распознавания речи способствовало активное развитие сначала многопоточных, а затем и многопроцессорных систем, в том числе и многоядерных с общей памятью.

В течении длительного времени использование систем автоматического распознавания больших параллельных потоков речи было ограничено в виду недостаточного быстродействия оборудования, а именно - невозможности обработки online. Для работы в реальном времени системам, оперирующим с непрерывными потоками речи, приходилось находить компромисс между объемом словаря (а значит, и потенциальной сферой применения), сложностью грамматики и точностью распознавания. Таким образом, повышение скорости работы распознавателя будет положительным образом сказываться на объеме тех задач, где возможна работа в реальном времени, а также на точности

распознавания. Хорошим примером может служить работа сотовой станции или call – центра, где на обработку одновременно может приходить огромное количество заявок, требующих процессинга в реальном времени.

Предметом настоящего исследования является задача распознавания языка и диктора, которая является частным случаем задачи распознавания образов, в которую также входят задачи классификации, регрессии и восстановления эмпирических зависимостей по историческим данным [5].

Целью данной работы является

1. Создание метода идентификации диктора по речевому сообщению для создания человеко - машинного интерфейса.
2. Разработка дикторонезависимых признаков речевого сигнала и методов их получения для решения задачи идентификации языка.
3. Анализ численных методов решения задач идентификации языка и диктора.
4. Построение параллельных алгоритмов решения задач идентификации языка и диктора.
5. Программная реализация указанных методов и исследование их практической применимости.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Исследование моделей акустических сигналов, применяемых в системах распознавания языка и диктора.
2. Разработка математический модели дикторонезависимых акустических признаков на основе 4-х параметрического семейства распределений.

3. Модификация метода опорных векторов для решения задачи идентификации диктора по речевому сообщению фиксированной длины с целью повышения качества распознавания.

4. Модификация метода симуляции отжига для повышения быстродействия системы и увеличения качества признаков, применяемых для распознавания языка.

5. Анализ предложенных и существующий моделей и методов для сравнения их быстродействия и точности распознавания.

Основные положения, выносимые на защиту:

1. Проведён анализ существующего состояния в сфере распознавания языка и диктора.

2. Выявлены дикторонезависимые признаки, основанные на 4-х параметрическом распределении, и доказана их оптимальность.

3. Разработана модификация алгоритма симуляции отжига, увеличивающия быстродействие системы при получении дикторонезависимых признаков.

4. Разработана и теоретически обоснована модифицикация метода опорных векторов, основанная на применении фишеровских ядер, которая позволяет увеличить точность распознавания диктора.

5. Проведён сравнительный анализ алгоритмов оптимизации для получения дикторонезависимых признаков по скорости и точности.

6. Разработаны и теоретически обоснованы методы и алгоритмы получения параметров классификатора для решения задач идентификации языка и диктора.

7. Создана программная реализация разработанной системы идентификации языка и диктора, фрагменты который внедрены на производстве.

8. Проведены экспериментальные исследования по оценке точности распознавания и быстродействию системы идентификации языка и диктора, которые показали преимущества разработанных методов по сравнению с применяемыми ранее.

Научная новизна:

1. Изучены информационные признаки идентификации языка и диктора на основе физиологических особенностей человеческого языка и дикции с учетом механизма восприятия звука человеком при распознавании речи.

2. Впервые предложена система характерных признаков для распознавания языка на основе 4-х параметрического семейства распределений (расширение Грам-Шарлье).

3. Разработана и обоснована теоретически модификация метода опорных векторов, основанная на применении фишеровских ядер, которая позволяет увеличить точность распознавания диктора.

4. Впервые проведён сравнительный анализ алгоритмов оптимизации для вычисления акустических дикторонезависимых признаков по скорости и точности.

5. Разработана модификация алгоритма симуляции отжига увеличивающая быстродействие системы при получении дикторонезависимых признаков за счет введения в алгоритм параллельно выполняющихся циклов.

6. Разработаны и теоретически обоснованы методы и алгоритмы получения параметров классификатора для решения задач идентификации языка,

основанные на использовании метода опорных векторов, повышающие точность распознавания.

7. Проведены экспериментальные исследования по оценке точности распознавания и быстродействию системы идентификации языка и диктора, которые показали преимущества разработанных методов по сравнению с применяемыми ранее.

Теоретическая значимость. Теоретическая значимость заключается в следующем.

1. Впервые разработаны методы идентификации диктора, основанные на методе опорных векторов с применением Фишеровских ядер.
2. Впервые была предложена и теоретически обоснована модель акустических дикторонезависимых признаков, использующая 4-х параметрическое распределение (расширение Грам-Шарлье) для моделирования речевых признаков, которая была использована для аутентификации и работе правоохранительных служб.
3. Впервые разработана модификация алгоритма симуляции отжига увеличивающая быстродействие системы при получении дикторонезависимых признаков за счет введения в алгоритм параллельно-выполняющихся циклов.

Практическая значимость. Полученные автором результаты имеют большое научное и народно-хозяйственное значение (имеется акт о внедрении) при создании человеко-машинных интерфейсов и идентификации личности и языка в работе различных государственных служб и органов.

Степень достоверности полученных результатов обеспечивается использованием строгих математических методов теории вероятностей,

математической статистики и распознавания образов. Достоверность подтверждается моделированием и проведенными вычислительными экспериментами с использованием реальных и симулированных данных, а также путём сопоставления результатов, полученных в диссертации, с результатами, доступными в открытой печати.

Апробация работы. По материалам диссертации опубликовано 5 статей (3 из которых в журналах из списка ВАК, одна в международном реферируемом журнале), 6 тезисов на международных конференциях. Результаты настоящего исследования были представлены на следующих конференциях и семинарах: Конференции студентов, аспирантов и молодых специалистов МИЭМ в 2010 г; Конференции студентов, аспирантов и молодых специалистов МИЭМ в 2011 г; Международной конференции «Моделирование нелинейных процессов и систем» (СТАНКИН 2011 г.); 5-я Международной Конференции «Распределённые вычисления и Грид-технологии в науке и образовании» (GRID - 2012) (Дубна Московская обл. 2012 г.); X Международной научно-технической конференции «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (Курск 2012); The First International Conference on Modern Manufacturing Technologies in Industrial Engineering “ModTech – 2013”, (Румыния, Синая 2013 г.); International Conference on Mathematic Modeling and Computing in Physics (MMCP’2013) (Дубна Московская обл., 2013 г.); XI Международной научно-технической конференции «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (Курск 2013).

Личный вклад. Во всех работах с соавторами вклад автора диссертации является определяющим.

Публикации. Основные результаты по теме диссертации изложены в 11 печатных изданиях [6–16], 3 из которых изданы в журналах, рекомендованных ВАК [6–8], одна работа [9] опубликована а международном реферируемом журнале, 6 — в тезисах докладов [11–15].

Объем и структура работы. Диссертация состоит из введения, четырёх глав и заключения. Полный объем диссертации составляет 134 страницы с 26 рисунками и 5 таблицами. Список литературы содержит 81 наименование.

Глава 1

Методология обработки речевого сигнала

1.1 Общая схема обработки речевого сигнала

Целью данного раздела является описание общей схемы обработки речевого сигнала для идентификации языка и диктора.

На рис. 1.1 изображена упрощённая схема речевого аппарата человека. Речевой сигнал получается прохождением воздуха через так называемый речевой тракт.

Определение. Речевым трактом называют часть речевого аппарата человека, которая располагается между голосовой щелью и губами.

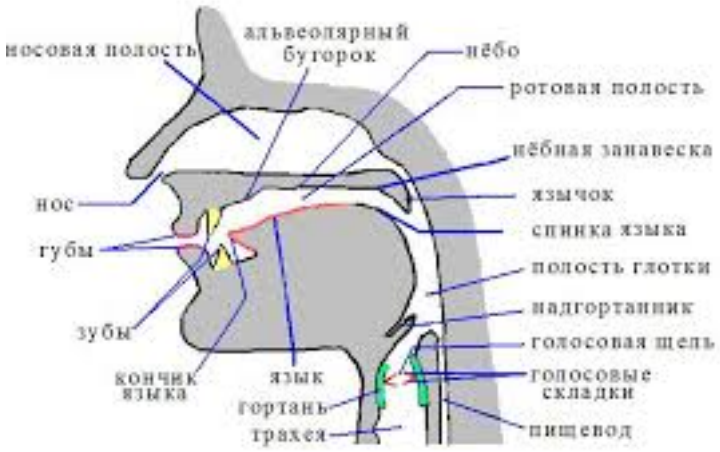


Рис. 1.1: Схема речевого аппарата человека [17].

Генерация речевого сигнала происходит следующим образом. Создаваемый легкими поток воздуха за счет вибраций голосовых связок модулируется в гортани, форма которой является важным для формирования звуков. Способность голосовых связок изменять свою форму и колебаться по частям в процессе голосообразования приводит к разнообразию издаваемых человеком звуков. При движении вдоль речеового тракта могут изменяться характеристики воздушного потока, что и приводит к преобразованию звукового сигнала в акустический речевой сигнал. Описание базовых акустистических речевых сигналов будет дано ниже.

В речевом сообщении содержится все информация, необходимая для его распознования, однако из-за сильной изменчивости сигнала необходимо проводить предварительную обработку для выделение признаков с целью последующего анализа.

На рис. 1.2 приведена упрощенная схема обработки речевого сигнал для идентификации языка и диктора.

На первом этапе обработки из речевого сигнала удаляют шум, производят усиление и выравнивают сигнал в спектральной области. Цель этого этапа заключается в том, чтобы сделать сигнал как можно более чистым. Стоит отметить, что свойства речевого сигнала медленно меняются со временем, то есть, он является квази-стационарным. Если рассматривать его на коротких временных интервалах (5-100 мс), то характеристики остаются постоянными. Поэтому на этапе предобработки речевой сигнал нарезают на участки, называемые фреймами, с помощью движущегося окна.

На втором этапе происходит выделение акустических признаков. Известно большое количество таких признаков, наиболее популярными из которых являются коэффициенты линейного предсказания [18] и кепстральные

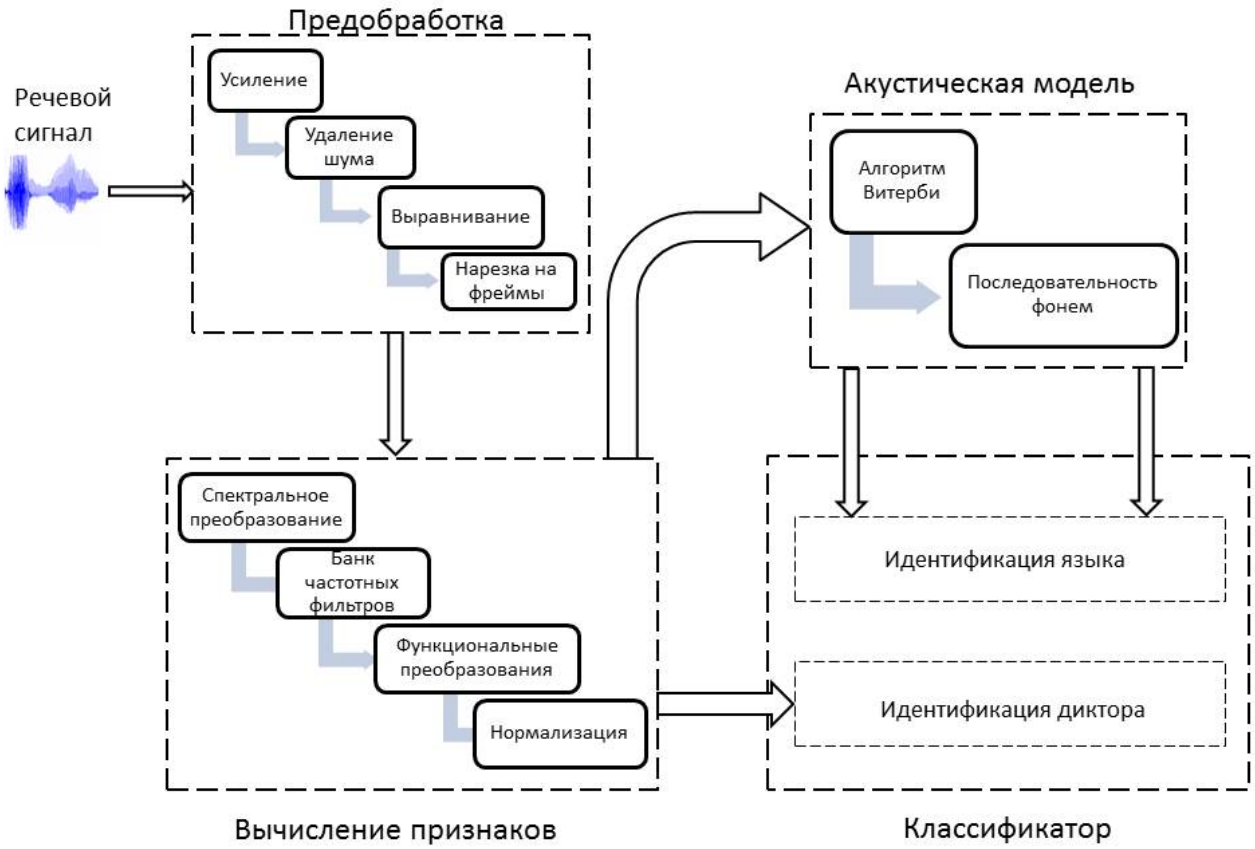


Рис. 1.2: Общая схема обработки речевого сигнала для идентификации языка и диктора.

коэффициенты [19]. Для получения признаков на каждом фрейме над сигналом проводят следующие операции.

- Спектральное преобразование (например, с помощью Быстрого Преобразования Фурье).
- Фильтрация с помощью банка фильтров. Пример подобного банка фильтров приведён на рис. 1.3.
- Функциональное преобразование. Например, логарифмирование.
- Нормализация. Например, центрирование на нулевое среднее и единичную дисперсию.

В зависимости от решаемой задачи вычисленные признаки либо непосредственно используются для классификации (например, для решения

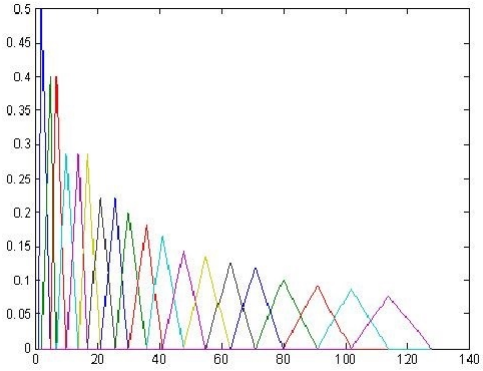


Рис. 1.3: Банк треугольных фильтров

задачи идентификации диктора), либо подаются на вход акустической модели, результат работы которой используется в дальнейшем (например для транскрибирования речи, либо для распознавания языка). В качестве акустической модели обычно используются Скрытые Марковские Модели.

1.2 Акустические характеристики и особенности речевых сигналов

Целью данного раздела является изложение особенностей физических аспектов акустических сообщений, используемых при идентификации речевых сигналов.

Физические свойства звука могут быть описаны в виде суперпозиции волн с разным звуковым давлением, которые распространяются в некоторой физической среде, например, такой как воздух. В настоящей работе будут исследоваться только продольные звуковые волны [20], то есть такие, где молекулы среды движутся относительно их средней позиции в направлении, совпадающем с направлением распространения волны. Распространение волны приводит к тому, что молекулы, располагающиеся на расстояние в

половину волны друг от друга, вибрируют в противоположных направлениях, что приводит к появлению сменяющих друг друга регионов сжатия и разряжения. Следовательно, давление звука, определяемое как разность между мгновенным и статическим давлением, представляет собой функцию позиции и времени.

В дальнейшем будет предполагаться, что звуковые волны распространяются исключительно в воздухе и среда распространения обладает следующими свойствами:

1. Гомогенность, то есть однородность структуры.
2. Изотропность, то есть независимость свойств среды от направления.
3. Стационарность, то есть независимость свойств среды от времени.

Среда, в которой возможно распространение звука, обладает свойствами массы и эластичности. Эластичность идеального газа определяется дилатацией объёма и сжатием объёма.

Сжатие объёма или отрицательная дилатация идеального газа определяется как

$$-\Delta = -\frac{\delta V}{V},$$

где V - объём, V - изменение объёма.

Эластичность идеального газа определяется объёмным модулем

$$\kappa = \frac{\delta p}{-\Delta},$$

где δp - изменение давления.

Распространение звука представляет собой адиабатический процесс, так как расширение и сжатие продольных волн происходят быстрее, чем распространение тепла. Обозначим через C_p и C_v теплоемкости при

постоянном давлении и постоянном объёме соответственно. Тогда объёмный модуль можно приблизить с помощью адиабатической экспоненты $\gamma = \frac{C_p}{C_v}$:

$$\kappa \approx \gamma p.$$

Скорость звука в направлении, противоположном от источника, была измерена Лапласом в условиях адиабатического процесса $c = \sqrt{\frac{\kappa}{\rho}}$, где ρ – плотность воздуха. Скорость звука в воздухе зависит главным образом от атмосферных условий (в основном от температуры и в меньшей степени от влажности). В предположении о том, что воздух представляет собой идеальный газ, давление воздуха не играет роли на скорость звука, так как давление и плотность влияют на скорость одинаково, и, как следствие, эти два эффекта компенсируют друг друга.

Чтобы дать определение интенсивности звука, введём понятие потенциала скорости. В консервативном и односвязанном векторном поле скорость потока может быть представлена как градиент от скалярной функции, которая и называется потенциалом скорости.

Определение. Интенсивность звука или акустическая интенсивность есть произведение звукового давления p на потенциал скорости ϕ: $I_{\text{звук}} = p\phi$.

Утверждение. Интенсивность звука обратно пропорциональна квадрату расстояния до источника [21].

Доказательство. Решение волнового уравнения [22] может быть представлено как суперпозиция исходящей и входящей звуковых волн:

$$p = \frac{A}{r} e^{\jmath\omega t - \jmath kr} + \frac{B}{r} e^{\jmath\omega t + \jmath kr},$$

где A, B - силы источников. Используя соотношение между звуковым давлением и потенциалом скорости [20] $p = \rho_0 \frac{\partial \phi}{\partial t}$, интенсивность звука представляется как

$$I_{\text{звук}} = \frac{p^2}{c\rho_0}.$$

Из этих двух уравнений получаем требуемое.

1.3 Особенности описания речевых сигналов для их идентификации

Цель этого раздела показать особенности представления речевого сигнала, описать методы разбиения его на разные фонетические единицы и изложить методы, характеризующие статистические свойства речевого сигнала.

1.3.1 Модель речеобразования

Известно, что речь состоит из звуковых волн, созданных прохождением воздуха через речевой тракт. Квазипериодическое открытие и закрытие речевых складок приводит к произношению звонких звуков, таких как гласные, отличающиеся периодичностью и большими значениями энергии, и некоторых согласных. В случае, когда речевые складки не вибрируют, образуются согласные звуки. Дополнительное разделение речевого сигнала на звонкие и глухие звуки очень важно, так как эти звуки имеют разные характеристики как в спектральной, так и временной областях.

Физиологические особенности речевого тракта приводят к тому, что речь каждого человека обладает уникальными параметрами, такими как высота тона, скорость произношения, акцент и др. При произношении гласных звуков форма и длина речевого тракта оказывают влияние на расположение и высоту спектральных пиков, называемых формантами. Форманты в свою очередь формируют спектр.

Моделирование речеобразования сводится к моделированию фонем, базовых лингвистических единиц, за образование которых отвечают два

фактора: случайный шум или возбуждающие импульсы и форма речевого тракта. При моделировании можно считать, что эти факторы независимы [23].

Процесс речеобразования обычно моделируют, используя линейную динамическую систему [20]. Пример такой модели приведён на рис. 1.4. Здесь

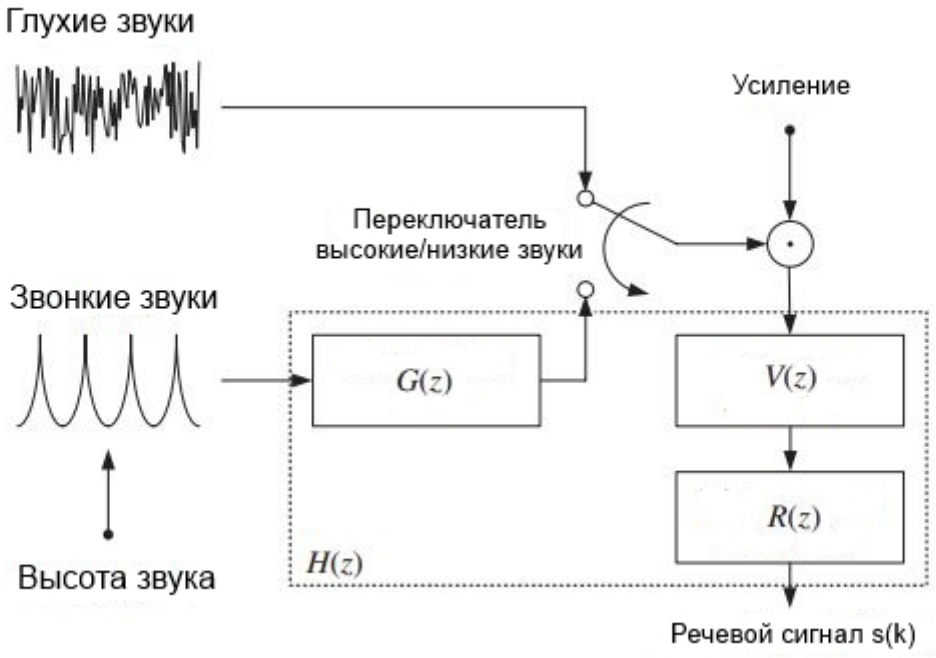


Рис. 1.4: Линейная динамическая система речеобразования

через фильтр речевого тракта $V(z)$ и фильтр губного испускания $R(z)$ проходит либо последовательность возбуждающих импульсов, либо зашумленный сигнал с плоским спектром. Фильтр речевого тракта $V(z)$ имеет плоский спектральный тренд, но при этом локальные резонансы и антирезонансы могут присутствовать. Губы в данной модели представляют собой фильтр высоких частот $R(z)$, с усилением 6 ДБ на октаву. Для моделирования звонких звуков возбуждающие импульсы имеют высоту звука p, с наложенным фильтром низких частот второго порядка $G(z)$, имеющим усиление, которое убывает на 12 ДБ на октаву. Этот фильтр моделирует прохождение звука через голосовую щель.

Для описания речи используются различные схемы. Примером такой схемы является фонемная. При этом фонемой называется элементарная лингвистическая единица, достаточная для различения двух слов.

Акустической реализацией фонемы является фон.

В соответсвии с Международным Фонетическим Алфавитом [20] фонемы могут быть разделены на два главных класса: гласные и согласные. Согласные звуки могут быть дальше классифицированы на лёгочные и не лёгочные. Дальнейшая классификация согласных звуков может быть произведена следующим образом.

- Носовые звуки.
- Взрывные звуки.
- Фрикативные звуки.

Классы гласных и согласных звуков могут быть расширены путем включения переходных классов, например, аппроксимантов и дифтонгов. Аппроксиманты - это звонкие звуки, лежащие между гласными и согласными. Дифтонги представляют собой комбинацию гласного звука и перехода от этого гласного звука к другому гласному звуку.

1.3.2 Статистические свойства речевого сигнала

Речевой сигнал представляет собой нестационарный процесс [24], то есть, его статистические свойства меняются со временем. Вместе с тем представляется возможным так “нарезать“ речевой сигнал на сегменты некоторой длины (такие сегменты называются фреймами), чтобы в пределах одного сегмента характеристики процесса менялись не слишком сильно. Таким образом, представляется возможным использование методов теории случайных процессов для моделирования речевых сигналов.

Статистические свойства речевого сигнала важны как для вычисления признаков, используемых для распознавания, так и для самого распознавания. На практике широко используются признаки, основанные на моментах второго порядка: спектр и автокорреляционная функция. В последнее время (см., например, [25, 26]) начали использоваться моменты более высокого порядка, таких как ассиметрия и эксцесс. Мотивацией этому служит явная негауссовость распределения речевого сигнала, как во временной области, так и в частотной. На рисунке 1.5 изображена гистограмма наблюдений амплитуды речевого сигнала с подогнанными распределениями. Поэтому настоящая работа

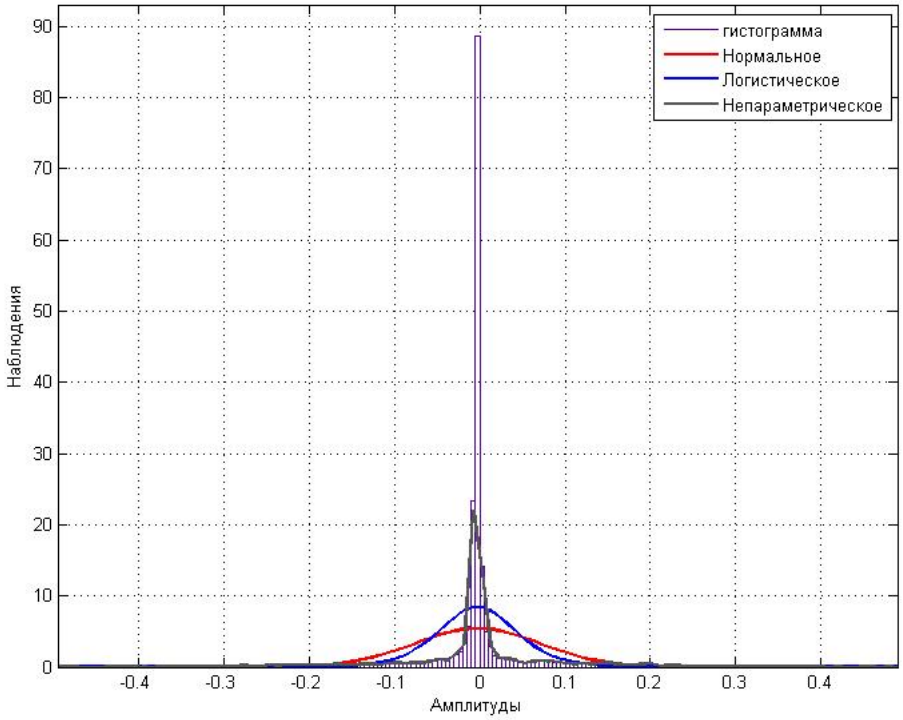


Рис. 1.5: Гистограмма значений амплитуды речевого сигнала.

базируется на использовании и моделировании моментов высокого порядка.

1.4 Анализ методов распознавания речи, языка и диктора

Существуют различные принципы распознавания речи. В настоящей работе будет использоваться классификация на основе [24], где выделяют следующие подходы к автоматическому распознаванию речи:

- Акустико-фонетический подход.
- Подход с точки зрения распознавания образов.
- Подход с точки зрения искусственного интеллекта.

Кратко рассмотрим эти подходы по отдельности.

1.4.1 Акустико-фонетический подход

Акустико-фонетический подход использует для распознавания речи последовательное декодирование сигнала, представленного в виде наблюдаемых акустических признаков, используя известные взаимосвязи акустических и фонетических символов [24]. В этом подходе постулируется, что в разговорном языке существуют различимые фонетические единицы, которые могут быть описаны с помощью набора характеристик, наблюдаемых в речевом сигнале. Кроме того, предполагается, что, несмотря на то, что эти характеристики могут значительно изменяться не только от диктора к диктору, но и между соседними фонетическим единицами, представляется возможным описать и применить на практике правила, описывающие эти изменения.

Таким образом, первым шагом в акустико-фонетическом подходе к распознаванию речи является разбиение (сегментация) речевого сигнала на дискретные во времени участки, в которых акустические свойства сигнала представлены одной фонетической единицей, с последующей разметкой этих участков с помощью фонетических символов (меток). На этом этапы

получается так называемая фонетическая решётка. Получение надёжной фонетической решетки уже представляет собой достаточно сложную проблему. На втором шаге происходит непосредственно распознавание, при котором определяются целые слова из последовательностей символов, полученных ранее. Также следует иметь в виду, что полученные слова должны удовлетворять заданным ограничениям, например, присутствовать в априори заданном словаре, иметь семантический смысл и т.д.

К сожалению, такой подход обладает многими серьёзными недостатками, которые ведут к слабым нестабильным результатам в системах распознавания речи. Среди таких недостатков можно отметить:

1. Необходимость объемных знаний об акустических свойствах фонетических единиц.
2. Качество распознавания зависит от выбранных признаков и определяется разработчиком системы, так как нет никакой методики их выбора.
3. Отсутствие оптимальной схемы классификации звуков.
4. Зависимость разметки речевого сигнала от человека, проводящего эту разметку, с учётом восприятия данного речевого сигнала.

1.4.2 Подход с точки зрения распознавания образов

Подход с точки зрения распознавания образов основан на том, что речевые паттерны используются без предварительного определения признаков и сегментации [24]. Как и в большинстве подходов основанных на распознавании образов, процесс распознавания происходит в два этапа. На первом этапе (обучение) система приобретает знания о речевых паттернах. Предполагается, что если системе будет предоставлено достаточное число экземпляров паттерна, то она сможет охарактеризовать его акустические

свойства. Таким образом, обучение системы заключается в том, чтобы находить устойчивые акустические свойства, которые повторяются на всём обучающем множестве. Распознавание здесь происходит путём сравнения паттерна из неизвестного сегмента речи со всеми возможными паттернами, которые получены за время обучения. Далее система соотносит неизвестный сегмент с тем паттерном, который наиболее близок (в заданной метрике) к этому сегменту. К достоинствам и недостаткам этого подхода можно отнести следующее:

1. Зависимость от объема имеющегося обучающего множества: чем больше объем обучающего множества, тем выше точность распознавания системы.
2. Система получается чувствительной к способу передачи сигнала и характеристикам среды, в виду того, что для описания паттернов часто используются спектральные характеристики.
3. Отсутствие использования информации, специфичной для данного сегмента речи: выбор словаря, тематической области, семантики.
4. Вычислительная нагрузка на обучение и классификацию зависит линейно от объёмов входных данных, следовательно, для больших задач обучение и классификация может занять слишком много времени.
5. Возможность включения ограничений на распознаваемые элементы прямо в распознающую структуру.

1.4.3 Подход с точки зрения исусственного интеллекта

Подход с точки зрения искусственного интеллекта представляет собой гибрид акустически-фонетического подхода и подхода с точки зрения

распознавания образов [24]. Идея заключается в том, чтобы собрать и объединить знания из разных источников и применить их к решаемой задаче. В этом случае делается попытка автоматизировать процесс распознавания, например, для сегментации и разметки речевого сигнала, обучения и адаптации к новым данным и т.д.

К различным источникам информации о речевом сигнале можно отнести:

- Акустическую информацию – информацию о том, какие фонетические единицы (например, звуки) были произнесены на основе спектрального анализа сигнала.
- Лексическую информацию – информацию о том, как звуки образуют слова.
- Синтаксическую информацию – информацию о том, как согласно заданной модели языка, слова образуют более сложные элементы языка (фразы или предложения).
- Семантическую информацию - информацию о предметной области, которая используется для проверки того, как уже распознанные фразы соотносятся с предметной области.

Существует несколько способов интегрировать указанные источники внутри распознавателя [24]. Наиболее часто используется подход, в котором низкоуровневая обработка сигнала (определение признаков, фонетическое декодирование) происходит непосредственно перед применением моделей языка и других средств обработки высокого уровня. Альтернативным подходом является использования модели языка для формулирования слов – гипотез, соответствующих данному речевому сигналу.

1.5 Методы выделения акустических признаков

Все три описанных подхода к автоматическому распознаванию речи (акустико-фонетический, с точки зрения распознавания образов и с точки зрения искусственного интеллекта) объединяет необходимость выявления признаков, описывающих волновой речевой сигнал с помощью набора параметров с целью его дальнейшего анализа и обработки.

Наиболее часто используемым параметрическим представлением речи является спектральный анализ. Примером получения спектральной информации о сигнале является Быстрое Преобразование Фурье. К сожалению, такая спектрограмма оказывается слишком сложным представлением речевого сигнала и требует некоторого изменения для практического применения.

Рассмотрим основные модели спектрального анализа, позволяющие характеризовать спектр в более сжатом виде: модель спектра на основе банка фильтров, модель на основе коэффициентов линейного предсказания (Linear Predictive Coding - LPC) и аудиторная модель.

1.5.1 Модель банка фильтров

Одним из методов для сжатого представления спектра является метод банка фильтров. Идея метода заключается в том, чтобы разделить интересующий диапазон частот на полосы и измерять общую энергетику в каждой полосе. Фильтры могут перекрываться по частотам.

Обозначим через $s(t)$ исходный речевой сигнал, тогда сигнал на выходе из фильтра представляет собой краткосрочное спектральное представление исходного сигнала в момент времени t. Очевидно, что в этой модели каждый фильтр обрабатывает речевой сигнал независимо.

Пройдя через банк фильтров, входной сигнал представляется в виде:

$$\begin{aligned} s_i(t) &= s(t) * h_i(t) \\ &= \sum_{m=0}^{M_i-1} h_i(m)s(t-m), \qquad 1 \le i \le N \end{aligned}$$

где $h_i(m)$ импульсный отклик i-го фильтра длительностью M_i, форма которого будет представлена позже.

Цель данного метода – представить энергию входного речевого сигнала в данной частотной полосе, поэтому каждый полосовой сигнал обрабатывается нелинейной функцией (например, полуволновым выпрямителем). Такое воздействие приводит к тому, что спектр сигнала смещается в полосу нижних частот и одновременно создаётся высокочастотный образ. Для удаления высокочастотной составляющей используются фильтры нижних частот. В результате получается набор набор сигналов $u_i(n), 1 \le i \le N$, которые представляют оценку энергии речевого сигнала в каждой полосе.

Важным вопросом является выбор частот, на которых будут располагаться фильтры. Существует несколько способов выбора.

1. Равномерное расположение. Центральная частота f_i i-го фильтра определяется как

$$f_i = \frac{F}{N}, \qquad 1 \le i \le N,$$

где F частота дискретизации, а N – количество фильтров необходимых для покрытия всего диапазона частот. Следует отметить, что использование равномерно расположенных фильтров не оптимально, так как при распознавании гласных их форманты будут попадать в $1, 2, 3$ банк фильтров, что приводит к тому, что почти все гласные попадут в один класс.

2. Логарифмическая шкала. Центральная частота и ширина полосы пропускания строятся следующим образом:

$$\begin{aligned} b_1 &= C, \\ b_i &= \alpha b_{i-1}, \qquad 2 \le i \le N, \\ f_i &= f_1 + \sum_{j=1}^{i-1} b_j + \frac{b_i - b_1}{2}, \end{aligned}$$

где C, f_1 – произвольные значения полосы пропускания и центральной частоты первого фильтра, а α - логарифмический фактор роста. В этом случае проблема с неудачной классификацией гласных пропадает, так как форманты будут попадать в разные участки частот.

3. Расположение в соответствии с особенностями человеческого восприятия. В этом случае выбор центральных частот фильтров производится так, чтобы фильтры давали одинаковый вклад в артикуляцию речи.

Следует отметить, что перед началом анализа с помощью банка фильтров необходимо произвести предварительную обработку речевого сигнала. Её цель заключается в том, чтобы сделать сигнал как можно более чистым: убрать шум, убрать долгосрочные спектральные тренды и выровнять сигнал в спектральной области. Предварительная обработка сигнала может включать в себя большой набор операций, среди которых необходимо отметить следующие:

1. Предварительное усиление. Усиление применяется для того, чтобы выровнять присущую речи неравномерности в спектре.

2. Удаление шума.

1.5.2 Коэффициенты линейного предсказания

Для представления модели речевого сигнала в распознавании речи широко используются коэффициенты линейного предсказания (LPC [18]) благодаря следующим особенностям:

1. Модель демонстрирует особенно хорошие свойства [24] для тех участков сигнала, где имеется звуковые элементы. На тех участках, где находиться тишина или переходящие участки речи, модель менее эффективна.

2. Применение LPC ведёт к тому, что происходит хорошее разделение источника и вокального тракта, что приводит к простому представлению его характеристик.

3. LPC представляет собой математически точную и достаточно простую в реализации модель.

Основная идея, используемая в модели линейного предсказания заключается в том, что сигнал в данный момент времени можно представить в виде линейной комбинации предыдущих значений:

$$s(t) \approx \sum_{i=1}^{p} a_i s(t-i),$$

где коэффициенты $\{a_i\}_{i=1}^{p}$ считаются постоянными на протяжении времени фрейма. Также можно добавить в уравнение возбуждающий сигнал $Gu(t)$:

$$s(t) = \sum_{i=1}^{p} a_i s(t-i) + Gu(t), \tag{1.1}$$

где $u(t)$ – нормализованный возбуждающий сигнал и G – усиление возбуждающего сигнала. Это уравнение интерпретируется в следующем виде [24]. Речевой сигнал является результатом прохождения возбуждающего сигнала через систему, чья передаточная функция имеет только полюса, определяемые коэффициентами $\{a_i\}_{i=1}^{p}$.

Главная проблема при использовании модели линейного предсказания заключается в том, что необходимо определять набор коэффициентов из речевого сигнала таким образом, чтобы спектральные свойства цифрового фильтра совпадали со свойствами данного участка звуковой волны в пределах данного окна наблюдений. Сложность этого проявляется в том, что спектральные характеристики речи изменяются во времени. Поэтому коэффициенты следует оценивать, основываясь на данных из короткого временного интервала перед данным моментом времени t. Чтобы выбрать необходимый временной интервал используют взвешивание на основе движущегося окна. Сами оценки можно получить методом наименьших квадратов [27].

Для обоснования применения метода наименьших квадратов к уравнению (1.1) докажем следующую теорему, введя определения состоятельности и асимптотической нормальности.

Определение. Пусть $\{X_i\}_{i=1}^n$ - выборка из распределения, зависящего от параметра $\theta \in \Theta$. Тогда оценка $\hat{\theta}$ называется состоятельной, если

$$\hat{\theta} \to \theta, \text{ по вероятности при } n \to \infty$$

Определение. Пусть $\{X_i\}_{i=1}^n$ - выборка из распределения, зависящего от параметра $\theta \in \Theta$. Тогда оценка $\hat{\theta}$ называется асимтотически нормальной с дисперсией σ^2, если

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathbb{Z}, \text{ по распределнию при } n \to \infty,$$

где $\mathbb{Z}$ - нормальная случайная величина с дисперсией σ^2 и средним 0.

Утверждение. Оценки коэффициентов $\{a_i\}_{i=1}^p$ из уравнения (1.1), полученные методом наименьших квадратов, состоятельные и асимптотически нормальные.

Доказательство. Введём матричные обозначения:

$$
\begin{aligned}
\alpha^T &= (a_1 \dots a_p) - \text{вектор искомых параметров,} \\
s_\tau &= (s(\tau-1) \dots s(\tau-p)), \\
y_t &= \begin{pmatrix} s(t) \\ \vdots \\ s(p+1) \end{pmatrix} - \text{вектор наблюдений,} \\
u_t &= \begin{pmatrix} u(t) \\ \vdots \\ u(p+1) \end{pmatrix} - \text{вектор значений возбуждающего сигнала,} \\
S_\tau &= \begin{pmatrix} s_\tau \\ \vdots \\ s_p \end{pmatrix} - \text{матрица из } s_\tau.
\end{aligned}
$$

Тогда уравнение (1.1) можно переписать в виде

$$y_t = S_t\alpha + Gu_t.$$

Отсюда согласно методу наименьших квадратов получаем, что

$$\hat{\alpha} = (S_t^T S_t)^{-1} S_t^T y_t = \alpha + G(S_t^T S_t)^{-1} S_t^T u_t, \tag{1.2}$$

заметим, что по закону больших чисел $\frac{S_t^T u_t}{T} \to \mathbb{E} s_t u_t = 0$. Откуда и получаем состоятельность.

Далее, для доказательства асимптотической нормальности, перенесём α из правой части (1.2) и умножим на $\sqrt{T}$. Получим, что

$$\sqrt{T}(\alpha - \hat{\alpha}) = \sqrt{T} G(S_t^T S_t)^{-1} S_t^T u_t.$$

Далее опять по закону больших чисел $\frac{1}{T} S_t^T S_t \to \mathbb{E} s_\tau s_\tau^T = Q$. По центральной предельной теореме

$$\frac{G}{\sqrt{T}} S_t^T u_t \to \mathcal{N}(0, G^2 Q)$$

Применяя теорему Слутского [27], получаем, что

$$\sqrt{T}(\alpha - \hat{\alpha}) = \mathcal{N}(0, G^2 Q^{-1}).$$

Что и требовалось доказать. Существует проблема, связанная с ситуацией, когда значения речевого сигнала будут выбираться так, чтобы они попадали в сегмент, в этом случае в сигнал будут вноситься сильные искажения.

Действительно, ведь это эквивалентно действию высокочастотного шума в начале и в конце окна. Это приводит к тому, что энергия сигнала будет размываться по прилегающим частотам (так называемая утечка спектра).

Для решения этой проблемы предлагается дифференцировано работать со значениями речевого сигнала. Для этого входной сигнал умножается на оконную функцию, в качестве которой обычно выступает функция Хемминга:

$$w(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right),$$

где $\alpha = 0.54,\ \beta = 1 - \alpha$.

1.6 Кепстральные коэффициенты

1.6.1 Строение человеческого уха

Для разработки и создания устойчивых методов анализа и представления речи были разработаны методы спектрального анализа, основанные на знании того, как функционирует человеческая акустическая система. Предполагается, что чем лучше исследователи понимают, как происходит обработка речевых сигналов в человеческой акустической системе, тем лучшую систему, воспринимающую значение и содержание речи, можно разработать.

Для понимания дальнейшего изложения необходимо рассмотреть физиологическую модель человеческого уха [17]. Оно состоит из трех различных участков: наружного, среднего и внутреннего уха. Наружное ухо

состоит из ушной раковины и слухового прохода. Звуковые волны проходят через наружное ухо к среднему уху. Среднее ухо состоит из барабанной перегородки, с которой сталкивается звуковая волна и заставляет её колебаться, и из трёх слуховых косточек (молоточка, наковальни и стремечка), которые преобразуют звуковые колебания в механические, одновременно усиливая их. Внутреннее ухо состоит из преддверия, улитки и полукружных каналов. Улитка представляет собой заполненный жидкостью канал, разделённый базальной мембраной. Механические колебания, сталкиваясь с овальным окном в преддверии, создают стоячие волны в жидкости внутри улитки, что заставляет базальную мембрану колебаться на частотах, соизмеримых с частотами входного сигнала, в местах связанных с этими частотами. Применение модели человеческого уха будет описано в главе 2.

1.6.2 Методы шкалирования полос

Использование банка фильтров обосновывается необходимостью моделирования процессов, происходящих в улитке внутреннего уха, которая ведёт себя как набор пересекающихся частотных фильтров. Полоса пропускания каждого фильтра называется критической полосой [28]. Два чистых звука называют лежащими в одной критической полосе, если они расположены так близко друг к другу, что существует значительное пересечение в их амплитудных огибающих в базальной мембране.

Для построения моделей перцепции, основанных на физиологии акустического восприятия человеческим ухом, были предложены две шкалы, позволяющие ввести некоторые характеристики, используемые для идентификации речи.

Шкала барков была предложена Эберхардом Цвикером и названа в честь Генриха Баркхаузена [28]. Эта шкала является попыткой описать

эффект возникновения этих критических полос. Расположение центральных критических полос является нелинейным, оно может быть описано с помощью психоакустической шкалы, предложенной Цвикером:

$$f_{bark} = 13 \arctan(0.00076 f) + 3.5 \arctan\left(\frac{f}{7500}\right)^2,$$

где f - частота в герцах. Ширина критической полосы может быть приближена следующей формулой:

$$b_{Hz} = \frac{52548}{f_{bark}^2 - 52.56 + 690.39}$$

Шкала мелов (название происходит от английского melody), предложенная Стивенсом [29] основана на альтернативном шкалировании частотного диапазона, с использованием нелинейных особенностей восприятия высоты тона человеческим ухом. Частота в мелах m может быть приблизительно получена из частоты в герцах f следующим образом:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

Эти шкалы могут быть применены для построения нелинейного банка фильтров. Банк фильтров в мелах может быть представлен как банк из M треугольных фильтров, усредняющих спектральную энергию вокруг каждой центральной частоты:

$$H_m[k] = \begin{cases} 0, & k < f_{m-1} \\ \frac{2(k - f_{m-1})}{(f_{m+1} - f_{m-1})(f_m - f_{m-1})}, & f_{m-1} \le k \le f_m \\ \frac{2(f_{m-1} - k)}{(f_{m+1} - f_{m-1})(f_m - f_{m-1})}, & f_m \le k \le f_{m+1} \\ 0, & k > f_{m+1} \end{cases}$$

где

$$f_m = \frac{N}{f_{\text{дискр}}} f^{-1}\left(f_{\text{низ}} + m \frac{f_{\text{низ}} - f_{\text{выс}}}{M+1}\right)$$

– функция самой низкой ($f_{\text{низ}}$) и самой высокой частот ($f_{\text{выс}}$) в банке фильтров, частоты дискретизации ($f_{\text{дискр}}$) и набора частот в линейной частотной области N. Для ширины пропускания треугольных фильтров назначаются такие значения, что середины центральных частот фильтров расположены на расстоянии в 3 дБ. Треугольные фильтры построены таким образом, чтобы количество энергии входного сигнала было равномерно распределено по каждому их них как показано на рис. 1.6).

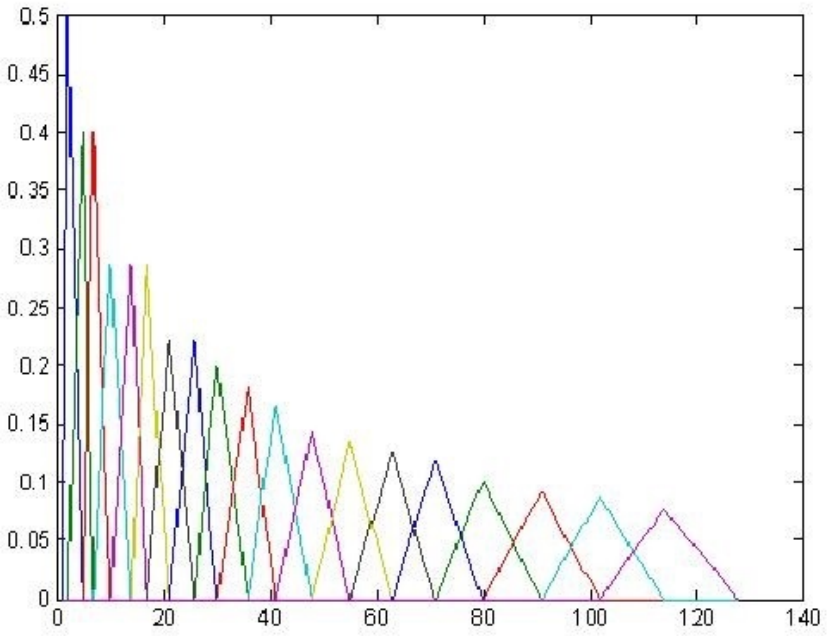


Рис. 1.6: Банк треугольных фильтров

ERB (Equivalent Rectangular Bandwidth) [30] - ещё одна шкала, используемая в психоакустике для моделирования акустических фильтров. Для данного акустического фильтра существует эквивалентный ему прямоугольный фильтр, который и называется ERB. ERB равен полосе пропускания совершенного прямоугольного фильтра, чья передаточная способность в его

полосе пропускания равна максимальной пропускной способности данного фильтра, пропускающего такую же мощность белого шума. Использование ERB объясняется тем, что она может объяснить некоторые существующие особенности восприятия, такие, например, как псевдо-логарифмический рост ширины критической полосы с ростом частоты и логарифмический закон восприятия интервалов частот [21].

К преимуществам ERB относят [31]: невосприимчивость к биениям и интермодуляциям между сигналом и маскером. ERB -шкала определяется как количество ERB лежащих ниже заданной частоты [31]:

$$ERBS(f) = 21.4 \lg(0.00437f + 1),$$

где f - частота в Гц.

1.6.3 Спектральные огибающие

Использование энергетического спектра для задач автоматического распознавания речи представляется неудобным, по причине того, что спектральные пики и впадины одинаково отображаются на спектре. Это, в свою очередь, ведёт к нежелательным воздействиям на фундаментальную частоту (то есть на первую основную (низшую) собственную частоту) и эффектам внешнего шума. Энергетический спектр подавляет фундаментальную частоту и её гармоники в речи, и, следовательно, приводит к плохим оценкам функции отклика вокального тракта. Кроме того, шум в логарифмической шкале лучше всего заметен на впадинах, а значит, моделирование впадин представляется бесполезным. С другой стороны, именно в спектральных пиках содержится наиболее важная для распознавания информация, и кроме того, спектральные пики меньше подвержены аддитивным искажениям.

Спектральная огибающая представляет собой функцию в осях энергия – частота, изображающая резонансы вокального тракта, как показано на

рис. 1.7. Спектральная огибающая обычно подвергается сглаживанию, чтобы избавиться от влияния спектральных эффектов шумов. Спектральная огибающая хорошо представляет пики в спектре и хуже – впадины.

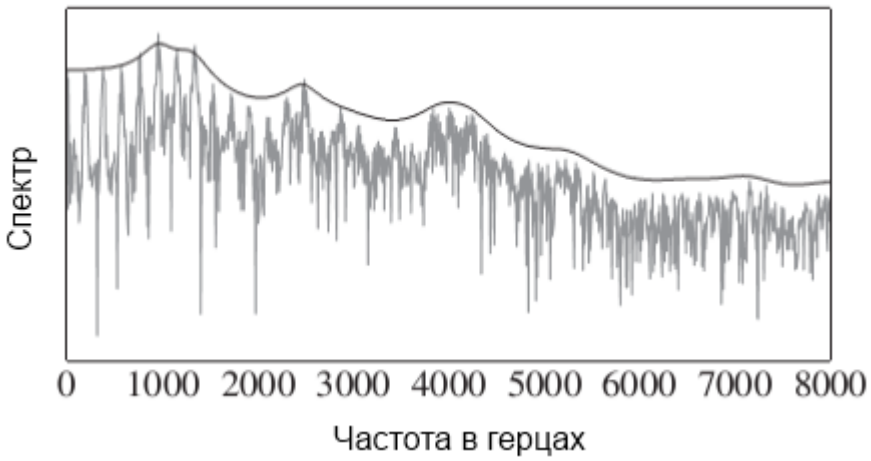


Рис. 1.7: Спектральная огибающая.

Спектральная огибающая может быть вычислена с помощью коэффициентов линейного предсказания [20]. Обычно используется эффективный метод вычисления коэффициентов линейного предсказания, основанный на автокорреляциях и использующий алгоритм Левинсона – Дарбина [32]. В задаче распознавания речи часто применяются коэффициенты линейного предсказания, вычисленные на основе степенного закона человеческого слуха [24]. Отличия от простого вычисления коэффициентов линейного предсказания заключаются в следующем. Во-первых, применяется шкала Барков и логарифмическая компрессия амплитуды до применения алгоритма Левинсона – Дарбина. Во-вторых, для соответствия степенному закону человеческого слуха мощность спектральных компонент возводится в степень 0.33. Эта модификация применяется в частотной области, что приводит к тому, что автокорреляционные коэффициенты не могут быть вычислены непосредственно, следовательно, необходимо применение дополнительного преобразования Фурье.

1.6.4 Кепстральная обработка речевого сигнала

Применение кепстрального анализа к системам распознавания речи было предложено в работе [19]. Изложим основы его использования, для чего рассмотрим стационарную последовательность x_n, чье z - преобразовние [33] обозначим через $X(z)$.

Определение. Комплексным спектром последовательности называется стационарная последовательность $\hat{x}_n$, обладающая таким свойством, что $\hat{X}(z) = \log X(z)$, где $X(z)$ z - преобразование последовательности, и логарифм понимается в комплексном смысле [34].

Для дальнейшего изложения автору необходимо доказать утверждение о свойствах комплексного спектра.

Утверждение. Вычисление комплексного спектра эквивалентно вычислению обратного преобразования Фурье от комплексного логарифма z - преобразования x_n на единичном круге в комплексной плоскости.

Доказательство. По определению обратного z-преобразования получаем

$$\hat{x}_n = \frac{1}{2\pi\jmath}\oint_C \hat{X}(z)z^{n-1}\mathrm{d}z = \frac{1}{2\pi\jmath}\oint_C \log X(z)z^{n-1}\mathrm{d}z.$$

При этом контур интегрирования должен лежать в области сходимости. Так как предполагается, что x_n – стационарна, то можно параметризовать контур интегрирования как $C = \{z = e^{j\omega} \,|\, \omega \in (-\pi, \pi)\}$, откуда:

$$\hat{x}_n = \frac{1}{2\pi j}\oint_C \log X(z)z^{n-1}\mathrm{d}z = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log X(\omega)e^{j\omega n}\mathrm{d}\omega.$$

Что и значит, что комплексный кепстр эквивалентен обратному преобразованию Фурье $\log X\left(e^{j\omega}\right)$.

Таким образом, дано объяснение применения спектра для обработки речевых сигналов, а также связь спектра и действительного кепстра. Отсюда можно определить **действительный кепстр** как

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)|e^{j\omega n}\mathrm{d}\omega.$$

1.6.5 Анализ акустических вариаций в речевых сообщениях

Во многих случаях зависимость систем автоматического распознавания языка от диктора основана на том, что используются такие признаки речевого сигнала, которые меняются от диктора к диктору. Разнообразие длин речевого тракта или его формы является основной причиной в изменчивости речевых признаков от диктора к диктору [35].

Тем не менее стоит упомянуть, что различие в длине и форме речевого тракта (см. раздел 1.3.1) не является единственной причиной того, что речевой сигнал является дикторозависимым. В качестве причин подобного различия можно отметить не только физиологические особенности диктора, но и лингвистические, такие как акцент, диалект, громкость голоса, скорость произношения и другие. Кроме того, изменения в речевом сигнале могут появляться даже у одного и того же диктора, в зависимости от его эмоций (гнев, страх) и физических кондиций (усталость, потеря дыхания из-за физической нагрузки) [36, 37].

Подобные различия могут приводить к следующим эффектам.

- Структура речевого сигнала может меняться под воздействием физиологических и эмоциональных факторов, к которым можно отнести болезнь, курение или обстановку, в которой было произнесено сообщение.
- Долговременные параметры речевого сигнала могут быть изменены диктором намеренно, например, для того, чтобы передать какие-либо эмоции или подчеркнуть информационную значимость сообщения.
- Акустическая реализация фонем может варьироваться из-за коартикуляции, замены фонем из-за акцента, или их подавления при спонтанной речи.

Проанализируем источники различий в речевом сигнале более подробно.

В работе [38] был проведен анализ разнообразия речевых сигналов между дикторами методом главных компонент. Было выяснено, что двумя главными компонентами, отвечающими за изменчивость речевого сигнала, являются пол и акцент. Кроме того, для точности распознавания является крайне важным выяснить на каком языке было произнесено речевое сообщение [39]. Распознавание языка искаженного сообщения будет рассмотрено позже. При этом необходимо отметить, что акцент приводит к сдвигу в пространстве признаков [40], при этом сдвиг является весьма значительным именно в случае, когда речевое сообщение было произнесено не носителем языка. Размер самого сдвига зависит и от родного языка диктора, и от степени владения языком.

В обычной речи (или в случае нехватки времени) диктор может нечётко произносить некоторые фонемы или даже слоги. При этом часто плохое произношение попадает на те участки речи, которые несут меньшую информацию, и наоборот, на тех участках речи, в которых имеется важная информация, произношение является очень чётким [41].

Другим важным аспектом при отображении акустического сигнала в фонемы является скорость речи. При этом может происходить как нечеткое произношение (или “проглатывание”) фонем (аналогично случаю спонтанной речи), так и изменение во временной структуре (сжатие или расширение). Скорость речи по-разному влияет на различные фонемы, например, различие в длительности гласных является более сильным, чем для согласных, при переходе от медленной речи к быстрой [42].

Эмоциональное состояние диктора при произношении речевого сообщения также может сильно влиять на спектральные характеристики речевого сообщения [43], что в свою очередь будет приводить к изменению речевых признаков, а следовательно, и точности их распознавания.

1.6.6 Способы компенсации длины речевого тракта

Наиболее популярными используемыми входными речевыми признаками являются Мел-Частотные Кепстральные Коэффициенты (рассмотрены в разделе 2.1.2).

Длина речевого тракта зависит как от пола человека, так и от других физиологических параметров, например, роста, и может изменяться от 13 см у женщин до 18 см у взрослых мужчин , поэтому происходит сдвиг частот центральных формант, который может достигать 25%. Из-за этого различия первоначально обученная модель может плохо распознавать сообщения нового диктора, то есть система становится дикторозависимой.

Рассмотрим механизм подобного поведения. Изменение в длине речевого тракта приводит к сдвигу спектра на мел-частотных осях. При этом, MFCC признаки генерируются с помощью дискретного косинус - преобразования, базисные функции которого не могут быть сдвинуты при изменении в длине речевого тракта. Действительно, подогнанный максимум косинуса к пику, соответствующему форманте при данной длине речевого тракта, не может быть сдвинут таким образом, чтобы соответствовать новому пику, образовавшемуся после сдвига в результате изменения длины речевого тракта. Таким образом, изменение в длине речевого тракта приводит к изменению величины всех кепстральных коэффициентов. Тем не менее информация о длине речевого тракта по - прежнему будет присутствовать в MFCC, что ведёт к увеличению времени обучения и размера обучающего множества.

Один из способов решения этой проблемы – применение так называемой нормализации длины речевого тракта (Voice tract length normalization, VTLN [44]), в ходе которой происходит преобразование исходного звукового сигнала, таким образом, чтобы центральные форманты находились на одной частоте. Например, можно предварительно сдвигать мел фильтры с помощью

линейного [36] или нелинейного преобразования [45] таким образом, чтобы расположение формант для конкретного диктора было бы близко к среднему их расположению. Пример подобного преобразования приведён на рис. 1.8.

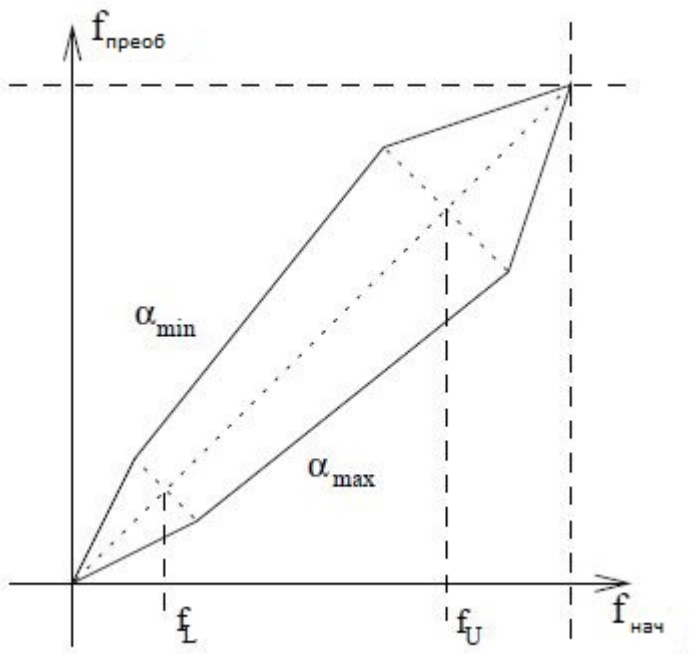


Рис. 1.8: Пример преобразования мел фильтров.

К сожалению, сложность этого подхода в том, что необходимо предварительно оценивать параметры этого преобразования для каждого конкретного диктора, что не всегда представляется возможным, в случае, когда объем речевого материала недостаточно велик. Это является следствием того, что зависимость длины речевого тракта и конкретного значения MFCC очень сложна, и изменение в длине речевого тракта приводит к изменению всех коэффициентов.

1.7 Выводы

В данной главе были описаны физические аспекты звука, а также характеристики звукового сигнала. В главе получены следующие результаты:

- Проведён анализ физических свойств звукового сигнала.
- Рассмотрены общие принципы генерации и восприятия звукового сигнала.
- Проанализирована модель генерации речевого сигнала в виде линейной динамической системы.
- Рассмотрены подходы к описанию речевого сигнала, такие как спектральный, кепстральный и прозодический.
- Приведены парадигмы распознавания речи.
- Проанализированы источники различий в звуковом сигнале, приводящие к зависимости системы распознавания речи от диктора.

Глава 2

Математические методы и алгоритмы, используемые для распознавания речи и диктора

В данной главе рассмотрены различные математические модели, использующиеся для построения систем распознавания речи (таких как Скрытые Марковские Модели), особое внимание уделяется методам, применяемым для разработки системы распознавания речи, точность идентификации которых не зависит от диктора. Для реализации этого рассматривается возможность предварительной идентификации диктора, а также способ построения дикторонезависимых признаков для описания речевого сигнала, опирающийся на психоакустическую модель восприятия человеком речевого сообщения. Кроме того, исследуются практические и теоретические аспекты построения указанных признаков, а также предлагается способ их вычисления на основе параллельных алгоритмов.

2.1 Скрытые Марковские Модели

Процессы, протекающие в реальной жизни, обычно характеризуются наблюдениями, которые можно рассматривать как сигналы. Эти сигналы

могут быть как дискретными (например, символы какого-либо алфавита), так и непрерывными (музыка, температура, речь). Сигналы могут быть стационарными (то есть, их статистические свойства не меняются во времени) или нестационарными. Сигналы могут быть чистыми (например, приходящими строго от одного источника) или могут быть испорчены каким-либо иным источником сигнала (шумом) или искажениями при передаче, реверберациями и т.д.

Модель прохождения сигналов может лежать в основе теоретического описания системы. Здесь основная проблема – построить описание сигнала в терминах модели его прохождения. Существует несколько причин, из-за которых применение таких моделей представляется удобным:

1. Модель прохождения сигнала может быть использована для обработки сигнала с целью получения желаемого результата. Например, если пользователи заинтересованы в улучшении качества речевого сигнала, который был испорчен шумом и/или искажениями при передаче. В этом случае можно использовать модель прохождения сигналов для создания системы, которая уменьшит шум и искажения оптимальным образом.

2. Модель прохождения сигналов позволяет определить характеристики источника сигнала при отсутствии самого источника. Это свойство особенно важно, когда получение сигнала непосредственно из источника очень дорого, например, сопровождено с большими затратами денег или требует большого количества времени. В этом случае представлятся возможным построить модель и с помощью симуляций выяснить свойства источника.

3. Модели прохождения сигналов хорошо работают на практике, а следовательно, позволяют эффективно создавать важные с практической

точки зрения предсказательные, распознающие и идентифицирующие системы.

Существует несколько способов выбора типа модели прохождения сигналов для описания характеристик данного сигнала. Выделяют два основных типа моделей: детерминистические и стохастические.

Детерминистические модели обычно используют некоторые известные свойства сигнала, например, представление сигнала синусоидальной волной или суммой экспонент и т.д. В этом случае спецификация модели достаточно проста: необходимо лишь оценить параметры сигнала – амплитуду, частоту, фазу и т.д.

В стохастических моделях как правило используются только статистические свойства сигнала. Примерами подобных моделей могут служить Гауссовы процессы, Пуассоновские процессы, Марковские процессы (в том числе и скрытые). В основе стохастических моделей лежит предположение о том, что сигнал может быть хорошо описан как параметрический случайный процесс и что его параметры могут быть оценены достаточно точно.

Скрытая Марковская Модель (HMM – Hidden Markov model) определяется как двойной случайный процесс. Лежащий в основе случайный процесс представляет собой однородную Марковскую цепь с конечным числом состояний. Последовательность состояний не наблюдается и поэтому называется скрытой. Эта цепочка состояний влияет на другой случайный процесс, который и производит последовательность наблюдений. Скрытые Марковский модели представляют собой важный класс моделей, которые успешно используются во многих отраслях знаний, например, при моделировании речи. Базовая теория по Скрытым Марковским Моделям будет дана ниже.

Можно выделить следующие преимущества использования скрытых Марковских моделей при использовании в задаче распознавания речи:

- HMM обладают простой математический структурой.
- Структура HMM позволяет моделировать сложную цепочки наблюдений.
- Параметры модели могут быть автоматически выбраны таким образом, чтобы описать имеющийся набор данных для обучения.

В системах распознавания речи Скрытые Марковские Модели обычно применяются для представления фонем или целых слов. Каждое скрытое состояние представляет часть фонемы или слова. В каждый момент времени состояние, в котором находится система, может быть изменено в соответствии с набором переходных вероятностей, связанных с данным состоянием. Схематично это представлено на рисунке 2.1.

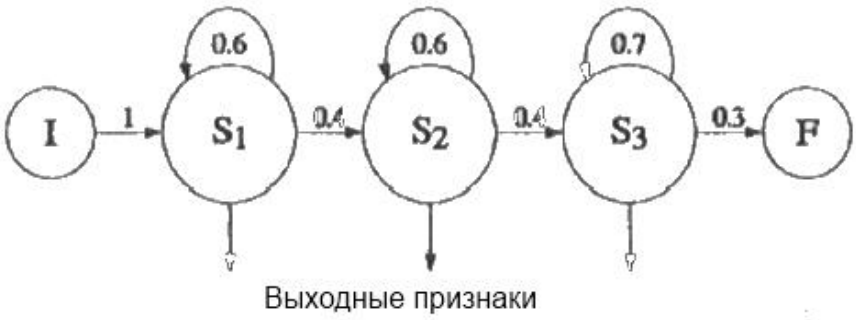


Рис. 2.1: Скрытая Марковская Модель с 5 состояниями. Символами I и F обозначены начальное и конечное состояния соответственно, $\{S_i\}_{i=1}^{3}$ - генерирующие состояния, дугами обозначены возможные переходы между состояниями, цифры над дугами обознают вероятности переходов между соответствующими состояниями.

Для Марковской модели первого порядка переходные вероятности зависят только от предыдущего состояния и не зависят от состояний в более ранние моменты времени.

Следует отметить, что существуют расширения HMM, в которых время нахождения системы в данном состоянии может моделироваться любым распределением. Такие модели носят название Скрытые Полумарковские Модели [46].

Когда состояние активно, оно может генерировать последовательность векторов признаков, один вектор признаков в каждый момент времени. Эти вектора признаков имеют ту же форму, что и вектора признаков, которые подучаются, когда распознаётся сказанное слово. Однако невозможно узнать точно последовательность состояний, пройденных системой для генерации данного набора наблюдаемых векторов признаков, так как каждое состояние дополнительно к переходным вероятностям определяется и плотностью распределения вероятности генерации векторов признаков. Она может быть использована для вычисления вероятности того, что вектор признаков был сгенерирован в данном состоянии. В качестве плотностей распределений обычно используются смеси гауссовых плотностей, каждая со своим средним, дисперсией и весом, например, как показано на рис. 2.2.

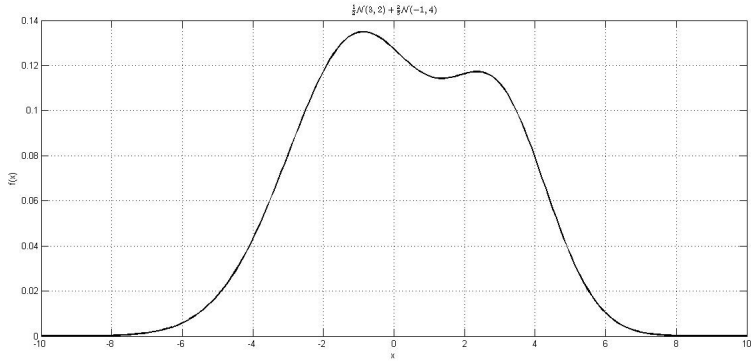


Рис. 2.2: Плотность смеси гауссовских распределений $\frac{3}{7}\mathcal{N}(3,2)+\frac{4}{7}\mathcal{N}(-1,4)$

Под обучением HMM понимают определение оценок параметров модели: переходных вероятностей, параметров плотности распределения и их весов. Эти параметры оптимизируются в соответствии с алгоритмом Баума-Уэлша [47]. Стоит отметить, что обычно для обучения требуется большой набор данных, при этом размер обучающего множества зависит от объёма словаря и параметров дикторов.

Распознавание производится посредством нахождения такой последовательности состояний, которая с наибольшей вероятностью сгенерировала последовательность векторов признаков. Такая последовательность находится с помощью алгоритма Витерби [48]. Зная последовательность состояний, можно просто определить соответствующую модельную последовательность – последовательность фонем или слов.

В связи с её использованием в работе, рассмотрим HMM более подробно.

2.1.1 Математическое описание Скрытых Марковских Моделей

Пусть имеется марковская цепь в дискретном времени с набором состояний $S = 1, \ldots, M$. Через регулярные промежутки времени в системе происходит переход из одного состояния в другое (возможно, назад в предыдущее состояние). Последовательность состояний обозначим через $S_{1:T} = S_1, \ldots, S_M$, где $S_t \in S$ - состояние в момент времени t. Реализацию $S_{1:T}$ обозначим s_{1T}. Полное вероятностное описание системы требует задания текущего состояния в момент времени t и всех предшествующих состояний.

В частном случае дискретной Марковской цепи первого порядка описание выглядит следующим образом:

$$P(q_t = S_j \,|\, q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \,|\, q_{t-1} = S_i).$$

В дальнейшем предполагается, что вероятности перехода не зависят от времени. Обозначим $a_{ij} = P(q_t = S_j \,|\, q_{t-1} = S_i), 1 \leq i, j \leq M$. При этом, $a_{ij} \geq 0, \sum_{j=1}^{M} a_{ij} = 1$.

Указанный случайный процесс может быть назван наблюдаемой Марковской моделью, так как выходные значения процесса в каждый момент времени представляют собой состояния процесса. В случае если состояния процесса в каждый момент времени не наблюдаемы, то модель носит название Скрытой Марковской.

Определение. Случайный процесс (Скрытая Марковская Модель), используемый в работе, задается следующими компонентами:

1. Количеством скрытых состояний N. Множество состояний модели обозначается $S = \{S_1, \ldots, S_N\}$. Состояния соединенны таким образом, что любое состояние S_i может быть достигнуто из любого другого состояния S_j за конечное число шагов (эргодическая модель).

2. Размером выходного алфавита M. Набор символов выходного алфавита обозначается через $V = \{v_1, \ldots, v_M\}$. Речевыми символами являются вектора из $\mathbb{R}^n$.

3. Матрицей переходных вероятностей $A = (a_{ij})$, где

$$a_{ij} = P(q_t = S_j \,|\, q_{t-1} = S_i), \qquad i, j = 1, \ldots, M$$

4. Распределением вероятности выходных символов $B = \{b_j(k) : j = 1, \ldots, N,\ k = 1, \ldots, M\}$ для данного состояния j, где k -порядковый номер символа v_k, а $b_j(k) = P(v \in V \,|\, q_t = S_j), \quad j = 1, \ldots, N,\ k = 1, \ldots, M$, то есть, $b_j(k)$ - вероятность того, что в момент времени t система, находясь в состоянии S_j, выдаст символ v_k.

5. Вероятностью нахождения в состоянии i в начальный момент времени π_i, формирующие начальное распределение Π.

Набор компонент A, B, Π, задающих марковскую модель, обозначается $\lambda = \{A, B, \Pi\}$. Последовательность наблюдений, сгенерированных марковской моделью за время T, обозначают $O = O_1, O_2, \ldots, O_T$.

Справедлива следующая теорема.

Теорема. Пусть Скрытая Марковсая Модель задаётся набором компонент $\lambda = \{A, B, \Pi\}$. Тогда для любого состояния S_k $P(q_{t+1} = S_k, \ldots, q_{t+T-1} = S_k, q_{t+T} \neq S_k \,|\, q_t = S_k) = a_{kk}^T(1 - a_{kk})$, то есть, время нахождения цепи в состоянии S_k распределено экспоненциально.

Доказательство. Обозначим через $\{S_i\}_{i=1}^M$ - множество состояний Марковской Модели, вероятности перехода $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$, $1 \leq i, j \leq M$. Тогда вероятность того, что Марковская Модель будет находиться в состоянии k T периодов времени, при условии, что она уже находится в этом состоянии, записывается как

$$P(q_{t+1} = S_k, \ldots, q_{t+T-1} = S_k | q_t = S_k). \tag{2.1}$$

Применяя формулу произведения вероятностей к 2.1 и пользуясь марковским свойством получаем требуемое:

$$\begin{aligned}
&P(q_{t+1} = S_k, \ldots, q_{t+T-1} = S_k, q_{t+T} \neq S_k \,|\, q_t = S_k) = \\
&= P(q_{t+T-1} = S_k, q_{t+T} \neq S_k \,|\, q_{t+T-2} = S_k, \ldots, q_t = S_k) \cdot \\
&\cdot P(q_{t+T-2} = S_k \,|\, q_{t+T-3} = S_k, \ldots, q_t = S_k) \cdot \ldots \cdot P(q_{t+1} = S_k \,|\, q_t = S_k) = \\
&= P(q_{t+T-1} = S_k, q_{t+T} \neq S_k \,|\, q_{t+T-2} = S_k) \cdot \ldots \cdot P(q_{t+1} = S_k \,|\, q_t = S_k) \\
&= a_{kk}^T(1 - a_{kk}).
\end{aligned}$$

Что и требовалось доказать.

2.1.2 Основный задачи, решаемые с помощью Скрытых Марковских Моделей

Существуют три основные задачи, которые представляют интерес при решении практических задач.

1. Как при заданной последовательности символов наблюдений $O = O_1, O_2, \ldots, O_T$ и модели $\lambda = \{A, B, \Pi\}$ вычислить вероятность наблюдения данной последовательности $P(O|\lambda)$ при условии, что она была сгенерирована моделью λ? Можно рассматривать эту проблему с точки зрения того, насколько хорошо данная модель соотносится с наблюдаемой последовательностью наблюдений: при наличии нескольких моделей, решение этой задачи позволяет выбрать модель, которая лучше соответствует данным.

2. Как при заданной последовательности символов наблюдений $O = O_1, O_2 \ldots, O_T$ и модели $\lambda = \{A, B, \Pi\}$ вычислить соответствующую последовательность состояний $Q = q_1, q_2, \ldots, q_T$, оптимальную в некотором смысле? Очевидно, что кроме вырожденных случаев не существует единственно «правильной» последовательности состояний, поэтому следует использовать критерий оптимальности для выбора последовательности состояний.

3. Как вычислить оптимальные с точки зрения максимизации $P(O \mid \lambda)$ параметры $\lambda = \{A, B, \Pi\}$?

На практике широко используется следующее определение.

Определение. Последовательность наблюдений, используемая для оптимизации параметров HMM, называется обучающим множеством [49].

Решение первой задачи позволит выбрать лучшую модель для объяснения имеющихся данных.

2.1.3 Алгоритмы решения основных задач, связанных с HMM

Решением первой задачи является метод, основанный на так называемом алгоритме прямого и обратного хода [50]. Опишем суть этого алгоритма.

Определение. Переменными прямого хода называются вероятность наблюдения частичной последовательности $O = O_1, O_2, \ldots, O_t$ и состояния S_i в момент времени t при заданной модели λ:

$$\alpha_t(i) = P(O = O_1, O_2, \ldots, O_t, q_t = S_i \mid \lambda).$$

Утверждение. Вероятность $P(O|\lambda)$ наблюдения последовательности $O = O_1, O_2, \ldots, O_T$ при условии, что она была сгенерирована моделью λ вычислятся через переменные прямого хода [50] как:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Доказательство. Алгоритм нахождения переменных прямого хода состоит из трёх последовательных шагов.

Шаг 1. Инициализация:

$$\alpha_1(i) = \pi b_1(O_1), \qquad 1 \leq i \leq N.$$

Шаг 2. Индукция:

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}.$$

Интерпретация этой формулы достаточно проста. Состояние S_j в момент времени $t+1$ может быть достигнуто из N возможных состояний s_i, $1 \leq i \leq N$, в которых система могла находиться в момент t. Из определения $\alpha_t(i)$ следует, что произведение $\alpha_t(i)a_{ij}$ есть совместная вероятность того, что наблюдалась последовательность $O = O_1, O_2, \ldots, O_t$ и состояние S_j было достигнуто в момент времени $t+1$ из состояния S_i. Суммируя эти

вероятности по всем возможным состояниям, получаем вероятность того, что система находиться в состоянии S_j и наблюдалась последовательность $O = O_1, O_2, \ldots, O_t$. Осталось принять во внимание, что в момент времени $t+1$ будет наблюдаться O_{t+1} в состоянии S_j. Для этого необходимо умножить предыдущее на $b_j(O_{t+1})$.

Шаг 3. Терминация:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

По определению $\alpha_T(i) = P(O = O_1, O_2, \ldots, O_T, q_T = S_i \mid \lambda)$, следовательно, для вычисления $P(O|\lambda)$ нужно лишь просуммировать все $\alpha_T(i)$.

Аналогично можно определить переменные обратного хода.

Определение. Переменной обратного хода называется совместная вероятность наблюдения последовательности, начиная с момента $t+1$ до конца, при заданном в момент t состоянии S_i и модели λ:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T \mid q_t = S_i, \lambda).$$

Утверждение. Переменные обратного хода выражаются рекурсивно по формуле:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \ldots, 1, \ 1 \le i \le N$$

Доказательство. Алгоритм нахождения переменных обратного хода состоит из двух последовательных шагов.

Шаг 1. Инициализация:

$$\beta_t(i) = 1, \qquad 1 \le i \le N.$$

Значения для $\beta_t(i)$ выбираются произвольно.

Шаг 2. Индукция:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \ldots, 1, \ 1 \leq i \leq N$$

Для того, чтобы в момент времени t находиться в состоянии S_i и учитывая всю последовательность наблюдений, начиная с момента времени $t+1$, необходимо рассмотреть все состояния S_j в момент $t+1$ и вероятности переходов в эти состояния, вероятность наблюдения O_{t+1} в момент $t+1$ и оставшуюся часть наблюдений из состояния j.

В отличие от решения первой задачи, при нахождении оптимальной последовательности символов необходимо уточнить критерий оптимальности. В качестве возможного критерия может выступать количество индивидуально наиболее вероятных состояний. Такой критерий обладает следующим недостатком. Если некоторая переходная вероятность $a_{ij} = 0$, то найденная оптимальная последовательность состояний может не быть допустимой. Эта проблема возникает потому, что алгоритм определяет наиболее вероятное состояние в данный момент времени и не учитывает вероятности появления последовательностей символов.

Наиболее часто встречаемый критерий заключается в том, чтобы найти единственную лучшую последовательность наблюдений, то есть максимизации $P(q_1, \ldots, q_T \,|\, O_1, \ldots, O_T, \lambda)$, что в силу теоремы Байеса эквивалентно максимизации $P(q_1, \ldots, q_T O_1, \ldots, O_T \,|\, \lambda)$. Алгоритм, решающий указанную задачу, называется алгоритмом Витерби [48].

Для нахождения лучшей последовательности состояний $Q = q_1, \ldots, q_T$ определим величину

$$\delta_t(i) = \max_{q_1, \ldots, q_{t-1}} P(q_1, \ldots, q_t = i, O_1, \ldots O_t \,|\, q_t = S_i, \lambda).$$

Тогда для нахождения $\delta_{t+1}(j)$ нужно взять максимальную (то есть, наиболее вероятную) $\delta_t(i)$ с предыдущего шага и умножить на вероятность наблюдения

символа O_{t+1} в состоянии S_j:

$$\delta_{t+1}(j) = b_j(O_{t+1}) \max_i \delta_t(i) a_{ij}.$$

Чтобы определить искомую последовательность символов, необходимо сохранять $\psi_t(i) = \arg\max \delta_t(i)$ для каждого i.

Теперь полная процедура нахождения оптимальной последовательности состояний (алгоритм Витерби) может быть записан следующим образом.

Шаг 1. Инициализация:

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(O_1), \qquad 1 \le i \le N, \\ \psi_1(i) &= 0. \end{aligned}$$

Шаг 2. Рекурсия:

$$\begin{aligned} \delta_t(j) &= b_j(O_{t+1}) \max_{1 \le i \le N} \delta_t(i) a_{ij}, \qquad 1 \le i \le N, \\ \psi_t(i) &= \arg\max_{1 \le i \le N} \delta_t(i) a_{ij}, \qquad 2 \le t \le T \end{aligned}$$

Шаг 3. Терминация:

$$\begin{aligned} \hat{P} &= \max_{1 \le i \le N} \delta_T(j), \\ \hat{q}_T &= \arg\max_{1 \le i \le N} \delta_T(j). \end{aligned}$$

Шаг 4. Определение последовательности состояний:

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}), \qquad t = T-1, \ldots, 1.$$

Стоит отметить, что алгоритм Витерби очень похож на вычисление переменных прямого хода за исключением того, что вместо суммирования по всем предыдущим состояниям, происходит максимизация.

Решение третьей задачи представляется самым сложным, так как нет аналитического решения максимизационной задачи по нахождению

оптимальных параметров модели. Представляется возможным найти такие параметры модели $\lambda = \{A, B, \Pi\}$, которые дают локальный максимум $P(O \mid \lambda)$. Поиск локального максимума может быть осуществлен с помощью итеративного алгоритма Баума-Велша.

Обозначим вероятность нахождения в состоянии S_i в момент времени t и в состоянии S_j в момент $t+1$ при данной модели и последовательности наблюдений через $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda)$.

Из определения переменных прямого и обратного хода следует, что

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}.$$

Обозначим через $\gamma_t(i)$ вероятность нахождения в состоянии S_i в момент времени t при заданной последовательности наблюдений и модели. Тогда $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$. Более того, используя $\gamma_t(i)$ можно подсчитать количество переходов из состояния S_i как $\sum_{t=1}^{T-1} \gamma_t(i)$. Кроме того, $\sum_{t=1}^{T-1} \xi_t(i,j)$ – ожидаемое количество переходов из состояния S_i в состояние S_j.

Запишем формулы, которые необходимо будет использовать при переоценке параметров модели A, B, Π:

$$\begin{aligned} \hat{\pi}_j &= \gamma_1(j), \\ \hat{a}_{ij} &= \frac{\sum_{j=1}^{N} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \\ \hat{b}_k &= \frac{\sum_{j=1, o_t = v_k}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}. \end{aligned}$$

В работе Баума [47] было показано, что либо параметры начальной модели λ являются критическими для функции правдоподобия, либо существует другая модель с параметрами $\hat{A}, \hat{B}, \hat{\Pi}$, которая является более вероятной в том смысле, что $P(O \mid \hat{\lambda}) > P(O \mid \lambda)$. Таким образом, если итеративно использовать модель $\hat{\lambda}$ вместо λ и повторять переоценку параметров, то на каждом шаге увеличивается вероятность того, что данная последовательность наблюдений

была получена из текущей модели. Будем повторять эту процедуру, пока не достигнем какой-либо точки останова. Финальный результат процедуры переоценивания называется оценкой максимального правдоподобия.

2.2 Методы распознавания диктора

Для организации акустического интерфейса и идентификации биометрических признаков оператора могут использоваться дикторозависимые признаки. При этом должно иметься большое количество речевого материала и количество дикторов (не очень большое) определено априори. Системы такого типа обычно применяются для распознавания речи в правоохранительных органах или военных целях. Например, при слежке за подозреваемым или получении разведывательной информации от конкретного лица. В последнее время такие системы становятся всё более актуальны в мобильных приложениях при распознавании речи обладателя устройства.

Таким образом, представляется крайне важным создание системы, способной производить предварительную идентификацию диктора для последующей подстройки её к конкретному лицу.

Здесь задача идентификации диктора по звуковому сообщению (см. рис. 1.2 блок с классификатором) является частным случаем задачи распознавания образов, для решения которой требуется построить статистический критерий принадлежности нового звукового сообщения к одному из классов, задаваемых «обучающими» сообщениями.

Задача идентификации решалась в следующей постановке. Пусть $\mathbb{X}$ - пространство объектов, $\mathbb{Y}$ - множество ответов, $f : \mathbb{X} \to \mathbb{Y}$ - целевая зависимость. Пусть $\mathbb{X}^t \in \mathbb{X} \times \mathbb{Y}$ - обучающее множество, то есть множество пар (X_i, y_i), где $y_i = f(X_i)$. По известному обучающему множеству требуется построить $\hat{f} : \mathbb{X} \to \mathbb{Y}$ аппроксимирующую f на всем $\mathbb{X}$.

В настоящее время существует два подхода к идентификации диктора: закрытый и открытый [49]. В первом случае классификации предполагается, что новое сообщение принадлежит одному из рассматриваемых дикторов, во втором – сообщение может относиться и к неизвестному диктору. Задача идентификации рассматривается как построение статистического критерия разделения полученных точек для конечного числа простых гипотез в случае закрытой задачи или для конечного числа простых и одной сложной гипотезы (сообщение неизвестного диктора) в случае открытой задачи.

2.2.1 Метод распознавания диктора, основанный на SVM

В качестве метода идентификации используется алгоритм, основанный на «машине опорных векторов» (Support Vector Machines, SVM) [51] – являющийся базовым инструментом для распознавания образов на основе статистической теории обучения. SVM широко используется в естественно-языковых приложениях при обработке речевых сигналов. Например, задачи распознавания языка искаженного текста и идентификация диктора. Одним из ключевых аспектов применения SVM к распознаванию речи является нахождения способа работы с предложениями переменной длины [52]. Возможный способ решения этой проблемы – применение Фишеровских ядер [52]. Вообще применение ядерных методов является стандартной техникой в случае, когда представляется невозможным решить задачу с помощью линейных методов. В качестве ядер могут быть использованы гауссианы, полиномы, сигмоидные функции и т.п. Фишеровское ядро задает скалярное произведение с помощью функции потерь из задачи максимального правдоподобия.

Идея SVM основана на следующих предпосылках. Предположим, что существуют два класса объектов в некотором n-мерном пространстве, которые

можно разделить гиперплоскостью так, что с одной стороны от гиперплоскости должны находиться вектора одного класса, с другой – второго. Очевидно, что такая гиперплоскость может быть не единственной. Построим две такие параллельные гиперплоскости. Для лучшего разделения классов требуется, чтобы расстояние между плоскостями было как можно больше. Обычно для нахождения параллельных разделяющих гиперплоскостей с максимальным расстоянием в методе опорных векторов минимизируется квадратичная функция с линейными ограничениями. Решение такой задачи выражается через координаты обучающих векторов, лежащих на краю разделяющей полосы – так называемых опорных векторов. В случае, когда классы линейно неразделимы в исходном пространстве, строится отображение (необязательно линейное) в пространство большей размерности, образы классов в котором линейно разделимы. Это пространство называется пространством вторичных признаков.

2.2.2 Базовая модель SVM

Алгоритм SVM обладает следующими преимуществами:

1. В силу решения задачи минимизации выпуклой функции алгоритм гарантирует получение единственного решения. Это является серьёзным преимущество перед нейронными сетями, в которых решением может быть локальный минимум или ответ может быть неопределенным.

2. В связи с тем, что алгоритм робастен к зашумленности исходного сигнала, он хорошо приспособлен для распознания речи.

3. Алгоритм позволяет работать с данными очень больших размерностей, что важно при распознавании речи, где размерность вектора признаков может достигать многих сотен или тысяч.

В задаче идентификации диктора кепстральные коэффициенты обычно используются как опорные вектора. Далее предполагается их модификация при помощи Фишеровского ядра. Для формализации задачи обучения SVM, обозначим вектора признаков как $\{X_n\}_{n=1}^N$, а линейную функцию $(\mathcal{W}, X)+b = 0$, где $(\cdot, \cdot)$ - скалярное произведение в $\mathbb{R}^k$. Обозначим разделяемые классы через A и B и введем метки классов:

$$y_i = \begin{cases} 1, \ X_i \in A, \\ -1, \ X_i \in B. \end{cases}$$

Будем искать $\hat{f}$ в виде $\hat{f}(X) = sign(\mathcal{W}^T X + b)$, используя метод опорных векторов, разработанный В. Н. Вапником [51].

Утверждение. Максимизация ширины разделяющей полосы эквивалентна минимизации нормы $\mathcal{W}$.

Доказательство. По построению алгоритма необходимо найти такую гиперплоскость $(\mathcal{W}, X) + b = 0$, которая на векторах из класса A принимает значение 1, а на векторах из класса B - значение -1, и расстояние между гиперплоскостями, полученными параллельным переносом искомой, было максимальным. Первое условие эквивалентно

$$(\mathcal{W}, X) + b = 1, X \in A, \tag{2.2}$$

$$(\mathcal{W}, X) + b = -1, X \in B. \tag{2.3}$$

Выразим ширину разделяющей полосы через $\mathcal{W}$. Заметим, что расстояние от начала координат до гиперплоскости $(\mathcal{W}, X) + b = 0$ вычисляется как $\frac{b}{||\mathcal{W}||}$, тогда расстояние между 2.2 и 2.3 есть $\frac{2}{||\mathcal{W}||}$. Отсюда следует, что минимизация нормы $||\mathcal{W}||$ и максимизация ширины разделяющей полосы эквивалентны.

Таким образом задача обучения SVM имеет следующий вид:

$$\frac{1}{2}(\mathcal{W}, \mathcal{W}) \to \min_{\mathcal{W}, b} \tag{2.4}$$

$$y_i((\mathcal{W}, X_i) + b) \geq 1, \qquad i = 1, \ldots, N$$

Здесь нужно минимизировать квадратичную функцию $\frac{1}{2}(\mathcal{W},\mathcal{W})$ в выпуклом многограннике в пространстве пар $(\mathcal{W},b)$, заданном линейными неравенствами. Прямая $\mathcal{W}=0$ с этим многогранником не пересекается, так что минимум будет строго положителен.

Необходимо отметить, что для решения поставленной задачи ищется минимум выпуклой вниз функции f на выпуклом множестве, а значит, справедлива следующая теорема [53].

Теорема Куна-Такера. Пусть вогнутые функции $f, h_i, \quad i = 1,\dots,l$, определенные на открытом множестве D, непрерывно дифференцируемы.

Определим $U = D \cap \{x \mid h_i \geq 0, i = 1,\dots,l\}$.

Пусть выполняется условие Слейтера: $\exists \bar{x} \in U : h_i(\bar{x}) \geq 0,\ i = 1,\dots,l$.

Тогда

$$x' = \arg\max_D f \iff$$

1. $\exists \lambda \in \mathbb{R}^l : Df(x') + \sum_{i=1}^{l} \lambda_i Dh_i(x') = 0,$
2. $\begin{cases} \lambda_i \geq 0, \\ \lambda_i h_i = 0. \end{cases}$

Утверждение. Решение задачи (2.4) выражается через вектора, для которых $y_i(((\mathcal{W},X_i)+b)-1)=0$, то есть, лежащих на разделяющей полосе [51].

Доказательство. Применим теорему Куна - Такера к 2.4. Лагранжиан записывается следующим образом:

$$L(\mathcal{W},b,\Lambda) = \frac{1}{2}(\mathcal{W},\mathcal{W}) - \sum_{i=1}^{N} \lambda_i (y_i((\mathcal{W},X_i)+b)-1).$$

Множество $U = D \cap \{X \mid h_i \geq 0,\ i = 1,\dots,N\}$, и ограничения на множители Лагранжа принимают вид:

$$\begin{cases} \lambda_i \geq 0, \\ \lambda_i (y_i((\mathcal{W},X_i)+b)-1) = 0,\ i = 1,\dots,N. \end{cases}$$

Тогда:

$$\frac{\partial L}{\partial \mathcal{W}} = \mathcal{W} - \sum_{i=1}^{N} \lambda_i y_i X_i = 0, \tag{2.5a}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \lambda_i y_i = 0. \tag{2.5b}$$

Из уравнения (2.5a) следует, что $\mathcal{W} = \sum_{i=1}^{N} \lambda_i y_i X_i$, а из второго уравнения следует, что в эту сумму с ненулевыми коэффициентами λ_i входят только вектора, лежащие на разделяющих гиперплоскостях: для них $y_i(((\mathcal{W}, X_i)+b)-1) = 0$. Такие вектора принято называть опорными.

Для опорных векторов $b = y_i - (\mathcal{W}, X_i)$, так как $y_i \in \{1, -1\}$.

Теперь определим λ_i. Для этого подставим (2.5a) и (2.5b) в лагранжиан. После упрощений лагранжиан принимает вид:

$$L(\mathcal{W}, b, \Lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j (X_i, X_j).$$

Для нахождения λ_i нужно найти критические точки лагранжиана при ограничениях:

$$\begin{aligned}
&\frac{\partial L(\mathcal{W}, b, \Lambda)}{\partial \lambda_i} = y_i((\mathcal{W}, X_i) + b) - 1 \geq 0, \\
&\frac{\partial L(\mathcal{W}, b, \Lambda)}{\partial \lambda_i} \lambda_i = 0, \\
&\sum_{i=1}^{N} \lambda_i y_i = 0, \\
&\lambda_i \geq 0, \qquad i = 1, \ldots, N.
\end{aligned}$$

А это в свою очередь эквивалентно максимизации

$$\begin{aligned}
&\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j (X_i, X_j) \to \max_{\Lambda} \\
&\sum_{i=1}^{N} \lambda_i y_i = 0, \\
&\lambda_i \geq 0, \qquad i = 1, \ldots, N,
\end{aligned}$$

то есть отрицательно определенной квадратичной функции от $\Lambda = \lambda_{ii}^N$ в пересечении положительного октанта с гиперплоскостью.

Решение задачи (2.4) называется обучением классификатора.

2.2.3 Метод SVM с ядрами

В общем случае линейное разделение векторов может быть невозможно. Для решения этой задачи можно преобразовать имеющиеся пространство таким образом, чтобы вектора классов после преобразования стали линейно разделимыми. Рассмотрим теперь, как будет ставиться и решаться задача в нелинейном случае.

Пусть ϕ произвольное отображение пространства признаков в гильбертово пространство H. От отображения требуется, чтобы образы обучающих векторов были линейно разделимы в H (оно называется пространством вторичных признаков). Тогда с подстановкой $\phi(X_i)$ вместо X_i получается классифицирующий алгоритм SVM [51]. Для его настройки и применения нужно знать не само отображение ϕ, а только функцию $K : X \times X \to \mathbb{R}$, вычисляющую скалярное произведение в H образов пары векторов признаков $K(X_i, X_j) = (\phi(X_i), \phi(X_j))$.

Такая функция K называется ядром, поскольку при наличии меры, в частности при $H = \mathbb{R}$, она является ядром интегрального оператора

$$f \to K(F) = \int_F K(\cdot, Y) f(Y) dY.$$

Наиболее часто используются следующие ядра:

1. Линейное:

$$K(X, Y) = a(X, Y) + c \tag{2.6}$$

2. Полиномиальное:

$$K(X, Y) = ((X, Y) + 1)^d + c \tag{2.7}$$

3. Гауссово:

$$K(X,Y) = e^{-\gamma||X-Y||} \tag{2.8}$$

Этих ядер обычно бывает достаточно для разделения любого набора векторов. Действительно, полиномиальное ядро переводит вектор X в набор всех мономов (то есть одночленов) степени не большей N от координат X, то есть сводит разделимость к полиномиальной и гарантирует разделение не менее чем $N+1$ вектора.

Если же ϕ задано гауссовым ядром, то для любого конечного набора векторов $X_1, \ldots, X_N$ функции $\phi(X_1), \ldots, \phi(X_N)$ линейно независимы, что и обеспечивает линейную разделимость.

После преобразования построение оптимальной разделяющей полосы производится таким же способом, что и в случае (2.4), за исключением того, что все X_i заменяются на $\phi(X_i)$, а скалярные произведения (X_i, X_j) - на $K(X_i, X_j)$.

2.2.4 Метод SVM со штрафами

После обучения может оказаться так, что полученный классификатор не способен к обобщению. То есть, он очень хорошо классифицирует обучающие вектора, но на произвольном тестовом наборе он показывает плохие результаты. Такой классификатор называется неспособным к обобщению (результатов обучения) [51]. Другое название – переобучение (overfitting). Переобучение происходит потому, что классификатор настраивается на шумы и помехи в данных.

Решение этой проблемы может быть следующим.

Вместо системы запретов вводится система штрафов за нарушение. В этом случае обучения классификатора сводится к решению следующей задачи:

$$\begin{aligned} \frac{1}{2}(\mathcal{W},\mathcal{W}) + C\sum_{i=1}^{N} p(e_i) &\rightarrow \min_{\mathcal{W},b} \\ y_i((\mathcal{W},X_i)+b) &\geq 1-e_i, \\ e_i &\geq 0, \\ i &= 1,\ldots,N \end{aligned} \tag{2.9}$$

где $p(e)$ - неотрицательная, монотонно неубывающая функция, такая, что $p(0) = 0$, а $C > 0$ - эмпирически подобранный коэффициент.

Идеальный штраф $p(e) = \theta(e-1)$, где $\theta(t)$ - функция Хевисайда, при котором $\sum_{i=1}^{N} p(e_i)$ представляет собой количество неправильно классифицированных векторов, оказывается неудобен из-за своей разрывности. На практике применяется непрерывный штраф, иногда квадратичный, а чаще – линейный.

В постановке задачи (2.4) векторам запрещено находиться на той стороне от гиперплоскости, которая соответствует другому классу, разделяющая полоса носит называние “жесткой“. В общем случае, задача решается в пространстве вторичных признаков, и вместо “жесткой“ разделяющей полосы рассматривается штраф за нарушение ограничения. Использование пространства вторичных признаков помогает справиться с линейной неразделимостью, а использование штрафов – с возможным «переобучением» (overfitting). В результате преобразований исходной задачи, получаем задачу квадратичного программирования с линейными ограничениями:

$$\begin{aligned} \frac{1}{2}(\mathcal{W},\mathcal{W}) + C\sum_{i=1}^{N} p(e_i) &\rightarrow \min_{\mathcal{W},b} \\ y_i((\mathcal{W},\phi(X_i))+b) &\geq 1-e_i, \\ e_i &\geq 0, \\ i &= 1,\ldots,N \end{aligned} \tag{2.10}$$

Коэффициент штрафа C подбирают так, чтобы и количество векторов, попавших на неправильную сторону разделяющей полосы, было небольшим, и общее количество опорных векторов тоже было невелико, поскольку, чем меньше слагаемых с ненулевыми коэффициентами в формуле (2.5a), тем быстрее работает классификатор.

Однако на практике приходиться решать задачу разделения на три и более классов. Для её решения можеть быть применён метод, называемый “каждый против каждого“ (One vs. One [54]). Суть метода заключается в следующем. На этапе обучения рассматривается $\frac{q(q+1)}{2}$ классификаторов SVM, различающих пары классов, где q - количество классов. Каждый из классификаторов обучается только на векторах, принадлежащих двум, соответствующим данному классификатору классам, поэтому время обучения и количество опорных векторов получаются меньше, чем у SVM типа “каждый против всех“. Для каждого распознаваемого вектора рассчитаем все значения классифицирующих функций, отделяющих i-й класс от j-го, затем вычислим q сумм $f(\phi, X) = \sum_{i=j} \phi(f_{ij}(X))$, где ϕ - некоторая монотонно неубывающая функция, (например, $sign$) и выберем из них наибольшую. Соответствующий класс, для которого получена максимальная оценка, и будет ответом распознавателя.

Следует отметить, что метод имеет следующие особенности:

1. Вычислительная сложность обычных алгоритмов, решающих задачу квадратичного программирования (таких как метод Ньютона), делает задачу обучения SVM крайне трудоемкой для больших наборов данных.

2. В каждом конкретном случае решение задачи подбора ядра требует предварительного изучения.

3. Для определения величины параметра штрафа C также необходимо предварительное исследование.

2.2.5 Подбор параметров распознавателя

В задаче обучения классификатора алгоритм обычно зависит от параметров, которые определяют форму решающей функции или способы поиска самой этой функции. Поэтому подбор параметров позволяет уменьшить ошибку классификации на тестовой выборке. В свою очередь оценка ошибки может быть выполнена с помощью таких методов как кросс – валидация [55] или jack-knife тестирование [56]. Кроме того, она может быть ограничена мажорантой, значение которой получено с помощью теоретического анализа.

Необходимо заметить, что на практике приходиться подбирать несколько параметров для оптимизации, при этом, ошибка классификатора имеет неявную зависимость от оптимизируемых параметров. Вследствие этого, невозможно осуществить простой перебор всех возможных значений параметров. В рассматриваемом случае использовались следующие параметры оптимизации:

- Параметр гауссова ядра γ в формуле (2.8).
- Параметр штрафа C в функции (2.9).

Оптимизация параметров функции-распознавателя производилась следующим образом. В качестве критерия для подбора параметров использовалась средняя ошибка при 5 кратной кросс-валидации, которая заключается в следующем. Обучающее множество делится на 5 частей: на одной из этих частей проводится обучение модели, которая затем тестируется на 4 оставшихся подмножествах, при этом ошибка кросс-валидации получается путём усреднения ошибок по этим 4 подмножествам. Далее решается задача

минимизации средней ошибки распознавателя в зависимости от параметров γ и C для поиска указанных параметров.

Ниже приведена схема алгоритма поиска оптимальных параметров C и γ.

Вход: Набор векторов $\{X_i\}_{i=1}^N$

Шаг 1. Для фиксированного k представить обучающее множество $\mathcal{X} = \{X_i\}_{i=1}^N$ как $\mathcal{X} = \bigcup_{j=1}^k \mathcal{X}_j$. Зафиксировать точность решения задачи ϵ.

Шаг 2. Выбрать начальное значение $x_0 = (C_0; \gamma_0) \in \mathbb{R}^2$ и величину шага Δ_0. **Шаг 3.** Выполнять пока $||x_k - x_{k+1}|| > \epsilon$

Подшаг 1. Решить задачу обучения SVM при $C = C_k, \gamma = \gamma_k$ и $X_i \in \mathcal{X}_1$.

Подшаг 2. Определить функцию $f(t) = \frac{1}{k-1}\sum_{j=2}^k E_j(t)$, где $E_j(t) = \frac{1}{|\mathcal{X}_j|}\sum_{X_i \in \mathcal{X}_j} \mathbb{I}\{\tilde{y}_{X_i}(t) \neq y_{X_i}\}$, где $\tilde{y}_{X_i}(t)$ - предсказанная метка вектора X_i, y_{X_i} - его настоящая метка.

Подшаг 3. Для $\forall t \in P_k = \{x_k \pm \Delta_k e_i : i = 1, 2\}$ вычислить $f(t)$

Подшаг 4. Если $\exists \hat{t} : f(\hat{t}) < f(x_k)$ установить $x_{k+1} = \hat{t}, \Delta_{k+1} = \Delta_k$;
иначе $x_{k+1} = x_k, \Delta_{k+1} = \frac{\Delta_k}{2}$.

Выход: оптимальные значения параметров классификатора $\hat{C}, \hat{\gamma}$.

В качестве преимуществ такого подхода можно отметить следующие:

- В виду того, что зависимость средней ошибки прогноза принадлежности данного высказывания определённому диктору от выбранных параметров классификатора является неявной, то нет оснований считать, что эта функция будет выпуклой. Следовательно, существует вероятность попадания в локальный минимум.
- Задача решалась в параллельных процессах, так как сама процедура кросс - валидации может быть выполнена параллельно, поскольку каждая

итерация может выполняться независимо от других, и нет никаких зависимостей по данным. Следовательно, весь процесс может быть легко проведен на многопроцессорных машинах. Кроме того, предложенный метод для решения поиска параметров C и γ тоже был реализован параллельно. Кроме того, данный метод может стартовать независимо и параллельно из несколько разных начальных точек, с последующим сравнением результатов для выбора наилучшего.

Таким образом, вся система может быть реализована на кластере или в монолитной многопроцессорной системе, с поддержанием многопоточности. На каждом процессоре алгоритм работает со своими начальными значениями, при этом вычисление функции $f(t)$ производится многопоточно.

2.2.6 Фишеровские ядра

Тем не менее задача идентификации диктора может быть решена с использованием классификаторов, которые напрямую строят разделяющую поверхность в пространстве признаков. В качестве таких классификаторов обычно используются Gaussian Mixture Models (GMM [57]) или линейный дискриминант Фишера [58]. Их недостаток заключается в том, что в целевую функцию не входит некоторая информация из сообщения. Таким образом, необходимая для классификации сообщений информация может быть потеряна, что негативно скажется на точности распознавания.

Для устранения этого недостатка применяется метод, основанный на Фишеровских ядрах [52], которые отображают всё озвученное диктором предложение целиком (полное высказывание) в единственную точку, что позволяет проводить их разделение. Однако, чтобы представить высказывание в виде одной точки, оно должно находиться в пространстве большой

размерности. Это не вызывает затруднений, поскольку SVM и предназначен для решения задач высокой размерности.

Идея разработанной модификации метода заключается в использовании в качестве ядра функции потерь, вычисленной с помощью апостериорных вероятностей наблюдений, которые получены от порождающей модели, в качестве которых могут выступать либо Скрытые Марковские модели либо GMM.

Теорема. Пусть $P(X|\theta)$ апостериорная вероятность, полученная от порождающей модели. Зададим в пространстве всех возможных $P(X|\theta)$ скалярное произведение как $U_X^T F^{-1} U_X$, где $F = \mathbb{E}_X U_X U_X^T$ - матрица информации Фишера и $U_X = \nabla \ln P(X|\theta)$ фишеровская функция потерь. Тогда функция

$$K(X_i, X_j) = U_{X_i}^T F^{-1} U_{X_j}. \tag{2.11}$$

является ядром.

Доказательство. Для доказательства достаточно проверить симметричность и положительную полуопределенность функции.

Докажем симметричность. $K(X,Y) = U_X^T F^{-1} U_Y = (U_X^T F^{-1} U_Y)^T = U_Y^T F^{-1} U_X = K(Y,X)$. Докажем положительную полуопределённость функции. Матрица информации Фишера является положительно полуопределённой формой, причем она принимает значение равное 0, только в том случае, когда, плотность вероятности сосредоточена в подпространстве меньшей размерности, чем размерность вектора X. Тогда $K(X,X) = U_X F^{-1} U_X$ также является положительно полуопределённой формой, так как взятие обратной матрицы не влияет на знакоопределённость.

В ряде случаев, для простоты можно использовать не саму матрицу Фишера, а её приближение в виде единичной — тогда 2.11 преобразуется в обычное скалярное произведение в евклидовом пространстве.

На практике вторым шагом работы алгоритма является преобразование Фишеровского ядра в гауссовское или любое другое. Так как не существует единственного теоретически определённого метода выбора ядра, то главное, что требуется от ядра – хорошая разделимость признаков.

Таким образом, получается два отображения пространства первичных признаков: на основе Фишеровского ядра, и с помощью классических ядер (гауссовского или полиномиального).

Схема алгоритма приведена ниже.

Вход: Набор векторов $\{X_i\}_{i=1}^N$, оценки параметров порождающей модели θ и параметра γ.

Шаг 1. Для $\forall i \in 1, \ldots, N$ вычислить U_{X_i}.

Шаг 2. Получить оценку матрицы информации $\hat{F} = \frac{1}{N}\sum_{i=1}^{N} U_{X_i} U_{X_i}^T$ из порождающей модели и вычисление её обратной. На практике обычно используется её приближение в виде единичной.

Шаг 3. Для $\forall i, j \in 1, \ldots, N$ вычислить $K(X_i, X_j) = U_{X_i}^T F^{-1} U_{X_j}$.

Шаг 4. Вычислить $\hat{K}(X_i, X_j) = e^{\gamma K(X_i, X_j)}$

Выход: Значения ядра $\hat{K}(X_i, X_j)$ на всех парах векторов X_i, X_j.

2.3 Метод, основанный на дикторонезависимых признаках

2.3.1 Auditory Image Model

Другой подход заключатся в использовании признаков, которые не меняются от диктора к диктору. В качестве таких признаков можно использовать признаки из Auditory Image Model (AIM) [59], которая была разработана в лаборатории Роя Петерсона из Кембриджского университета с

целью моделирования человеческой психоакустики. AIM – функциональная модель человеческой слуховой системы, которая принимает во внимание биологическую информацию. Модель состоит из трёх последовательных модулей:

1. Банка фильтров, состоящего из гамматоных фильтров. Банк фильтров применяется для моделирования колебаний базальной мембраны.
2. Двумерного адаптивного порогового механизма. Этот механизм используется для вычисления Neural Activity Pattern (NAP).
3. Стробируемого интегратора. Стробируемый интегратор применяется к NAP представлению для синхронизации периодов между максимумами NAP.

Этапы обработки речевого сигнала для получения AIM изображены на рис. 2.3. Рассмотрим схему функционирования модели. Банк фильтров

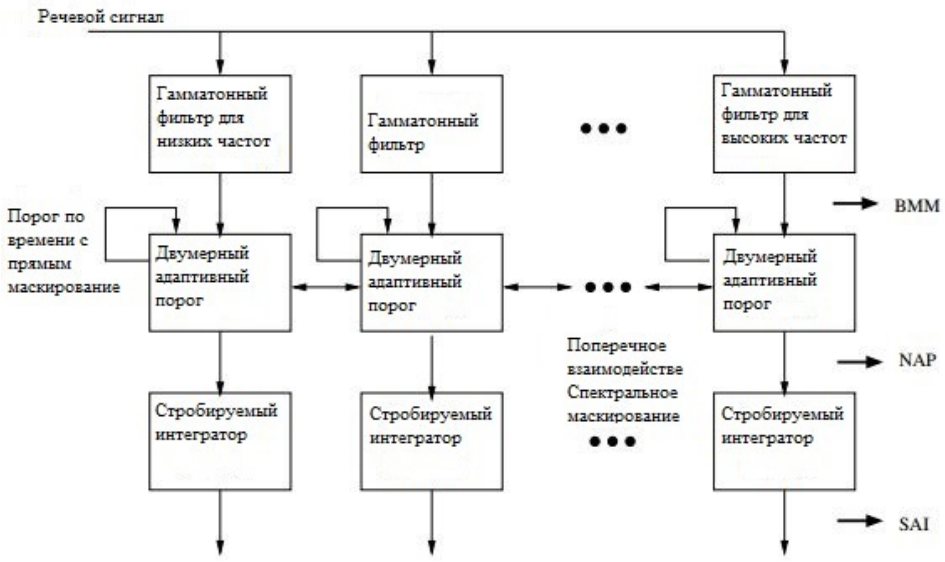


Рис. 2.3: Структура AIM.

сконструирован согласно ERB - масштабированию, рассмотренному в разделе 1.6.2. Выходной сигнал банка фильтров соответствует колебаниям базальной мембраны внутреннего уха. Отклик банка фильтров вычисляется с помощью гамматонных фильтров [60]. Выходной сигнал с этого этапа обозначен на рис. 2.3 как BMM.

На втором этапе сигнал, полученный на выходе банка фильтров, пропускается через полупериодный выпрямитель и проходит через сжимающий логарифмический нелинейный канал. После этого применяется двумерный (по времени и частоте) адаптивный пороговый механизм. Временной порог основан на использовании краткосрочной памяти прошлой активности: если предыдущая по времени активность была низкой, то большая часть движения базальной мембраны пропускается через порог. Спектральный порог основан на взаимодействии близких частотных каналов: высокая активность в канале будет частично подавлять активность в менее активном канале. Выходной сигнал на этом этапе схож с нейронной моделью деятельности (Neural Activity Pattern - NAP) слухового нерва, который соединяет улитку внутреннего уха с ядрами ствола мозга. В дальнейшем для анализа будут использованы NAP признаки.

Для получения признаков использовалась следующая методика. Активность в каждом канале NAP сглаживается при помощью фильтра высоких частот с частотой среза 100 Гц, после этого NAP нарезается на фреймы длиной 10 мс. Окончательно NAP профиль формируется суммированием активности внутри фрейма. Далее полученный профиль нормализуется таким образом, чтобы с ним можно было работать как с плотностью распределения. Для описания полученной плотности было использовано расширение Грам-Шарлье.

2.3.2 Расширение Грам-Шарлье

Если истинная плотность распределения случайной величины z неизвестна, то разумно представить её в виде:

$$g(z) = p_n(z)\psi(z), \tag{2.12}$$

где $\psi(z)$ – плотность стандартного нормального распределения, а $p_n(z)$ выбрана таким образом, чтобы $g(z)$ имела те же моменты, что и истинная плотность z. Такая аппроксимация носит название расширения Грам-Шарлье. Идея, заложенная в такое представление, заключается в том, чтобы построить характеристическую функцию одной случайной величины на основе другой, например, нормальной, с известными свойствами.

При построении функции $p_n(z)$ обычно используется метод, основанный на полиномах Эрмита.

Полиномы Эрмита образуют ортогональный базис относительно скалярного произведения, порожденного математическим ожиданием, взятым по плотности стандартного нормального распределения. Это свойство позволяет использовать многочлены Эрмита H_i в функции $p_n(z)$:

$$p_n(z) = 1 + \sum_{i=1}^{n} c_i H_i(z).$$

Параметры c_i имеют близкое отношение к кумулянтам распределения. Таким образом представление 2.12 представляется удобным, так как даёт возможность моделировать моменты высокого порядка, что может быть важным при распознавании речи. Действительно, в работах [25, 26] было отмечено, что моделирование моментов высоких порядков может увеличить точность распознавания.

К сожалению, полученная функция не является в строгом смысле плотностью: для некоторых значений параметров функция может принимать

отрицательные значения. Для решения этой проблемы предложено [61] использовать положительную плотность:

$$g(z) = \psi(z)\frac{(1+\sum_{i=1}^{n} c_i H_i(z))^2}{k}, \tag{2.13}$$

где $k = 1 + \sum_{i=1}^{n} c_i^2 i!$. Подобная плотность удобна не только с теоретической точки зрения, но и с практической – при оценке параметров методом максимального правдоподобия логарифмическая функция правдоподобия получается разделяемой и содержит логарифмы положительных выражений, что упрощает численную оптимизацию:

$$\hat{\ell}(z_i) = \ln(\psi(z_i)) + \ln(1+\sum_{j=1}^{n} c_j H_j(z_i))^2 - \ln(1+\sum_{j=1}^{n} c_j^2 j!) \tag{2.14}$$

Для получения оценок неизвестных параметров воспользуемся методом максимального правдоподобия. Для этого придется решать следующую оптимизационную задачу:

$$\ell(z,\theta) = \frac{1}{N}\sum_{i=1}^{N} \hat{\ell}(z_i) \to \max_{\theta}, \tag{2.15}$$

$$f(\theta) \leq 0, \tag{2.16}$$

где $\ell(z_i)$ задано в (2.14), θ - вектор неизвестных параметров, и $f(\theta)$ - функция ограничений, которая может быть добавлена для того, чтобы значения параметров удовлетворяли каким-либо априорно заданным ограничениям. Например, оценки некоторых параметров должны быть положительны.

Теорема. Решение задачи (2.15) дает состоятельные и асимптотически нормальные оценки параметра θ.

Доказательство. Запишем условие первого порядка для (2.15):

$$\frac{\partial \ell(z,\theta^*)}{\partial \theta} = 0. \tag{2.17}$$

Разложим функцию из 2.17 в окрестности истинного параметра θ^* по формуле Тейлора до первого порядка. В нашем случае взятие производной по

параметру возможно по всей области определения $\ell(z,\theta)$, так целевая функция представляет собой многочлен по компонентам θ. Разложение выглядит следующим образом:

$$\frac{\partial \ell(z,\theta)}{\partial \theta} = \frac{\partial \ell(z,\theta^*)}{\partial \theta} + \frac{\partial^2 \ell(z,\tilde{\theta})}{\partial \theta^T \partial \theta}(\theta - \theta^*) = 0, \tag{2.18}$$

где точка $\tilde{\theta}$ лежит покомпонентно между θ^* и θ. Теперь выразим разность $\theta - \theta^*$ из уравнения (2.18):

$$(\theta - \theta^*) = \left(-\frac{\partial^2 \ell(z,\tilde{\theta})}{\partial \theta^T \partial \theta}\right)^{-1} \frac{\partial \ell(z,\theta^*)}{\partial \theta}. \tag{2.19}$$

Заметим, что $\frac{\partial \ell(z,\theta)}{\partial \theta} = \frac{1}{N}\sum_{i=1}^N \frac{\partial \ell(z_i,\theta)}{\partial \theta}$ и $\frac{\partial^2 \ell(z,\theta)}{\partial \theta^T \partial \theta} = \frac{1}{N}\sum_{i=1}^N \frac{\partial^2 \ell(z_i,\theta)}{\partial \theta^T}$. Подставим эти выражения в (2.19) и умножим на $\sqrt{N}$. Получим, что

$$\sqrt{N}(\theta - \theta^*) = \left(-\frac{1}{N}\sum_{i=1}^N \frac{\partial^2 \ell(z,\tilde{\theta})}{\partial \theta^T \partial \theta}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{\partial \ell(z_i,\theta^*)}{\partial \theta}.$$

При этом по равномерному закону больших чисел $-\frac{1}{N}\sum_{i=1}^N \frac{\partial^2 \ell(z,\tilde{\theta})}{\partial \theta^T \partial \theta}$ сходится к $H = \mathbb{E}\left[-\frac{\partial^2 \ell(z,\theta^*)}{\partial \theta^T \partial \theta}\right]$, а в силу теоремы о непрерывным преобразованиях над случайными величинами $\left(-\frac{1}{N}\sum_{i=1}^N \frac{\partial^2 \ell(z,\tilde{\theta})}{\partial \theta^T \partial \theta}\right)^{-1}$ сходится к H^{-1}. В свою очередь, $\frac{1}{\sqrt{N}}\sum_{i=1}^N \frac{\partial \ell(z_i,\theta^*)}{\partial \theta}$ по центральной предельной теореме сходится к стандартной нормальной случайной величине со средним ноль и дисперсией $I = \mathbb{E}\frac{\partial \ell(z,\theta^*)}{\partial \theta}\left(\frac{\partial \ell(z_i,\theta^*)}{\partial \theta}\right)^T$. Выражение для I есть матрица информации Фишера.

Таким образом получаем, что распределение оценок нормально:

$$\sqrt{N}(\theta - \theta^*) \to \mathcal{N}(0, H^{-1} I H^{-1}), \tag{2.20}$$

откуда и следует состоятельность. Вместе с тем, в случае, если z_i имеют распределение, задаваемое (2.13), то асимптотическая дисперсия упрощается в I^{-1}. Состоятельность также может быть представлена в следующем виде. Для

любых $\delta > 0$ и $\epsilon > 0$ существует такое количество наблюдений $T = T(\delta, \epsilon)$, что для всех $T \geq T(\delta, \epsilon)$ оценка θ_T, полученная по T наблюдениям обладает следующим свойством:

$$P(|\theta_T - \theta^*| < \epsilon) > 1 - \delta \tag{2.21}$$

2.3.3 Алгоритм получения признаков

Для того, чтобы на практике получить значения параметров θ функции (2.14) необходимо численно решить оптимизационную задачу (2.15).

Существует множество методов численного решения задичи (2.15). Эти методы можно разделить на градиентные и безградиентные. Подробный анализ применимости различных методов к задаче (2.15) приведён к разделе 4.2.1. В данном же разделе представлены модификации алгоритма симуляции отжига [62], которые используют параллельные вычислительные процессы, и могут увеличить как точность, так и скорость работы алгоритма. Первая модификация алгоритма, являющаяся более вычислительно затратной, приведена ниже. Суть модификации заключается в независимом старте k процессов отжига из разных начальных точек.

Вход: набор значений $\{z_i\}_{i=1}^N$

Шаг 1. Сгенерировать k начальных значений параметров $\{\theta_i\}_{i=1}^k$.

Шаг 2. К каждому значению θ_i применить алгоритм симуляции отжига, получив k финальных оценок $\{\hat{\theta}_i\}_{i=1}^k$.

Шаг 3. Вычислить $\ell(z, \hat{\theta}_i)$ для каждого $\hat{\theta}_i$, $i = 1, \ldots, k$.

Шаг 4. $\tilde{\theta} = \max_{i=1,\ldots,k} \hat{\theta}_k$.

Выход: Оптимальное значение $\tilde{\theta}$

При этом шаги 2-4 выполняются параллельно.

Вторая модификация представляет собой гибридную схему: случайный поиск предшествует процессу отжига. Алгоритм приведён ниже.

Вход: набор значений $\{z_i\}_{i=1}^N$

Шаг 1. Сгенерировать k начальных значений параметров $\{\theta_i\}_{i=1}^k$.

Шаг 2. Вычислить $\ell(z, \hat{\theta}_i)$ для каждого $\hat{\theta}_i$, $i = 1, \ldots, k$.

Шаг 3. Вычислить $\tilde{\theta} = \arg\max_{i=1,\ldots,k} \ell(z, \hat{\theta}_i)$.

Шаг 4. Применить алгоритм симуляции отжига к $\tilde{\theta}$, получив финальную оценку $\hat{\theta}$

Выход: Оптимальное значение $\hat{\theta}$

При этом шаги 2-3 выполняются параллельно.

2.4 Выводы

В данной главе проанализированы математические модели, применяющиеся для построения систем распознавания речи. В главе получены следующие результаты:

1. Произведён анализ применимости Скрытых Марковских Моделей для систем распознавания речи.
2. Рассмотрены методы построения дикторонезависимых систем распознавания речи.
3. Представлены модификации алгоритма опорных векторов с использованием Фишеровских признаков.
4. Доказана возможность использования Фишеровских признаков в качестве ядерного преобразования.

5. Доказано существование и единственность решения задачи обучения машины опорных векторов с Фишеровскими признаками.

6. Представлена вероятностная модель речевых дикторонезависимых признаков.

7. Доказана состоятельность и асимптотическая нормальность оценок предложенных признаков, полученных с помощью метода максимального правдоподобия.

Глава 3

Реализация системы идентификации языка и диктора

В данной главе рассматриваются различные аспекты реализации системы идентификации языка и диктора с применением предложенных методов. Последовательно рассматриваются следующие этапы: обработка речевого сообщения, вычисление признаков, поиск наиболее вероятной цепочки состояний, а также идентификация языка и диктора. Предлагается программная реализация системы идентификации языка и диктора.

3.1 Общий вид системы идентификации языка и диктора

На рис. 3.1 приведен предложенный автором вид системы идентификации языка и диктора, являющийся частным случаем схемы обработки речевого сигнал для задач идентификации, приведённой на рис. 1.2.

Рассмотрим подробно все этапы, представленные на рисунке.

Цель первого этапа (рис. 3.1 Предобработка) заключается в том, чтобы сделать сигнал как можно более чистым. На этом этапе обработки речевого сигнала, который представляет собой файл с фонограммой с расширением .wav, из речевого сигнала удаляется шум с помощью адаптивного винеровского фильтра [63]. Затем полученный очищенный сигнал s_t усиливают, путём

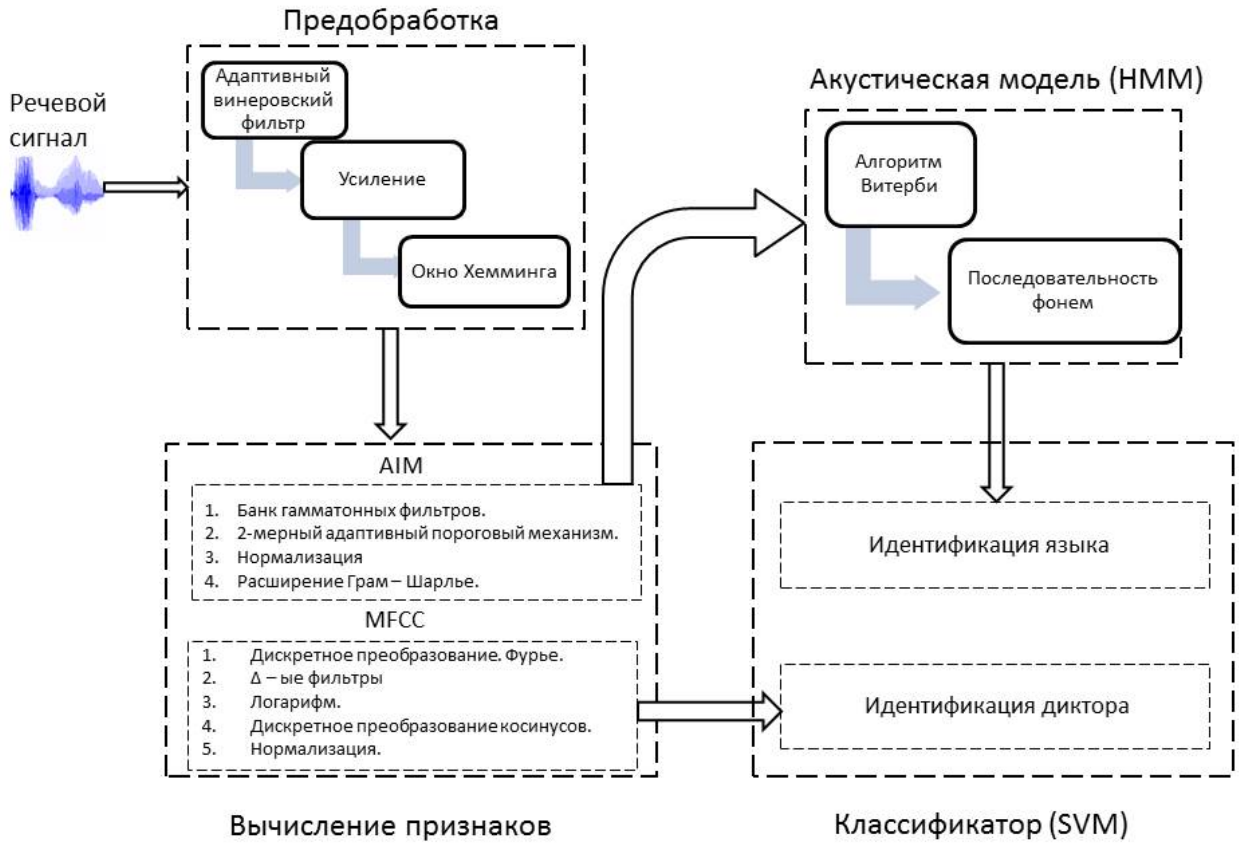


Рис. 3.1: Общий вид системы идентификации языка и диктора.

применения разностного уравнения первого порядка

$$s'_t = s_t - ks_{t-1},$$

где s'_t - усиленный сигнал, а k - коэффициент усиления. Напомним, что свойства речевого сигнала медленно меняются со временем, то есть, он является квази-стационарным. Если рассматривать его на коротких временных интервалах (5-100 мс), то характеристики остаются постоянными. Таким образом, на этапе усиления речевой сигнал нарезают на участки, называемые фреймами, с помощью движущегося окна Хемминга (см. раздел 1.5.2 уравнение (1.3)).

На втором этапе (рис. 3.1 Вычисление признаков) происходит выделение акустических признаков. В предложенной системе используются мел - частотные кепстральные коэффициенты (см. раздел 2.1.2) и

модифицированные признаки из AIM (см. раздел 2.3.1). Для получения мел - частотных кепстральных коэффициентов на каждом фрейме над сигналом проводят следующие операции.

- К фрейму применяется дискретное преобразование Фурье.
- С помощью треугольного фильтра приводиться к частоте в мелах.
- На каждой частоте берётся логарифм.
- Производиться дискретное преобразование косинусов.
- Полученный спектр нормализуется.

Для получения признаков из AIM выполняются следующие действия

- К фрейму применяется банк гамматоных фильтров.
- Применение полупериодного выпрямителя.
- На каждой частоте берётся логарифм.
- Применяется двумерный пороговый механизм: по времени и частоте.
- Нормализация

В зависимости от решаемой задачи просходит использование вычисленных признаков. Для решения задачи идентификации диктора используются мел - частотные кепстральные коэффициенты, к которым применяется Фишеровское ядро по алгоритму, описанному в разделе 2.2.6. Преобразованные признаки используются для идентификации диктора с помощью предварительно обученного классификотора на основе машины опорных векторов.

В случае решения задачи распознавания языка используются признаки из AIM, которые сначала подаются на вход акустической модели, основанной на HMM (см. раздел 2.1). Для получения списка фонем, которые

используются для распознавания языка используется алгоритм Витерби (см. раздел 2.1.2). Полученные фонемы также используются для идентификации языка с помощью предварительно обученного классификотора на основе машины опорных векторов.

3.2 Архитектура программной реализации

В данном параграфе рассматривается архитектура реализованной системы идентификации языка и диктора на языке UML [64] в виде диаграм классов и последовательностей. Диаграма классов отражает статическую структуру системы. Она состоит из описания классов и взаимосвязей между ними. Диаграма последовательностей отражает динамические связи в системе, например, последовательность вызовов.

На рис. 3.2 представлена диаграмма классов сущностей, которые являются объектными представлениями данных, которыми управляет система идентификции.

Абстрактный класс Features предназначен для хранения и вычисления признаков входного речевого сигнала. Класс состоит из массива объектов FeatureValue и метода получения Extract, выполняющего извлечение признаков из полученного на вход речевого сигнала. Наследниками класса являются классы, вычисляющие мел - частотные кепстральные коэффициенты, модифицированные признаки из AIM.

Абстрактный класс Classifier предназначен для реализации классифицирующего алгоритма опорных векторов. Класс состоит из методов Train и Classify, а также объекта Parameters, который содержит все необходимые для работы классификатора параметры. Метод Train принимает на вход словарь, в котором ключём является метка класса, а значением - объект типа Features, и

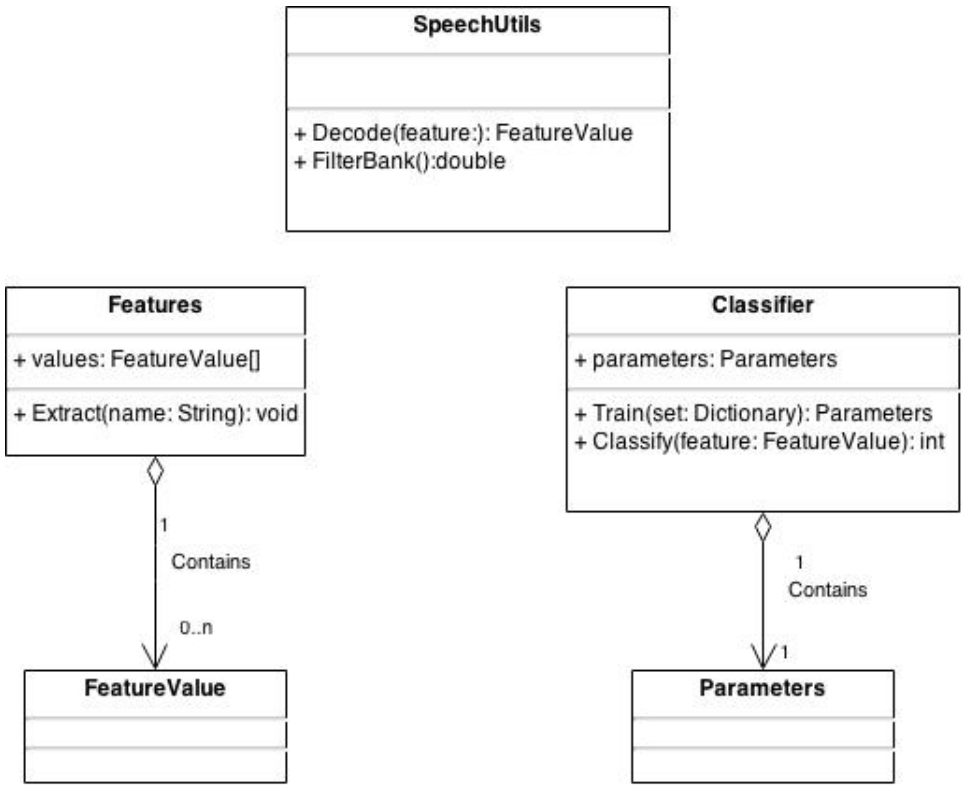
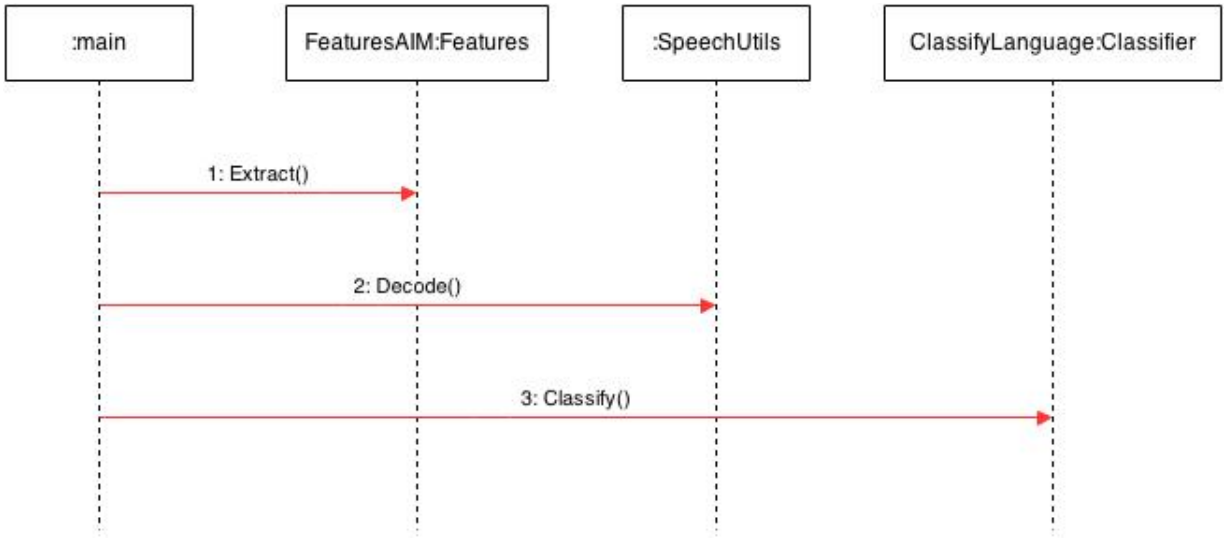


Рис. 3.2: Диаграмма классов - сущностей.

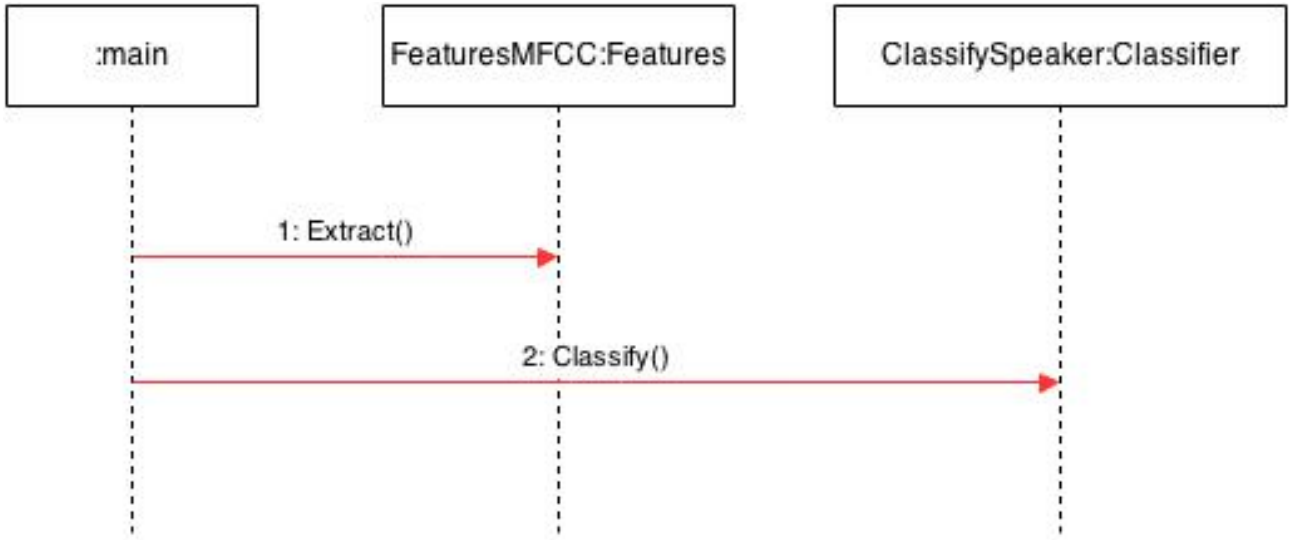
возвращает объект Parameters. Метод Classify принимает объект FeatureValue и возвращает значение решающей функции, а также метку класса - решения.

Класс SpeechUtils содержит вспомагательные методы, необходимые для вычисления признаков и классификации, такие как, например, вычисление выхода банка фильтров и алгоритм Витерби.

Последовательность вызовов методов классов для идентификации языка представлена на рис. 3.3a, а для идентификации диктора - на рис. 3.3b . Сначала вызывается метод Extract у классов FeaturesMFCC и FeaturesAIM, которые являются наследниками класса Features. После этого вызываются метод Classify класса ClassifySpeaker, на вход которому подаётся объект FeaturesMFCC.FeatureValue, и метод Decode класса SpeechUtils, который реализует алгоритм Витерби, который принимает объект FeatureValue и возвращает объект Phonems, являющийся наследником FeatureValues. После этого



(a)



(b)

Рис. 3.3: Диаграммы вызовов методов для идентификации языка (3.3a) и диктора (3.3b)

происходит вызов метода Classify класса ClassifyLanguage, на вход которому подаётся объект Phonems.

Результатом последовательности вызовов являются метки класса диктора и языка, к которым классификатор отнес входной речевой сигнал.

3.3 Применение параллельных вычислений в задаче идентификации языка и диктора

Необходимо заметить, что дизайн параллельной архитектуры для системы идентификации языка и диктора должен быть максимально гибким в том смысле, что полученное улучшение в качестве работы последовательного распознавателя (применение новых признаков речи, вычисление вероятностей и т.д.) не должно требовать перестройки всей системы. Возможные способы использования параллелизма для идентификации языка и диктора включают в себя примение специальных аппаратных средств [65], решение задачи распознавания отдельных слов [66], распараллеливание отдельных этапов идентификации языка и диктора (например, вычисление правдоподобия).

Рассмотрим подробно те стадии идентификации языка и диктора, на которых применение параллельных технологий может дать значительные улучшения, как в скорости работы, так и в точности распознавания.

На начальном этапе обработка речевых сигналов, поступающих из различных источников, выполняется на матричных (векторных) компьютерах. В этом случае каждый узел независимо обрабатывает свой речевой сигнал. При этом, над каждым речевым сигналом производятся одинаковые действия: усиление, нарезка на фреймы, вычисление признаков и собственно распознавание.

Отдельно рассмотрим этап вычисления признаков. Исследования (например, [67]) показывают, что применение различных наборов признаков для описания речевого сигнала приводит к увеличению точности распознавания как для задач с большим объемом словаря, так и для задач, где размер словаря мал. Таким образом, на этом этапе распараллеливанию подвергается вычисление различных типов признаков:

кепстральные коэффициенты, аудиторные признаки (подробное описание дано в разделе в 2.3.1), прозодические признаки [68] и т.д. При этом каждый процесс, который вычисляет признаки, должен иметь доступ к речевому сигналу для его обработки. Таким образом, представляется возможным описать этап вычисления признаков в виде системы с общей памятью. Подобная система позволяет использовать общие ресурсы, такие как код и данные.

Все процессоры в системе с общей памятью используют одно и то же адресное пространство общей памяти через сеть с внутрисистемной коммутацией (interconnection network), роль которой обычно [69] исполняет шина, но в случае больших систем в целях улучшения производительности используют сети. Для измерения производительности подобной системы важно знать как количество обращений к памяти в единицу времени, которую система может поддерживать, так и временную задержку между запросом доступа к памяти и получением доступа. В подобной системе может также использоваться несколько модулей памяти. При этом следует иметь в виду, что при одновременном доступе к памяти возможны коллизии: изменения состояния памяти одним процессором, в то время, как остальные процессоры используют устаревшие данные. Таким образом, контроль за синхронизацией является важным этапом разработки подобных систем.

Очевидно, что при использовании различных признаков важна синхронизация потоков вычислений. Действительно, в зависимости от используемых признаков время их вычисления может сильно различаться, а во – вторых, для предотвращения скачкообразных изменений признаков на границах фреймов, их вычисление часто производится с перекрытиями (см. рисунок 3.4), следовательно, для эффективного использования ресурсов необходимо организовать хранение и доступ к уже вычисленным признакам.

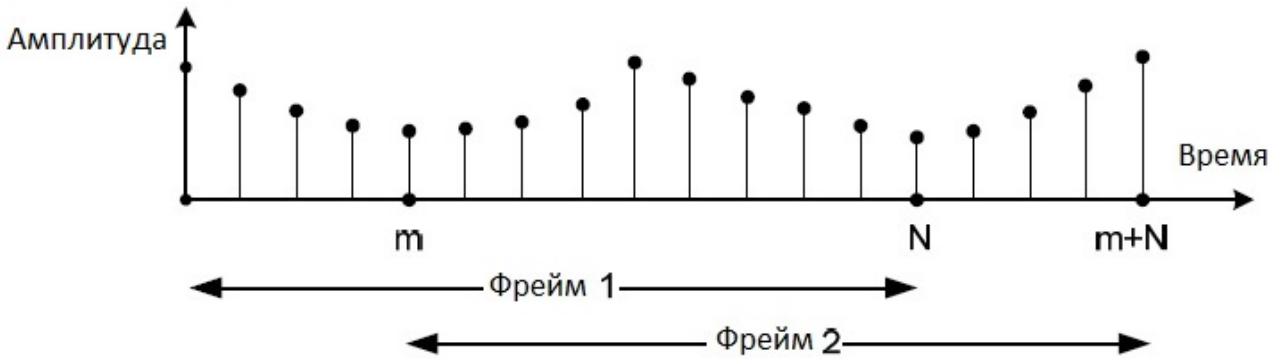


Рис. 3.4: Перекрывание фреймов. Схематически изображено разделение фреймов с перекрытиями для вычисления признаков, N – длина фрейма, m – величина временного сдвига при формировании нового фрейма.

На последнем шаге происходит распознавание, то есть нахождение наиболее вероятной цепочки состояний с помощью алгоритма Витерби. В алгоритме Витерби вероятность последовательности слов вычисляется итеративно, и зависит от вероятностей, вычисленных на предыдущих шагах алгоритма, таким образом, на каждой итерации приходиться сохранять множество состояний, соответствующих различным возможным интерпретациям входного речевого сигнала.

Для каждого входящего фрейма можно выделить 3 фазы распознавания [70]:

1. Вычисление вероятности наблюдения. На этом шаге вычисляется вероятности $b(O_i; m_k)$ появления признаков O_i, вычисленных на фрейме m_k, согласно акустической модели (обычно это смесь гауссовых распределений). Эта фаза является самой вычислительно затратной.

2. Вычисление вероятности $\psi_t(s_j; w_{jt})$ нахождения в состоянии s_j в момент t при условии того, что до этого была распознана последовательность слов w_{jt}:

$$\psi_t(s_j; w_{jt}) = \max_i(\psi_{t-1}(s_j; w_{jt-1})a_{ij}b(O_t; m_k)),$$

где a_{ij} - вероятность перехода из состояния s_i в состояние s_j .

3. Вычисление терминальных вероятностей, как произведение переходных вероятностей a_{ij} и вероятности предыдущей цепочки

$$\psi_t(s_j; w_{jt}) = \max_i(\psi_{t-1}(s_j; w_{jt-1})a_{ij})$$

Полученные терминальные вероятности используются для определения оптимальной цепочки состояний, которые соответствуют фонемам.

Стоит отметить, что, несмотря на то, что процесс распознавания представляет собой итеративную процедуру, являясь, таким образом, последовательной операцией, представляется возможным производить вычисления в рамках каждой фазы параллельно. При этом производительность системы зависит от того, насколько эффективно вычисления будут положены на архитектуру SIMD.

3.4 Особенности конвейерной обработки речевого сигнала

Заметим, что для получения признаков необходима только однократная обработка речевого сигнала и для обработки используется только часть сигнала в пределах окна. Таким образом, представляется возможным использовать конвейерные вычисления для их получения.

Рассмотрим особенности обработки речевого сигнала с помощью конвейерных вычислений на примере получения мел - частотных кепстральные коэффициентов. Обозначим через $f_i, i = 1, \ldots, 5$ этапы получения мел - частотных кепстральные коэффициентов, приведёнными на странице 85.

Конвейерная схема обработки речевого сигнала представлена на рис. 3.5.

Входной речевой сигнал будет поступать на первый прибор (процессор), на котором выполняется первая функция f_1. После того, как некоторая, достаточная для начала выполнения функции f_2, часть входного сигнала будет обработана функцией f_1, эта часть потока (обозначенная p_1) поступает на

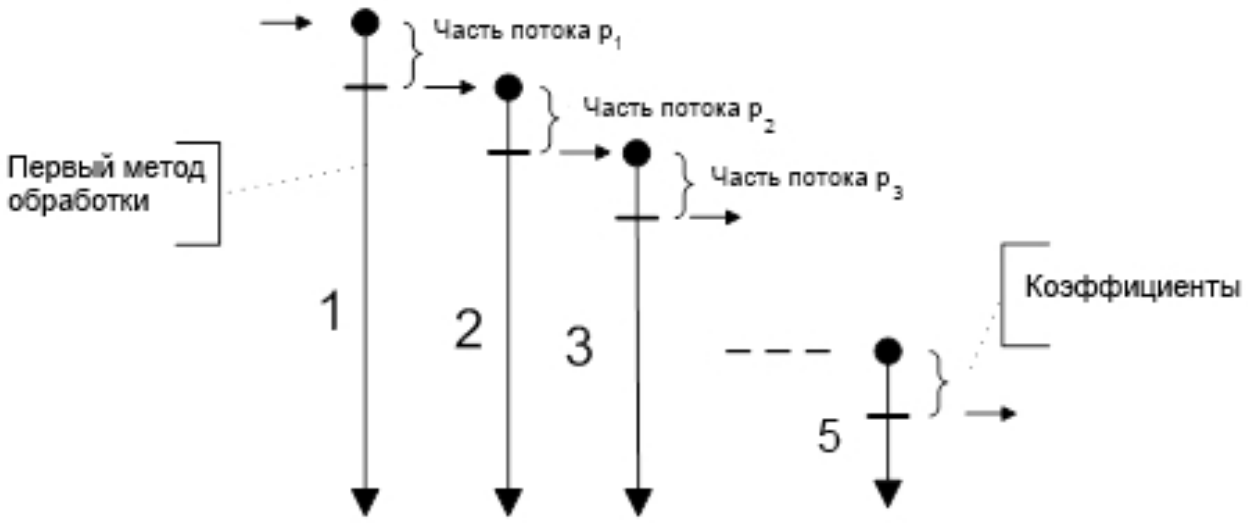


Рис. 3.5: Конвейерная схема процессов обработки речевого сигнала

второй прибор, где выполняется функция f_2. Аналогично первому прибору, после обработки части сигнала, обозначаемой p_2 и необходимой для работы третьей функции f_2, он поступает на третий прибор, и т.д. Конечным этапом обработки речевого потока будет набор мел - частотных кепстральные коэффициентов.

Вычисление признаков из AIM может выполняться аналогично.

3.5 Архитектура вычислительного комплекса

Принимая во внимание, что существуют области приминения систем идентификации языка и диктора, в которых системе приходится обрабатывать большой поток входящих заявок. При этом, такая обработка не может быть выполнена путём использования более высокопроизводительной техники, а требует специальных аппаратно - программных решений. Примером подобной ситуации может служить сотовая станция, в которой осуществляется

выборочное прослушивание звонков. В этом случае необходимо определить по голосу диктора, нужно ли прослушивать звонок.

Ясно, что для удовлетворения требований к обработке речевых сигналов в реальном масштабе времени при высоком, практически неограниченном по скорости потоке, необходимо снижать эту скорость до такого значения, при котором становится реальной его обработка на отдельном компьютере. Такое снижение скорости потока может быть достигнуто следующим путем. Пусть имеется поток речевой информации, имеющий скорость в V логических единиц в секунду. Разбивая этот поток на части, и циклически отправляя каждую часть на отдельное устройство (компьютер), мы можем снизить скорость этого потока пропорционально числу используемых устройств.

Вычислительный комплекс, осуществляющий указанную процедуру, представляет собой компьютерную систему, где каждый элемент независимо обрабатывает свой речевой поток (рис. 3.6). Подобная организация системы позволяет обрабатывать множество независимо приходящих сообщений. Здесь обученная акустическая модель и классификатор находятся либо на каждой машине, либо в случае большого объема хранятся в независимой памяти, доступ к которой осуществляется через сеть с внутрисистемной коммутацией (interconnection network). Такая организация памяти позволяет производить более чем одну операцию чтения из памяти одновременно.

В ряде случаев система работает в режиме реального времени (пример с сотовой станцией), следовательно, увеличение времени ожидания обработки заявки недопустимо. Таким образом, узлы системы обрабатывают приходящие потоки речевых сообщений независимо. Это вызвано тем, что если вновь прибывшая заявка будет обрабатываться одним из уже задействованных узлов системы, то при большой плотности прихода новых заявок накладные расходы на переключение контекстов и синхронизацию различных узлов системы

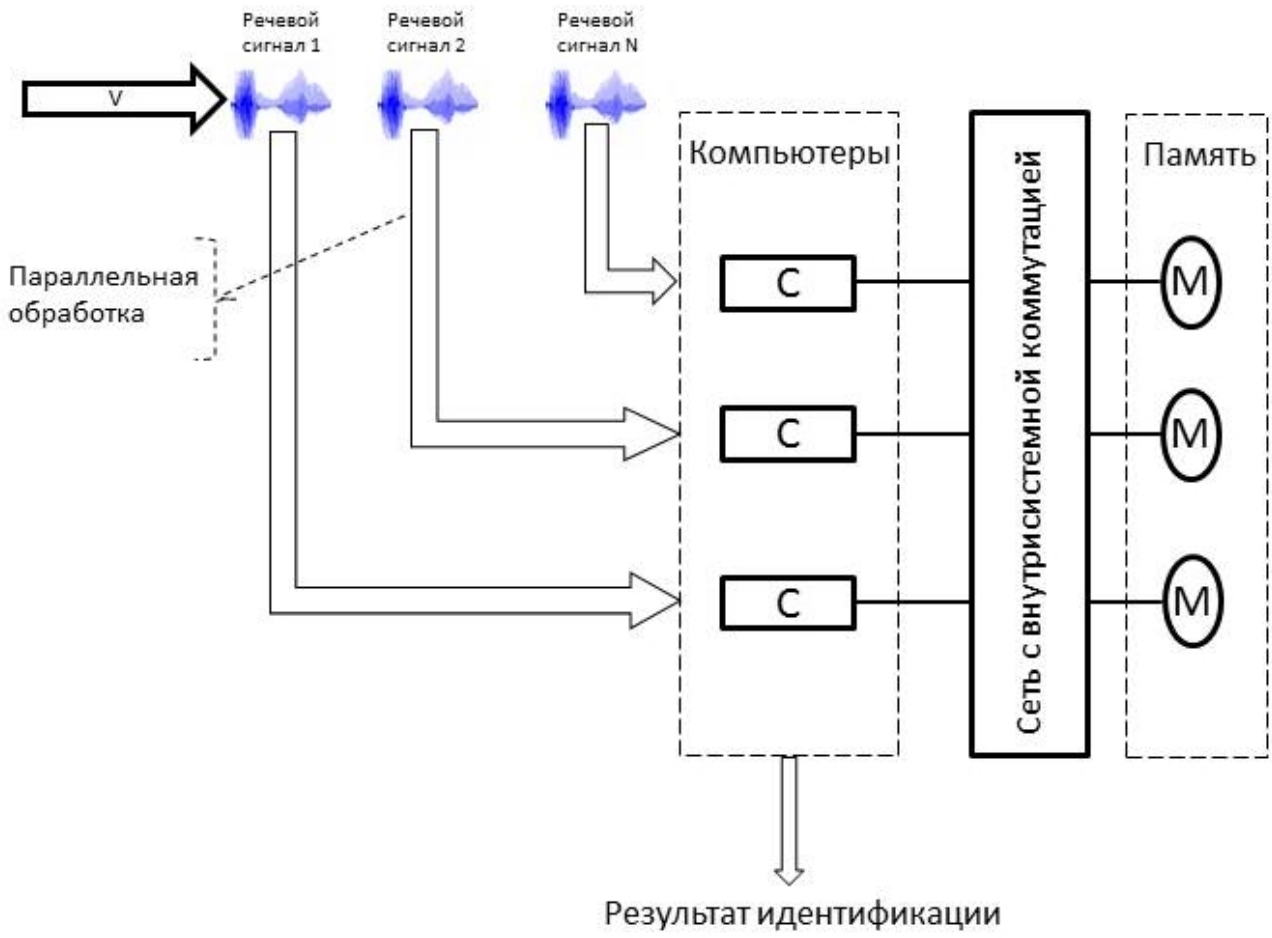


Рис. 3.6: Архитектура вычислительного комплекса. Изображена общая схема независимой обработки входящих речевых сигналов. На рисунке символом С обозначен компьютер, обрабатывающий речевой поток, символом М - независимая память

превышают выигрыш от использования дополнительных вычислительных мощностей на обработку этой заявки. Кроме того, синхронизация процессов на разных узлах системы и перенос данных между узлами увеличивает нагрузку на сеть.

Каждый узел комплекса представляет собой реализацию системы, изображенной на рис. 3.1. Следует отметить, что элемент системы, вычисляющей признаки, является в свою очередь сложным асимметричным мультипроцессором с неоднородным доступом к памяти (NUMA [69]), в котором каждый процессор предназначен для вычисления своих собственных признаков (кепстральные и признаки из AIM). Такая организация вычислений

позволяет справиться с особенностями вычисления признаков, описанными в предыдущем параграфе.

Действительно, в NUMA системе каждый процессор может обращаться к своей собственной памяти, а для доступа к другим областям памяти используется механизм передачи сообщений, интерфейс которого является независимым протоколом. Таким образом, доступ к памяти не является симметричным.

При этом, мультипроцессор может состоять из неоднородных процессоров, выполняющих каждый свою задачу.

Реализация акустической модели, осуществляющей поиск фонем, является аналогом [70].

После распознавания каждый элемент компьютера через сеть записывает результат распознавания во внутреннюю память. Запись в общую базу происходит в то время, когда нагрузка на систему минимальна. Такую информацию получают, собрав статистическую информацию о плотности потока заявок (речевых сообщений) по времени.

3.6 Выводы

В данной главе была рассмотрена реализации системы идентификации языка и диктора. Последовательно рассматрены элементы системы, выполняющие следующие действия: обработка речевого сообщения, вычисление признаков, поиск наиболее вероятной цепочки состояний, а также идентификация языка и диктора. Рассмотрена архитектура программной реализация системы идентификации языка и диктора, а также архитектура вычислительного комплекса, обрабатывающего большой поток входящих речевых сигналов в реальном масштабе времени.

1. Представлен общий вид системы идентификации языка и диктора.
2. Изложена архитектура программной реализации.
3. Проанализированы подходы к применению параллельных вычислительных процессов для идентификации языка и диктора.
4. Преставлена конвейерная схема обработки речевого сигнала для получения признаков.
5. Представлена архитектура вычислительного комплекса, способного обрабатывать большой поток входящих речевых сигналов в реальном масштабе времени.

Глава 4

Результаты экспериментов по распознаванию диктора и моделированию речевых признаков

В данной главе будут представлены результаты экспериментов с реальными и симулированными данными с применением моделей, описанных в главе 2. Даётся объяснение полученных результатов с точки зрения теории оптимизации, минимизации структурного риска, а также геометрических соображений.

4.1 Данные и описание экспериментов моделирования на Фишеровских признаках

При проведении экспериментов в качестве входных данных использовалась база речевых отрезков различной длительности. В ней содержатся данные по 15 дикторам, записанные с помощью обычного телефона, телефона GSM и микрофона. Характеристики входного сигнала для каждого канала: битрейт 16, частота дискретизации 8 кГц, соотношение сигнал – шум в среднем 15дб. Следует отметить, что в случае микрофонного канала при записи использовались микрофоны с очень разными АЧХ, поэтому фонограммы

сильно отличаются друг относительно друга. В следствие этих факторов распознавание в этом случае становится более сложной задачей.

На основе звукового сообщения длиной 250 мс вычислялись кепстральные коэффициенты (MFCC) без производных, которые по изложенному в главе 2 алгоритму, сначала подвергались трансформации с помощью Фишеровских, а затем гауссовских ядер.

Подбор параметров классификатора осуществлялся по алгоритму, описанному в разделе 2.2.5

4.1.1 Обсуждение результатов.

Сначала приведем результаты классификатора без применения Фишеровский ядер. В Таблице 4.1 показаны результаты классификации в микрофонном канале. В качестве исследуемой характеристики (точности) использовалось отношение высказываний, с правильно указанной принадлежностью к диктору, к их общему количеству.

Таблица 4.1: Сравнение точности распознавания диктора в различных каналах.

Длит., с	Точность, %					
	Микрофон		Телефон		GSM	
	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.
5	0.33	16.33	44.98	64.91	45.88	55.93
10	0	26.17	80.46	86.13	82.98	88.70
100	4.5	46.45	87.68	82.02	93.80	96.97

Результаты экспериментов, представленные в Таблице 4.1, показывают, что традиционные методы, основанные на SVM практически не работают. Действительно, даже при достаточно большом времени звучания классификатору не удается разделить дикторов. Небольшое увеличение точности на малом времени звучания можно объяснить тем, что в пространстве

вторичных признаков вектора случайно расположись таким образом, что несколько векторов попали на нужную сторону гиперповерхности, разделяющей классы. В пользу этой гипотезы можно отнести и то, что при большем количестве векторов ни один не попал необходимым для успешной классификации образом.

Теперь посмотрим на результат классификатора, с использованием Фишеровских признаков. В этом случае, как видно из Таблицы 4.1, применение Фишеровских признаков значительно увеличивает точность работы классификатора. Очевидно, что даже на небольшой выборке классификатор на новых признаках работает лучше, чем на MFCC признаках. Это можно объяснить, используя размерность Вапника-Червоненкиса (VC - размерность) [71].

Получаемое количество опорных векторов может служить хорошим приближением к VC – размерности. В случае применения Фишеровских признаков, опорных векторов получается больше, поэтому VC – размерность выше, а значит, большее количество точек может быть разделено гиперплоскостью. Дальнейшее отображение с помощью гауссова – ядра преобразует линейное разделение (с помощью гиперплоскости) к нелинейному (с помощью гиперповерхности).

Результаты исследований классификатора, использующего Фишеровские признаки, в более легких для распознавания телефонном и GSM каналах с фонограммами, записанными в одинаковых условиях, также представлены в Таблице 4.1.

В этих случаях легко заметить, что в этих каналах точность классификации достаточно высока. При этом зависимости от типа канала отсутствует, так как в обоих каналах точность примерно одинакова.

Результаты исследования показали увеличение точности распознавания при возрастании длительности речи. К сожалению, добиться хороших результатов на коротких (5 секунд) длительностях пока не удалось. Следует отметить, что применение Фишеровских ядер существенно улучшает точность распознавания в сильно зашумленном канале.

Таким образом, из приведённых результатов экспериментов можно сделать следующие выводы.

1. В тех системах распознавания речи, в которых не является возможным получение достаточно продолжительного речевого сигнала (менее 10 секунда), методы идентификации диктора не могут быть использованы для надежной предварительной настройки системы.

2. В тех же случаях, когда качество канала является приемлемым (такими случаями являются телефонный и GSM канал), и есть возможность получение продолжительного речевого сигнала (от 10 секунд), можно добиться хорошей точности идентификации диктора, и, как следствие, настройки всей системы под конкретного пользователя, тем самым повысив точность распознавания речи.

3. Применение SVM для решения задачи идентификации диктора в сильно зашумлённом канале в значительной мере зависит от используемых при распознавании признаков. При использовании стандартных MFCC признаков не удаётся добиться хороших результатов распознавания, при этом модификации MFCC признаков сначала с помощью Фишеровских, а потом гауссовских ядер позволяет значительно увеличить точность распознавания.

4. На точность системы идентификации диктора оказывает влияние качество канала - при малом соотношении сигнал/шум точность распознавания

оказывается очень низкой, при этом, почти нет различий в точности распознавания между телефонным и GSM каналом.

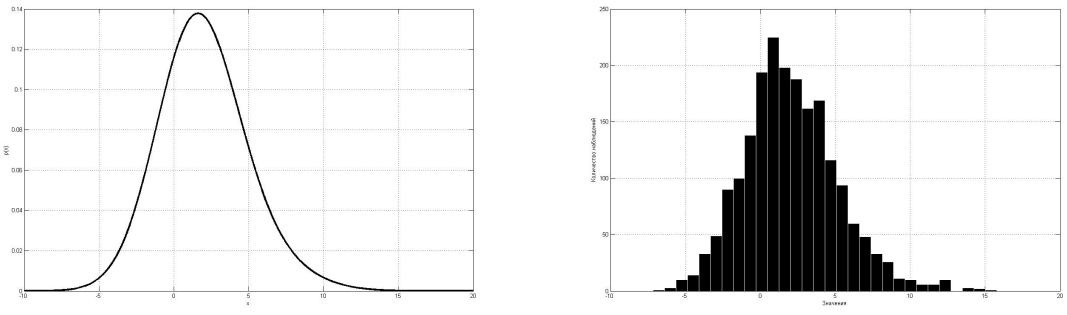
4.2 Результаты экспериментов по AIM

4.2.1 Монте - Карло эксперименты

В качестве предварительного анализа проведём исследование плотности распределения (2.13), используя метод Монте - Карло. Такой подход позволяет проанализировать различные методы получения оценок, зная априори из какого семейства распределений приходят наблюдения, тем самым уменьшая модельный риск.

В качестве примера плотности распределения используем расширение Грам - Шарлье нормальной плотности с параметрами $c_1 = 2, c_2 = 3, c_3 = 6, c_4 = 10$. Распределение с такими параметрами будет обладать средним равным c_1 и дисперсией c_2, но быть скошенным и иметь более тяжелые хвосты, чем нормальное распределение.

График плотности изображен на рис. 4.1a, на рис. 4.1b представлена гистограмма наблюдений из распределения с указанной плотностью. Из



(a) (b)

Рис. 4.1: График плотности расширения Грам-Шарлье нормального распределения с кумулянтами $_1 = 2, _2 = 3, _3 = 6, _4 = 10$ (4.1a) и выборка из этого распределения (4.1b)

приведённых рисунков видно, что распределение действительно скошенно

вправо и имеет более тяжёлые хвосты, чем нормальное. Таким образом, представляется оправданным моделирование речевых признаков с помощью распределений такого рода.

Для того, чтобы получить выборку из распределения с данной плотностью представляется удобным воспользоваться методом Монте-Карло по схеме марковской цепи (Monte-Carlo Markov's Chain, MCMC, см. ниже). Этот метод позволяет получить выборку из сложных (возможно многомерных) распределений. Метод основан на построении такой марковской цепи, предельным распределением которой является такое распределение, выборку из которого необходимо получить.

Наиболее популярными способами для порождения марковской цепи является:

1. Алгоритм Метрополиса - Хастингса [72], использующий вспомогательное распределение для симуляции случайного блуждания и метода для фильтрации полученных значений. Метод обладает тем недостатком, что существует необходимость предварительно выбрать вспомогательное распределения таким образом, чтобы эффективно получать наблюдения из исследуемого распределения.

2. Алгоритм Гиббса [73]. Чаще всего используется при моделировании многомерных распределений. Представляет собой модификацию алгоритма Метрополиса - Хастингса. Идея метода заключается в том, чтобы получать выборку из условных распределений, так как это проще, чем получение маржинальных распределений, путем интегрирования совместного. Недостатком этого метода может служить то, что нам нужно уметь получать выборку из нестандартного распределиния, что может быть сложно.

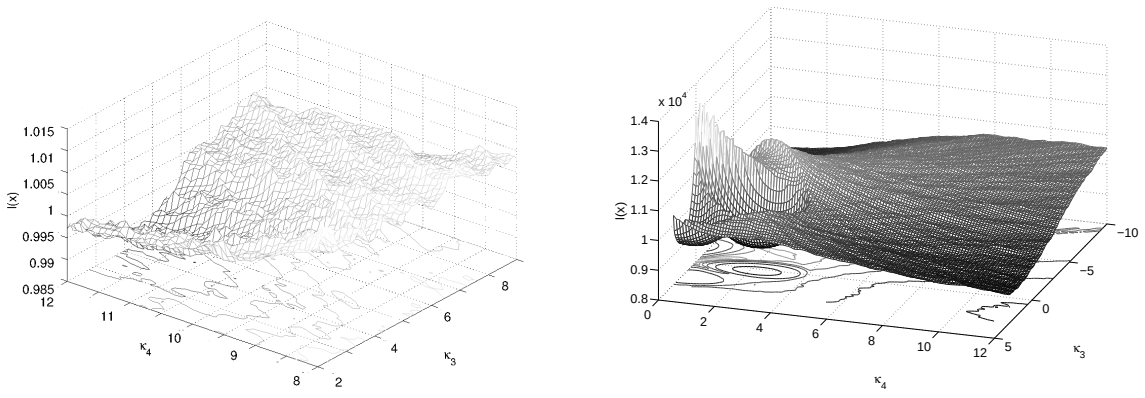
3. Алгоритм срезов [74]. Идея метода состоит в последовательном равномерном сэмплировании точек из области под плотностью. Таким образом, можно последовательно получать выборки сначала в “вертикальном“ направлении, а потом “горизонтальном“, соответствующем “вертикальному“ срезу.

При применении методов Монте-Карло по схеме марковской цепи следует обращать внимания на следующие сложности, возникающие из-за того, что нам бы хотелось получать такие наблюдения, которые обладают свойствами независимости и одинаковой распределённости.

- Эффективность метода заключается в том, насколько быстро построенная цепь сойдётся к своему предельному распределению. Поскольку первые полученные значения будут представлять плохое приближение к исследуемому распределению, то необходимо выбросить из рассмотрения некоторое количество наблюдений. Сколько именно - зависит от эффективности алгоритма генерации марковской цепи. Кроме того, важные участки распределения (например, разные моды у мультимодальных распределений) будут пропущены.

- Сгенерированные наблюдения могут быть автокоррелированы. Для устранения этого обычно для исследований оставляют только одно из нескольких последовательных наблюдений.

Для получения выборки воспользуемся алгоритмом срезов. Это представляется удобным, так как нам известен параметрический вид плотности того распределения, из которого генерируется выборка. В ходе симуляций из выборки были исключены первые 5000 наблюдений, для того, чтобы марковская цепь сошлась к своему стационарному распределению. Для устранения автокорреляций в выборке, сгенерированной марковской цепью,



(a) (b)

Рис. 4.2: Вид отрицательной функции правдоподобия при фиксированных $c_1 = 0, c_2 = 1$ (4.2a) и $c_1 = 2, c_2 = 3$ (4.2b)

в итоговую выборку попадало каждое 5-ое сгенерированное значения. Всего было сгенерировано 2000 наблюдений.

Для отыскания оценок методом максимального правдоподобия необходимо найти решение задачи (2.15). В данном случае эта задача об отыскании экстремума функции 4 переменных. Особый интерес представляют оценки параметров c_3 и c_4, описывающие ассиметрию и эксцесс распределения. Из представленных на рис. 4.2 графиков видно, что в случае некорректных значений параметров c_1 и c_2 целевая функция имеет очень много локальных минимумов и не является ни гладкой, ни выпуклой книзу. При этом, даже при верных значениях c_1 и c_2 (рис. 4.2b) функция имеет локальные экстремумы. Стоит также отметить, что в реальном случае, когда наблюдения будут зашумлены, целевая функция может иметь гораздо более сложную структуру.

Теперь посмотрим на оценки параметров, полученные различными оптимизационными алгоритмами.

Первый алгоритм, из исследованных в работе, это алгоритм градиентного спуска. Выбор этого алгоритма является обычным при решении задачи оптимизации и применяется во многих сферах. Метод градиентного спуска основывается на численной оценке градиентов целевой функции. Следствием

этого является то, что качество работы алгоритма может быть улучшено, если градиент может быть вычислен аналитически. К преимуществам алгоритма можно отнести то, что в случае гладкой и выпуклой целевой функции он приведёт к оптимальному значению за небольшое число итераций. К сожалению, последовательность значений может не сойтись к глобальному оптимуму, так как алгоритм не в состоянии отличить глобальный оптимум от локального. Более того, алгоритм может не дать финальные оценки параметров за разумное число шагов или дать бесконечные значения параметров. Таким образом, алгоритм не может применяться из-за нелинейности функции или ограничений, ведущих к не дифференцируемости.

Второй алгоритм - симплекс алгоритм Нелдера - Мида [75]. Данный алгоритм не использует знания о градиенте, таким образом, целевые функции с отклонениями от гладкости и даже функции с разрывами могут быть оптимизированы с помощью этого метода. С другой стороны, алгоритм делает большое количество вычислений значения функции, что может быть важно в случае, когда вычисление значения функции является затратным с точки зрения вычислительных ресурсов.

Третий алгоритм - метод симуляции отжига [62]. Этот метод также не требует вычислений градиента для отыскания оптимального значения целевой функции. Идея алгоритма заключается в использовании аналогии между процессом охлаждения металла с последующим его застыванием в точке с минимальной энергией и процессом поиска минимума в более общих системах. Для работы алгоритма нам нужно определить два распределения: одно, порождающее новые точки - кандидаты, и второе - принимающее или отвергающее данную точку в качестве приближения к оптимуму. Оба этих распределения зависят от температуры, то есть меры размаха. Таким образом, в случае, если алгоритм оказался в локальном экстремуме, то существует такая

вероятность $p > 0$, что алгоритм покинет эту точки, в то же время, всегда есть ненулевая вероятность того, что данная реализация траектории поиска ведёт к точке, доставляющей целевой функции значение, сильно отличающееся от оптимального. Это значит, что поверхность целевой функции исследуется в двух направлениях, а не только в одном (в случае максимизационной задачи - увеличения значения функции). Поэтому алгоритм в меньшей степени зависит от начальных данных. Он обладает тем же недостатком, что и алгоритм Нелдера - Мида. В отличие от двух предыдущих методов, метод симуляции отжига является стохастическим алгоритмом, то есть для него необходимо проанализировать распределение полученных оценок.

Рассмотрим результаты численной оптимизации. В таблице 4.2 приведены оценки параметров, полученные с помощью методов градиентного спуска, Нелдера-Мида и симуляции отжига (при ограничении в 5000 вычислений значения функции). В качестве начальных значений во всех случаях использовались эмпирические кумулянты выборочного распределения. Видно, что все алгоритмы хорошо справились с оцениванием первых двух кумулянтов: истинные значения параметров лежат в соответствующих доверительных интервалах и стандартные ошибки (указаны в скобках) достаточно маленькие.

Таблица 4.2: Оценки параметров, полученные разными численными методами

Параметр	Метод градиентного спуска	Метод Нелдера - Мида	Метод симуляции отжига
$c_1 = 2$	2.04 (0.07)	2.02 (0.07)	1.97 (0.07)
$c_2 = 3$	3.01 (0.05)	3.01 (0.05)	2.94 (0.05)
$c_3 = 6$	5.4 (0.84)	5.38 (0.85)	5.35 (0.84)
$c_4 = 10$	3.82 (5.1)	6.03 (5.12)	9.65 (5.84)

Ситуация несколько хуже с третьим кумулянтом. Истинное значение по-прежнему в доверительном интервале, но стандартные ошибки сильно

увеличились. Стоит отметить, что в данном случае для вычисления стандартных ошибок использовался обратный гессиан отрицательной функции правдоподобия в качестве асимптотической матрицы ковариаций в 2.20. Кроме того, гессиан был получен численно.

Из таблицы видно, что алгоритмы достаточно плохо справляются с оцениванием c_4: стандартные ошибки очень велики и только метод симуляции отжига дал оценку близкую к истинному значению параметра. Более того, оценки получились незначимыми на 95% уровне. Следует отметить, что в реальных ситуациях, когда наблюдения будут искажены шумами, целевая функция может иметь гораздо более сложную структуру, и результаты оценивания могут быть ещё более ненадёжными.

Результаты численной оптимизации подтверждают, что подобные функции правдоподобия крайне плохо оптимизируются с помощью численных методов. Большие стандартные ошибки могут быть следствием неустойчивости гессиана, и могут быть усилены тем, что сам гессиан плохо обусловлен и применение обратной операции даёт неточные результаты. Всё это говорит в пользу использования аналитически вычисленной формы гессиана для получения стандартных ошибок.

Для того, чтобы глубже исследовать свойства полученных оценок, полученных методом симуляции отжига, построим их эмпирические функции распределения. Для этого повторим процесс оценивания 1000 раз. Эмпирические функции распределения представлены на рис. 4.3.

Полученные эмпирические функции распределения находятся в соответствии с результатами одиночного оценивания. Средние значения параметров c_1 и c_2 находятся близко к настоящим значениям параметров с маленькими стандартными ошибками. Эти оценки обладают гораздо лучшими свойствами, чем оценки двух других параметров. Снова получились

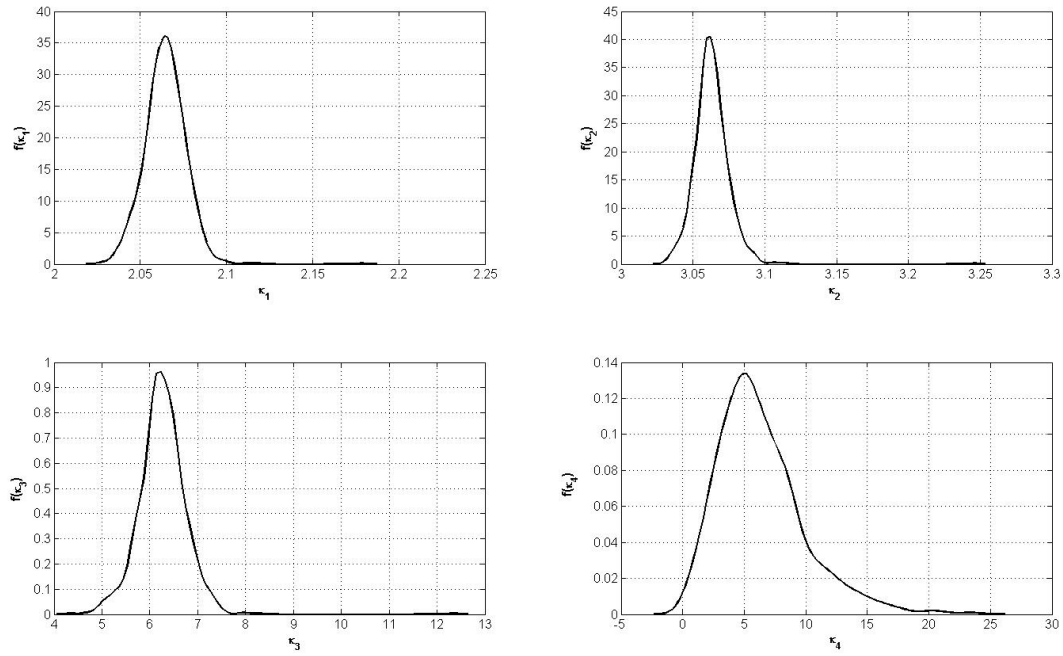


Рис. 4.3: Эмпирические функции распределения оценок параметров, полученных методом симуляции отжига.

сравнительно большие стандартные ошибки для параметра c_3, в то же время его среднее близко к настоящему значению параметра. Стоит отметить тот факт, что даже в случае большого количества перезапусков мы не можем получить надёжные оценки параметра c_4.

Вместе с тем, оценивание моментов высоких порядков всегда сопряженно с трудностями. В нашем случае, вероятно, не представляется возможным различить две плотности с различными значениями параметра c_4. Мотивацией этому служит тот факт, что параметр c_4 характеризует куртозис распределения и для оценки параметра достаточное количество наблюдений должно попасть в хвосты распределения.

4.2.2 Эксперименты с реальными данными

Монте - Карло эксперименты позволяют проанализировать качество работы алгоритма и выявить особенности его применения. Вместе с тем крайне важно протестировать работу алгоритма на реальных данных.

Будем использовать методику, предложенную в работе [76], в которой авторы использовали смесь гауссовских распределений (GMM, Gaussian Mixture Models) для моделирования профилей из AIM. Авторы предлагают использовать параметры смесей как входные признаки для системы распознавания речи.

В диссертации был модифицирован метод [76] в том, чтобы использовать смесь распределений Грам-Шарлье нормального распределения для моделирования NAP профиля.

Воспользуемся методами этих работ и проведём эксперимент и анализ модифицированного алгоритма симуляции отжига с использованием параллельных вычислений по следующим схемам, подробно описанным в части 2.3.3:

- Первая модификация предполагает использование гибридной схемы оптимизации. Процесс отжига начинается из разных точек в пространстве поиска и дальше независимо строятся пути поиска для каждой начальной точки. После того, как каждая ветвь вычислений будет закончена, сравниваются полученные результаты, и выбирается те значения, которые дают лучшее значение целевой функции при заданных ограничениях.

- Во второй модификации предлагается предварительное вычисление начальных точек. Таким образом, представляется возможным получить лучшую начальную точку, и последовательность значений может сойтись к решению быстрее.

На рис. 4.4 приведена типичная подгонка NAP профиля с помощью распределения Грам-Шарлье. Можно заметить, что подгонка почти идеальна: все пики корректно определены. Кроме того, видно, что поведение на хвостах нашей смеси распределений близко к тому, что наблюдается у профиля. Этот факт может служить подтверждением тому, что несмотря на относительно плохую работу алгоритма в Монте-Карло симуляциях, полученные оценки параметров достаточно точные, чтобы хорошо приблизить профиль.

Также было проведено сравнение две предложенные модификации алгоритма по скорости выполнения и точности подгонки.

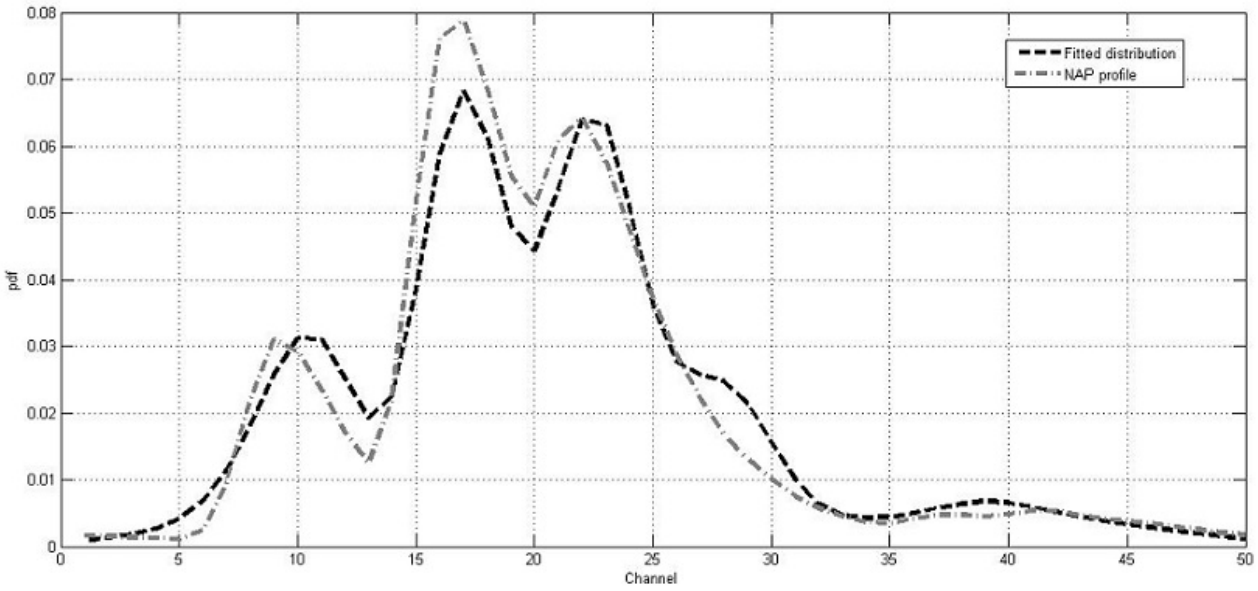


Рис. 4.4: Подгонка распределения Грам-Шарлье к NAP профилю.

В таблице 4.3 приведена функция времени исполнения обоих алгоритмов в зависимости от количества процессоров (вычисления выполнялись на машине с процессорами Intel Core i7 с частотой 3.2 Гц). Как и следовало ожидать, первый алгоритм оказался более затратным по времени. При этом, зависимость времени работы алгоритмов от количества процессоров почти линейная. Таким образом, при имплементации алгоритма на графических картах, даже первый алгоритм, может работать в реальном времени. Тем не менее, не следует ожидать, что эта тенденция сохраниться в случае увеличения

Таблица 4.3: Время работы алгоритмов.

Количество процессоров	Время 1го алгоритма, с	Время 2го алгоритма, с
1	9756	15
3	4465	13
6	2463	11

количества процессоров. С увеличением их количества придётся принимать во внимание зависимость между путями поиска разных веток алгоритма для того, чтобы запретить алгоритму проходить один и тот же путь несколько раз. Таким образом, в случае большого количества процессоров будет проявляться зависимость по данным. Но в случае второго алгоритма этой зависимостью можно пренебречь. Вследствие этого на первый план выходит точность подгонки распределения.

Представляется также важным сравнение точности подгонки двух предложенных алгоритмов. В качестве меры используется расхождение Кульбака - Лейблера [77] и значение логарифмической функции правдоподобия.

Расхождением Кульбака-Лейблера $D_{KL}(P||Q)$ между двумя распределениями P и Q с плотностями $p(x)$ и $g(x)$, которая вычисляется по формуле

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} \ln \frac{p(x)}{g(x)} p(x) \mathrm{d}x$$

Значения логарифмической функции правдоподобия напрямую показывает, насколько близки к оптимуму полученные значения. Результаты сравнения приведены в таблице 4.4.

Таблица 4.4: Точность подгонки алгоритмов.

Мера подгонки	Результат 1го алгоритма	Результат 2го алгоритма
Расхождение Кульбака - Лейблера	0.117	0.140
Лог-правдоподобие	-32935	-34289

Как видно из таблицы 4.4, первый алгоритм даёт лучшую степень подгонки: расхождение Кульбака-Лейблера меньше и значение логарифмической функции правдоподобия больше. Тем не менее, преимущество первого алгоритма нельзя назвать большим.

Из сравнения результатов работ по алгоритму 1 и 2 и соображений относительно возможного ухудшения производительности, указанного выше, можно сделать вывод о том, что второй алгоритм, хотя и слегка менее точный, является более предпочтительным при реальном применении.

4.3 Способы определения языка по искаженному сообщению

В разделе 1.6.5 отмечалось, что на точность работы систем распознавания речи оказывает сильное влияние акцент диктора, и в особенно то, является ли язык сообщения родным для диктора. Из приведённых результатов исследований можно сделать вывод о важности распознавания языка сообщения при разработке систем распознавания речи, например, для использования “родного“ распознавателя, то есть такого распознавателя, который обучался на речевых сообщениях того же языка, что и тот, на котором говорит диктор.

В этой части описывается методология идентификации языка на примере распознавания языка текста и применение её к идентификации языка при речевом сигнале.

4.3.1 Использования SVM для идентификации языка

При проведении исследований был предложен способ применения SVM для идентификации языка искаженного текста [6]. Приведём основные результаты этой работы.

Идентификации языка по своей формулировке похожа на задачу распознавания диктора: требуется построить статистический критерий разделения векторов в пространстве признаков для конечного числа простых гипотез в случае закрытой задачи (то есть, количество языков известно априорно) или для конечного числа простых и одной сложной гипотезы (сообщение неизвестного языка) в случае открытой задачи.

Рассматривается закрытая задача идентификации английского, испанского, польского и французского языков (предполагается, что тестируемый текст написан на одном из рассматриваемых языков). Эксперименты проводятся с текстами рассказов А. Конан-Дойля о Шерлоке Холмсе, записанными в латинице без пробелов и знаков препинания в одном регистре. Тексты разбиваются на два подмножества: обучающее (около 150000 символов для каждого языка) и тестовое (около 120 векторов длиной 1000 символов для каждого языка).

Тексты подвергались искажениям типа ”замена”. С вероятностью $P_{\text{замена}}$ буква исходного текста заменяется любой другой буквой латинского алфавита (выбор замещающей буквы осуществляется случайно равновероятно).

В качестве признаков используются относительные частоты встречаемости символов и биграмм текста. Методология использования SVM подробно изложена в части 2.2.1. При проведении экспериментов применялось гауссовское ядро (см. 2.8), где параметр γ подвергался оптимизации методом градиентного спуска на обучающем множестве с помощью подхода кросс-валидации [78]. Как показали исследования (см., например, [79, 80]), использование гауссова ядра является предпочтительным по сравнению с линейным и полиномиальным.

Переход к многоклассовой классификации осуществлялся методом “каждый против каждого” (one vs. one), как и в случае с дикторами. В пограничных

случаях, когда несколько классов набрали одинаковое количество голосов распознавателей, победителем признается класс с наименьшим номером. Такой подход не является самым лучшим, но очень прост и часто используется на практике.

4.3.2 Результаты экспериментов. Тексты

Под точностью идентификации понимается отношение количества текстов с правильно идентифицированным языком к общему количеству текстов. Результаты экспериментов приведены на графиках зависимости точности идентификации от вероятности искажения. Под точностью идентификации понимается отношение количества текстов с правильно идентифицированным языком к общему количеству текстов. Результаты экспериментов приведены на графиках зависимости точности идентификации от вероятности искажения.

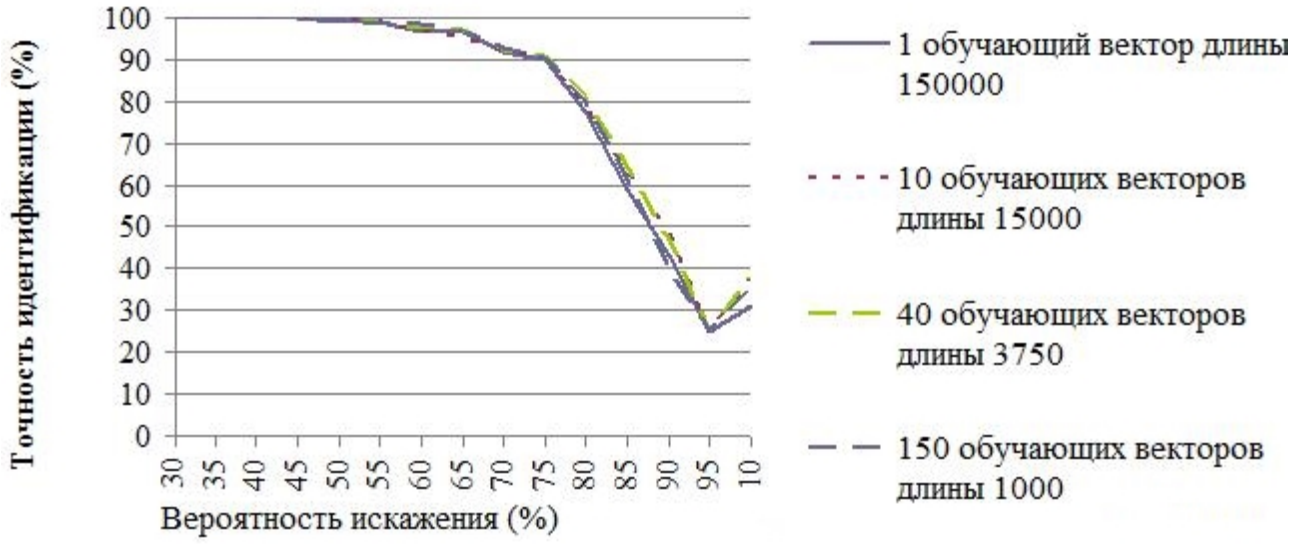


Рис. 4.5: Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z и количества обучающих векторов. Значковые статистики

Как видно из рис. 4.6, в рамках проведенных экспериментов на точность идентификации оказывает существенное влияния выбор количества обучающих векторов (и, соответственно, их длины). Объяснение этого

Рис. 4.6: Точность идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения $P_{\text{замена}}$ и количества обучающих векторов. Биграммные статистики.

результата следующее. Рассмотрим обучающие вектора как точки в пространстве признаков и предположим для простоты, что классов всего два и они линейно разделимы. Точки образуют области сложной формы. Малое количество обучающих векторов не позволяет судить о форме области, поэтому разделяющая гиперплоскость может быть проведена неоптимальным образом, как показано на рис. 4.7.

Увеличение числа обучающих векторов предоставляет больше информации о формах областей, позволяя проводить разделяющую гиперплоскость с их учётом, что увеличивает точность идентификации. При этом следует иметь в виду, что при ограниченном обучающем множестве такой подход приводит к уменьшению размера фрагментов текста, по которым формируются обучающие векторы. На маленьких фрагментах текста слабее проявляются статистические свойства, присущие данному языку. Это, в свою очередь, приводит к падению точности идентификации, что наблюдается в результатах экспериментов, показанных на рис. 4.6.

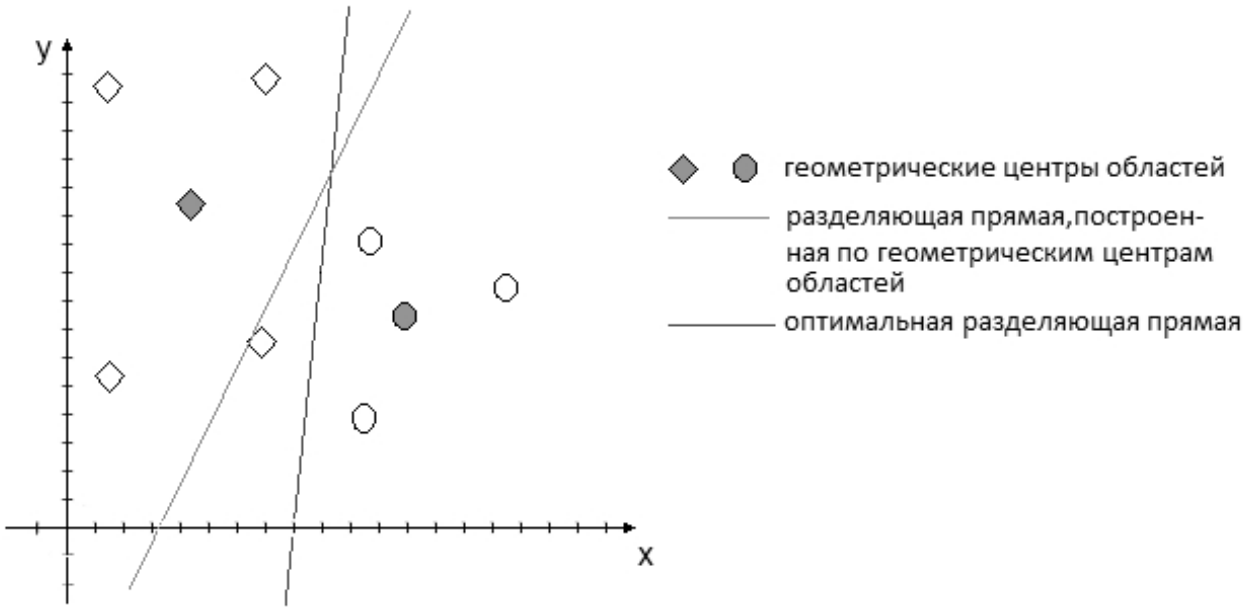


Рис. 4.7: Разделяющие прямые, построенные по обучающим векторам и геометрическим центрам областей.

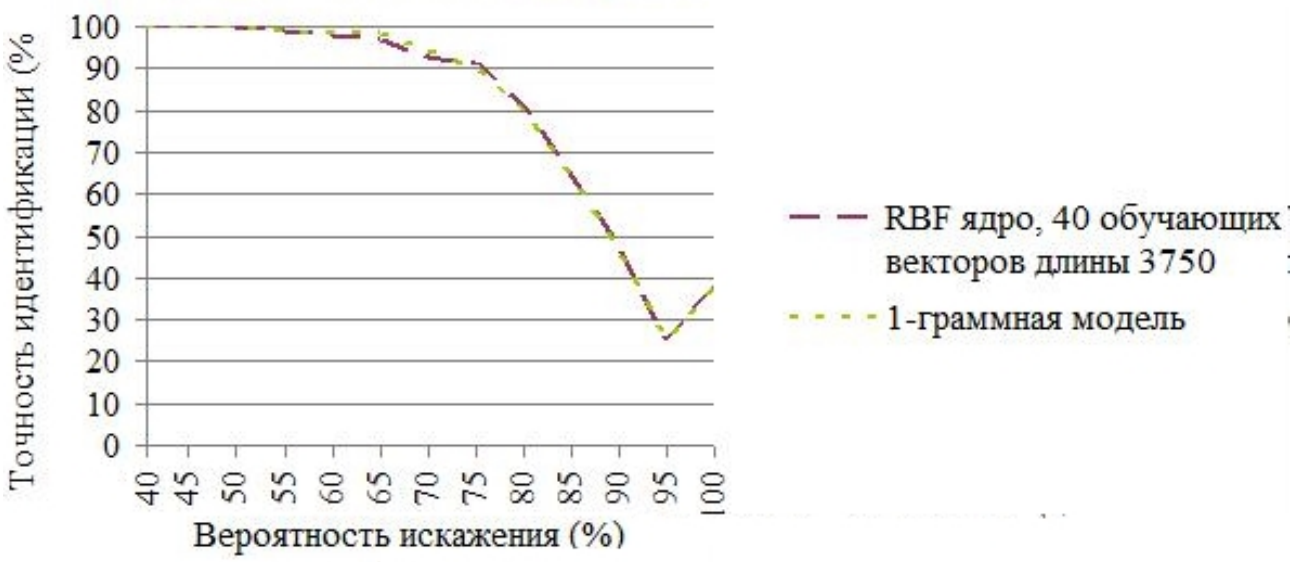


Рис. 4.8: Сравнение точности идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z, полученные с применением модели SVM и мультиграммных моделей. Значковые статистики.

Результаты, полученные с использованием SVM и мультиграммных моделей [81], представлены на рис. 4.8 и 4.9. Можно заметить, что на биграммных статистиках использование SVM может привести к увеличению точности идентификации, на значковых статистиках результаты сопоставимы.

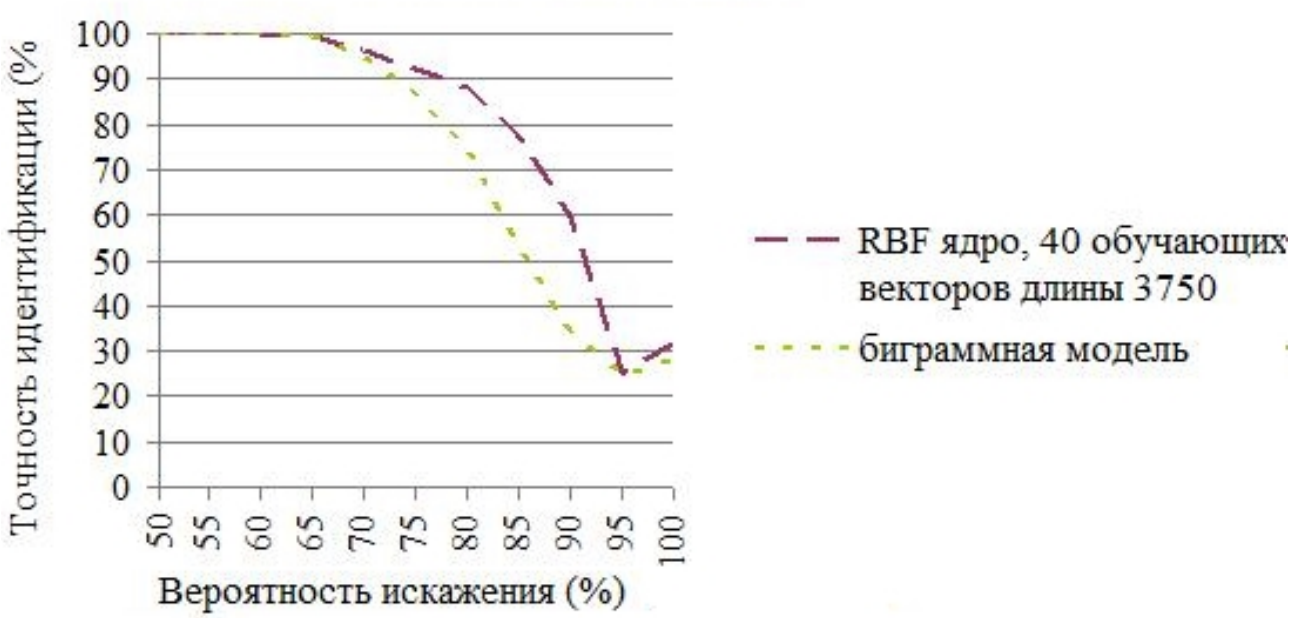


Рис. 4.9: Сравнение точности идентификации языка фрагментов текста длиной 1000 символов в зависимости от вероятности искажения P_z, полученные с применением модели SVM и мультиграммных моделей. Биграммные статистики.

В результате проведённой работы было установлено, что на биграмных статистиках SVM более эффективен, чем мультиграммные модели.

Кроме того, было установлено, что при идентификации языка искаженного художественного текста использование SVM с RBF ядром на биграммных статистиках даёт большую точность, чем мультиграммные модели.

4.3.3 Результаты экспериментов. Речь

Изложенный в предыдущей части подход был применён к распознаванию языка по предварительно распознанной речи следующим образом. Сначала к звуковому сообщению применялся распознаватель, в качестве которого использовался стандартный алгоритм на Скрытых Марковских Моделях (см. 2.1). В результате работы распознавателя получалась последовательность фонем, к которой и применялся описанный ранее алгоритм.

При использовании метода на текстах нам приходилось искусственно вводить искажения, в случае речевого сигнала подобные искажения появляются

вследствии работы распознавателя, который может неправильным образом распознать фонему.

При проведении экспериментов по распознаванию языка использовалась речевая база, содержащая данные по языкам (арабскому, английскому, мандарину, польскому, русскому и турецкому), записанным в телефонном канале. Характеристики входного сигнала: битрейт 16, частота дискретизации 8 кГц, соотношение сигнал – шум в среднем 15дб. Продолжительность речи на каждом языке составляло примерно 120 минут. Тестовое множество для каждого языка состояло из 500 векторов.

Для построения признаков также использовались биграммные статистики, при этом каждый вектор строился по 100 фонемам. В таблице 4.5 приведены зависимости точности распознавания от длины речевого сообщения (то, есть от величины обучающего множества).

Таблица 4.5: Точность распознавания языка (в процентах).

	Длительность, с							
	5		10		20		40	
	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.
Точность	34.1	48.4	22.7	46.9	28.4	33.7	25.7	36.9

Результаты экспериментов показывают преимущество в точности распознавания при применении предложенных методов.

4.4 Выводы

В данной главе представлены результаты экспериментов с реальными и симулированными данными. В главе получены следующие результаты:

1. Представлен практический способ подбора параметра распознавателя для модели идентификации диктора.
2. Представлены и объяснены результаты применения алгоритма опорных векторов с Фишеровскими признаками для построения дикторонезависимых систем в различных каналах связи.
3. Произведено сравнение различных алгоритмов генерации выборок по схеме марковской цепи.
4. Рассмотрены результаты Монте-Карло экспериментов по сравнению различных численных методов для получения оценок максимального правдоподобия.
5. Построены эмпирические функции распределения оценок, полученные методом симуляции отжига.
6. На основе реальных данных доказана практическая применимость параллельных модификаций алгоритма симуляции отжига для получения дикторонезависимых признаков.
7. Представлен способ определения языка искаженного сообщения на основе алгоритма опорных векторов.
8. Дана геометрическая интерпретация эмпирической зависимости точности распознавания от величины обучающего множества.

Заключение

Основные результаты работы заключаются в следующем.

1. Проведён анализ существующего состояния в сфере распознавания языка и диктора.
2. Выявлены дикторонезависимые признаки, основанные на 4-х параметрическом распределении, и доказана их оптимальность.
3. Разработана модификация алгоритма симуляции отжига, увеличивающия быстродействие системы при получении дикторонезависимых признаков.
4. Разработана и теоретически обоснована модифицикация метода опорных векторов, основанная на применении фишеровских ядер, которая позволяет увеличить точность распознавания диктора.
5. Проведён сравнительный анализ алгоритмов оптимизации для получения дикторонезависимых признаков по скорости и точности.
6. Разработаны и теоретически обоснованы методы и алгоритмы получения параметров классификатора для решения задач идентификации языка и диктора.
7. Создана программная реализация разработанной системы идентификации языка и диктора, фрагменты который внедрены на производстве.
8. Проведены экспериментальные исследования по оценке точности распознавания и быстродействию системы идентификации языка и

диктора, которые показали преимущества разработанных методов по сравнению с применяемыми ранее.

Список рисунков

Список таблиц

Литература

1. Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition. 1989. P. 257–286.

2. Larochelle H., Erhan D., Courville D. An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation // International Conference on Machine Learning. 2007.

3. Mermelstein P. Distance measures for speech recognition, psychological and instrumental // Pattern recognition and artificial intelligence. 1976. Vol. 116. P. 374–388.

4. Grerzl F. Probabilistic and bottle-neck features for the BN features LVCSR of meetings // In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2007. P. 4729–4732.

5. Tou J., Gonzalez R. Pattern Recognition Principles. Addison Wesley, 1974.

6. Ермилов А.В. Распознавание языка искаженного текста методом опорных векторов // Вестник РУДН. Серия Математика, Информатика, Физика. 2012. Т. 2. С. 126–130.

7. Ермилов А.В. Моделирование речевых признаков с помощью алгоритма симуляции отжига // Вестник РУДН. Серия Математика, Информатика, Физика. 2014. Т. 2. С. 354–358.

8. Гостев И.М., Ермилов А.В. О применении Фишеровских ядер в задаче распознавания диктора // Известия Юго-Западного Государственного Университета. Серия Вычислительная Техника, Информатика, Медицинское приборостроение. 2011. Т. 2. С. 15–20.

9. Ermilov A. V. Speech Technologies in human computer interactions // International Journal of Modern Manufacturing Technologies. 2013. Vol. 4. P. 52–57.

10. Ермилов А.В. Параллельные технологии в задаче максимизации правдоподобия // Труды 5-ой Международной конференции ”Распределенные вычисления и грид-технологии в науке и образовании”. 2012. С. 302–305.

11. Ermilov A. V. Parallel Technologies in maximum likelihood estimation // Book of Abstracts of 5th International Conference “Distributed Computing and Grid-Technologies in Science and Education”(GRID-2012). 2012. p. 99.

12. Ermilov A. V. Fisher Kernels for speaker recognition // Book of Abstracts of Second International Scientific Symposium “Modeling Of Nonlinear Processes And Systems (MNPS-2011)”. 2011. p. 291.

13. Ermilov A. V. Speech technologies in human-computer interactions // Book of Abstracts of the First International Conference on Modern Manufacturing Technologies in Industrial Engineering “ModTech - 2013”. 2013. p. 197.

14. Ermilov A. V. Modeling of speech features via simulated annealing algorithm // Book of Abstracts of the international Conference ”Mathematical Modeling and Computational Physics - 2013”(MMCP’2013). 2013. p. 76.

15. Ермилов А.В. Применение расширения Грам-Шарлье для моделирования речевых признаков // Сборник материалов X Международной научно-технической Конференции “Оптико-электронные приборы и устройства в

системах распознавания образов, обработки изображений и символьной информации. Распознавание - 2012". 2012. с. 97.

16. Ермилов А.В. Математическая модель параллельных вычислений в системе автоматического распознавания речи // Сборник материалов XI Международной научно-технической Конференции "Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации. Распознавание - 2013". 2013. с. 252.

17. Батуев А.С. Физиология высшей нервной деятельности и сенсорных систем: Учебник для вузов. СПб.: Питер, 2005.

18. Makhoul J. Linear prediction: A tutorial review // Proceedings of the IEEE. 1975. Vol. 63, no. 4. P. 561–580.

19. Davis S., Mermelstein P. Experiments in syllable-based recognition of continuous speech // IEEE Transcactions on Acoustics, Speech and Signal Processing. 1980. Vol. 28. P. 357–366.

20. Wölfel M., McDonough J. Distant speech recognition. John Wiley & Sons, 2009.

21. Howard D., Angus J. Acoustics and psychoacoustics. Taylor & Francis, 2009.

22. Тихонов А. Н., Самарский А. А. Уравнения математической физики. Изд-во Моск. ун-та М., 1999.

23. Deller J., Proakis J., Hansen J. Discrete-time processing of speech signals. Wiley, 2000.

24. Rabiner L., Juang B.-H. Fundamentals of speech recognition. Prentice-Hall, Inc., 1993.

25. Nemer E., Goubran R., Mahmoud S. Speech enhancement using fourth-order cumulants and optimum filters in the subband domain // Speech Communication. 2002. Vol. 36, no. 3. P. 219–246.

26. Salavedra J., Masgrau E., Moreno A. Robust coefficients of a higher order AR modelling in a speech enhancement system using parameterized Wiener filtering. 1994. P. 69–72.

27. Rao C. R. Linear Statistical Inference and Its Applications. Second edition. Wiley, 1973.

28. Fletcher H. Auditory patterns // Reviews of modern physics. 1940. Vol. 12, no. 1. p. 47.

29. Stevens S., Volkmann J., Newman E. A scale for the measurement of the psychological magnitude pitch // The Journal of the Acoustical Society of America. 1937. Vol. 8. p. 185.

30. Moore B. Frequency selectivity in hearing. Academic Press London, 1986.

31. Moore B., Glasberg B. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns // The Journal of the Acoustical Society of America. 1983. Vol. 74. p. 750.

32. Franke J. A Levinson-Durbin recursion for autoregressive-moving average processes // Biometrika. 1985. Vol. 72, no. 3. P. 573–581.

33. Oppenheim A. V., Schafer R. W. Discrete-Time Signal Processing. Prentice Hall, 2009.

34. Churchill R., Brown J. Complex Analysis and Applications. 1990.

35. Mokhtari P. An acoustic-phonetic and articulatory study of speech-speaker dichotomy // In proceeding of 3rd European Conference on Speech Communi-

cation and Technology, EUROSPEECH 1998 - INTERSPEECH 1998. 1998. P. 1555–1558.

36. Tuerk C., Robinson T. A new frequency shift function for reducing inter-speaker variance. // In Proceedings of 3rd European Conference on Speech Communication and Technology, EUROSPEECH 1993. 1993. p. 351–354.

37. Nolan F. The phonetic bases of speech recognition. Cambridge University Press, 2009.

38. Huang C. et al. Analysis of speaker variability. // In proceeding of 4th European Conference on Speech Communication and Technology, EUROSPEECH 2001 - INTERSPEECH 2001. 2001. P. 1377–1380.

39. Lawson A., Harris D., Grieco J. Effect of foreign accent on speech recognition in the NATO n-4 corpus. // In proceeding of 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003. 2003. P. 1505–1508.

40. van Compernolle D. Recognizing speech of goats, wolves, sheep and… non-natives // Speech Communication. 2001. Vol. 35, no. 1. P. 71–79.

41. Lindblom B. Explaining phonetic variation: A sketch of the H&H theory // Speech production and speech modelling. 1990. P. 403–439.

42. Kuwabara H. Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. // In proceeding of 2nd European Conference on Speech Communication and Technology, EUROSPEECH 1997 - INTERSPEECH 1997. 1997. P. 1355–1358.

43. Roddy C., Randolph R. Describing the emotional states that are expressed in speech // Speech communication. 2003. Vol. 40, no. 1. P. 5–32.

44. Cohen J., Kamm T., Andreou A. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability // The Journal of the Acoustical Society of America. 1995. Vol. 97. p. 3246.

45. Ono Y., Wakita H., Zhao Y. Speaker normalization using constrained spectra shifts in auditory filter domain // In Proceedings of 3rd European Conference on Speech Communication and Technology, EUROSPEECH 1993. 1993. P. 3037–3040.

46. Yu S. Hidden semi-Markov models // Artificial Intelligence. 2010. Vol. 174, no. 2. P. 215–243.

47. Baum L., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains // The annals of mathematical statistics. 1966. Vol. 37, no. 6. P. 1554–1563.

48. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm // Information Theory, IEEE Transactions on. 1967. Vol. 13, no. 2. P. 260–269.

49. Аграновский А.В., Леднов Д.А. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. М.: Радио и связь, 2004.

50. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach. 3rd edition. Prentice Hall, 2009.

51. Cortes C., Vapnik V. Support-vector networks // Machine learning. 1995. Vol. 20, no. 3. P. 273–297.

52. Jaakkola T., Haussler D. et al. Exploiting generative models in discriminative classifiers // Advances in neural information processing systems. 1999. P. 487–493.

53. Kuhn H., Tucker A. Nonlinear programming // Proceedings of the 2nd Berkeley symposium on mathematical statistics and probability. 1951. Vol. 5. P. 481–492.

54. Hsu C.-W., Lin C. A comparison of methods for multiclass support vector machines // IEEE Transactions on Neural Networks. 2002. Vol. 13, no. 2. P. 415–425.

55. Lawera M. Predictive Inference: An Introduction // Technometrics. 1995. Vol. 37, no. 1. P. 121–121.

56. Efron B. Bootstrap methods: Another look at the jackknife // Annals of Statistics. 1979. Vol. 7.

57. McLachlan G., Peel D. Finite mixture models. Wiley, 2004.

58. Friedman J. Regularized discriminant analysis // Journal of the American statistical association. 1989. Vol. 84, no. 405. P. 165–175.

59. Munich M., Lin Q. Auditory image model features for automatic speech recognition. // In Proceedings of 9th European Conference on Speech Communication and Technology, Interspeech 2005 - Eurospeech 2005. 2005. P. 3037–3040.

60. Patterson R., Allerhand M., Giguere C. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform // The Journal of the Acoustical Society of America. 1995. Vol. 98. p. 1890.

61. Niguez T., Perote J. Multivariate semi-nonparametric distributions with dynamic conditional correlations // International Journal of Forecasting. 2011. Vol. 27, no. 2. P. 347–364.

62. Kirkpatrick S., Gelatt D., Vecchi M. P. Optimization by simmulated annealing // Science. 1983. Vol. 220, no. 4598. P. 671–680.

63. Simon H. Adaptive filter theory. 2002. Vol. 2. P. 478–481.

64. Rumbaugh J., Jacobson I., Booch G. The Unified Modeling Language Reference Manual. Pearson Higher Education, 2004.

65. Wen K., Wang J. Efficient computing methods for parallel processing: An implementation of the Viterbi algorithm // Computers & Mathematics with Applications. 1989. Vol. 17, no. 12. P. 1511–1521.

66. Noda H., Shirazi M. A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM // In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 94). 1994. Vol. 1. p. 597.

67. Bourlard H., Dupont S. A new ASR approach based on independent processing and recombination of partial frequency bands // In Proceedings of Fourth International Conference on Spoken Language (ICSLP 96). 1996. Vol. 1. P. 426–429.

68. Shriberg E., Stolcke A. Prosody modeling for automatic speech recognition and understanding // Mathematical Foundations of Speech and Language Processing. Springer, 2004. P. 105–114.

69. Gebali F. Algorithms and Parallel Computing. Wiley, 2011.

70. You K. et al. Parallel scalability in speech recognition // Signal Processing Magazine, IEEE. 2009. Vol. 26, no. 6. P. 124–135.

71. Vapnik V., Chervonenkis A. On the uniform convergence of relative frequencies of events to their probabilities // Theory of Probability & Its Applications. 1971. Vol. 16, no. 2. P. 264–280.

72. Hastings W. Monte Carlo sampling methods using Markov chains and their applications // Biometrika. 1970. Vol. 57, no. 1. P. 97–109.

73. Gelfand A., Smith A. Sampling-based approaches to calculating marginal densities // Journal of the American Statistical Association. 1990. Vol. 85, no. 410. P. 398–409.

74. Neal R. Slice Sampling // Annals of Statistics. 2003. Vol. 31, no. 3. p. 705–767.

75. Lagarias J., Reeds J., Wright M. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions // SIAM Journal of Optimization. 1998.

76. Monaghan J. et al. Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition // Journal of the Acoustical Society of America. 2008. Vol. 123, no. 5. p. 3066.

77. Kullback S., Leibler R. On information and sufficiency // The Annals of Mathematical Statistics. 1951. Vol. 22, no. 1. P. 79–86.

78. Kohavi R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 95). 1995. Vol. 14, no. 2. P. 1137–1145.

79. Joachims T. Text categorization with support vector machines: Learning with many relevant features. Springer, 1998.

80. Jalam R., Teytaud O. Kernel-based text categorisation // Proceedings of International Joint Conference on Neural Networks (IJCNN'01). 2001. Vol. 3. P. 1891–1896.

81. Кулай А.Ю., Мельников С.Ю. О точности идентификации языка искаженного текста в зависимости от степени искажения. // Вестник Московского Государственного Лингвистического Университета. Серия Языкознание. 2009. Т. 57. С. 200–209.