ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
**«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»**
(Н И У « Б е л Г У »)

ИНСТИТУТ ИНЖЕНЕРНЫХ ТЕХНОЛОГИЙ И ЕСТЕСТВЕННЫХ НАУК
КАФЕДРА ИНФОРМАЦИОННО-ТЕЛЕКОММУНИКАЦИОННЫХ
СИСТЕМ И ТЕХНОЛОГИЙ

ИССЛЕДОВАНИЕ ПРОСТРАНСТВ ПРИЗНАКОВ И МЕР БЛИЗОСТИ В ЗАДАЧАХ РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

Выпускная квалификационная работа
обучающегося по направлению подготовки
11.04.02 Инфокоммуникационные технологии и системы связи,
магистерская программа «Системы и устройства радиотехники и связи»
очной формы обучения, группы 07001636
Кравченко Данила Николаевича

Научный руководитель
кандидат технических наук, доцент,
доцент кафедры ИТСиТ Прохоренко Е.И.

Рецензент

БЕЛГОРОД 2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

По мере развития компьютерных систем, в современном мире, становится все более очевидным, что использование этих систем в повседневной деятельности человека будет расширяться. Немаловажным фактором, для развития компьютерных систем является возможность использования человеческой речи как интерфейса для работы с компьютером: управление персональным компьютером голосом в реальном времени, а также ввод и вывод информации в виде устной речи.

В настоящий момент, повсеместно, ведутся работы по созданию систем обработки устной речи, среди которых особенное место занимает распознавание речи. Система распознавания речи получает информацию об акустических колебаниях воздуха через микрофон, сравнивает полученные данные с имеющимися в системе и, в случае совпадения идентифицирует участок сигнала. Для учета вариативности и обучения моделей фонем и слов требуются большие объемы текста и речевого материала, подготовка которых требует огромных трудозатрат. Современные системы распознавания обучения на ограниченных речевых корпусах обладают рядом недостатков.

Сегодня существуют два основных вида технологий распознавания речи. Один из них — это распознавание речи, зависящей от диктора, т. е. пользователь должен сначала научить систему распознавать его голос, и только после этого система может функционировать. Второй — это распознавание речи, не зависящее от диктора, т. е. система способна распознать любую речь, независимо от того, кто говорит. Системы распознавания изолированных слов работают с дискретными словами — в этом случае требуется пауза между словами.

Существующие системы распознавания созданы для работы с английским, немецким, испанским и другими популярными языками и малоприменимы в работе с русской речью. Это связано с тем, что русский

язык принципиально отличается от других языков не только фонетически, но и свободным порядком слов в предложении, что значительно усложняет математическое языковой модели. Важной задачей при разработке систем распознавания речи, является выделение таких признаков, которые бы обладали 1) свойством инвариантности на отрезках, полученных при произнесении одних и тех же звуков русской речи, 2) и вариативности на отрезках, содержащих разные звуки. Кроме того, требуются подходы к сравнению данных признаков – т.е. определение адекватных мер близости.

В основе многих из разработанных подходов [25,26,49,50] используются частотные представления, так как порождаемые звуками речи отрезки РС обладают свойством концентрации энергии в достаточно узких полосах частотной оси. В связи с этим можно упомянуть рассматриваемое в литературных источниках разбиение частотной полосы на так называемые критические полосы слуха, которые опосредованно отражаются на частотных свойствах РС.

Необходимо отметить, что предлагаемые в настоящее время методы распознавания РС на основе анализа их частотных свойств, в качестве признакового пространства, либо не отражают свойства концентрации энергии, либо недостаточно точно отображают характер изменения энергии в речевом сигнале.

Целью работы является определение важных, с точки зрения решения задачи распознавания речи – признаков речевых сигналов и мер их близости.

1. Для достижения цели необходимо решить следующие задачи:

2. Проанализировать особенности обработки речевых сигналов в задачах распознавания речи;

3. Изучить существующие методы представления речевых сигналов в задачах распознавания и провести их сравнительный анализ;

4. Изучить меры близости, применяемые для сравнения признаков речевых сигналов в системах распознавания речи;

5. Определить важные для задачи распознавания речи характеристики речевого сигнала.

ГЛАВА 1 АНАЛИЗ ПРОБЛЕМЫ РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

1.1 Современное состояние направления распознавания речевых сигналов

С начала развития человечества до момента экспансии современных медиа технологий, человеческая речь остается основным видом коммуникационный деятельности людей и используется для обмена информации и социального взаимодействия каждый день. В виду процессов компьютеризации и информатизации общества, речь используется во многих технологических аспектах социального и информационного взаимодействия человека будь то: телефонные переговоры, радио, телевидение, Интернет. Это говорит о безусловной важности речевых технологий для общества, так и о ведущем месте речи – как средства коммуникации для человека в целом.

На данный момент наблюдается необходимость развития методов и алгоритмов распознавания речевых сигналов для широкого спектра задач и приложений: 1) речевые командные системы (управление транспортом, роботами, и т.п.), 2) системы перевода речи в текст, 3) системы синхронного перевода, 4) системы помощи людям с ограниченными возможностями; 5) умная телефония – голосовой набор, голосовое меню; и многое другое.

Процесс распознавания речи можно описать следующей моделью, представленной на рисунке 1.1. Сообщение (W) сформированное источником сообщений передается на источник речевого сигнала, где формируется акустический сигнал (x), соответствующий кодируемому сообщению. Речевой сигнал (x) передается через коммуникационный канал (в рассматриваемом случае – идеальный) поступает на приемник, где происходит его обработка с целью получить сообщение (W), которое, должно соответствовать переданному сообщению (W). [50]

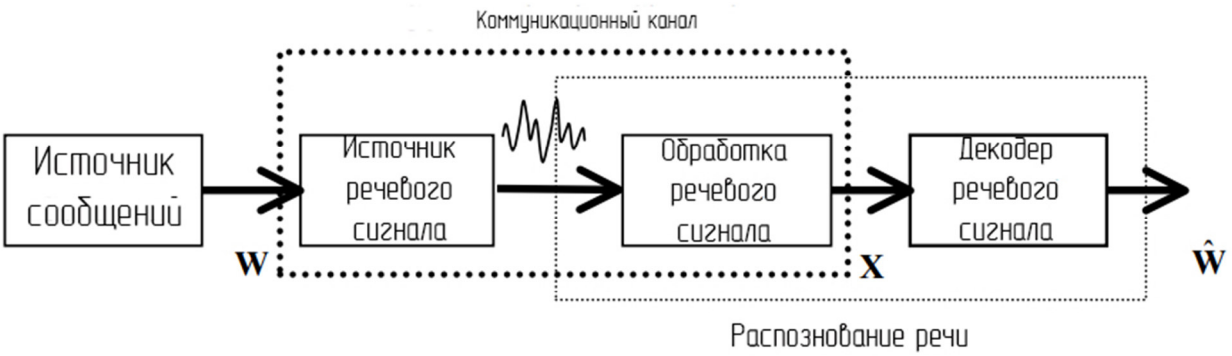


Рисунок 1.1 – Общая схема распознавания речевых сообщений

Реализация системы распознавания речи может быть представлена следующей блок-схемой, изображенной на рисунке 1.2. Речевой сигнал подается на блок обработки сигнала, где выполняются следующие операции: оцифровка сигнала с заданной частотой дискретизации; представление сигнала в виде некоторого вектора признаков. Далее, на декодере, выполняется операции по сопоставлению входящего речевого сигнала с некоторыми образцами речевых сообщений, класс которых заранее известен.

Акустическая модель служит для определения фонем, из которых состоит речевой сигнал, путем их классификации на основании подходов теории вероятностей и математической статистики. При этом фонема представляет собой минимальный элемент речи в языке, который может служить для определения разницы одного слова от другого.

Языковая модель служит для построения слов и предложений из, полученных в результате работы акустической модели, транскрипцией, используя заложенные в ней правила грамматики и лингвистики.

Таким образом, для распознавания речи необходимо выполнить следующие этапы: 1) оцифровка речевого сигнала, 2) представление сигнала в виде некоторого вектора признаков, 3) акустический анализ отрезков на основании выбранных признаков; 4) языковой анализ полученной транскрипции с целью формирования слов и имеющих смысл словосочетаний и предложений. [50]

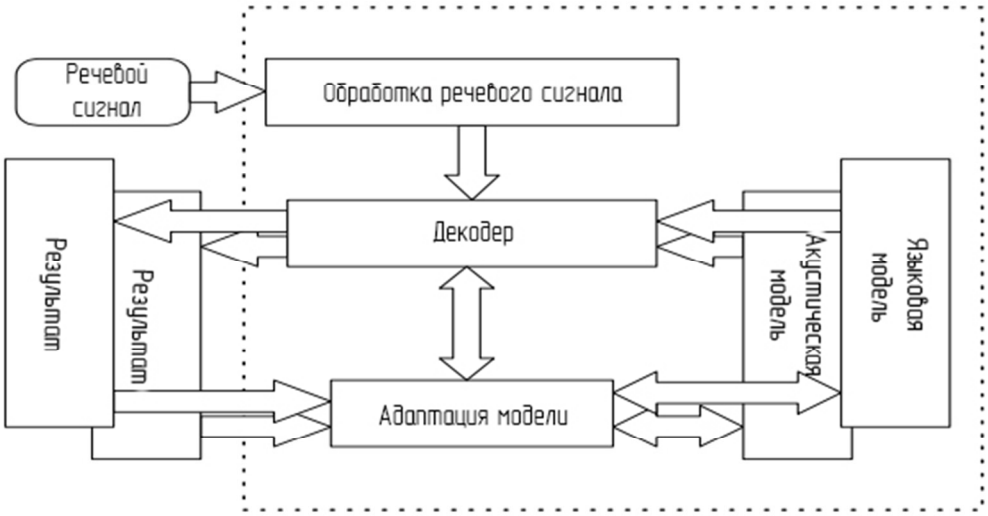


Рисунок 1.2 – Схема реализации системы распознавания речи

Эффективность методов как распознавания речи в целом, прямо зависит от качества полученного речевого сигнала (x), а также от типа представления речевого сигнала как некоего набора его признаков, обеспечивающих максимальное расстояние между разными классами, т.е. от выбора признакового пространства, для которого формируется решающая функция.

Речь представляет собой последовательность звуков, следующих друг за другом и разделенных паузами. Информация в речевом сигнале представлена в виде акустического колебания сложной формы. В традиционных, или узкополосных телефонных разговорах, предел звуковых частот находится в диапазоне от 300 до 3400 Герц. Для передачи одного канала голосовой частоты, включая защитную полосу частот, обычно выделяют полосу пропускания 4 кГц, допускающую частоту дискретизации 8 кГц.

Исходя из теоремы Котельникова, известно, что, частота дискретизации в 16 кГц является достаточной для кодирования речевых колебаний с частотой до 8кГц. Снижение частоты дискретизации влечет за собой увеличение ошибок при распознавании речи в виду снижения разборчивости из-за потери части энергетических составляющих звуков,

расположенных в области высоких частот, относящимся к следующим буквам русского алфавита: “с”, “ч”, “ш”, “ф”.

Проведенные американскими учеными исследования [46] доказывают, что оптимальной, с точки зрения распознавания речи является частота дискретизации 16 кГц, при глубине дискретизации 16 бит.

Таблица 1.1 – Зависимость разборчивости слов от частоты дискретизации оцифрованного речевого сигнала [50]

Частота дискретизации	Уровень разборчивости слов
8 кГц	опорный уровень
11 кГц	+ 10 %
16 кГц	+ 10 %
22 кГц	+ 0 %

Как видно из представленной таблицы, выигрыш в распознавании по сравнению с опорным уровнем в 8 кГц дает использование частоты дискретизации 11 кГц дает выигрыш в 10 %. Увеличение частоты дискретизации до уровня 16 кГц, дает дополнительный выигрыш в 10%, но дальнейшее увеличение частоты дискретизации не дает выигрыша в разборчивости речи. [50]

Для выбора признакового пространства при разработке алгоритма идентификации речевых данных необходимо изучить их структуру, природу формирования, а также модель восприятия речи человеком.

Как было отмечено ранее, при распознавании речевых сигналов, как правило, оперируют не с аналоговым речевым сигналом, а с так называемым описанием речевого сигнала, экономно представляющим речевой сигнал и содержащим смысловую и биометрическую информацию. Речь идет об обработке последовательностей отсчетов, регистрируемых через определенные промежутки времени:

$$\vec{x} = (x_1, x_2, .., x_N)^T, \quad (1.1)$$

$$x_i = x(i\Delta t),\ i = 1 \ldots N \quad , \tag{1.2}$$

где N – длительность сигнала в отсчетах; Δt – период дискретизации, равный величине обратной частоте дискретизации f_d:

$$\Delta t = 1/f_d \quad . \tag{1.3}$$

При анализе речевого сигнала, он подвергается сегментации на участки определенной длительности – фреймы. Длительность фрейма составляет около 10-30 мс, данное значение выбрано в виду стационарности речевого сигнала на участках данной длительности.

Далее фреймы преобразуют из временной области в частотную с помощью преобразования Фурье :

$$X(z) = \sum_{k=1}^{N} x_k e^{-j \cdot z \cdot (k-1)} \quad , \tag{1.4}$$

где x_k – отсчеты анализируемого отрезка сигнала; N – длительность окна анализа; j – мнимая единица ($j^2 = -1$).

Так чтобы близость фреймов относительно Евклидова расстояния:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{1.5}$$

соответствовала близости участков сигнала во временной области.

Далее производится операция выделения векторов признаков, и установление соответствия между фреймами определенного прецедента и исследуемого речевого сигнала для осуществления классификации. Проблема данного этапа заключается в том, что длительности одного и того же слова (слога или фонемы) отличаются степенью сжатия или растяжения во времени.

Возможности компьютерных систем по распознаванию речи ограничены, в виду того что человек использует для распознавания не только акустическую информацию. Для распознавания речи человек использует семантическую и синтаксическую связь слов, а также визуальную

информацию (чтение по губам, мимика и др.) – общеизвестно что, находясь в шумной обстановке речь собеседника легче распознавать если следить за его губами.

Фонетические модели, которые используются при создании алгоритмов идентификации речи, не являются точными, так как невозможно охватить все многообразие факторов. При задании фонетических прецедентов используют статистические методы, которые зачастую предполагают, что параметры фонем распределены по нормально закону. Но в действительности, точная модель прецедентов звуков и слов должна включать в себя множество всевозможных вариантов, что не осуществимо в реальных системах.

Кроме того, существует проблема дикторозависимости систем идентификации, будучи обученными по эталонам одного диктора системы показывают менее надежные результаты при распознавании речевых сигналов, произнесенных другим голосом.

Вышеперечисленные факты, говорят о том, что распознавание отрезков речевых сигналов с помощью ЭВМ обладает ограниченной надежностью, которую невозможно существенно повысить ни путем совершенствования алгоритмов, ни путем увеличения вычислительных мощностей. Имея в виду данное утверждение, можно сделать вывод о том, что разработка новых подходов к идентификации должна основываться на четко поставленной задаче, отталкиваясь от которой необходимо совершить выбор пространства признаков для сравнения, а также алгоритмов и методов, применяемых для выявления тождественности между прецедентом и объектом идентификации.

В качестве признаков в задачах распознавания могут использоваться временные, частотные и прочие характеристики речевого сигнала, которые, в различной степени, отражают особенности речевого аппарата, характерные для конкретного диктора. Чтобы получить представление о таких особенностях, необходимо ознакомиться с особенностями восприятия речи человеком.

1.2 Особенности речеобразования и восприятия речи человеком

Одной из основных тем, связанных со звуком, является восприятие звука человеком. Является нецелесообразным говорить о распознавании речи, не затронув озвученный вопрос.

1.2.1 Речевой аппарат

В соответствии с теорией речеобразования речь представляет собой струю воздуха, которая излучается системой органов: легкими, бронхами и трахеей, а затем преобразуется голосовым трактом в набор звуков.

В речевой аппарат человека входят ротовая и носовая полости с придаточными полостями, глотка (верхние резонаторы), гортань с голосовыми складками, трахея и бронхи (нижний резонатор), лёгкие, грудная клетка с дыхательными мышцами и нервная система организует их функции в единый, целостный процесс звукообразования, являющийся сложным психофизическим актом. [1, 2]

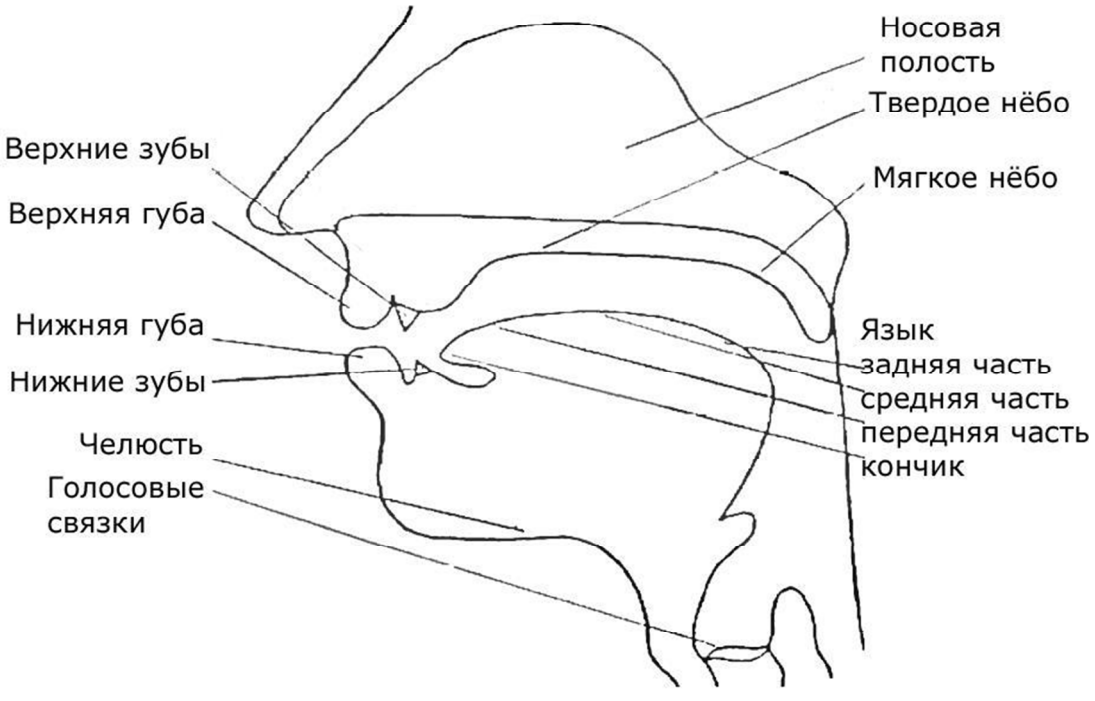


Рисунок 1.3 – Речеобразующие органы человека

Согласно рисунку 1.3, речевой тракт начинается с голосовых связок и заканчивается ротовой щелью, его длина у взрослого человека составляет примерно 17 см, площадь поперечного сечения тракта, которая определяется исходя из положения языка, губ, челюстей и небной занавески, может изменятся от 0 до 20 $см^2$.

Речевой аппарат человека способен порождать акустические колебания различной формы (шумоподобные или невокализированные и квазипериодические или вокализированные). Характер порождаемых акустических колебаний зависит от анатомии различных артикуляторов человека и их точек касания речевого тракта. [40]

Легкие – это источник воздуха и давления в процессе речеобразования.

Голосовые связки: когда голосовые связки находятся на маленьком расстоянии друг от друга и колеблются друг относительно друга в процессе речи, говорят, что звук – вокализованный. Если же связки не колеблются, то говорят, что звук – невокализованный.

Мягкое нёбо: работает как заслонка, которая открывает проход воздуху в носовую полость. Твердое нёбо: длинная, относительно твердая поверхность верхней стенки ротовой полости, в сочетании с языком позволяет произносить согласные звуки.

Язык: гибкий артикулятор. При отдалении от нёба позволяет произносить гласные звуки, при приближении к нёбу – согласные. Зубы: в сочетании с языком используются при произношении некоторых согласных звуков. Губы: могут округляться или растягиваться, изменяя звучание гласных звуков, либо смыкаться для остановки воздушного потока при произношении некоторых согласных звуков.

Основным различием между звуками является их разграничение на вокализованные и невокализованные звуки.

Вокализованные звуки в своей частотной и временной структуре имеют квазипериодическую составляющую. Она вносится, когда при произношении звука участвуют голосовые связки, вибрирующие с различной

частотой (от 60 Гц у взрослого мужчины до 300 Гц или выше у девушки или ребенка). Частота вибрации голосовых связок называется основной частотой звука, так как она является базовой частотой для остальных высокочастотных гармоник, создаваемых в гортанной и ротовой полости. Также, основная частота больше, чем какой-либо другой фактор влияет на основной тон речи. [38]

На рисунке 1.4 изображены этапы цикла состояния голосовых связок человека при прохождении через них воздушного потока. На стадии (а), голосовая щель сомкнута, и воздушный поток останавливается перед голосовыми связками. [38]

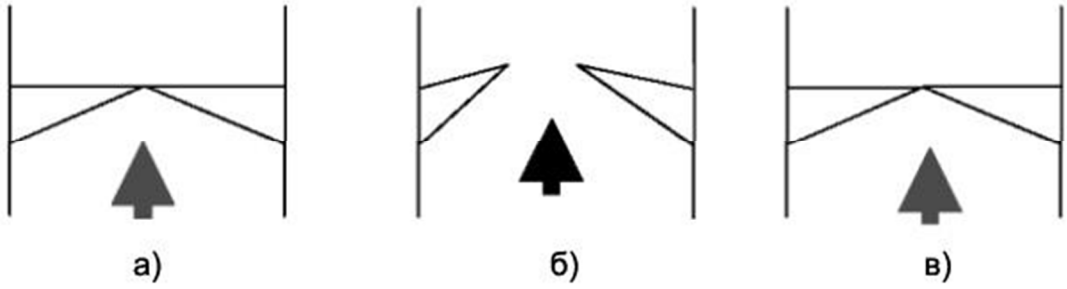


Рисунок 1.4 – Цикл сокращения голосовых связок. а) перекрытие голосовой щели, нарастание давления воздушного потока; б) открытие щели под воздействием давления; в) закрытие голосовой щели за счет выравнивания давления и эластичности тканей.

В какой-то момент (стадия б), давление воздуха перед связками преодолевает барьер, и воздух вырывается наружу через голосовую щель.

Тем не менее, ткани и мускулы голосовых связок, благодаря природной эластичности, возвращаются в исходное состояние, закрывая голосовую щель (стадия в). Таким образом, создается последовательность звуковых колебаний, которая является источником энергии для всех вокализованных звуков.

При произношении невокализованных звуков голосовые связки либо расслаблены, либо сильно напряжены, вследствие чего не производят звуковых колебаний. Воздух свободно проходит из легких в ротовую и/или

носовую полость речевого тракта. В результате взаимодействия воздуха с различными артикуляторами происходит преобразование воздушного потока, что приводит к произношению того или иного звука. [38]

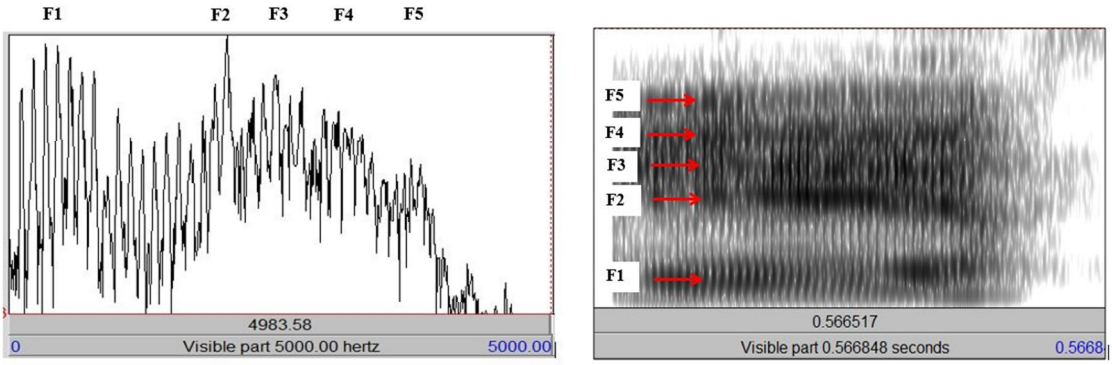


Рисунок 1.5 – Сигналограмма (слева) и спектрограмма (справа) речевого сигнала, соответствующего произнесенному звуку “А“

Таким образом, речевой сигнал представляет собой периодический процесс, состоящий из основной частоты F_0 и некоторого количества гармоник. Отдельно стоить отметить усиленные гармоники – форманты (F_1, F_2, F_3, F_4, F_5). Форманта представляет собой определенную частотную область (рисунок 1.5), в которой вследствие резонанса усиливается некоторое число гармоник тона, производимого голосовыми связками, то есть в спектре звука форманта является достаточно отчетливо выделяющейся областью усиленных частот. [40] Фактически феномен форманты есть проявление работы активного полосового фильтра в составе речевого тракта. Исходя из вышеперечисленного, речевой сигнал представляется естественным изображать как сумму синусоидальных колебаний.

1.2.2 Восприятие речевого сигнала человеком

В системе восприятия речи есть две основных составляющих части: внешние слуховые органы и слуховой отдел мозга. Ухо обрабатывает сигнал, который несет в себе звуковая волна, путем преобразования его в механическую вибрацию барабанной перепонки и последующим

отображением этой вибрации в последовательность импульсов, передаваемых слуховым нервом. Полезная информация извлекается в различных участках слухового отдела мозга человека. Рисунок 1.6 изображает анатомию уха человека. [40]

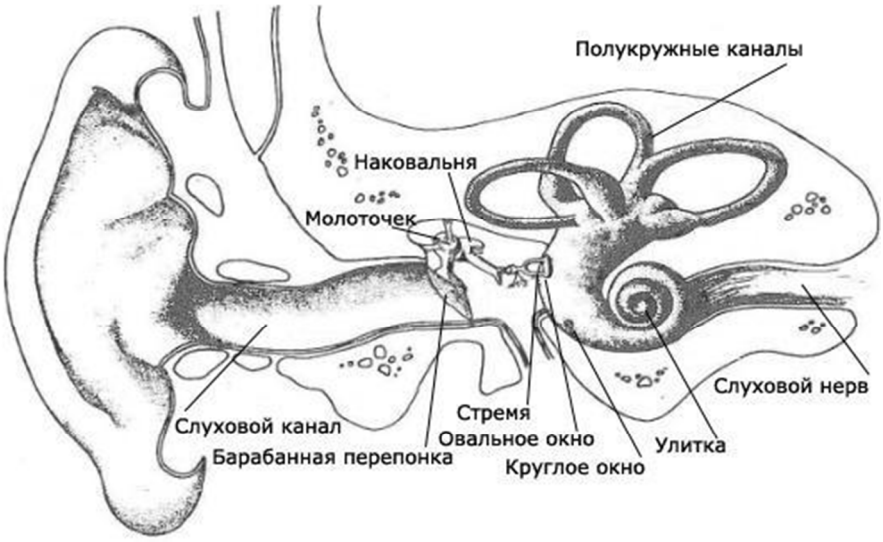


Рисунок 1.6 – Строение группы слуховых органов [40]

Ухо человека состоит из трех отделов: наружное ухо, средне ухо и внутреннее ухо.

Наружное ухо состоит из ушной раковины, которая концентрирует звуковые колебания, и внешнего слухового канала. Звуковая волна, попадая в ушную раковину, проходит дальше по слуховому каналу (длина ~3 см, диаметр ~0.5 см) и попадает в среднее ухо, где ударяется о барабанную перепонку, представляющую собой тонкую мембрану. Барабанная перепонка преобразует звуковую волну в вибрации, при этом усиливается эффект от слабой волны и ослабляется от сильной. Эти вибрации передаются по присоединённым к барабанной перепонке специальным косточкам (молоточку, наковальне и стремечку) во внутреннее ухо, которые представляет собой завитую трубку с жидкостью диаметром ~0.2 мм и длиной ~4 см. Эта трубка называется улиткой, и она сообщается непосредственно со слуховым нервом. Внутри нее находится базилярная мембрана, представляющую узкую ленту длинной ~32 мм, вдоль которой

располагаются нервные окончания (около 20 тысяч волокон). Базилярную мембрану иногда называют “овальным окном”, это своеобразный интерфейс между средним ухом и внутренним ухом (улиткой), так как остальная поверхность внутреннего уха состоит из костной ткани. [40]

Толщина базилярной мембраны в начале улитки и в ее вершине различны. В результате такого строения мембрана резонирует разными своими частями в ответ на звуки разной частоты. Звуковые колебания высокой частоты раздражают нервные окончания в начале улитки, а низкой – у ее вершины.

Продольная мембрана разделяет спираль улитки на две заполненных жидкостью части. Внутренняя поверхность улитки покрыта ресничковыми клетками-рецепторами, которые соединены напрямую со слуховым нервом и воспринимают информацию о давлении жидкости в определенной точке улитки.

Структура внутреннего уха устроена так, что при различных частотах начального сигнала, максимальная амплитуда изменения давления жидкости в улитке будет регистрироваться на определенном расстоянии от ее основания. Таким образом, улитку можно представить, как гребенку фильтров, выходной сигнал которой упорядочен по расстоянию от основания улитки. Фильтры, более близкие к основанию улитки отвечают за более высокие частоты. Таким образом, улитка представляет собой анализатор спектра. [3]

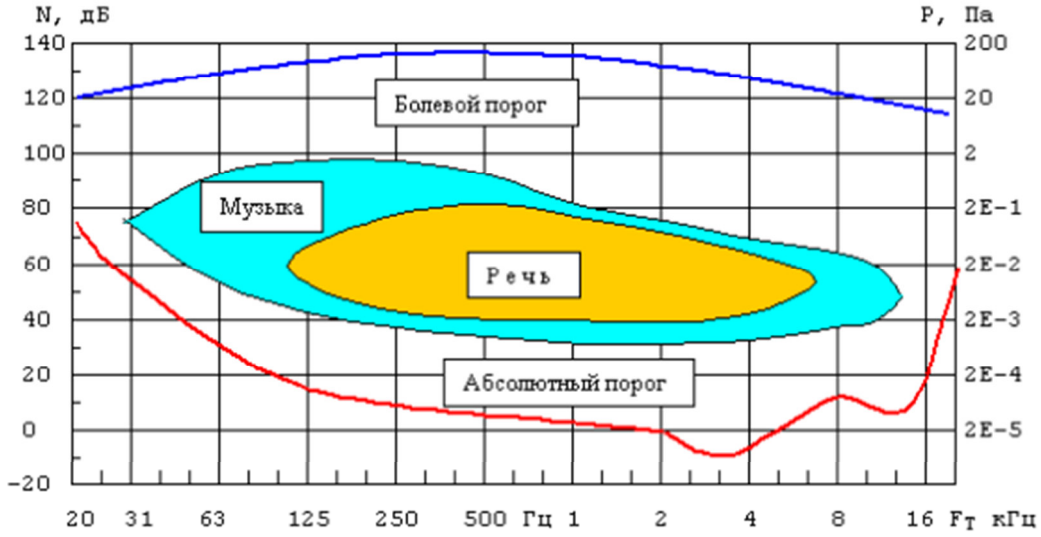


Рисунок 1.7 – График восприятия звука человеком [3]

Слуховой нерв представляет собой набор частотных каналов. В каждый частотный канал входит группа нейронов, соединенных с одним или соседними фильтрами улитки, то есть те, которые имеют одинаковые или близкие характеристические частоты. Этот набор признаков подается в качестве мгновенного изображения сигнала в мозг человека, в котором, посредством сложной нейронной сети, происходит выделение полезной информации из полученного сигнала. К сожалению, точных данных о том, как данная информация извлекается внутри человеческого мозга, нет. Есть только ряд теорий, которые по-разному описывают возможные нейронные структуры внутри мозга и их взаимодействие.

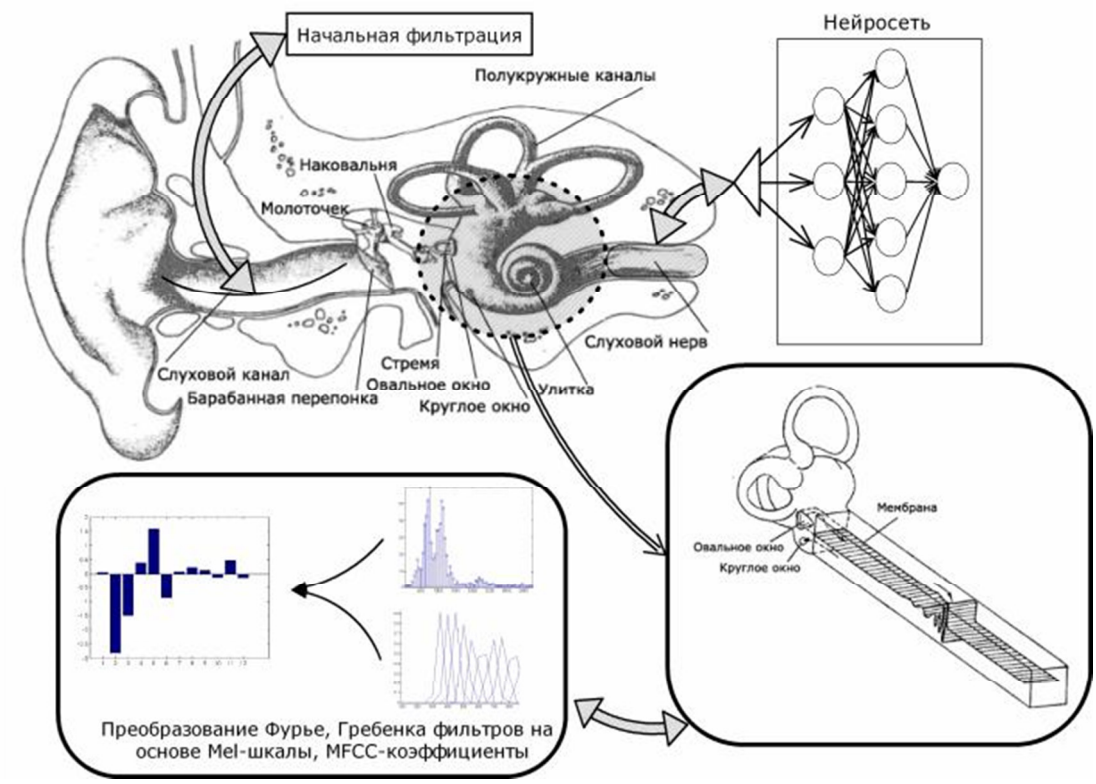


Рисунок 1.8 – Гипотетическая схема распознавания речи человеком

На рисунке 1.8 представлена упрощенная, гипотетическая схема распознавания речи человеком. Многие элементы различных систем распознавания речи основываются на слуховом тракте человека и пытаются имитировать механизмы его работы. Так, один из популярных на сегодняшний день характеристический признак речевого сигнала (Мел-кепстральные коэффициенты MFCC) основан на изучении методов преобразования сигнала во внутреннем ухе человека. Также, разработка и развитие нейросетевых алгоритмов связаны с исследованиями мозга человека. [23,24, 50]

Учеными были проведены исследования, для определения градации частот, которая моделировала бы естественную реакцию человеческой системы восприятия речи, в которой улитка действует как спектральный анализатор [40]. Сложный механизм внутреннего уха и слухового нерва предполагает, что свойства восприятия звуков на различных частотах не могут быть, очевидно, простыми или линейными. Широко известно, что музыкальный тон разделяется на октавы и полутона.

Частота f_1 выше частоты f_2 на октаву тогда и только тогда, когда $f_1=2f_2$. В 1 октаве 12 полутонов, следовательно, f_1 выше частоты f_2 на полутон тогда и только тогда, когда выполняется равенство (1.6):

$$f_1 = 2^{(1/12)} f_2 \tag{1.6}$$

В результате различных исследований [52], основывающихся на человеческих ощущениях звуков различных частот, был выведен ряд шкал, которые позволяли представить частоту звука в более близких человеческому восприятию величинах.

При распознавании человеческой речи получила распространение – мел-шкала [52], линейная при частотах ниже 1кГц и логарифмическая при частотах выше 1кГц. Мел-шкала была получена в результате экспериментов с образцовыми тонами (синусоидами) в которых с испытуемых требовалось разделить данные диапазоны частот на 4 равных интервала или настроить частоту требуемого тона так, чтобы он был в половину частоты исходного. 1 мел определяется как 1 тысячная уровня тона в 1 кГц. Как и в любых других попытках создать подобные шкалы, рассчитывается, что шкала мел более точно моделирует чувствительность человеческого уха. Вычисление мел-значений можно приблизительно представить следующей формулой (1.9):

$$m(f) = 2595 \cdot \log_{10}(1 + f/700), \tag{1.7}$$

Обратное преобразование можно осуществить с помощью формулы:

$$f(m) = 700 \cdot \left(10^{\frac{m}{2595}} - 1\right), \tag{1.8}$$

где f – частота звука в Гц, m – высота звука в мелах.

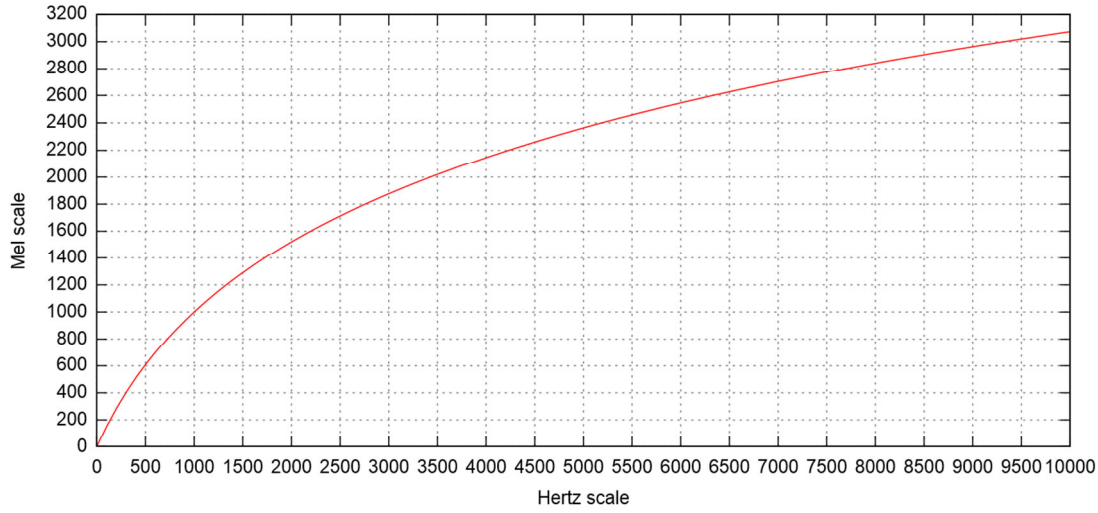


Рисунок 1.9 – Шкала Герц/Мел

Ряд современных техник обработки речевого сигнала основывается на применении таких шкал. На рисунке 1.9 представлена шкала соответствия единиц частоты звука мел единицам Герц

Целесообразно отметить некоторые основные параметры, присущие речевым сигналам, такие как основной тон, высота тона и тембр звука.

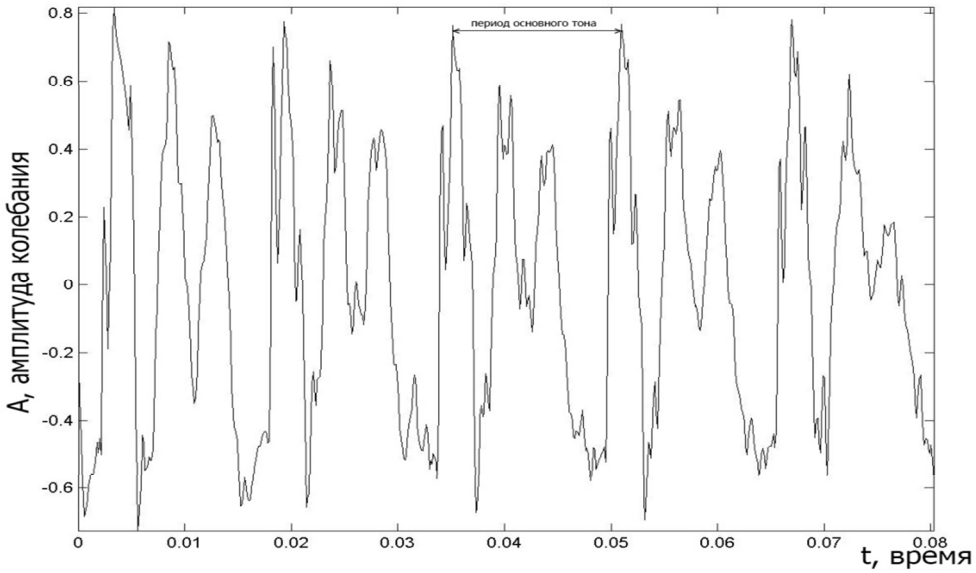


Рисунок 1.10 – Сигналограмма фонемы, соответствующей звуку “А” с выделенным периодом основного тона

В спектре речи зачастую присутствует наиболее выделяющаяся по амплитуде и периоду частотная составляющая (соответствует наименьшей частоте и наибольшему периоду в данном спектре), ее принято называть

основным тоном (ОТ). На рисунке 1.10 представлен период основного тона фонемы, соответствующей звуку “А”.

Высота звука – характеристика, условно распределяющая звуки по некоторой шкале от низких к высоким. На воспринимаемую высоту звука влияет частота основного тона, в тоже время форма и период звуковой волны могут также оказывать влияние на высоту звуку. Высота звука может быть определена слуховой системой человека и у сложных сигналов, но лишь в случае периодичности сигнала. [40]

В зависимости от соотношения амплитуд, частотных составляющих сигнала звук может приобретать различную окраску и воспринимается как тон или как шум. Если спектр сигнала имеет ярко выраженные пики или один пик, то этот участок является тоном, если же спектр звука является сплошным, то этот участок является шумовым. [40]

Тембр звука является субъективной характеристикой качества звука, благодаря которой звуки одной и той же высоты, и интенсивности можно отличить друг от друга. Основными объективными параметрами, определяющими оценку тембра музыкантами, являются спектр и характер переходного процесса основного тона и обертонов.

1.3 Методы цифровой обработки сигналов в задачах распознавания речевых сигналов

Существующие сегодня системы распознавания речи основываются на сборе информации, необходимой для распознавания слов. Тем не менее, в настоящее время даже при распознавании небольших сообщений нормальной речи, пока невозможно после получения разнообразных реальных сигналов осуществить прямую трансформацию в лингвистические символы. Первым шагом процесса распознавания речи является первоначальное трансформирование речевого колебания, поступающего на микрофоны, так

чтобы его можно бы подвергнуть компьютерному анализу, т.е. дискретизация и квантование. Следующим этапом является анализ речевого сигнала, который может быть выполнен как в базисе Фурье, так и с помощью вейвлетов или путем вычисления кепстра, что не только позволяет сжать информацию, но и дает возможность сконцентрироваться на важных, применительно к решаемой задаче, аспектах речевого сигнала.

1.3.1 Спектральный анализ в базисе Фурье

Для анализа оцифрованных сигналов используют чаще всего спектральный анализ в базисе Фурье [33]. Данный метод обработки позволяет характеризовать частотную составляющую анализируемого сигнала. Для реализации частотного анализа можно воспользоваться дискретным прямым (1.11) и обратным (1.12) преобразованиями Фурье. Для одномерного массива речевых данных *x(i)*, где *i=1,2,...,N* , коэффициенты ряда Фурье определяются следующим образом:

$$X(k) = \sum_{i=1}^{N} x(i) e^{-j\frac{2\pi}{N}(i-1)(k-1)}, \qquad (1.9)$$

где *k=1,2,...,N, j* – комплексная единица, $j = \sqrt{-1}$.

Обратное преобразование во временную область:

$$x(i) = \frac{1}{N} \sum_{k=1}^{N} X(k) e^{j\frac{2\pi}{N}(i-1)(k-1)} \qquad (1.10)$$

Вещественная часть трансформанты Фурье (косинусная) определяется так: $\mathrm{Re}(X(k)) = X\cos(k)$, мнимая (синусная) так: $\mathrm{Im}(X(k)) = X\sin(k)$.

Затем определяются следующие значения:

амплитуда $A(k) = \sqrt{\mathrm{Re}(X(k))^2 + \mathrm{Im}(X(k))^2}$ (1.11)

фаза $\Psi(k) = arctg\left(\frac{\mathrm{Im}(X(k))}{\mathrm{Re}(X(k))}\right)$ (1.12)

энергия $P(k) = \mathrm{Re}(X(k))^2 + \mathrm{Im}(X(k))^2$ (1.13)

Так как переменная *k* имеет смысл частоты, то можно анализировать частотные характеристики сигнала исследуя функции *A(k)* и *Ψ(k)*.

Основной недостаток преобразования Фурье заключается отсутствие возможности отличить стационарный сигнал от нестационарного сигнала, с тем же спектральным составом, что проиллюстрировано на рисунке 1.11. Данное явление накладывает ряд ограничений на применение преобразования Фурье при анализе сигналов.

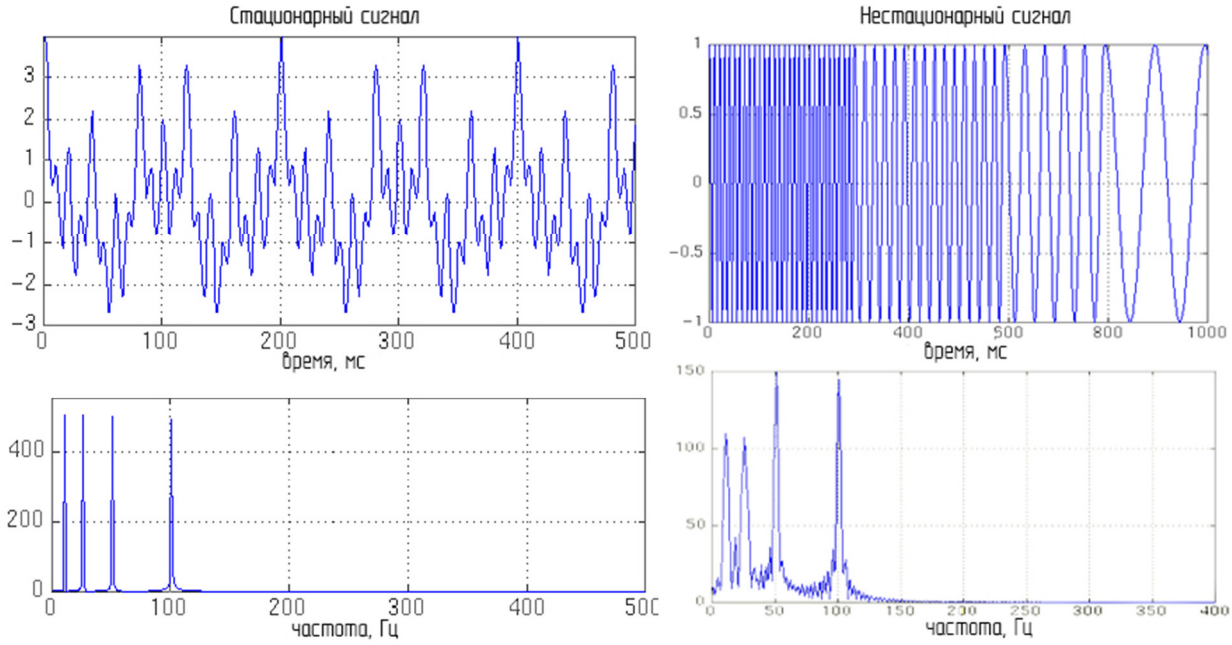


Рисунок 1.11 – Демонстрация преобразования Фурье для стационарного и нестационарного сигналов

1.3.2 Оконный анализ в базисе Фурье

Так как преобразование Фурье дает информацию только про частоту, которая присутствует в сигнале и не дает никакой информации про то, в какой промежуток времени эта частота присутствует в сигнале. Следовательно, для анализа кусочно-стационарного сигнала, каким и является речевой сигнал, необходимо использовать окно, достаточно узкое для того, чтобы сигнал внутри него выглядел стационарным, т.е. для речевого сигнала длительностью не менее одного периода основного тона.

Такой подход получил название оконного (или кратковременного) преобразования Фурье (ОПФ) [33]. При ОПФ сигнал делится на отрезки («окна»), в пределах которых его можно считать стационарным. Для этого к сигналу применяется оконная функция *w*, ширина которой должна быть равной ширине окна.

$$X(k) = \sum_{i=1}^{N} x(i) * w \cdot (i-k) e^{-j\frac{2\pi}{N}(i-1)(k-1)} \qquad (1.14)$$

Оконная функция и сигнал перемножаются. Затем произведение подвергается преобразованию Фурье. Если анализируемый отрезок стационарен, то полученный результат преобразования корректно отображает частотное наполнение текущего окна анализа.

Следующим шагом является сдвиг оконной функции на некоторое количество точек. Сдвинутая функция вновь умножается с сигналом, выполняется ПФ произведения. Эта процедура повторяется до достижения конца исходного сигнала.

Проблемы ОПФ имеют свои корни в явлении, которое называется принципом неопределенности. В основе принципа неопределенности лежит тот факт, что невозможно сказать точно какая частота присутствует в сигнале в данный момент времени (можно говорить только про диапазон частот) и невозможно сказать в какой точно момент времени частота присутствует в сигнале (можно говорить лишь про период времени).

В связи с этим возникает проблема разрешающей способности. Разрешающую способность оконного преобразования Фурье можно регулировать с помощью ширины окна. Проблема в данном случае связана с шириной использующейся оконной функции. Эта ширина называется еще носителем функции. Если окно достаточно узкое, то говорят о компактном носителе. Итак, чем уже окно, тем лучше временное разрешение, но хуже частотное. И наоборот.

Перед вычислением спектра [33] сигнала нужно выбрать отрезок сигнала *x(i), i=1,2,…,N*, на котором будет вычисляться спектр. Длина отрезка

N должна быть степенью двойки для работы Быстрого Преобразования Фурье (БПФ). Иначе сигнал надо дополнить нулями до нужной длины. После этого к выбранному участку сигнала применяют БПФ [33].

При вычислении спектра указанным образом возможен следующий нежелательный эффект. При разложении функции в ряд Фурье предполагается, что функция периодическая, с периодом, равным размеру БПФ. Вычисляется спектр именно такой функции (а не той, из которой извлекли фрагмент). При этом на границах периодов такая функция наверняка будет иметь разрывы (ведь исходная функция не была периодической). А разрывы в функции сильно отражаются на ее спектре, искажая его.

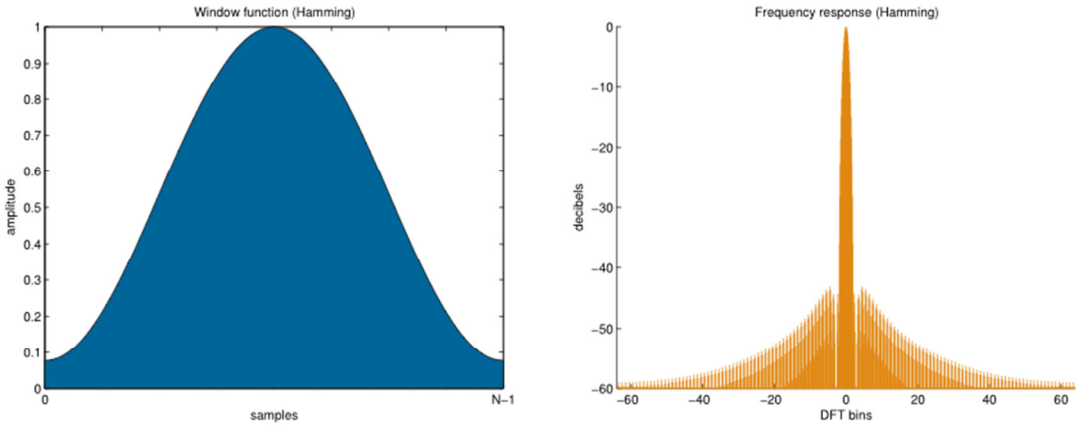


Рисунок 1.12 – Взвешивающее окно Хэмминга

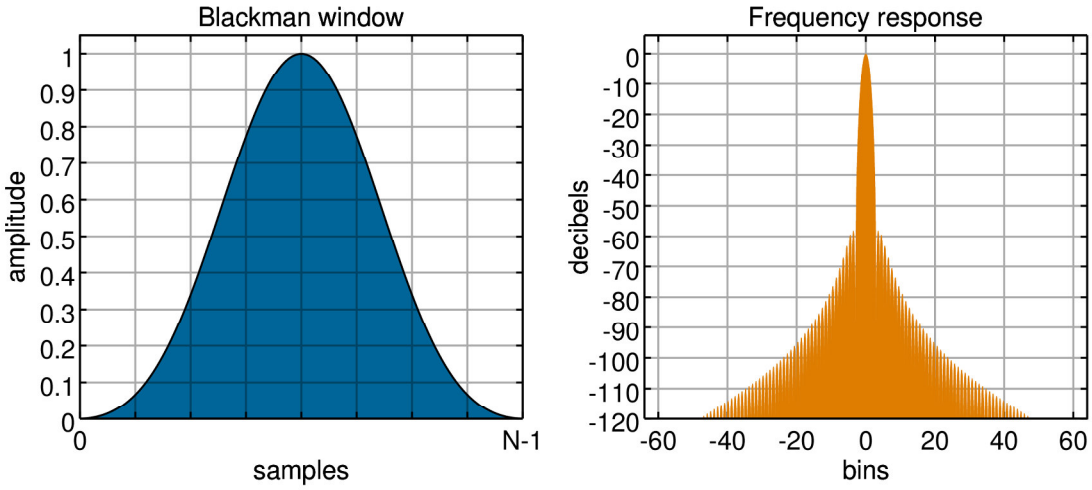


Рисунок 1.13 – Взвешивающее окно Блэкмана

Для устранения этого эффекта применяются так называемые взвешивающие окна. Они плавно сводят на нет функцию вблизи краев анализируемого участка. Весовые окна имеют форму положительной полуволны косинуса. Выбранный для анализа участок сигнала помножается на весовое окно, которое устраняет разрывы функции при «зацикливании» данного участка сигнала. [33]

«Зацикливание» происходит при ДПФ, так как алгоритм ДПФ полагает, что функция периодическая. Существует множество весовых окон, названных в честь их создателей. Все они имеют похожую форму и в значительной степени устраняют рассмотренные искажения спектра. Ниже приведены формулы двух окон: Хэмминга (Hamming window) и Блэкмана (Blackman window):

$$w_{Hamming}[n] = 0.54 - 0.46\cos\frac{2\pi n}{N} \tag{1.15}$$

$$w_{Blackman}[n] = 0.42 - 0.5\cos\frac{2\pi n}{N} + 0.08\cos\frac{4\pi n}{N}. \tag{1.16}$$

Здесь окно применяется к сигналу с индексами от 0 до N. Окно Хэмминга наиболее часто используется. Окно Блэкмана обладает более сильным действием по устранению рассмотренных искажений, однако имеет свои недостатки.

1.3.3 Вейвлет анализ

Вейвлет-преобразование (ВП) [10,29] обеспечивает частотно-временное представление сигналов. Сигнал пропускается через два фильтра – низкочастотный и высокочастотный, и процедура повторяется. Эта операция называется декомпозицией. Декомпозиция продолжается какое-то число раз. В конечном счете, получается множество вторичных сигналов, которое представляет исходный сигнал. Каждый вторичный сигнал соответствует какому-то диапазону частот. Можно построить трехмерный график, отложив по одной оси время, по второй – частоту и по третьей – амплитуду. Таким

образом, можно увидеть, какие частоты присутствуют в каждый интервал времени.

Таким образом, вейвлет-преобразования, в отличие от оконного преобразования Фурье, которое имеет постоянный масштаб в любой момент времени для всех частот, имеет лучшее представление времени и худшее представление частоты на низких частотах сигнала и лучшее представление частоты с худшим представлением времени на высоких частотах сигнала. [29]

На рисунке 1.14 хорошо видно, что полученное вейвлет-преобразование является более детализированным по времени в области высоких значений масштаба (низких частот) и менее детализирована в области низких значений масштаба (высоких частот).

Из этого следует, что вейвлет преобразования дает возможность уменьшить влияние принципа неопределенности на полученном частотно-временном представлении сигнала. С его помощью низкие частоты имеют более детальное представление относительно времени, а высокие — относительно частоты. [29]

Невозможно одновременно обеспечить хорошее разрешение по времени и по частоте. Чем уже окно, тем выше разрешение по времени и ниже разрешение по частоте. В отличие от постоянного разрешения по осям в преобразовании Фурье (рисунок 1.16).

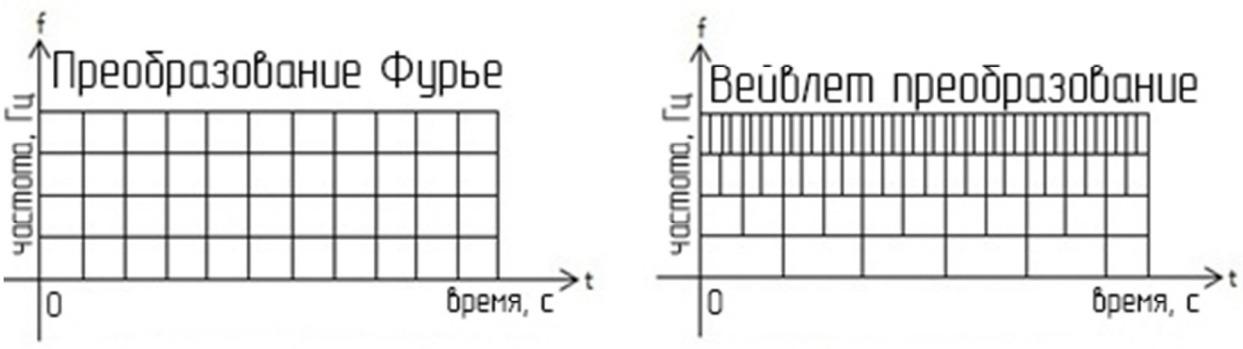


Рисунок 1.14 – Сравнение разрешающей способности Вейвлет-преобразования с преобразованием Фурье

Поэтому можно лишь говорить об интервале времени и о наблюдающейся в нем частотной полосе.

Вейвлет-преобразование является альтернативой преобразованию Фурье в тех случаях, когда сигнал не носит периодического характера. Различают непрерывное и дискретное Вейвлет-преобразования. Предполагается, что все интегралы, рассмотренные ниже, существуют.

Непрерывное вейвлет преобразование

Пусть имеется функция *f(x)* и некоторая функция *s(x)* - материнская функция. Рассмотрим числа вида

$$F(a,b) = \int_{-\infty}^{\infty} f(x)\bar{s}(\frac{x-b}{a})dx \qquad (1.17)$$

Если $s(x)=e^{ix}$, то в результате получаем обычное преобразование Фурье (параметр b не используется по понятной причине). Формула (1.17) определяет общее Вейвлет преобразование. Существует формула обратного преобразования, позволяющая в некоторых случаях восстановить исходную функцию по ее преобразованию:

$$a^{-1}\int_{-\infty}^{\infty} | s(\frac{x-b}{a}) |^2 \, dx \qquad (1.18)$$

$$u(a,b) = \frac{1}{\sqrt{a}} s(\frac{x-b}{a}), \qquad (1.19)$$

Это означает, что вектор, заданный функцией (1.17) имеет постоянную длину в смысле пространства L_2. Предположим, что удалось найти такие значения параметров, для которых $F(a,b)\sqrt{a}$ достигает локального максимума. Это означает, что проекция функции *f(x)* на соответствующую функцию *u(a,b)* имеет максимальное значение, поэтому графики этих функций аналогичны. Положив *g(x)=f(x)-u(a,b,x)*, получим невязку, для которой решается такая же задача. В результате получаем приближение

исходной функции функциями, порожденными с помощью функций *s(x)*. Это дает альтернативное описание исходной функции.

В зависимости от того, какого рода особенности требуется обнаружить, выбирают вид материнской функции. При цифровой обработке, когда исходная функция задана лишь в отдельных точках, используется дискретное преобразование. Оказалось, что и в общем случае удается построить теорию, напоминающую теорию преобразования Фурье.

На практике, в качестве материнской функции при указанном подходе часто используют функцию $m(t)=a(1-t^2)exp(-t^2/2)$. Константу а определяют из условия нормировки.

Рассмотрим множество функций L_2 на вещественной оси. Пусть $s(x) \in L_2$, причем функции $s(x-k)$, $k \in Z$ образуют ортонормированную систему. Это означает, что:

$$\int s(x-k)\bar{s}(x-m)dx = \delta_{k,m} \tag{1.20}$$

Такую функцию называют шкалирующей. Например, любая функция, имеющая носитель внутри единичного интервала и норму равную 1, удовлетворяет условию (1.11). Обозначим через

$$S(w) = \int s(x)e^{-2\pi iwx}\,dx \tag{1.21}$$

Имеет место формула

$$\sum_{k=-\infty}^{\infty} |\, S(u-k)\,|^2 = 1. \tag{1.22}$$

Важным примером материнской функции является функция, равная 1 на интервале [0,1] и 0 в остальных точках. Такую функцию обозначают через $e(x)$. Существует возможность анализа сигнала при помощи альтернативного подхода, имя которому – кратномасштабный анализ (КМА). КМА позволяет получить хорошее по времени (плохое по частоте) на

высоких частотах и хорошее разрешение по частоте (плохое по времени) на низких частотах. [29]

Дискретное Вейвлет-преобразование (ДВП) обеспечивает достаточно информации как для анализа сигнала, так и для его синтеза, являясь вместе с тем экономным, как по числу операций, так и по требуемой памяти. Истоки ДВП восходят к 1976, когда Кройцер, Эстебан и Галанд разработали метод декомпозиции дискретных сигналов. Кройхер, Веббер и Фланган в тот же год опубликовали аналогичную работу по кодированию речевых сигналов. Они назвали свой метод анализа субполосным кодированием. Основная идея – та же, что и при НВП. Масштабно-временное представление сигнала получается с использованием методов цифровой фильтрации. В дискретном случае для анализа сигнала на разных масштабах используются фильтры с различными частотами среза. Сигнал пропускается через древовидно соединенные ВЧ и НЧ фильтры [29].

1.3.4 Кепстральный анализ

Основой кепстрального анализа речевых сигналов является предположение, что речевой сигнал представляет собой сигнал на выходе линейной системы с постепенно меняющимися параметрами.

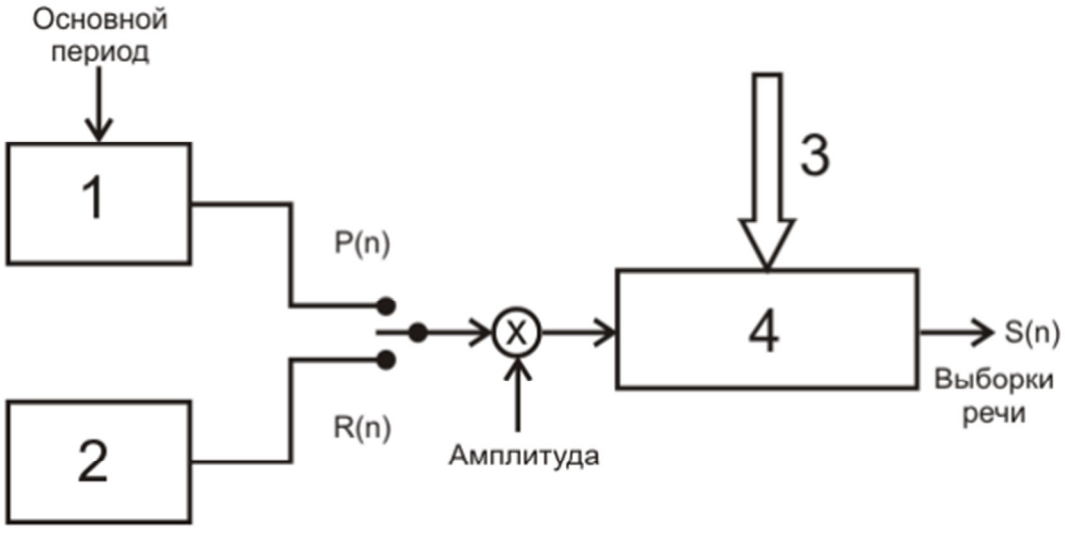


Рисунок 1.15 – Модель речевого аппарата как линейной системы:
1 – генератор импульсной последовательности, 2 – генератор ПСП ,

3 – коэффициенты цифрового фильтра (параметры голосового тракта),

4 – нестационарный цифровой фильтр

Таким образом, на коротких отрезках речевой сигнал можно рассматривать как сигнал на выходе линейной системы с постоянными параметрами, возбуждаемой последовательностью импульсов или случайным шумом. Если источники возбуждения и форма голосового тракта относительно независимы речевой аппарат можно представить в виде следующей совокупности элементов, изображенной на рисунке 1.15.

Тем самым, проблема анализа сигнала сводится к измерению параметров модели и оценке изменения параметров с течением времени. Известно, что сигнал возбуждения и импульсная характеристика фильтра взаимодействуют через операцию свертки. Анализ речи может рассмариваться как задача разделения компонент, участвующих в операции свертки, т.е. задача обратной свертки [31,33]. Один из способов решения поставленной задачи является кепстральный анализ.

В представленной модели фильтры имею постоянные характеристики на временном интервале ~10-20 мс. Поэтому на каждом интервале фильтр можно характеризовать импульсной или частотной характеристикой, или набором коэффициентов, если импульсная характеристика фильтра бесконечна. Такая модель позволяет применять для анализа речевых сигналов гомоморфную развертку. [33]

Пусть задан сигнал $S_{вых}(t)$ на выходе фильтра. Необходимо получить информацию о входном сигнале $S_{вх}(t)$ и самом фильтре, путем определения его импульсной характеристики $h(t)$.

Выходной сигнал определяется сверткой:

$$S_{вых}(t) = S_{вх}(t) \otimes h(t) \qquad (1.23)$$

Т.к. $S_{вых}(w) = S_{вх}(w)H(w)$ в частотной области, то прологарифмировав получаем выражение:

$$\ln[S^2{}_{вых}(w)] = \ln[S^2{}_{вх}(w)] + \ln[H^2(w)] \qquad (1.24)$$

Применим к нему обратное преобразование Фурье, можно получить выражение вида:

$$C(q) = C_s(q) + C_h(q) \tag{1.25}$$

из которого, методами линейной фильтрации, можно выделить некоторые характеристики $S_{вх}(t)$ и $h(t)$.

Также $C(q)$ может быт представлено в виде:

$$C(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \ln[S(w)]^2 e^{iwq}\, dw \tag{1.26}$$

Данное преобразование принято называть “кепстр”, данное название получено путем перестановки букв в слове “спектр”. Аргумент q имеет размерность времени, данное время является в некотором роде особенным, т.к. $C(q)$ в любой момент q зависит от функции $s(t)$ исходного сигнала со спектром $S(w)$ заданной при $-\infty<t<\infty$. Иногда q называются *“сачтота”* или *“кьюфренси”* (анаграммы от русского *“частота”* и английского *“фрекьюнси”*).

В виду того что в компьютерных и телекоммуникационных системах работают с дискретным представлением речевого сигнал, целесообразно привести запись кепстра в дискретной форме:

$$C(n) = \frac{1}{N}\sum_{k=0}^{N-1} \ln\left|X(k)\right|^2 e^{i\frac{2\pi}{N}kn}, 0 \le n \le N-1 \tag{1.27}$$

Одной из основных составляющих, позволяющей эффективно решать задачи распознавания и идентификации отрезков речевых сигналов, является выбор пространства признаков, которое в наибольшей степени отражает необходимые для распознавания отличительные свойства речевого сигнала. При этом важно сохранять приемлемый объем используемых признаков, чтобы не повышать вычислительную сложность алгоритмов выше необходимого значения.

1.4 Субполосный подход к обработке речевых сигналов

Исследования особенностей распределения энергии по частотным интервалам отрезков сигналов, соответствующих звукам русской речи, проведенных в [14,16,17], показали, что большая часть формант вокализованной русской речи сосредоточена в достаточно узком частотном диапазоне, а энергия шумовых звуков распределена по нескольким интервалом в высокочастотной части частотной оси. Оцифрованный речевой сигнал может быть представлен в виде:

$$\vec{x} = (x_1, x_2 .. x_{(m-1)}, x_m) \tag{1.28}$$

где x – речевой сигнал, m – количество отсчетов.

Далее необходимо выбрать длину окна анализа - N и количество частотных интервалов - R, исходя из необходимой разрешающей способности.

Речевой сигнал x подвергается разделению на окна длительностью N:

$$\vec{x}_n = (x_1, x_2 .. x_{(n-1)}, x_n), N = 1..n \tag{1.29}$$

Вычисление распределения долей энергий речевого сигнала происходит по формуле [14]:

$$P_r = \bar{x}_n A_r \bar{x}_n^T \tag{1.30}$$

где матрица $A_r = \{a_{ik}\}$ с элементами вида:

$$a_{ik} = \frac{\sin[v_{r+1}(i-k)] - \sin[v_r(i-k)]}{\pi(\mathrm{i - k})} \tag{1.31}$$

где $v = [0, \Delta, \Delta + 2\Delta, ...(n\Delta - 4\Delta) + 2\Delta, n\Delta]$ вектор, задающий границы частотных интервалов (рисунок 1.16), ширина которых должна удовлетворять выражению:

$$\Delta = \frac{\pi}{2R + 1} \tag{1.32}$$

Рисунок 1.16 – Частотная ось в субполосном анализе, разбитая на R интервалов

Основными преимуществами субполосного подхода являются:

- Обеспечивает минимальное просачивание энергии через границы частотных интервалов
- Обеспечивает расчет распределения долей энергий без перехода в частотную область
- Отсутствие проблемы частотно-временного разрешения.

ГЛАВА 2 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

2.1 Акустико-фонетический подход к распознаванию речевых сигналов

Любой участок РС можно, некоторым образом, характеризовать через определенный признак, соответствующий данному отрезку, или же набор таких признаков. Описанные в первой главе методы цифровой обработки РС позволяют в определенной степени отражать природу порождающего воздействия. Однако, в виду того что речевой сигнал представляет собой лишь отдельную реализацию передаваемой человеком информации, то невозможно характеризовать весь возможный набор таких реализаций всего лишь одним признаком. [25] Речевой сигнал зависит от множества факторов, как внешних (шумы, окружающая обстановка и т.п.), так и зависящих от источника информации – человека (темп речи, эмоциональная окраска, тембр, и др.), что также говорит о том, что необходимо подобрать такой набор признаков, который позволит выделять тождественные (в информационном смысле) реализации из речевого потока [12].

Речевой сигнал, представляющий собой акустическую реализацию передаваемой человеком информации, содержит в себе некоторое сообщение. Передаваемое одним и тем же человеком сообщение, например, предложение “Я иду в университет” при акустической реализации может принимать различные формы, которые могут отличаться по множеству параметров: громкость, длительность, эмоциональная окраска, высота т.п.

Одним из наиболее популярных подходов к распознаванию речевых сигналов является акустически-фонетический подход, принципиальная схема которого представлена на рисунке 2.1 [50]. Речевой сигнал подвергается процедуре выделения признаков, в результате которой формируются векторы признаков, которые позволяют классифицировать отрезки речевых сигналов.

Использование набора векторов признаков обусловлено необходимостью учитывать различные характеристики сигнала, которые невозможно детектировать всего лишь одним вектором признаков. Таким образом, для получения наиболее полного описания речевого сигнала в каждый момент времени естественно использовать совокупность признаков, которые позволят осуществлять точную идентификацию отрезков.

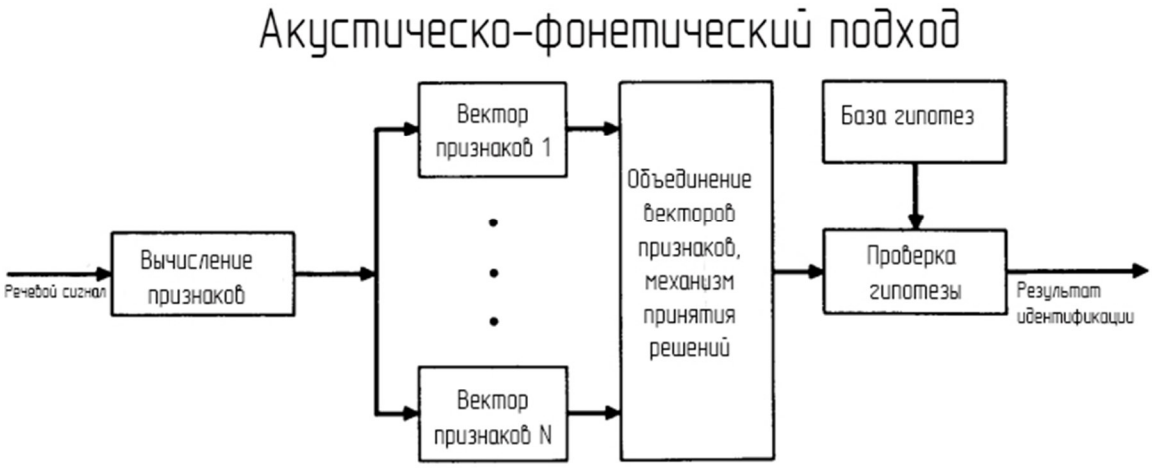


Рисунок 2.1 – Принципиальная схема идентификации отрезков речевых сигналов

Существует несколько подходов к получению векторов признаков из речевого сигнала. Все они задаются общей целью уменьшить избыточность сигнала и выделить наиболее релевантную информацию, и, в то же время, отбросить нерелевантную. Как правило, признаки, описывающие речевой сигнал с разных точек зрения, комбинируются в один вектор признаков, на основе которого происходит идентификации использованием выбранной обученной модели. Далее будут представлены наиболее популярные признаки, выделяемые из речевых сигналов. В данной работе будут исследованы следующие векторы признаков:

1- Декомпозиция сигнала банком фильтров;
2- Распределение мгновенных энергий сигнала;
3- Распределение долей энергии по частотным интервалам;
4- Распределение информационных интервалов;

5- Частота переходов через ноль;

6- Ширина частотной области, занимаемая сигналом;

7- Мел-кепстральные коэффициенты.

Как было отмечено в первой главе сам речевой сигнал во временной области не обладает достаточной репрезентативностью для проведения распознавания в виду высокой избыточности, следовательно, необходимо выбрать такое признаковое пространство, которое бы позволило решать поставленную задачу распознавания речевых сигналов с заданной точностью.

Признаки РС, соответствующих различным фонемам, имеют отличия в виду различий в природе порождающих их воздействий. Необходимо найти такие различия в признаках, которые могли бы обеспечить возможность объединения фонем, имеющих одну природу в один класс, а также максимально разделить фонемы соответствующие различным звукам.

Любой участок РС можно, некоторым образом, характеризовать через определенный признак, соответствующий данному отрезку, или же набор таких признаков. Описанные в первой главе методы цифровой обработки РС позволяют в определенной степени отражать природу порождающего воздействия. Однако, в виду того что речевой сигнал представляет собой лишь отдельную реализацию передаваемой человеком информации, то невозможно характеризовать весь возможный набор таких реализаций всего лишь одним признаком. [25] Речевой сигнал зависит от множества факторов, как внешних (шумы, окружающая обстановка и т.п.), так и зависящих от источника информации – человека (темп речи, эмоциональная окраска, тембр, и др.), что также говорит о том, что необходимо подобрать такой набор признаков, который позволит выделять тождественные (в информационном смысле) реализации из речевого потока [12].

Речевой сигнал, представляющий собой акустическую реализацию передаваемой человеком информации, содержит в себе некоторое сообщение. Передаваемое одним и тем же человеком сообщение, например, предложение “Я иду в университет” при акустической реализации может

принимать различные формы, которые могут отличаться по множеству параметров: громкость, длительность, эмоциональная окраска, высота т.п.

Одним из наиболее популярных подходов к распознаванию речевых сигналов является акустически-фонетический подход, принципиальная схема которого представлена на рисунке 2.1 [50]. Речевой сигнал подвергается процедуре выделения признаков, в результате которой формируются векторы признаков, которые позволяют классифицировать отрезки речевых сигналов. Использование набора векторов признаков обусловлено необходимостью учитывать различные характеристики сигнала, которые невозможно детектировать всего лишь одним вектором признаков. Таким образом, для получения наиболее полного описания речевого сигнала в каждый момент времени естественно использовать совокупность признаков, которые позволят осуществлять точную идентификацию отрезков.

Существует несколько подходов к получению векторов признаков из речевого сигнала. Все они задаются общей целью уменьшить избыточность сигнала и выделить наиболее релевантную информацию, и, в то же время, отбросить нерелевантную. Как правило, признаки, описывающие речевой сигнал с разных точек зрения, комбинируются в один вектор признаков, на основе которого происходит идентификации использованием выбранной обученной модели. Далее будут представлены наиболее популярные признаки, выделяемые из речевых сигналов. В данной работе будут исследованы следующие векторы признаков:

Как было отмечено в первой главе сам речевой сигнал во временной области не обладает достаточной репрезентативностью для проведения распознавания в виду высокой избыточности, следовательно, необходимо выбрать такое признаковое пространство, которое бы позволило решать поставленную задачу распознавания речевых сигналов с заданной точностью.

Признаки РС, соответствующих различным фонемам, имеют отличия в виду различий в природе порождающих их воздействий. Необходимо найти такие различия в признаках, которые могли бы обеспечить возможность

объединения фонем, имеющих одну природу в один класс, а также максимально разделить фонемы соответствующие различным звукам.

2.2 Вычислительные аспекты субполосного анализа речевых сигналов в задачах распознавания речевых сигналов

При субполосном анализе [14] распределения энергий отрезков РС область оси нормированных частот $[0, \pi]$ разбивается на ряд неперекрывающихся интервалов границы, которых определяет вектор следующего вида: $v = [0, \Delta, \Delta + 2\Delta, ...(n\Delta - 4\Delta) + 2\Delta, n\Delta]$, где ширина нулевого частотного интервала $[0, \Delta)$ определяется как: $\Delta = \frac{\pi}{2R+1}$; а R – количество интервалов, на которые разбивается ось частот, данная переменная находится в прямой зависимости от длительности окна анализа N: $R \le \left\lceil \frac{N}{4} \right\rceil$; последующая ширина частотных интервалов равна удвоенной ширине первого интервала.

Значение длительности окна при анализе РС целесообразно [50] брать равным такому количеству отчетов дискретного сигнала, которое бы соответствовало 10 - 20 мс аналогового сигнала.

Представляется целесообразным взять значение окна анализа N равным 256, которое соответствует отрезку времени $t = 16\,мс$.

Запишем вектор, задающий границы интервалов в следующем виде:

$$v = [v_r, v_{r+1}) \cup [v_r, v_{r+1}); 0 \le v_r < v_{r+1} \le \pi; \qquad (2.1)$$

где $r = 0 \ldots R$;

Из описанного в главе 1.6 следует, что распределение долей энергии отрезка сигнала в заданном частотном интервале может быть оценено на основе выражения [14]:

$$P_r(\vec{x}_N) = \vec{x}_N^T A_r \vec{x}_N, \qquad (2.2)$$

где $\vec{x}_N$ – анализируемый отрезок сигнала; $A_r = \{a_{ik}^r\},\ i,k = 1,..,N$ субполосная матрица с элементами вида [14]:

$$a_{ik} = \frac{\sin[v_{r+1}(i-k)] - \sin[v_r(i-k)]}{\pi(\text{i - k})} \tag{2.3}$$

В соответствии с вектором разбиения частотной полосы, элементы матрицы, соответствующие нулевому частотному интервалу, могут быть представлены в виде:

$$a_{ik}^0 = \frac{\sin[v_{r+1}(i-k)] - \sin[v_r(i-k)]}{\pi(\text{i - k})} = \frac{\sin[v_{r+1}(i-k)]}{\pi(\text{i - k})} \tag{2.4}$$

Тогда для других частотных интервалов, элементы матрицы могут быть представлены в виде:

$$a_{ik}^r = \frac{2 \cdot \sin[(v_{r+1} - v_r) \cdot {}^{(i-k)}\!/_{2}] \cdot \cos[(v_{r+1} - v_r) \cdot {}^{(i-k)}\!/_{2}]}{\pi(\text{i - k})} \tag{2.5}$$

С учетом вектора разбиения (2.1) и выражения (2.4) выражение (2.5) может быть переписано в виде:

$$a_{ik}^r = 2 \cdot a_{ik}^0 \cos(\omega_r (i-k)), \tag{2.6}$$

где $\omega_r = \frac{4\pi \cdot (r+1)}{N}$ - центральная частота интервала, $r = 0...R$

Субполосные матрицы являются симметричными и неотрицательно определенными. Поэтому они обладают полным набором ортонормальных собственных векторов и соответствующих им неотрицательных собственных чисел, для которых выполняются соотношения:

$$\lambda_{kr}\vec{q}_{kr} = A_r\vec{q}_{kr}, \tag{2.7}$$

$$(\vec{q}_{kr}, \vec{q}_{ir}) = \sum_{j=1}^{N} q_{jk}^r q_{ji}^r = \begin{cases} 1, i = k \\ 0, i \neq k \end{cases}, \tag{2.8}$$

Причем можно представить субполосную матрицу выражением:

$$A_r = \sum_{k=1}^{N} \lambda_{kr}\vec{q}_{kr}\vec{q}_{kr}^T = QLQ^T \tag{2.9}$$

где $Q = \{\vec{q}_{kr}\}, k = 1,.., N$; матрица собственных векторов, а $L = diag(\lambda_1^r, \lambda_2^r, .., \lambda_N^r)$ - диагональная матрица собственных чисел. Подставляя (2.9) в выражение (2.2):

$$P_r(\vec{x}_N) = \vec{x}_N^T A_r \vec{x}_N = \vec{x}_N^T \cdot \sum_{k=1}^{N} \lambda_{kr} \vec{q}_{kr} \vec{q}_{kr}^T \cdot \vec{x}_N =$$

$$= \sum_{k=1}^{N} \lambda_{kr} \vec{x}_N^T \vec{q}_{kr} \vec{q}_{kr}^T \vec{x}_N = \sum_{k=1}^{N} \lambda_{kr} \alpha_{kr}^2 , \qquad (2.10)$$

где α_{kr} – скалярные произведения (проекции):

$$\alpha_{kr} = (\vec{q}_{kr}, \vec{x}_N) \qquad (2.11)$$

Из выражения (2.10) видно, что свойства собственных чисел субполосной матрицы таковы, что количество не нулевых значений относительно мало, таким образом можно сократить количество вычислений, с помощью исключения из процесса расчета близких к нулю значений собственных чисел. Для этого собственные числа суболосной матрицы должны быть отсортированы по убыванию.

$$\lambda_1^r > \lambda_2^r > ... > \lambda_N^r \geq 0 . \qquad (2.12)$$

И выбираем J собственных чисел из отсортированного множества матрицы L_r, и соответствующие им собственные вектора $\vec{q}$.

$$J = 2[N \cdot a_{ii}^r / 2\pi] + 5 ; \qquad (2.13)$$

Из выбранных собственных чисел и векторов формируем матрицу следующего вида:

$$AA = \begin{pmatrix} \sqrt{L_1} Q_1^T \\ \sqrt{L_2} Q_2^T \\ \dots \\ \sqrt{L_R} Q_R^T \end{pmatrix} \tag{2.14}$$

где элементы соответствующие частному интервалу могут быть представлены как $AA_r = \sqrt{L_r} Q_r^T$

При этом операция извлечения квадратного корня обусловлена необходимостью соблюсти выражение (2.9), т.к. матрица АА используется дважды: для получения приближенных значений долей энергии и восстановления нужной составляющей сигнала.

Для получения приближенных значений частотных составляющих необходимо используя собственные вектора матрицы A_r получить вектор:

$$\vec{y}\vec{y} = AA\,\vec{x} \tag{2.15}$$

Чтобы получить частотную составляющую сигнала, советующую временному отрезку воспользуемся формулой:

$$\vec{y}_r = AA_r^T\,\vec{y}\vec{y}_k \tag{2.16}$$

где $k = [r \cdot j + 1] \dots [r + 1] \cdot j;\ r = 0 \dots R$

Словесное описание алгоритма предварительного расчета матрицы

1. Ввести значения длительности окна анализа N;
2. Расчет значений алгоритма R, Δ;
3. Построение вектора задающего границы частотных интервалов (2.1);
4. Сформировать матрицу $A_r = \{a_{ik}^r\},\ i,k = 1,..,N$;
5. Расчет количества сохраняемых собственных чисел J;

6. Определить собственные вектора $\vec{q}_k^r$ и числа λ_N^r субполосной матрицы A_r;

7. Сформировать матрицу собственных векторов $Q = \left\{ \vec{q}_k^r \right\}, k = 1,..,N$;

8. Сформировать матрицу $L = diag(\lambda_1^r, \lambda_2^r, .., \lambda_N^r)$ из собственных чисел, отсортированных по убыванию (2.12);

9. Сформировать матрицу AA (2.14);

10. Сохранить матрицу AA для дальнейших расчетов.

Матрица АА в дальнейшем используется как банк фильтров, который позволяет выделить необходимые признаки, например, распределение энергии сигнала по частотной оси. Предварительный расчет матрицы позволяет существенно снизить вычислительную нагрузку при вычислении векторов признаков.

На рисунке 2.2 представлена блок-схема алгоритма предварительного расчета фильтрующей матрицы АА. В дальнейшем, данная матрица будет применятся для анализа речевого сигнала и выделения векторов признаков.

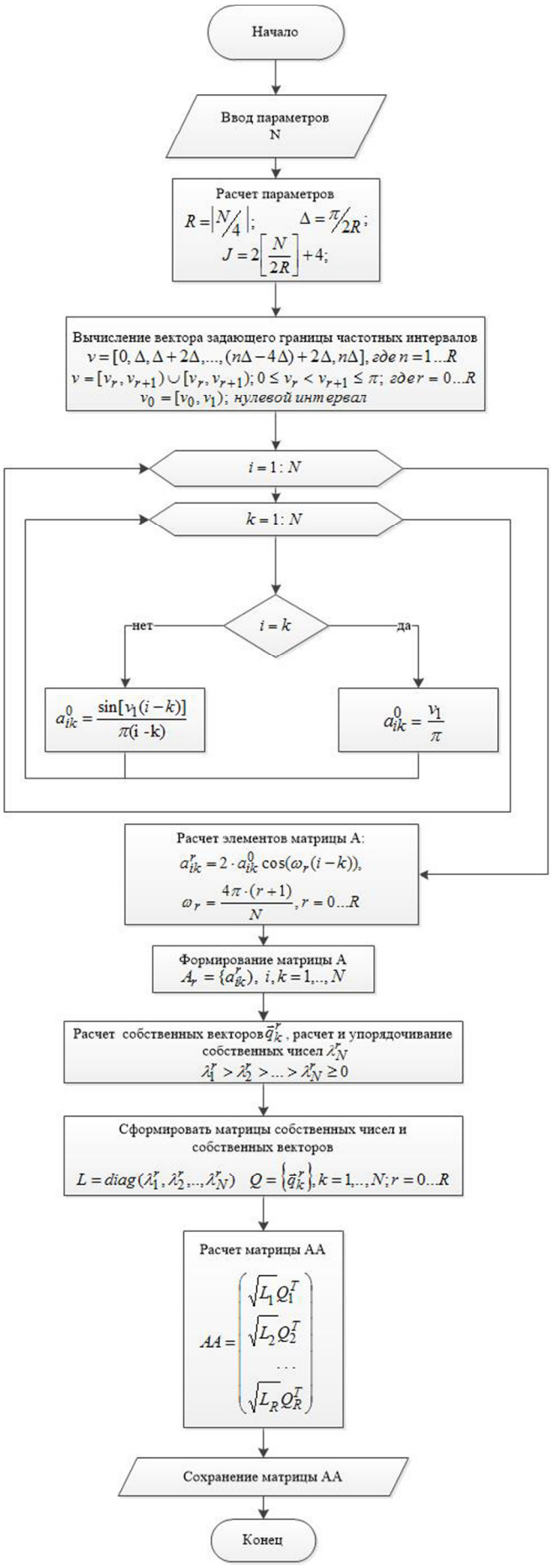


Рисунок 2.2 – Блок схема алгоритма вычисления матрицы AA

2.3 Исследование пространств признаков в задачах распознавания речевых сигналов

2.3.1 Декомпозиция сигнала банком фильтров

Одним из признаков сигнала, может служить декомпозиция сигнала на определенное количество частотных компонент. В данной работе декомпозиция сигнала будет выполняться субполосным банком фильтров, с помощью рассчитанной заранее матрицы AA.

Для получения приближенных значений частотных составляющих необходимо используя собственные вектора матрицы A_r получить вектор (2.15):

$$\vec{y}\vec{y} = AA\,\vec{x}$$

где $\vec{x} = (x_1 \ldots . x_N)$, речевой сигнал.

Чтобы получить нужную частотную составляющую сигнала, советующую временному отрезку воспользуемся формулой (2.16):

$$\vec{y}_r = AA_r^T\,\vec{y}\vec{y}_k$$

где $k = [r \cdot j + 1] \ldots [r + 1] \cdot j;\ r = 0 \ldots R$.

Длина вектора соответствует длине анализируемого сигнала.

На рисунках 2.3-2.5 представлены отрезки сигнала, соответствующие произнесенным словам “восемь” и “семь” и декомпозиции этих сигналов. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

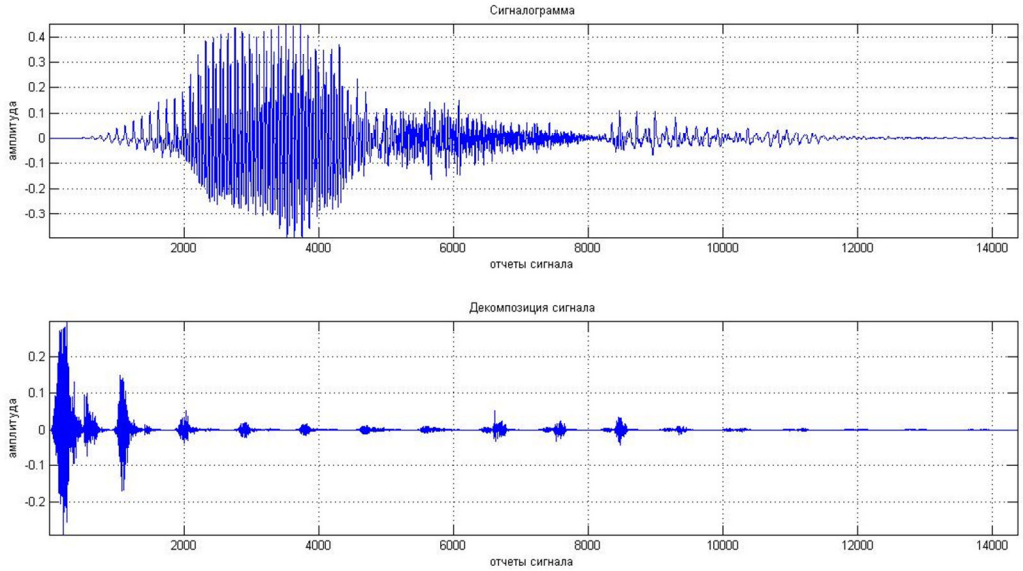


Рисунок 2.3 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – декомпозиция сигнала банком фильтров

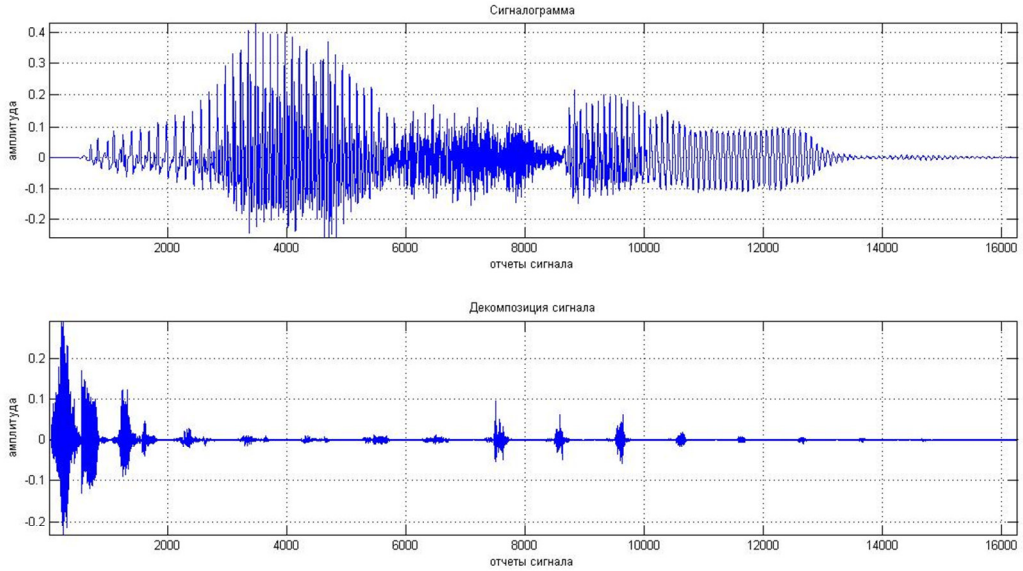


Рисунок 2.4 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – декомпозиция сигнала банком фильтров

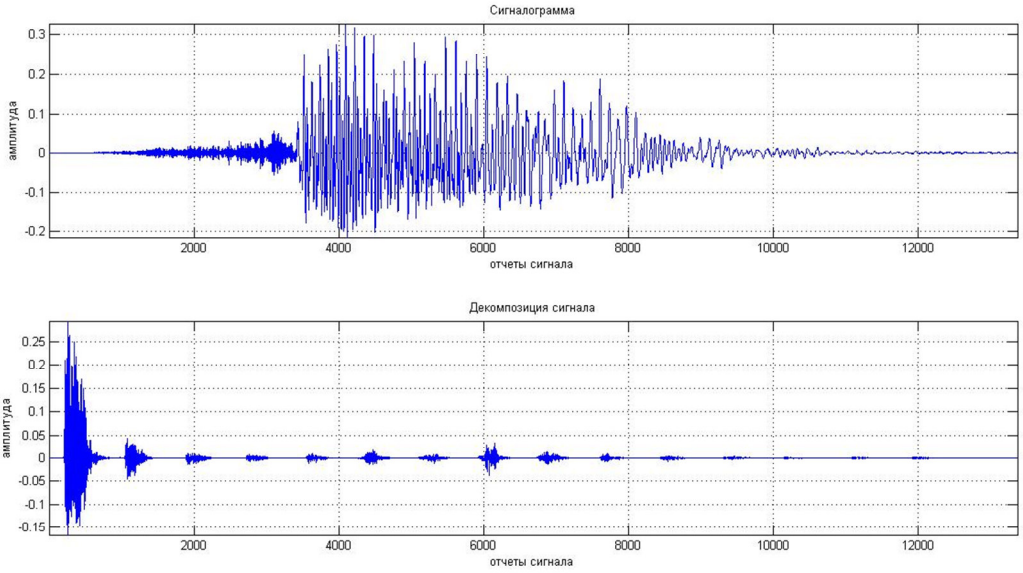


Рисунок 2.5 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “семь” и вектор признаков – декомпозиция сигнала банком фильтров

2.3.2 Распределение мгновенных энергий отрезка РС

Другим, немаловажным признаком может служить вектор распределения мгновенных энергий.

Для получения мгновенных энергий для частотного интервала r воспользуемся формулой:

$$\psi_r(i) = y_r^2(i) \tag{2.17}$$

Длина вектора соответствует длине анализируемого сигнала.

На рисунках 2.6-2.8 представлены отрезки сигнала соответствующие произнесенным словам “восемь” и “семь” и распределение мгновенных энергий этих сигналов. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

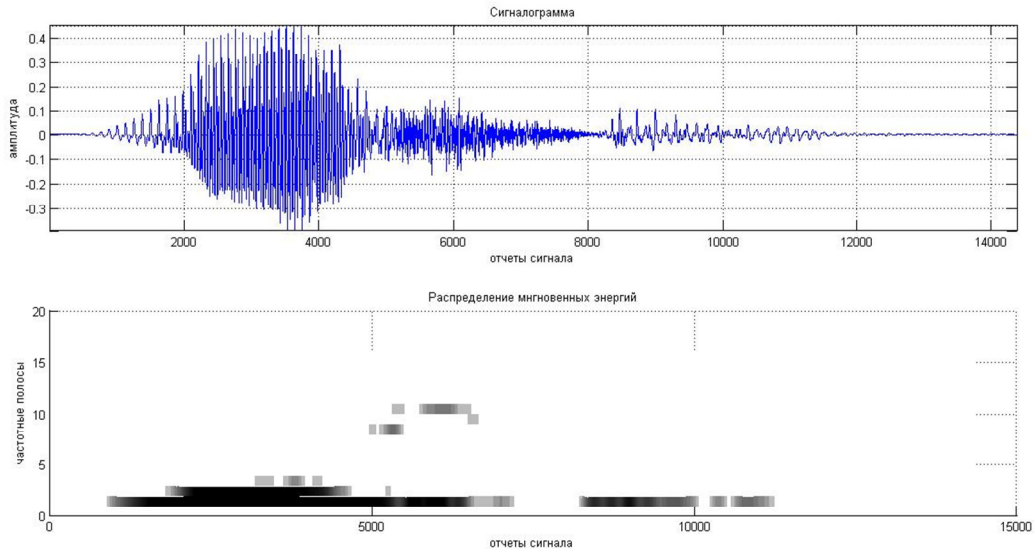


Рисунок 2.6 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение мгновенных энергий PC

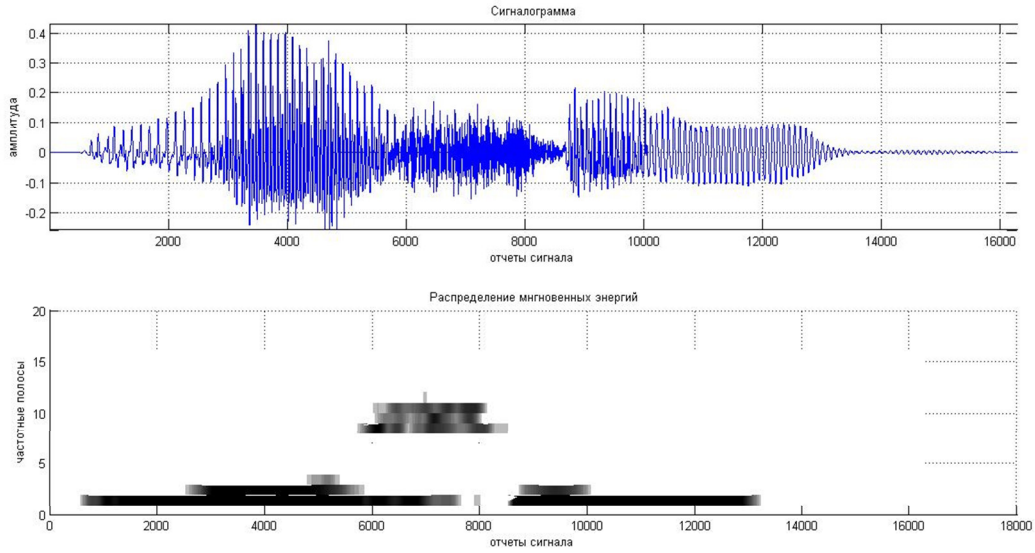


Рисунок 2.7 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение мгновенных энергий PC

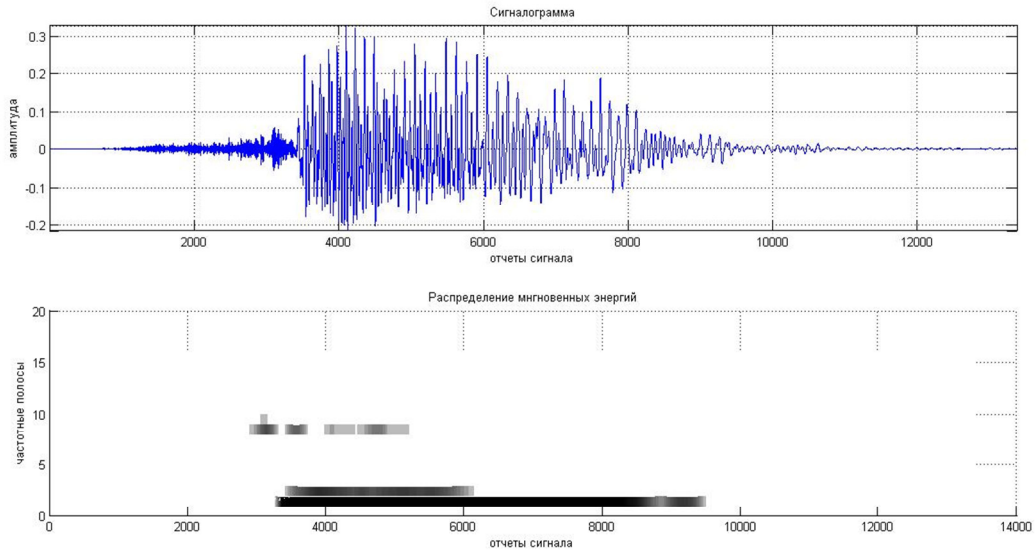


Рисунок 2.8 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение мгновенных энергий РС

2.3.3 Распределение долей энергии отрезка РС

Распределение долей энергии по частотным интервалам является основным, и одним из наиболее информативных признаков речевого сигнала.

Для получения распределения долей энергии воспользуемся ранее вычисленным признаком $\vec{y}_r$ и применив формулу:

$$\mathrm{P}_r = \sum_{i=1}^{N} y_{r_i}^2 \qquad (2.18)$$

где $i = 1....N$, отчеты сигнала $\vec{x} = (x_1....x_N)$.

В результате сформируем вектор предложенного признака в следующем виде:

$$\vec{\mathrm{P}} = (P_1...P_R);$$

где R, количество частотных полос.

Получим распределение долей энергии отрезка искомого сигнала. Длина вектора соответствует количеству частотных интервалов R.

На рисунках 2.9-2.11 представлены отрезки сигнала, соответствующие произнесенным словам “восемь” и “семь” и распределение долей энергии

этих сигналов. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

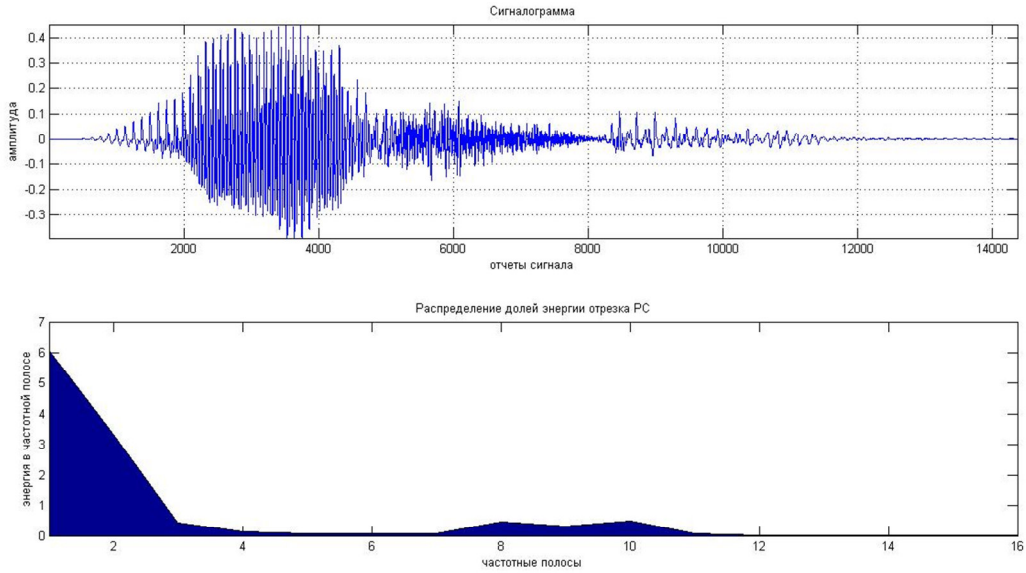


Рисунок 2.9 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение долей энергии PC

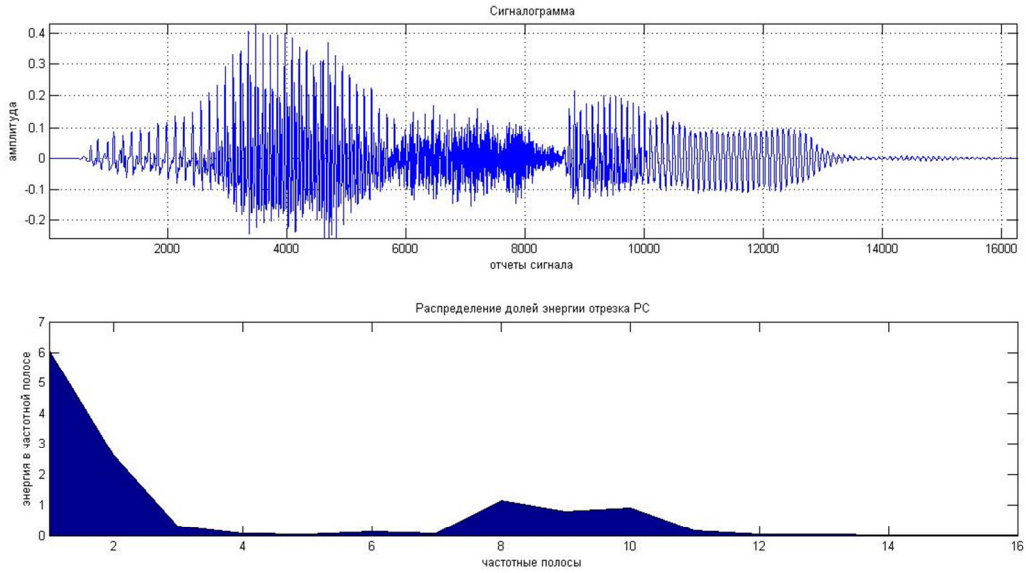


Рисунок 2.10 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение долей энергии PC

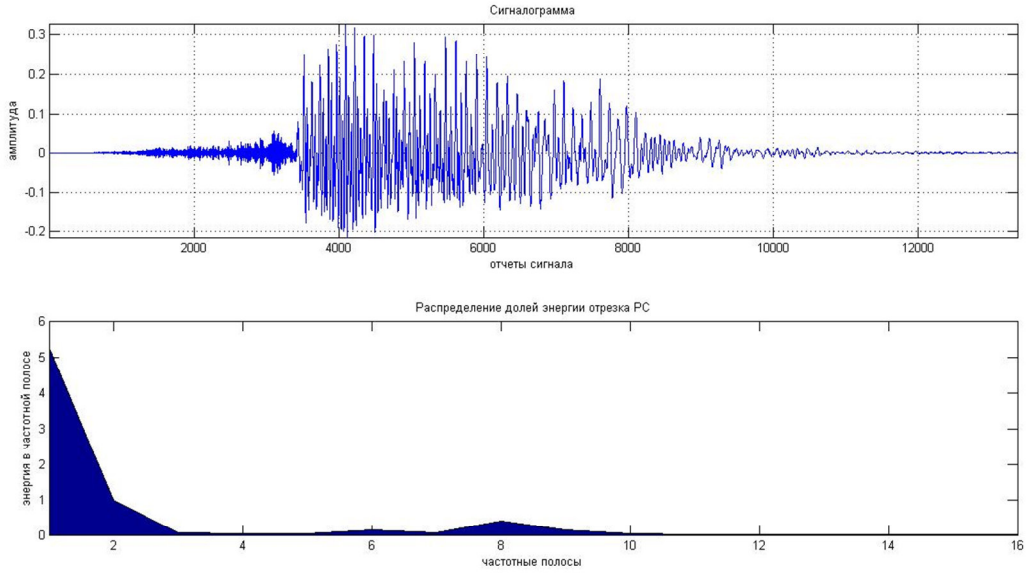


Рисунок 2.11 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “семь” и вектор признаков – распределение долей энергии РС

2.3.4 Распределение информационных интервалов отрезка РС

Для селекции наиболее информативных признаков данного вектора произведем расчет порога:

$$h_{\psi} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 / R \tag{2.19}$$

где $i = 1....N$, отчеты сигнала $\vec{x} = (x_1 ... x_N)$,

Сравнивания значения $\vec{\psi}_r$ с порогом выполняется следующее выражение:

$$\psi_r(i) = \begin{cases} \psi_r(i),\ \psi_r(i) \geq h_{\psi}, \\ 0,\ \psi_r(i) < h_{\psi} \end{cases} \tag{2.20}$$

Матрица значения для всех частотных интервалов:

$$\Psi = \begin{bmatrix} \vec{\psi}_1 \\ ... \\ \vec{\psi}_R \end{bmatrix} \tag{2.21}$$

где R, количество частотных полос.

Для получения нормированного распределения информационных интервалов необходимо суммировать значения, полученные в матрице (2.21) по столбцам:

$$\xi_r = \frac{1}{N}\sum_{i=1}^{N}\psi_r(i) \tag{2.22}$$

Результат – вектор распределения информационных интервалов:

$$\vec{\xi} = (\xi_1 .. \xi_R) \tag{2.23}$$

где R, количество частотных полос.

На рисунках 2.12-2.14 представлены отрезки сигнала, соответствующие произнесенным словам “восемь” и “семь” и распределение информационных интервалов этих сигналов. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

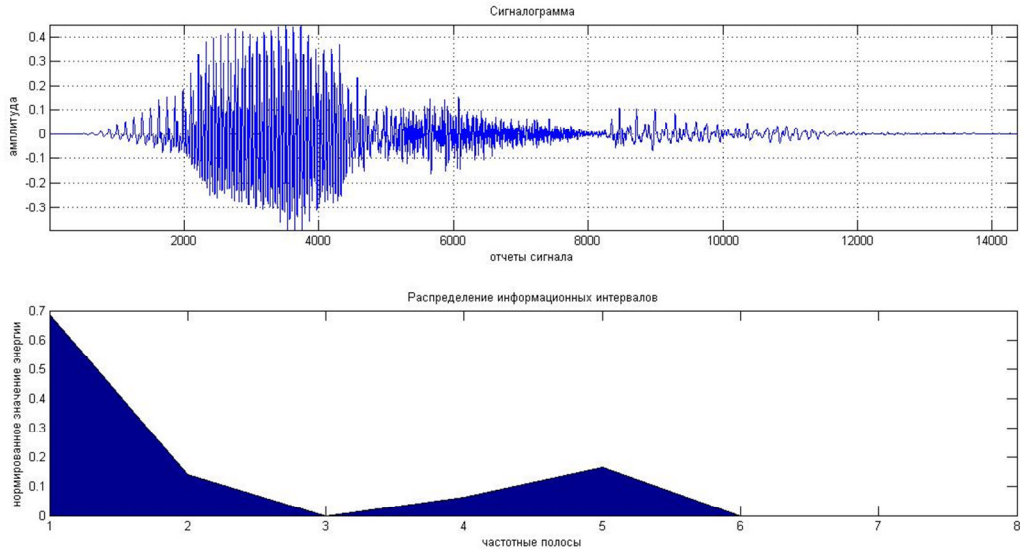


Рисунок 2.12 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение информационных интервалов

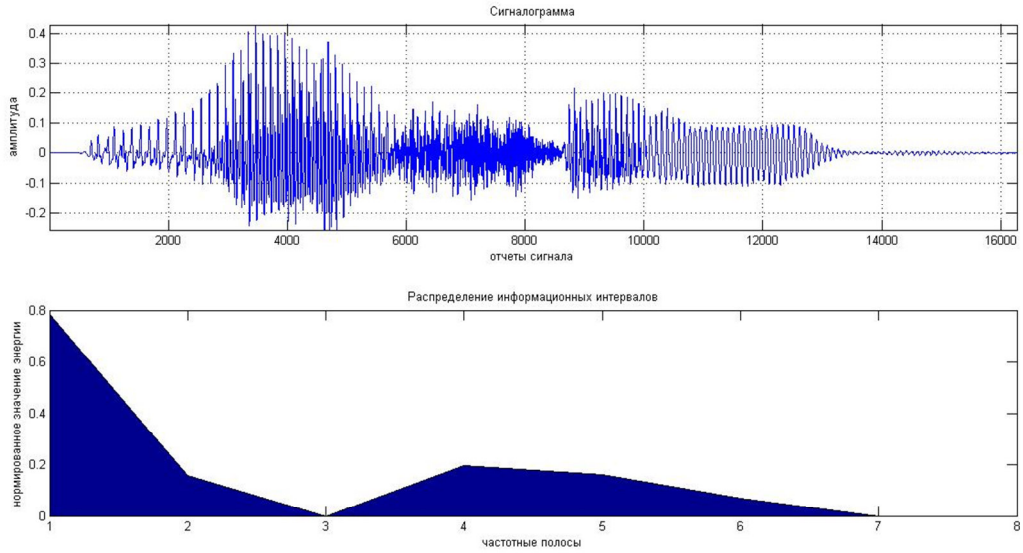


Рисунок 2.13 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – распределение информационных интервалов

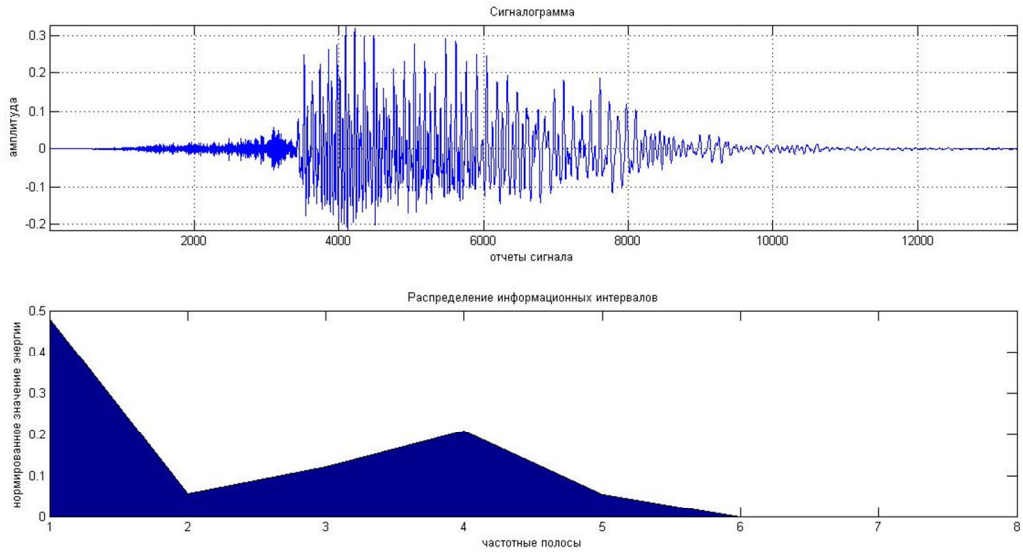


Рисунок 2.14 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “семь” и вектор признаков – распределение информационных интервалов

2.3.5 Частота переходов через ноль

При обработке сигналов в дискретном времени, считается, что если два последовательных отчета имеют различные знаки, то произошел переход

сигнала через ноль. Частота появления нулей в сигнале может служить одной из характеристик его спектральных свойств. Это наиболее справедливо для узкополосных синусоидальных сигналов. [33]

При анализе широкоплосных речевых сигналов частота переходов через ноль менее очевидно описывает спектральные свойства сигнала, однако даже грубые оценки, полученные таким образом являются достаточно весомым признаком.

Признак, показывающий частоту переходов сигнала через ноль, на заданном участке речевого сигнала может быть использован для выделения шумоподобных фонем, основным признаком которых является высокая частота подобных переходов. Результаты работы данной функции вкупе с таким признаком как ширина частотной полосы, занимаемая сигналом, позволяют детектировать участки, которые можно отнести к шумным согласным.

Определим среднее число переходов сигнала через ноль следующим образом:

$$\sigma_n = \sum_{i=1}^{N} \left| \mathrm{sgn}[x_i] - \mathrm{sgn}[x_{i-1}] \right| \cdot w(n-i), \tag{2.17}$$

где

$$\mathrm{sgn}[x_i] = \begin{cases} 1, & x_i \geq 0; \\ -1, & x_i < 0 \end{cases} \tag{2.18}$$

а также

$$w(n) = \begin{cases} 1/2N, 0 \leq n \leq N-1, \\ 0, в\ противном\ случае \end{cases} \tag{2.19}$$

Таким образом, можно сформировать вектор предложенного признака:

$$\vec{\omega} = (w_1 .. w_N); \tag{2.20}$$

где N, количество отчетов сигнала $\vec{x} = (x_1 .... x_N)$.

Модель речеобразования предполагает, что энергия вокализованных сегментов речевого сигнала концентрируется на частотах ниже 3 кГц, что

обусловлено убывающим спектром сигнала возбуждения, тогда как для невокализованных сегментов большая часть энергии лежит в областях высоких частот. Поскольку высокие частоты приводят к большому числу переходов через ноль, а низкие к малому, то существует жесткая связь между числом нулевых пересечений и распределением энергии по частотам. [33]

Можно предположить, что большому числу нулевых пересечений соответствует невокализованные сегменты, а малому числу – вокализованные сегменты. Однако в виду неопределенности меры в данном случае, возникает сложность в однозначном определении. Для уточнения результатов, полученной от расчета данной характеристики можно использовать коэффициент ширины частотной области занимаемой сигналом.

На рисунках 2.15-2.17 представлены отрезки сигнала, соответствующие словам “восемь” и “семь” и функция частоты переходов сигнала через ноль. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

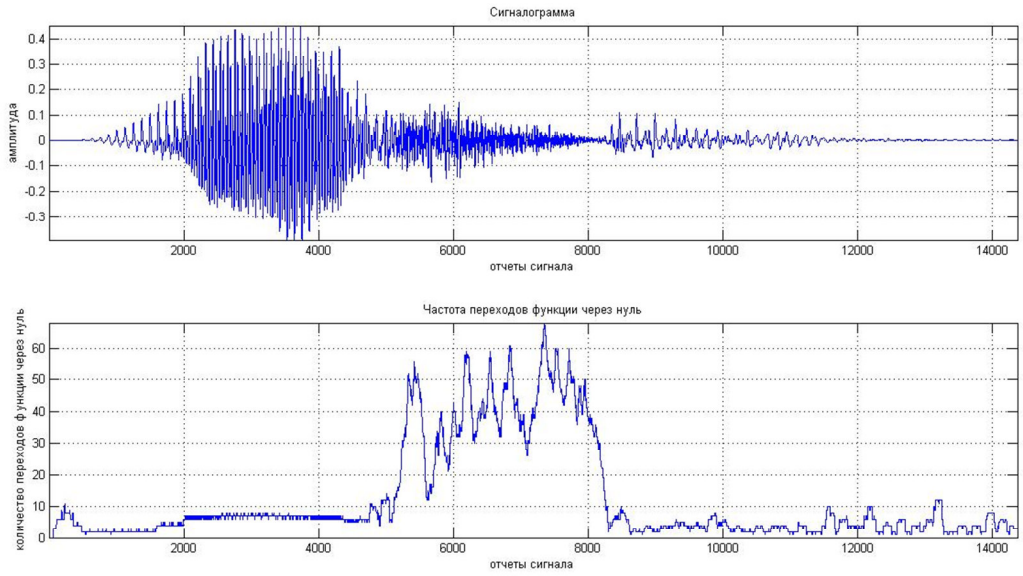


Рисунок 2.15 – Сигналограмма отрезка речевого сигнала, соответствующего слову “восемь” и вектор признаков – Частота переходов функции через ноль

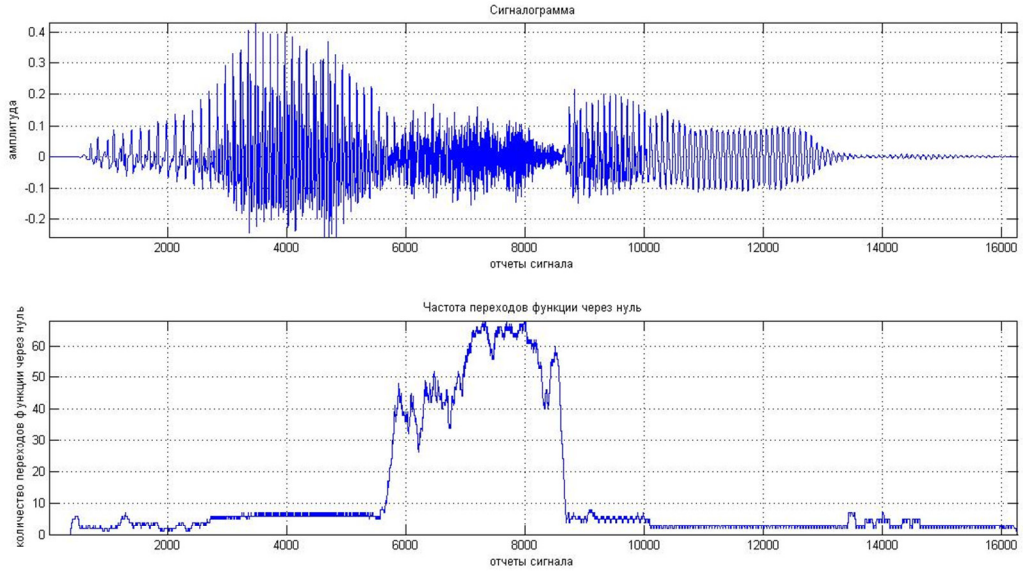


Рисунок 2.16 – Сигналограмма отрезка речевого сигнала, соответствующего слову “восемь” и вектор признаков – Частота переходов функции через ноль

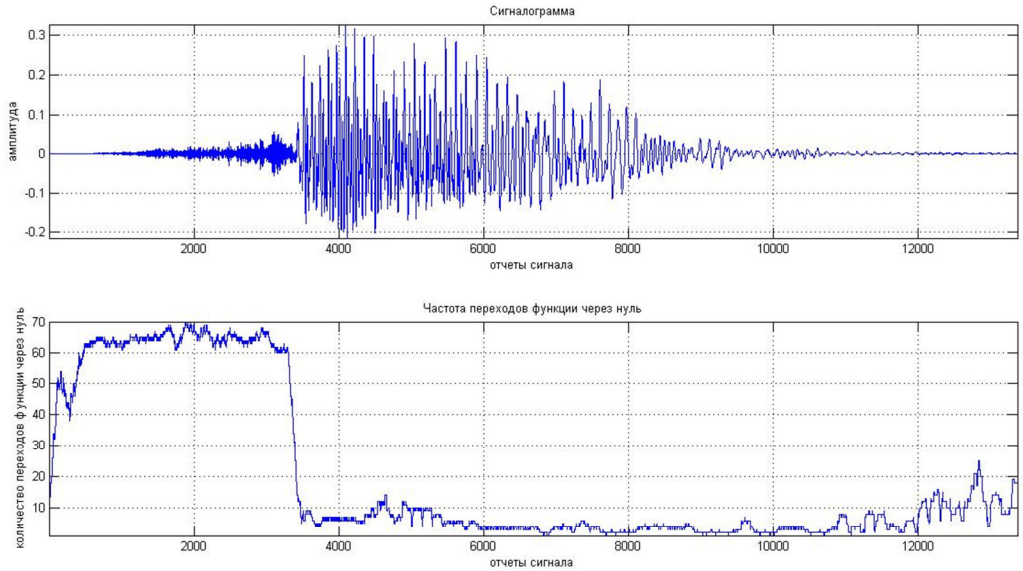


Рисунок 2.17 – Сигналограмма отрезка речевого сигнала, соответствующего слову “семь” и вектор признаков – Частота переходов функции через ноль

2.3.6 Ширина частотной области, занимаемая сигналом

Исходя из результатов исследования распределения информационных интервалов, для звуков характерной особенностью может служить ширина частотной области, в которой сконцентрированы информационные интервалы. Например, вокализованные звуки отличаются довольно плотной концентрацией информационных интервалов (~10% от общей полосы), в то время как невокализованные звуки (Ж, З, Ф, Х, Ц, Т, Ч, С, Ш, Щ) занимают порядка ~45-70% частотной полосы. Следовательно, можно сделать вывод, о том, что данный признак может служит для идентификации невокализованных отрезков речевого сигнала. [25,26]

Для оценки ширины частотной области, занимаемой сигналом, предлагается сформировать следующий вектор: $\vec{\upsilon}=(\upsilon_1..\upsilon_N)$, N – длительность анализируемого отрезка.

Для оценки ширины частотной области необходимо подвергнуть анализу распределение долей энергий сигнала полученные в выражении (2.20)

$$\Psi = \begin{bmatrix} \psi_{11} & \cdots & \cdots & \psi_{1N} \\ \psi_{21} & \cdots & \cdots & \psi_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{R1} & \cdots & \cdots & \psi_{RN} \end{bmatrix},$$

где R - количество частотных интервалов, а N - длина анализируемого сигнала.

Необходимо осуществить сортировку значений матрицы Ψ в каждом столбце N по убыванию:

$$\psi_1^n > \psi_2^n > ... > \psi_r^n \geq 0 \qquad (2.21)$$

Далее произведем вычисление суммы энергии для каждого отчета сигнала по всем интервалам:

$$\Sigma_n = \sum_{r=1}^{R} \psi_r^n \qquad (2.22)$$

В результате сформируем вектор $\vec{\Sigma}$:

$$\vec{\Sigma} = (\Sigma_1 .. \Sigma_N); \tag{2.23}$$

где N, количество отчетов сигнала $\vec{x} = (x_1 .... x_N)$.

Производим расчет ширины частотной области, занимаемой сигналом в соответствии с установленным коэффициентом концентрации энергии *K*, с помощью буферного вектора $\vec{\beta}_r$:

$$\vec{\beta}_r = \sum_{i=1}^{r} \vec{\psi}_i \tag{2.24}$$

Коэффициент концентрации энергии *K* несет в себе следующий смысл – относительное количество энергии, сосредоточенной в минимальном количестве частотных интервалов.

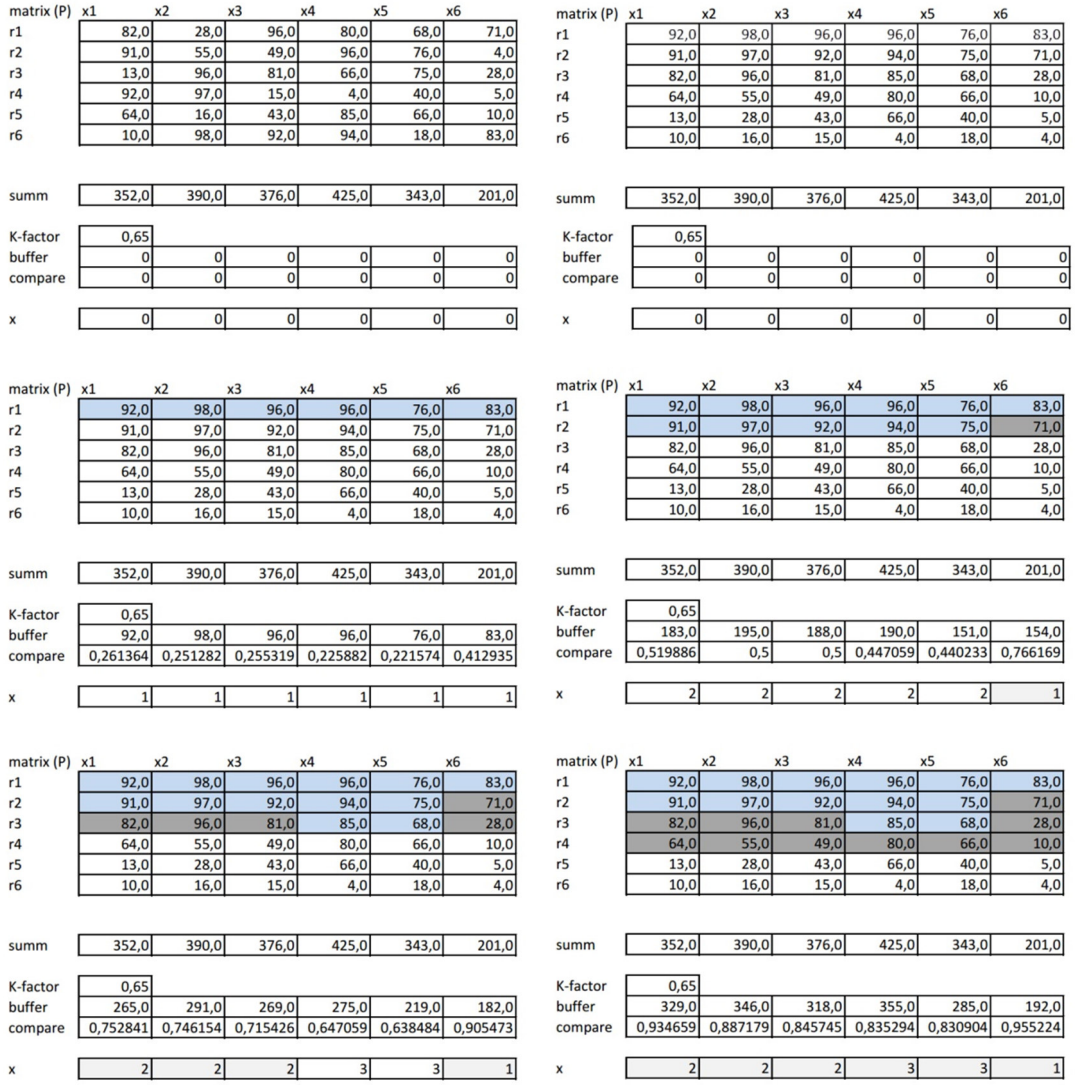
Далее производим последовательное суммирование вектора $\vec{\beta}_r$ для каждой строки и сравнение с вектором $\vec{\Sigma}$:

$$\upsilon(n) = \begin{cases} i, \text{ если } \beta_r(i) / \Sigma(i) \geq K \\ \upsilon(n), \text{ если } \beta_r(i) / \Sigma(i) < K \end{cases} \tag{2.25}$$

На выходе получаем вектор значений признака в следующей форме:

$$\vec{\upsilon} = (\upsilon_1 .. \upsilon_N) \tag{2.26}$$

На рисунке 2.18 представлена визуализация алгоритма расчет ширины частотной области.



matrix (P)	x1	x2	x3	x4	x5	x6
r1	82,0	28,0	96,0	80,0	68,0	71,0
r2	91,0	55,0	49,0	96,0	76,0	4,0
r3	13,0	96,0	81,0	66,0	75,0	28,0
r4	92,0	97,0	15,0	4,0	40,0	5,0
r5	64,0	16,0	43,0	85,0	66,0	10,0
r6	10,0	98,0	92,0	94,0	18,0	83,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	0	0	0	0	0	0
compare	0	0	0	0	0	0
x	0	0	0	0	0	0

matrix (P)	x1	x2	x3	x4	x5	x6
r1	92,0	98,0	96,0	96,0	76,0	83,0
r2	91,0	97,0	92,0	94,0	75,0	71,0
r3	82,0	96,0	81,0	85,0	68,0	28,0
r4	64,0	55,0	49,0	80,0	66,0	10,0
r5	13,0	28,0	43,0	66,0	40,0	5,0
r6	10,0	16,0	15,0	4,0	18,0	4,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	0	0	0	0	0	0
compare	0	0	0	0	0	0
x	0	0	0	0	0	0

matrix (P)	x1	x2	x3	x4	x5	x6
r1	92,0	98,0	96,0	96,0	76,0	83,0
r2	91,0	97,0	92,0	94,0	75,0	71,0
r3	82,0	96,0	81,0	85,0	68,0	28,0
r4	64,0	55,0	49,0	80,0	66,0	10,0
r5	13,0	28,0	43,0	66,0	40,0	5,0
r6	10,0	16,0	15,0	4,0	18,0	4,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	92,0	98,0	96,0	96,0	76,0	83,0
compare	0,261364	0,251282	0,255319	0,225882	0,221574	0,412935
x	1	1	1	1	1	1

matrix (P)	x1	x2	x3	x4	x5	x6
r1	92,0	98,0	96,0	96,0	76,0	83,0
r2	91,0	97,0	92,0	94,0	75,0	71,0
r3	82,0	96,0	81,0	85,0	68,0	28,0
r4	64,0	55,0	49,0	80,0	66,0	10,0
r5	13,0	28,0	43,0	66,0	40,0	5,0
r6	10,0	16,0	15,0	4,0	18,0	4,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	183,0	195,0	188,0	190,0	151,0	154,0
compare	0,519886	0,5	0,5	0,447059	0,440233	0,766169
x	2	2	2	2	2	1

matrix (P)	x1	x2	x3	x4	x5	x6
r1	92,0	98,0	96,0	96,0	76,0	83,0
r2	91,0	97,0	92,0	94,0	75,0	71,0
r3	82,0	96,0	81,0	85,0	68,0	28,0
r4	64,0	55,0	49,0	80,0	66,0	10,0
r5	13,0	28,0	43,0	66,0	40,0	5,0
r6	10,0	16,0	15,0	4,0	18,0	4,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	265,0	291,0	269,0	275,0	219,0	182,0
compare	0,752841	0,746154	0,715426	0,647059	0,638484	0,905473
x	2	2	2	3	3	1

matrix (P)	x1	x2	x3	x4	x5	x6
r1	92,0	98,0	96,0	96,0	76,0	83,0
r2	91,0	97,0	92,0	94,0	75,0	71,0
r3	82,0	96,0	81,0	85,0	68,0	28,0
r4	64,0	55,0	49,0	80,0	66,0	10,0
r5	13,0	28,0	43,0	66,0	40,0	5,0
r6	10,0	16,0	15,0	4,0	18,0	4,0
summ	352,0	390,0	376,0	425,0	343,0	201,0
K-factor	0,65					
buffer	329,0	346,0	318,0	355,0	285,0	192,0
compare	0,934659	0,887179	0,845745	0,835294	0,830904	0,955224
x	2	2	2	3	3	1

Рисунок 2.18 – Визуализация алгоритма определения ширины частотной области занимаемой сигналом

На рисунках 2.19-2.21 представлены отрезки сигнала, соответствующие произнесенным словам “восемь” и “семь” и функция ширины частотной области, занимаемой сигналом. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

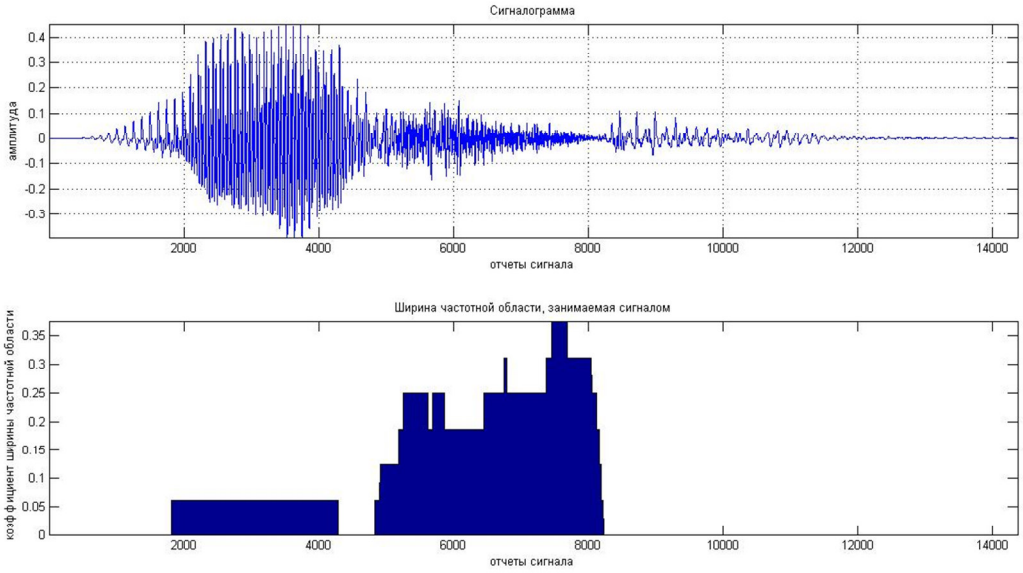


Рисунок 2.19 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – Ширина частотной области, занимаемая сигналом

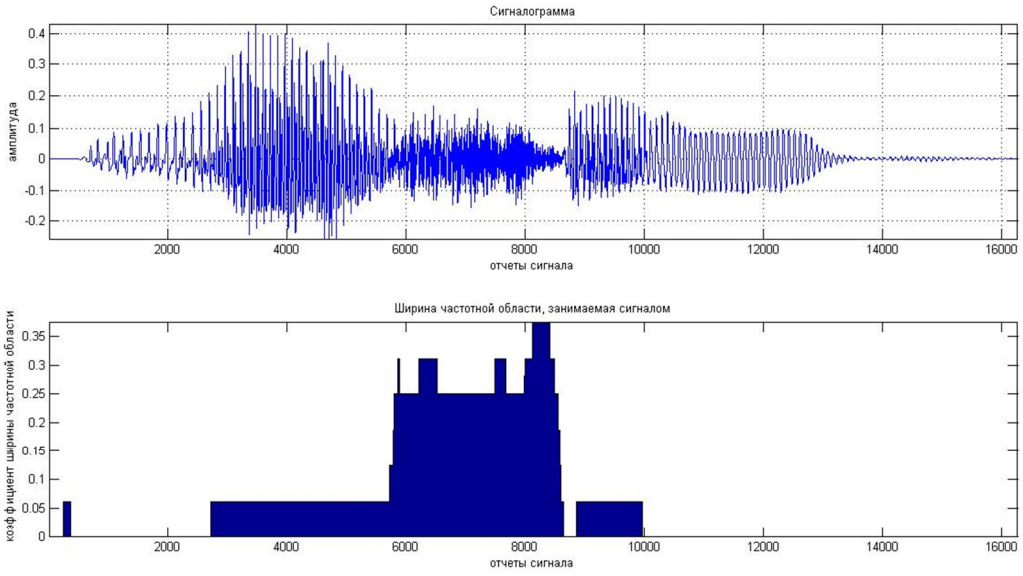


Рисунок 2.20 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – Ширина частотной области, занимаемая сигналом

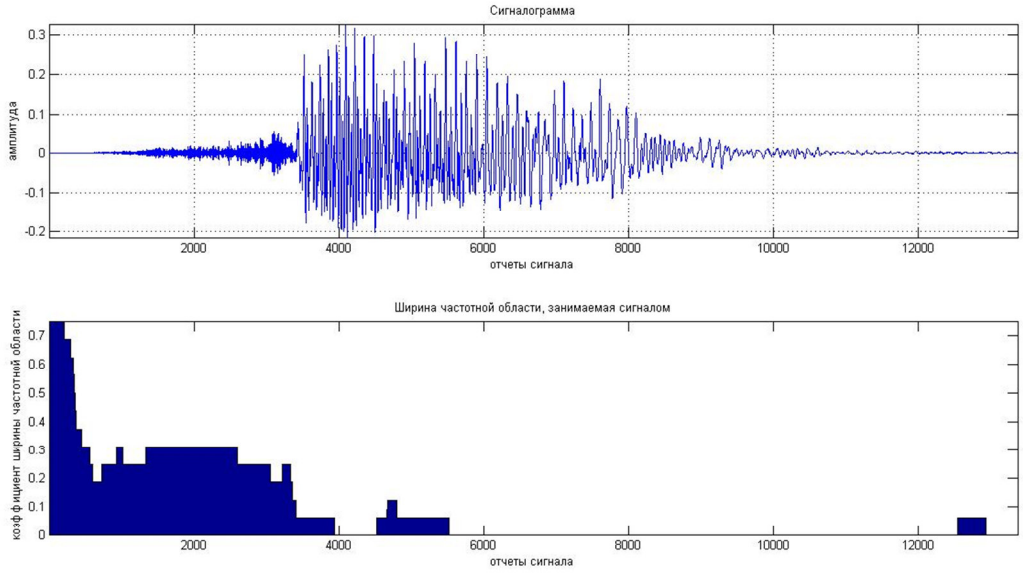


Рисунок 2.21 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “семь” и вектор признаков – Ширина частотной области, занимаемая сигналом

2.3.7 Мел-кепстральные коэффициенты речевого сигнала

Мел-кепстральные коэффициенты представляют собой такое представление сигнала, в котором частотные полосы распределены согласно мел шкале (рисунок 1.9). Это приводит к большей плотности фильтров в области низких частот и меньшей плотности в области высоких частот, что отражает чувствительность восприятия звуковых сигналов человеческим ухом. Таким образом, основная информация рассчитывается из аудио сигнала в области низких частот, что является наиболее репрезентативным признаком для задача распознавания речевых сигналов [25,26,49,50].

Чтобы рассчитать мел-кепстральные коэффициенты требуется совершить следующие действия :

Произвести расчет ДПФ для сигнала (1.9):

$$X(k) = \sum_{i=1}^{N} x(i) e^{-j\frac{2\pi}{N}(i-1)(k-1)}$$

где $k=1,2,\ldots,N$, j – комплексная единица, $j = \sqrt{-1}$.

Полученный спектр сигнала $X(k)$ подается на вход банка фильтров. Банк фильтров представляет собой перекрывающие друг друга треугольные фильтры, распределенные по мел-шкале (рисунок 1.9).

$$H_m = \begin{cases} 0 & \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & k < f[m-1] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m-1] \le k < f[m] \\ 0 & f[m] \le k \le f[m+1] \\ & k > f[m+1] \end{cases} \tag{2.27}$$

Частоты получаем из следующего равенства:

$$f[m] = \left(\frac{N}{F_s}\right) \cdot B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right) \tag{2.28}$$

где

$$B^{-1}(b) = 700(\exp(b/1125) - 1)$$

Далее найдем натуральный логарифм от суммы долей энергии сигнала по интервалам, после оконного преобразования Фурье:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 \cdot H_m(k)\right), \ 0 \le m < M \tag{2.29}$$

Применяя дискретное косинусное преобразование Фурье (ДКП) получим мел-кепстральные коэффициенты для отрезка РС

$$\mu(n) = \sum_{m=0}^{M-1} S(m)\cos(\pi n(m+1/2)/M), \ \ 0 \le n < M \tag{2.30}$$

На рисунках 2.22-2.24 представлены отрезки сигнала, соответствующие произнесенным словам “восемь” и “семь” и мел-кепстральные коэффициенты этих сигналов. Данные графики позволяют визуально оценить степень релевантности рассматриваемого признака.

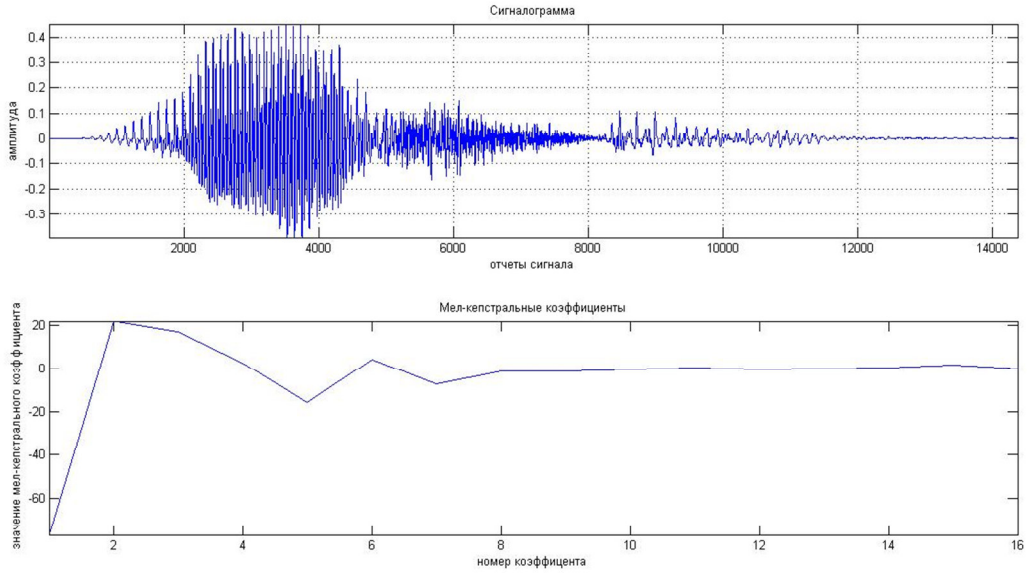


Рисунок 2.22 – Сигналограмма отрезка речевого сигнала, соответствующего произнесенному слову “восемь” и вектор признаков – Мел-кепстральные коэффициенты

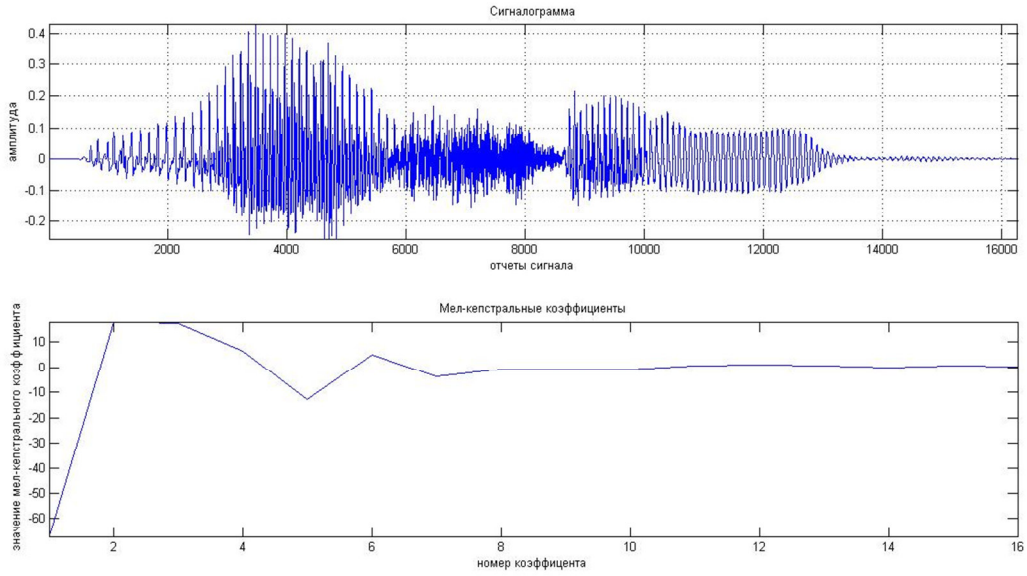


Рисунок 2.23 – Сигналограмма отрезка речевого сигнала, соответствующего слову “восемь” и вектор признаков – Мел-кепстральные коэффициенты

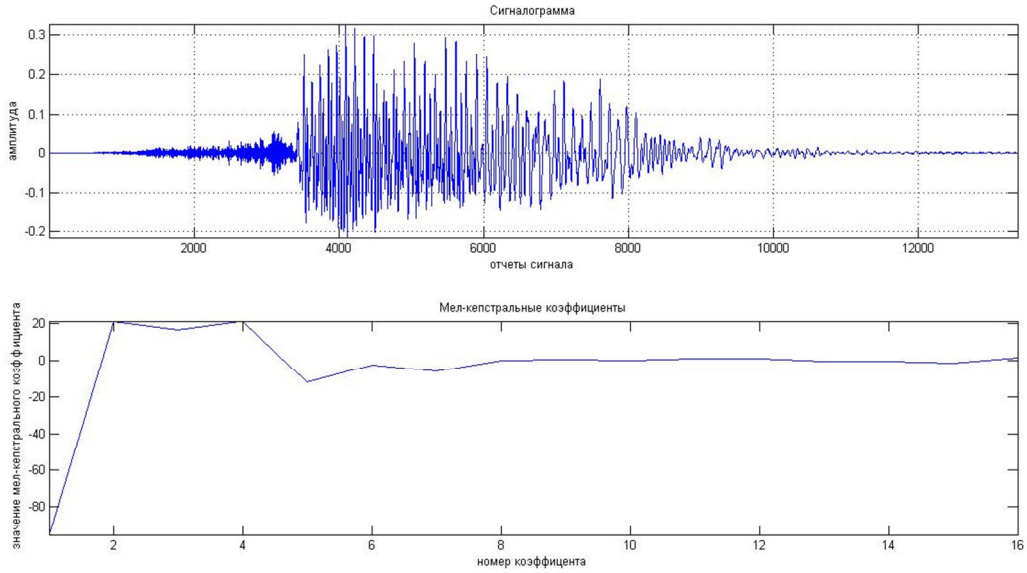


Рисунок 2.24 – Сигналограмма отрезка речевого сигнала, соответствующего слову “семь” и вектор признаков – Мел-кепстральные коэффициенты

2.4 Меры близости в задачах распознавания речевых сигналов

При определении гипотезы принадлежности сравниваемого объекта к тому или иному классу используются различные меры близости: Евклидово расстояние, расстояние Махаланобиса, корреляция, динамическая трансформация временной шкалы, среднеквадратическое отклонение и другие.

2.4.1 Евклидово расстояние

Евклидово расстояние является одной из наиболее популярных мер близости, которая позволяет вычислять расстояние ε между двумя i-ми точками последовательностей x и y в евклидовом пространстве [32], т.е. является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

$$\varepsilon = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{2.31}$$

где x_i - *i-ый* элемент вектора $\vec{x} = (x_1, x_2 ... x_{(i-1)}, x_i),\ i = 1...N$; y_i *i-ый* элемент вектора $\vec{y} = (y_1, y_2 ... y_{(i-1)}, y_i),\ i = 1...N$.

Евклидова мера расстояния может оказаться бессмысленной, если признаки измерены в разных единицах. Чтобы исправить положение, прибегают к нормированию каждого признака. Применение евклидова расстояния оправдано в следующих случаях:

1) свойства (признаки) объекта однородны по физическому смыслу и одинаково важны для классификации;

2) признаковое пространство совпадает с геометрическим пространством. [32]

2.4.2 Среднеквадратическое отклонение

Среднеквадратическое отклонение является показателем рассеивания значений одной случайной величины от другой и вычисляется следующим образом:

$$\delta = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{\sum_{i=1}^{N} x_i^2}} \tag{2.32}$$

где x_i - *i-ый* элемент вектора $\vec{x} = (x_1, x_2 ... x_{(i-1)}, x_i),\ i = 1...N$; y_i *i-ый* элемент вектора $\vec{y} = (y_1, y_2 ... y_{(i-1)}, y_i),\ i = 1...N$.

2.4.3 Расстояние Махаланобиса

Расстояние Махаланобиса [51] есть мера расстояния между векторами случайных величин, обобщающая понятие евклидова расстояния. С помощью расстояния Махаланобиса можно определять сходство неизвестной и известной выборки. Оно отличается от расстояния Евклида тем, что

учитывает корреляции между переменными и инвариантно к масштабу. Расчет расстояние Махаланобиса можно выполнить следующим образом:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})} \tag{2.33}$$

где $\vec{x},\vec{y}$ - исследуемые векторы, S- матрица ковариации.

$$S = \left(\sigma_{ij}\right), i, j = 1 \ldots n \tag{2.34}$$

$$\sigma_{ij} = \operatorname{cov}(x_i, x_j) = M(x_i - Mx_i)(x_j - Mx_j) \tag{2.35}$$

где M – математическое ожидание случайной величины.

2.4.4 Корреляция последовательностей

Корреляционный анализ занимается степенью связи между двумя переменными, *x* и *y*. Сначала предполагаем, что как *x*, так и *y* количественные переменные. Предположим, пара величин (*x*, *y*) измерена у каждого из n объектов в выборке. Мы можем отметить точку, соответствующую паре величин каждого объекта, на двумерном графике рассеяния точек. Обычно на графике переменную *x* располагают на горизонтальной оси, а *y* — на вертикальной. Размещая точки для всех n объектов, получают график рассеяния точек, который говорит о соотношении между этими двумя переменными. [34]

Коэффициент корреляции двух последовательностей может быть рассчитан следующим образом: даны две последовательности $\vec{x}$ и $\vec{y}$ длительностью N: $\vec{x} = (x_1, x_2, ..., x_N); \vec{y} = (y_1, y_2, ..., y_N).$

Необходимо произвести расчет средних значений каждой последовательности:

$$\overline{X} = \sum_{i=1}^{n} x_i; \overline{Y} = \sum_{i=1}^{n} y_i \tag{2.36}$$

Коэффициент корреляции можно рассчитать следующим образом:

$$\eta = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\cdot\left(y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}} \quad (2.37)$$

Коэффициент корреляции η изменяется в пределах от -1 до 1. В данном случае это линейный коэффициент корреляции, он показывает линейную взаимосвязь между $\vec{x}$ и $\vec{y}$: η равен 1, если связь линейна (или -1, если связи не линейна) [34].

2.4.5 Динамическая трансформация временной шкалы

Алгоритм динамической трансформации временной шкалы (DTW-алгоритм, от англ. dynamic time warping) — алгоритм, позволяющий найти минимальное расстояние между двумя последовательностями разной длины. Впервые применен в распознавании речи, где использован для определения того, как два речевых сигнала представляют одну и ту же исходную произнесённую фразу. Впоследствии были найдены применения и в других областях [45,48].

Измерения какой-либо физической величины во времени широко распространенный тип данных, встречающийся, фактически, в любой научной области, и сравнение двух последовательностей является стандартной задачей. Для вычисления отклонения бывает достаточно простого измерения расстояния между компонентами двух последовательностей (Евклидово расстояние). Однако часто две последовательности имеют различную длительность. Чтобы определить подобие между такими последовательностями, мы должны «деформировать» ось времени одной (или обеих) последовательности, чтобы достигнуть лучшего выравнивания. [45,48]

Рассмотрим алгоритм расчета данной меры близости.

Пусть даны две последовательности $\vec{x}$ и $\vec{y}$ длинной N и M соответственно: $\vec{x} = (x_1, x_2, ..., x_N); \vec{y} = (y_1, y_2, ..., y_M)$.

Далее необходимо произвести расчет пути наименьшей сходимости. Определим матрицу расстояний $\Omega^{N \times M}$ так, чтобы ее элемент $\omega_{ij}^{N \times M}$ соответствовал расстоянию между *i-ым* и *j-ым* элементами последовательностей $\vec{x}$ и $\vec{y}$, то есть соответствовал выравниванию между x_i и y_j. В качестве меры близости используется Евклидово расстояние:

$$\varepsilon_{i,j} = \sqrt{(x_i - y_j)^2} \tag{2.38}$$

Далее используя значения матрицы расстояний $\Omega^{N \times M}$ сформируем матрицу деформации $D^{N \times M}$ по значениям которой будет рассчитан путь трансформации последовательностей [94]. Причем элементы матрицы деформации определяются следующим образом:

$$d_{i,j} = \begin{cases} \varepsilon_{i,j} + \min(d_{i,j-1}, d_{i-1,j}, d_{i-1,j-1}), \text{ если } i > 1,\ j > 1 \\ \varepsilon_{i,j}, \text{ если } i = 1,\ j = 1 \end{cases} \tag{2.39}$$

Путь трансформации W - это набор смежных элементов матрицы $D^{N \times M}$, который устанавливает соответствие между $\vec{x}$ и $\vec{y}$, минимизируя расстояние между ними.

$$W = \{w_1, ... w_K\}, \tag{2.40}$$

где K – длина пути.

Определим *k-ый* элемент пути трансформации W как элемент матрицы $w_k = d_{i,j}$. Причем путь трансформации должен удовлетворять ряду условий:

Граничное условие: начало пути W – первый элемент матрицы деформации $w_1 = d_{1,1}$, а конец пути – последний, $w_K = d_{N,M}$. Это ограничение

гарантирует, что путь трансформации содержит все точки обоих временных рядов.

Условие непрерывности: любые два смежных элемента пути W, $w_k = \{w_i, w_j\}$ и $w_{k+1} = \{w_i, w_j\}$, удовлетворяют следующим неравенствам $w_i - w_{i+1} \le 1,\ w_j - w_{j+1} \le 1$. Данное условие обеспечивает ограничение на один шаг при выборе следующего элемента пути.

Условие монотонности: любые два смежных элемента пути $w_k = \{w_i, w_j\}$ и $w_{k-1} = \{w_{i-1}, w_{j-1}\}$, удовлетворяют следующим неравенствам $w_i - w_{i-1} \le 0,\ w_j - w_{j-1} \le 0$. Это ограничение гарантирует, что путь трансформации не будет возвращаться назад к пройденной точке. То есть оба индекса либо остаются неизменными, либо увеличиваются (но никогда не уменьшаются).

Существует большое количество путей трансформации, удовлетворяющих всем вышеуказанным условиям, однако нас интересуют только тот путь, который минимизирует DTW расстояние [94]. Искомый путь должен удовлетворять условию минимальной стоимости пути:

$$DTW(\vec{x}, \vec{y}) = \min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} w_k}\right\} \tag{2.41}$$

Следует отметить, что последнее рассчитанное значение $d_{i,j}$ матрицы деформации является искомым расстоянием (мерой близости) между последовательностями $\vec{x}$ и $\vec{y}$.

Приведенные меры близости используются для определения расстояния между признаками распознаваемых объектов, в дальнейшем полученные значения расстояния используются для принятия решения о принадлежности сравниваемого объекта тому или иному классу.

ГЛАВА 3 ИССЛЕДОВАНИЕ ПРИГОДНОСТИ ПРЕДСТАВЛЕНИЙ РЕЧЕВЫХ СИГНАЛОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ

3.1 Методика оценки методов распознавания речевых сигналов

Оценка методов распознавания речевых сигналов будет проводится с помощью следующей методики.

Задано множество объектов X – отрезков речевых сигналов, множество допустимых ответов Y, и существует целевая функция y^*: X→Y, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1,...,x_\ell\} \subset X$. Пары «объект–ответ» (x_i, y_i) называются прецедентами. Совокупность пар $X^\ell = (x_i, y_i)_{i=1}^\ell$ называется обучающей выборкой. Задача идентификации по прецедентам заключается в том, чтобы по выборке X^ℓ восстановить зависимость y^*, то есть построить решающую функцию α: X→Y, которая приближала бы целевую функцию $y^*(x)$, причём не только на объектах обучающей выборки, но и на всём множестве X. Решающая функция α должна допускать эффективную компьютерную реализацию; по этой причине будем называть её алгоритмом. [7]

В качестве обучающей выборки будем использовать базу речевых данных, состоящую из речевых сигналов записанных при произнесении числительных: “ноль”, “один”, “два”, “три”, “четыре”, “пять”, “шесть”, “семь”, “восемь”, “девять”. Каждый отрезок РС является объектом $x_i \in X^\ell$ и представляет собой запись отдельного слова-ответа $y_i \in Y^\ell$. Таким образом, прецедентом (x_i, y_i) будем считать слово и его акустическую реализацию.

Признак f объекта x — это результат измерения некоторой характеристики объекта. Формально признаком называется отображение f:X→D_f, где D_f — множество допустимых значений признака. Пусть имеется

набор признаков $f_1,...,f_n$. Вектор $f_1(x),...,f_n(x)$ называют признаковым описанием объекта $x \in X$.

Для каждого объекта в обучающей выборке $x_i \in X^\ell$ производилось вычисление определенного вектора признаков $f_n(x)$ с разными параметрами (длина окна анализа, количество частотных интервалов). При этом использовалась следующая выборка признаков с установленными параметрами – таблица 3.1.

Таблица 3.1 – Таблица выборки признаков

Признак	**Параметры расчета**	f_n
Декомпозиция речевого сигнала	N=128, R= 8 N=256, R=16 N=512, R= 32	f_1
Распределение мгновенных энергий речевого сигнала	N=128, R= 8 N=256, R=16 N=512, R= 32	f_2
Распределение долей энергии речевого сигнала	N=128, R= 8 N=256, R=16 N=512, R= 32	f_3
Распределение информационных интервалов речевого сигнала	N=128, R= 8 N=256, R=16 N=512, R= 32	f_4
Мел-кепстральные коэффициенты речевого сигнала	N=128, R= 8 N=256, R=16 N=512, R= 32	f_5
Частота переходов через нуль речевого сигнала	N=128 N=256 N=512	f_6
Ширина частотной области, занимаемая речевым сигналом	N=128 N=256 N=512	f_7

В дальнейшем мы не будем различать объекты из X и их признаковые описания, полагая X = $D_{f1} \times ... \times D_{fn}$. При этом значения m параметров содержатся в множестве $P=\{p_{1,...}p_m\}$.

Совокупность признаковых описаний всех объектов выборки X_ℓ, записанную в виде таблицы размера $\ell \times m$, называют матрицей объектов – признаков:

$$F = \left\| f_n(x_i) \right\|_{\ell \text{ x m}} = \begin{pmatrix} f_1(x_1) & ... & f_m(x_1) \\ ... & ... & ... \\ f_1(x_\ell) & ... & f_m(x_\ell) \end{pmatrix} \qquad (3.1)$$

Для каждого рассчитанного признака запишем матрицу объектов – признаков.

$$F_n = \begin{pmatrix} f_1(x_1) & ... & f_m(x_1) \\ ... & ... & ... \\ f_1(x_\ell) & ... & f_m(x_\ell) \end{pmatrix}$$

Каждый столбец матрицы содержит значения признака, рассчитанные с определенными параметрами, например, 3-й столбец матрицы F_1 содержит значения признака: f_1; рассчитанные с параметрами: N=256; R=16.

Природа множества допустимых ответов Y определяет тип задачи. Если $Y = \{1,...,M\}$, то это задача классификации на M непересекающихся классов. В этом случае всё множество объектов X разбивается на классы $K_y=\{x \in X: y^*(x) = y\}$, и алгоритм $\alpha(x)$ должен давать ответ на вопрос «какому классу принадлежит x?». [7]

В случае решения задачи идентификации количество классов ограничивается двумя: $Y = \{0,1\}$ и алгоритм $\alpha(x)$ должен давать ответ на вопрос «тождественны ли сравниваемые объекты x_q и x_s?». При этом объекты x_q и $x_s \in X$. Следовательно, мы можем определить следующую гипотезу:

H_0: *сравниваемые объекты* x_q *и* x_s *тождественно равны*.

Альтернативная гипотеза определяется следующим образом:

H_1: *сравниваемые объекты* x_q *и* x_s *не равны*.

Как утверждалось ранее, в задачах идентификации по прецедентам элементы множества X — это не реальные объекты, а лишь доступные данные о них. Данные могут быть неточными, поскольку измерения значений признаков $f_n(x)$ и целевой зависимости $y^*(x)$ обычно выполняются с погрешностями. Данные могут быть неполными, поскольку измеряются не все мыслимые признаки, а лишь физически доступные для измерения. В результате одному и тому же описанию x могут соответствовать различные объекты и различные ответы. В таком случае $y^*(x)$, строго говоря, не является функцией. Устранить эту некорректность позволяет вероятностная постановка задачи.

Вместо существования неизвестной целевой зависимости $y^*(x)$ предположим существование неизвестного вероятностного распределения на множестве $X \times Y$ с плотностью $p(x,y)$, из которого случайно и независимо выбираются ℓ наблюдений $X^\ell = (x_i, y_i)_{i=1}^\ell$. Такие выборки называются простыми или случайными одинаково распределёнными. Вероятностная постановка задачи считается более общей, так как функциональную зависимость $y^*(x)$ можно представить в виде вероятностного распределения $p(x, y) = p(x)p(y|x)$ [7], положив $p(y|x) = \delta(y - y^*(x))$, где $\delta(z)$ — дельта-функция.

Для оценки эффективности исследуемых мер близости, предлагается использовать функционал качества, который определяется через функцию потерь.

Функция потерь − это неотрицательная функция $L(a,x)$, характеризующая величину ошибки алгоритма α на объекте x. Если $L(a,x)=0$, то ответ $a(x)$ называется корректным. [7]

Функционал качества алгоритма a на выборке X^ℓ (3.2):

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(a, x_i) \qquad (3.2)$$

Функционал Q называют также функционалом средних потерь или эмпирическим риском [7], так как он вычисляется по эмпирическим данным $(x_i, y_i)_{i=1}^{\ell}$.

В данном исследовании, целесообразно применить бинарную функцию потерь, принимающую только значения 0 и 1. В этом случае $L(a, x)$=1 означает, что алгоритм α допускает ошибку на объекте x, а функционал Q называется частотой ошибок алгоритма α на выборке X^{ℓ}.

В данном исследовании будем использовать следующий идентификатор функции потерь:

$$L(a, x) = [a(x) \neq y^*(x)] = 1 \qquad (3.3)$$

В соответствии с принятой нулевой гипотезой можно разделить ошибки, возникающие при работе алгоритма α на два типа – таблица 3.2.

Таблица 3.2 – Таблица типов ошибок

		Нулевая гипотеза Сравниваемые РС тождественны	
		Верна	*Не верна*
Результат применения критерия	*Принята*	идентичность подтверждена	идентичность неверно подтверждена
	Отвергнута	идентичность неверно отвергнута	идентичность отвергнута

Следуя информации изложенной в таблице можно описать ошибки следующим образом:

I) Ошибками первого рода будем считать неверно отвергнутую тождественность.

II) Ошибками второго рода будем считать неверное подтверждение гипотезы.

Для подсчета количества ошибок первого и второго рода необходимо разделить исследуемую обучающую выборку на два непересекающихся

подмножества: $X^{\ell} \subset X^{\ell_1} \cup X^{\ell_2}$. Элементы подмножеств можно записать следующим образом:

$X^{\ell_1} = \{x_1^{\ell_1}, ..., x_\mu^{\ell_1}\}$ - объекты, на которых возникают ошибки 1 рода.

$X^{\ell_2} = \{x_1^{\ell_2}, ..., x_\eta^{\ell_2}\}$ - объекты, на которых возникают ошибки 2 рода/

Применение функции потерь (3.3) на каждое подмножество даст количество ошибок, которые дает алгоритм:

$$L(a(h), x^{\ell_1}) = \sum_{i=1}^{\mu} L_i[a(x_i^{\ell_1}) \neq y^*(x_i^{\ell_1})],$$

$$L(a(h), x^{\ell_2}) = \sum_{i=1}^{\eta} L_i[a(x_i^{\ell_2}) \neq y^*(x_i^{\ell_2})],$$

где *h* значение порога решающей функции.

Данные параметры необходимо использовать при поиске оптимального, с точки зрения уровня значимости, порога *h*, для решающего правила гипотезы, предварительно задав предел относительного количества ошибок на одном из подмножеств. В качестве алгоритмов α будем использовать меры близости, описанные в первой главе: 1) Евклидово расстояние; 2) Среднеквадратическое отклонение; 3) Расстояние Махаланобиса; 4) Корреляция; 5) Динамическая трансформация временной шкалы. Таким образом, мы можем определить наиболее адекватные применительно к решаемой задаче признаки объектов. Вариации сравниваемых признаков и мер близости представлены в таблице 3.3.

Таблица 3.3 – Вариации параметров эксперимента

Признак, F_n	**Мера близости, α**
Декомпозиция речевого сигнала	Евклидово расстояние, Среднеквадратическое отклонение, Расстояние Махалонобиса, Корреляция, Динамическая трансформация временной шкалы
Распределение мгновенных энергий речевого сигнала	
Распределение долей энергии речевого сигнала	
Распределение информационных	

Признак, F_n	**Мера близости, α**
интервалов речевого сигнала	
Мел-кепстральные коэффициенты	
Частота переходов через нуль	Расстояние Махалонобиса, Динамическая трансформация временной шкалы
Ширина частотной области, занимаемая сигналом	

3.2 Исследование подходов к распознаванию речевых сигналов

Для исследования подходов к распознаванию речевых сигналов запишем речевые сигналы при произнесении следующих числительных диктором мужчиной: “ноль”, “один”, “два”, “три”, “четыре”, “пять”, “шесть”, “семь”, “восемь”, “девять”. В количестве по 10 штук каждого числительного. Параметры записи: частота дискретизации 16 кГц, глубина дискретизации 16 бит, моно канал.

Требуется произвести эксперименты, с целью оценки вероятности ошибок I и II рода при использовании различных признаков и мер близости для их распознавания.

Результаты расчета мер близости для имеющейся выборки признаков можно представить в виде следующих двумерных графиков (рисунок 3.1 – 3.5 графики для мел-кепстральных коэффициентов с параметрами *R=16, N=256*). Сравнение объектов обучающего подмножества происходит по принципу “каждый с каждым”, следовательно, элементы, располагающиеся на главной диагонали, представляют собой результат сравнения одного и того же объекта x_s *(самого с собой)*. С помощью данных графиков возможно визуально оценить применения того или иного признакового описания объектов X.

После расчёта мер близости можно приступать к определению порога решающей функции, и оценки вероятностей ошибок I и II рода.

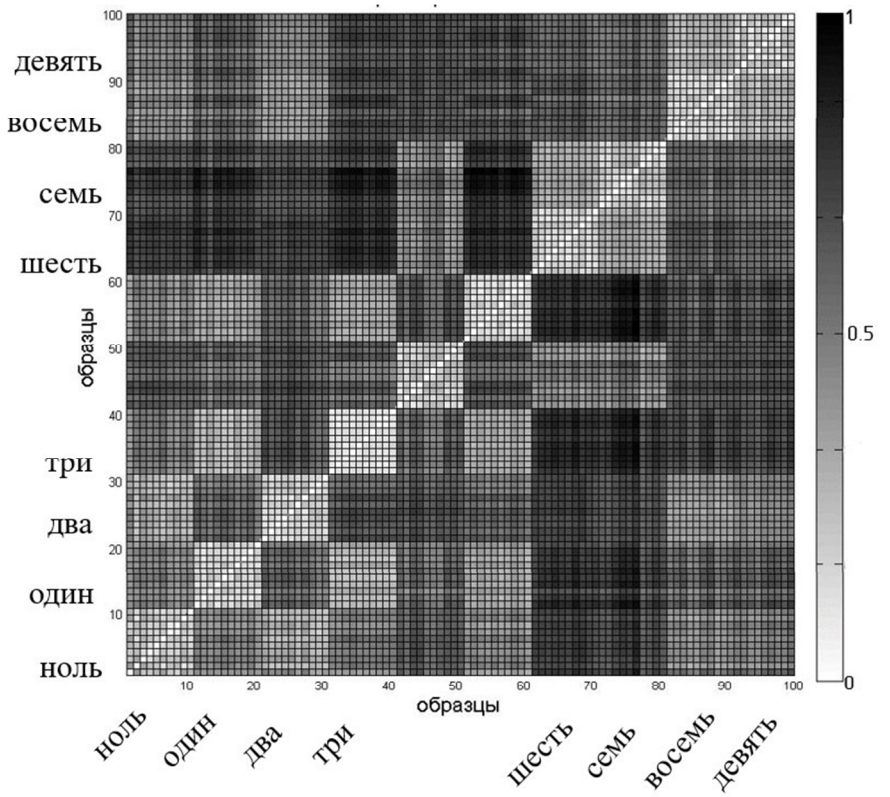


Рисунок 3.1 – Результат расчета динамической трансформации временной шкалы для Мел-кепстральных коэффициентов речевых сигналов

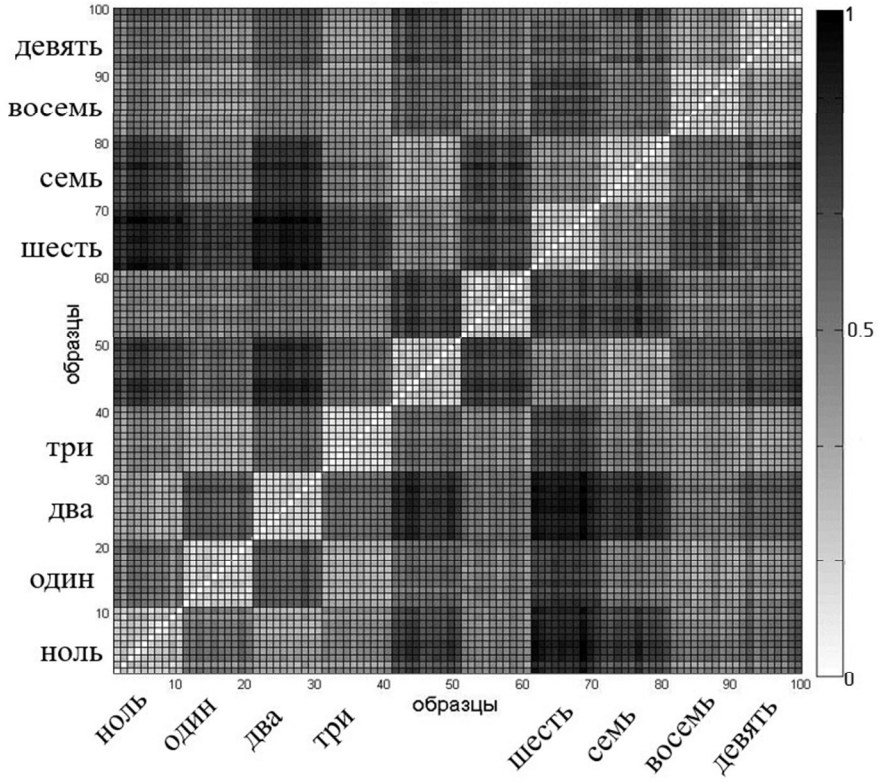


Рисунок 3.2– Результат расчета среднеквадратического отклонения для Мел-кепстральных коэффициентов речевых сигналов

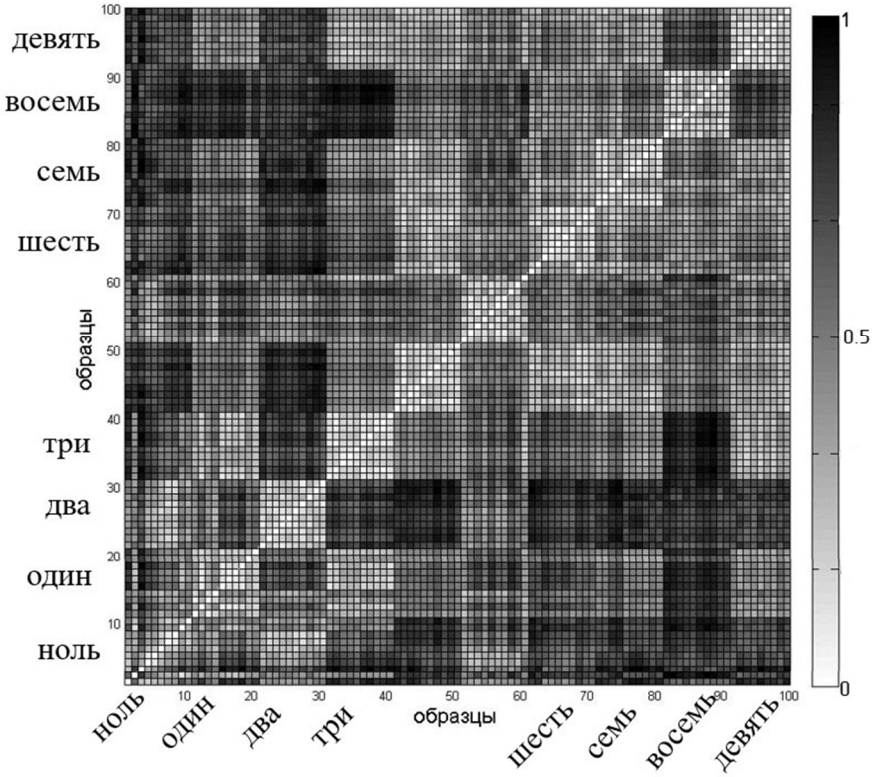


Рисунок 3.3– Результат расчета расстояния Махаланобиса для Мел-кепстральных коэффициентов речевых сигналов

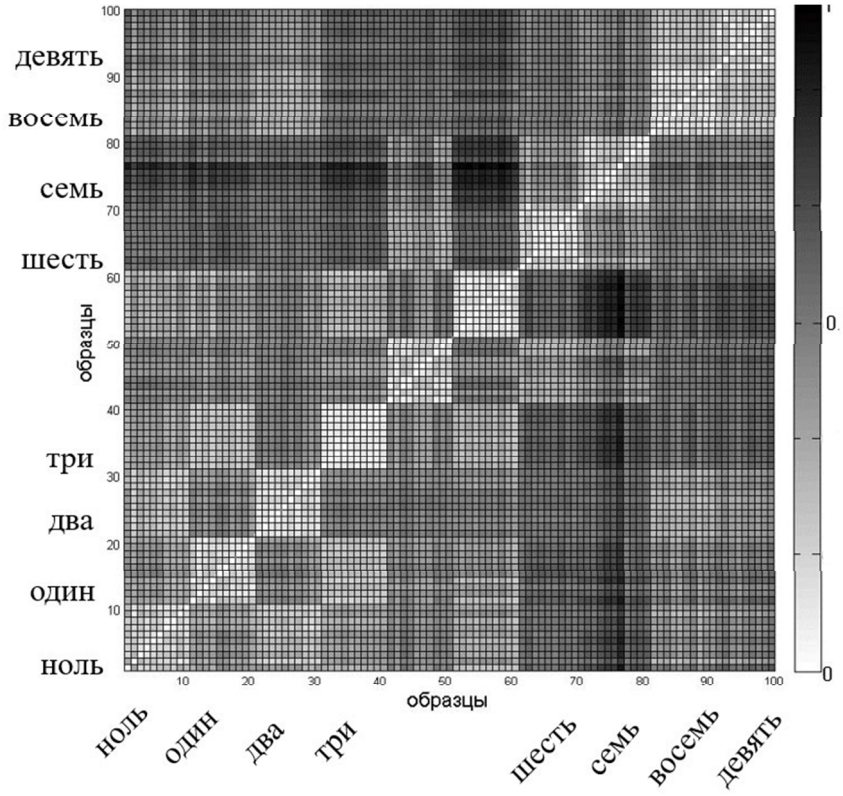


Рисунок 3.4 – Результат расчета Евклидова расстояния для Мел-кепстральных коэффициентов речевых сигналов

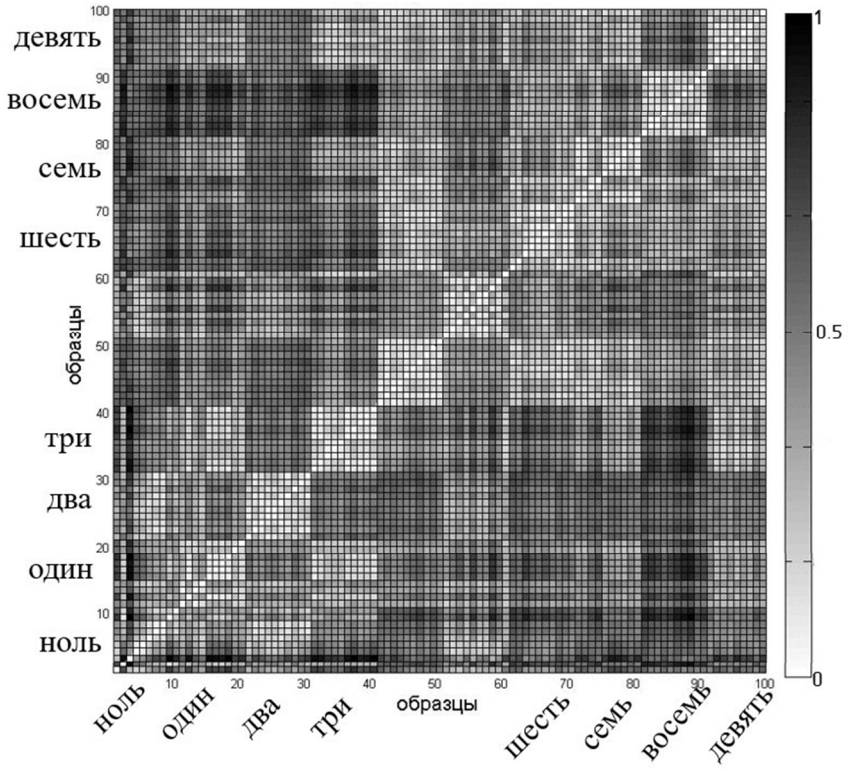


Рисунок 3.5 – Результат расчета корреляции для Мел-кепстральных коэффициентов речевых сигналов

Далее требуется произвести подбор значений порога для решающего критерия гипотезы H_0.

Подбор порога для гипотезы H_0 будем осуществлять следующим образом (3.4):

$$h = \min(L(a(h), x^{\ell_1}) + L(a(h), x^{\ell_2}) + \begin{cases} 99999, \mu / L(a(h), x^{\ell_1}) > \Delta \\ 0, \ \mu / L(a(h), x^{\ell_1}) \le \Delta \end{cases} \quad (3.4)$$

где Δ - заданная допустимая вероятность ошибок I рода. Перебор пороговых значений осуществляется в следующих пределах: $h = \{0, 0.01, 0.02, ..., 1\}$.

При подборе порога (3.4) осуществляется ограничение вероятности появления ошибок I рода значением в Δ=0.05. Модель установления порога с учетом оценки вероятностей ошибок I и II рода проиллюстрировано на рисунке 3.6. Порог решающей функции выставляется исходя из учета допустимого значения вероятности ошибок I рода.

При невозможности установить порог при заданной вероятности ошибок I рода Δ=0.05, пересчитаем порог для Δ=0.1. Если же не удается

установить порог при значении Δ=0.1, пересчитаем порог для Δ=0.25. Таким образом допустимые предельные значения вероятности ошибок I рода: Δ={0.05,0.1,0.25}.

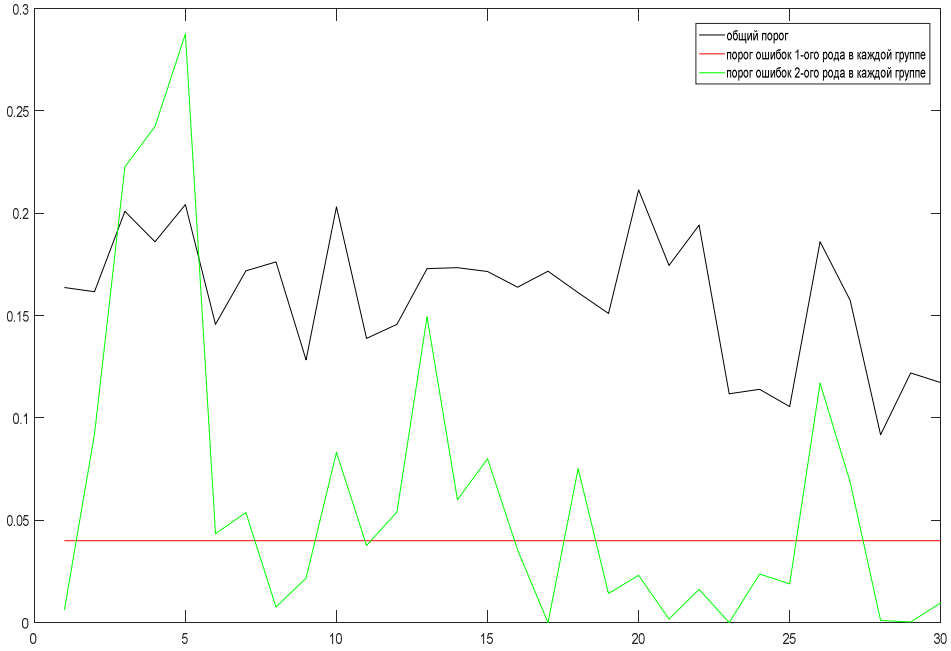


Рисунок 3.6 – Модель установления порога с учетом оценки вероятности ошибок I и II рода

Согласно описанному выше алгоритму, производился расчет функционала качества представленных алгоритмов (мер близости) для различных признаковых пространств. Основные результаты данного эксперимента представлены в таблицах 3.4 – 3.6. В таблицах представлены результаты оценки ошибок I и II рода для различных мер близости (Евклидово расстояние, среднеквдратическое отклонение, расстояние Махаланобиса, корреляция, динамическое сжатие временной последовательности) при их использовании с признаками: декомпозиция РС банком фильтров; распределение мгновенных энергий РС, распределение долей энергии РС, распределение информационные интервалов РС, мел-кепстральные коэффициенты РС, частота переходов через ноль, ширина частотной области, занимаемая сигналом.

Таблица 3.4 – Результаты оценки ошибок I и II рода, R=8, N=128

Вид признака	Тип ошибок	Меры близости, оценка вероятности ошибок				
		Евклидово расстояние	Среднеквадр. отклонение	Расстояние Махаланобиса	Корреля ция	DTW
Декомпоз. речевого сигнала	Ошибки I рода	0,25	0,24	0,25	0,25	0,23
	Ошибки II рода	0,48	0,40	0,46	0,42	0,39
Распр. мгновен. энергий РС	Ошибки I рода	0,24	0,22	0,24	0,24	0,19
	Ошибки II рода	0,46	0,41	0,44	0,47	0,38
Распр. долей энергии РС	Ошибки I рода	0,09	0,10	0,24	0,24	0,08
	Ошибки II рода	0,40	0,41	0,39	0,41	0,36
Распр. информаци онных интерв. РС	Ошибки I рода	0,05	0,05	0,10	0,10	0,04
	Ошибки II рода	0,27	0,28	0,34	0,39	0,22
Мел-кепстр. коэфф. РС	Ошибки I рода	0,05	0,04	0,10	0,10	0,04
	Ошибки II рода	0,25	0,19	0,39	0,34	0,16
Частота переходов через нуль	Ошибки I рода	0,25	0,24	0,25	0,25	0,21
	Ошибки II рода	0,47	0,44	0,52	0,49	0,45
Ширина частотной области, занимаем. сигналом	Ошибки I рода	0,24	0,24	0,25	0,25	0,24
	Ошибки II рода	0,42	0,41	0,48	0,46	0,36

Таблица 3.5 – Результаты оценки ошибок I и II рода, R=16, N=256

Вид признака	Тип ошибок	Меры близости, оценка вероятности ошибок				
		Евклидово расстояние	Среднеквадр. отклонение	Расстояние Махаланобиса	Корреля ция	DTW
Декомпоз. речевого сигнала	Ошибки I рода	0,24	0,24	0,25	0,25	0,18
	Ошибки II рода	0,46	0,39	0,42	0,37	0,36
Распр. мгновен. энергий РС	Ошибки I рода	0,22	0,22	0,22	0,24	0,17
	Ошибки II рода	0,41	0,44	0,45	0,46	0,32

Окончание таблицы 3.5

Вид признака	Тип ошибок	Меры близости, оценка вероятности ошибок				
		Евклидово расстояние	Среднеквадр. отклонение	Расстояние Махаланобиса	Корреляция	DTW
Распр. долей энергии РС	Ошибки I рода	0,08	0,07	0,21	0,22	0,08
	Ошибки II рода	0,29	0,30	0,29	0,31	0,26
Распр. информационных интерв. РС	Ошибки I рода	0,05	0,04	0,09	0,10	0,05
	Ошибки II рода	0,12	0,10	0,27	0,30	0,09
Мел-кепстр. коэфф. РС	Ошибки I рода	0,05	0,04	0,08	0,08	0,04
	Ошибки II рода	0,08	0,07	0,28	0,26	0,07
Частота переходов через нуль	Ошибки I рода	0,22	0,23	0,24	0,25	0,20
	Ошибки II рода	0,40	0,39	0,48	0,46	0,38
Ширина частотной области, занимаем. сигналом	Ошибки I рода	0,21	0,22	0,23	0,24	0,19
	Ошибки II рода	0,39	0,36	0,41	0,42	0,37

Таблица 3.6 – Результаты оценки ошибок I и II рода, R=32, N=512

Вид признака	Тип ошибок	Меры близости, оценка вероятности ошибок				
		Евклидово расстояние	Среднеквадр. отклонение	Расстояние Махаланобиса	Корреляция	DTW
Декомпоз. речевого сигнала	Ошибки I рода	0,25	0,24	0,24	0,24	0,24
	Ошибки II рода	0,49	0,45	0,42	0,47	0,44
Распр. мгновен. энергий РС	Ошибки I рода	0,25	0,23	0,22	0,23	0,21
	Ошибки II рода	0,47	0,40	0,44	0,45	0,39
Распр. долей энергии РС	Ошибки I рода	0,08	0,08	0,22	0,23	0,09
	Ошибки II рода	0,35	0,34	0,38	0,37	0,32
Распр. информационных интерв. РС	Ошибки I рода	0,05	0,05	0,09	0,10	0,04
	Ошибки II рода	0,22	0,18	0,31	0,35	0,15

Окончание таблицы 3.6

Вид признака	Тип ошибок	Меры близости, оценка вероятности ошибок				
		Евклидово расстояние	Среднеквадр. отклонение	Расстояние Махаланобиса	Корреляция	DTW
Мел-кепстр. коэфф. РС	Ошибки I рода	0,05	0,04	0,10	0,10	0,04
	Ошибки II рода	0,13	0,10	0,27	0,29	0,09
Частота переходов через нуль	Ошибки I рода	0,23	0,21	0,22	0,24	0,20
	Ошибки II рода	0,42	0,41	0,47	0,42	0,37
Ширина частотной области, занимаем. сигналом	Ошибки I рода	0,23	0,24	0,25	0,25	0,21
	Ошибки II рода	0,39	0,35	0,41	0,40	0,32

Проанализировав полученные в ходе экспериментов результатов, можно сделать следующие выводы:

I) Наиболее пригодными, для задач распознавания векторами признаков можно считать:

1. Мел-кепстральные коэффициенты;
2. Распределение информационных интервалов речевого сигнала;

II) При оценке данных векторов признаков наиболее пригодными, для решаемой задачи, мерами близости можно считать:

1. Динамическая трансформация временной шкалы;
2. Среднеквадратическое отклонение;
3. Евклидово расстояние.

Наименьшая вероятность ошибок I и II рода достигается при применении в качестве признаков Мел-кепстральные коэффициенты. В качестве меры близости лучший результат показывает DTW (динамическая трансформация временной шкалы).

Можно также утверждать, что в качестве окна анализа для получения признака целесообразно выбирать окно длительностью 256 отсчётов.

Для получения лучших результатов можно использовать сравнение объекта с некоторым усредненным значением признака по классу, т.к. расстояние между усредненным значением признака и отдельными векторами признаками будет меньше, чем в случае сравнения строго говоря уникальных объектов.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы было выявлено, что наиболее подходящими для задач распознавания признаками являются такие признаки, которые отражают свойства концентрации энергии и учитывают особенности восприятия слуха человека для получения адекватного природе речевого воздействия частотного разбиения. Данными признаками являются: мел-кепстральные коэффициенты и распределение информационных интервалов речевого сигнала.

Проведен анализ особенностей обработки речевых сигналов в задачах распознавания речи: приведены концептуальные схемы распознавания, даны сведения о восприятии и воспроизведении звука человеком. Изучены существующие методы представления речевых сигналов в задачах распознавания: частотные (например, частотное распределение) и временные (например, частота перехода сигнала через ноль) представления, проведен их сравнительный анализ. Для получения некоторых признаков (декомпозиция сигнала по банку фильтров, распределение мгновенных энергий, распределение долей энергии, распределение информационных интервалов) использовался субполосный подход, который позволяет точно выделять распределение долей энергий по частотным интервалам с минимальным просачиванием.

Изучены меры близости, применяемые для сравнения признаков: евклидово расстояние, среднеквадратическое отклонение, расстояние Махаланобиса, корреляция, динамическая трансформация временной шкалы. Наиболее подходящими для задач распознавания речевых сигналов являются:

динамическая трансформация временной шкалы; среднеквадратическое отклонение; евклидово расстояние.

Проведен сравнительный анализ применения различных мер близости и векторов признаков который показал, что наиболее пригодными, для задач

распознавания, векторами признаков можно считать: Мел-кепстральные коэффициенты; распределение информационных интервалов речевого сигнала, а для их сравнения использовать следующие меры близости: динамическую трансформация временной шкалы; среднеквадратическое отклонение; евклидово расстояние.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Аграновский, А.В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов [Текст]/ А.В. Аграновский, Д.А. Леднов – М.: Радио и связь, 2004. – 164 с.

2. Алдошина, И.А. Слуховые модели восприятия линейных и нелинейных искажений в музыке и речи. Часть1 [Текст] / И.А. Алдошина // Информационно-технический журнал «Звукорежиссер». - 2006. - №3. – С.38-44.

3. Ананьев, Б. Г. Теория ощущений. — Л., 1961. — С. 579. — 928 с.

4. Вапник, В. Н. Теория распознавания образов [Текст] / Вапник В. Н., Червоненкис А. Я. / М.: Наука, 1974

5. Винцюк, Т.К., Анализ, распознавание и интерпретация речевых сигналов [Текст]/ Винцюк Т.К. - Киев: Наук.думка, 1987. - 264с.

6. Воеводин, В.В. Матрицы и вычисления [Текст] / В.В. Воеводин, Ю.А. Кузнецов. – М.: Наука, 1984. – 318 с.

7. Воронцов К.В. Математические методы обучения по прецедентам [Текст]/ материалы лекций МФТИ – М., 2008

8. Герасимов, А.В. Применение метода модифицированного линейного предсказания к задачам выделения акустических признаков речевых сигналов [Текст] / А.В.Герасимов, О.А. Морозов, В.Р. Фидельман // Радиотехника и Электроника. – 2005. – том 50. №10. – С. 1287-1292.

9. Гребнов, С.В. Аналитический обзор методов распознавания речи в системах голосового управления [Текст]/ С.В. Гребнов // Вестник ИГЭУ. – 2009. – Вып.3. – С.83-85.

10. Гривен, В.Г. Введение в Вейвлет преобразование / АВТЭКС, Санкт-Петербург, 2009, С. 302

11. Губочкин, И.В. Разработка алгоритмов анализа и распознавания речи на основе адаптивной кластерной модели и критерия минимального информационного рассогласования [Текст]: автореф. дис. канд. техн наук / И.В. Губочкин – Нижний Новгород: НГЛУ, 2011. – 22с.

12. Гудонавичюс, Р.В. Распознавание речевых сигналов по их структурным свойствам [Текст]/Р.В. Гудонавичюс, П.П. Кемешис, А.Б. Читавичюс – Л.: «Энергия», 1977. – 64 с.

13. Деркач, М.Ф. Динамические спектры речевых сигналов [Текст]/ М.Ф. Деркач, Р.Я. Гумецкий, Б.М. Гура, М.Е. Чабан – Львов: Виша школа. Изд-во при Львов. ун-те, 1983. – 168 с.

14. Жиляков, Е.Г. Вариационные методы анализа и построения функций по эмпирическим данным: моногр. [Текст] / Е.Г. Жиляков. – Белгород: Изд-во БелГУ, 2007. – 160 с.

15. Жиляков, Е.Г. Методы обработки речевых данных в информационно-телекоммуникационных системах на основе частотных представлений [Текст]/ Е.Г. Жиляков, С.П. Белов, Е.И. Прохоренко. – Белгород: Изд-во БелГУ, 2007. - 136 с.

16. Жиляков, Е.Г. Модели распределения энергии звуков русской речи на основе частотных представлений [Текст] / Е.Г. Жиляков, А.В. Болдышев, А.А. Фирсова// XXIII Международной научной конференции Математические методы в технике и технологиях – Саратов. – 2010. – С.236-239.

17. Жиляков, Е.Г. Частотный анализ речевых сигналов [Текст] / Е.Г. Жиляков, Е.И. Прохоренко // Научные ведомости Белгородского государственного университета. Сер. Информатика и прикладная математика – 2006. – №2(31), выпуск 3. – С.201-208.

18. Засыпкин, А.В. О диктороне зависимой системе голосового телефонного номеронабирателя [Текст] / А.В. Засыпкин, А.Т. Мицевич, М.В. Овецкий, В.Ю. Шелепов// Труды международной конференции “Знание-Диалог-Решение”. – Ялта. – 1995. – С.427-430.

19. Кавальчук, А.Н. (2011), "Формула для перехода из области частот к шкале барков и обратно," А.Н. Кавальчук, Ал.А. Петровский // Информатика, 2011, 4(32), стр. 71-81

20. Каганов, А.Ш. Криминалистическая экспертиза звукозаписей. – М.: "Юрлитинформ", 2005. - 272с.

21. Кипяткова И.С. Автоматическая обработка разговорной русской речи: монография / И.С. Кипяткова , А.Л. Ронжин, А.А. Карпов. СПИИРАН – СПб.: ГУАП, 2013. – 314 с.

22. Колерс, П.А. Распознавание образов. Исследование живых и автоматических распознающих систем [Текст]/ П.А. Колерс, Е.Д. Мюрей, пер. Л.И. Титомира – М.: «Мир», 1970. – 288 с.

23. Ле, Н.В. Распознавание речи на основе искусственных нейронных сетей [Текст] / Н.В. Ле, Д.П. Панченко // Технические науки в России и за рубежом: материалы междунар. заоч. науч. конф.– Москва. – 2011. – С.8-11.

24. Леонович, А.А. Современные технологии распознавания речи [Текст] /А.А. Леонович // Материалы конференции «Диалог: Компьютерная лингвистика и интеллектуальные технологии». – Звенигород. – 2005.

25. Ли, У.А. Методы автоматического распознавания речи. [Текст] В 2-х книгах. Кн.1. / Пер. с англ./Под ред. У.Ли. – М.; Мир, 1983. –328 с.

26. Ли, У.А., Методы автоматического распознавания речи. [Текст] В 2-х книгах. Кн.2. /Пер. с англ. Под ред. У.Ли. – М.; Мир, 1983. – 392 с.

27. Мазуренко, И.Л. Компьютерные системы распознавания речи [Текст] / И.Л. Мазуренко // Интеллектуальные системы. – Москва. – 1998. – т.3. вып. 1-2. – С.117-134.

28. Мазуренко, И.Л. Одна модель распознавания речи [Текст] / И.Л. Мазуренко // Компьютерные аспекты в научных исследованиях и учебном процессе. – Москва – 1996 – С.107-112.

29. Малла, С. Вэйвлеты в обработке сигналов [Текст] / М.: Мир, 2005. — 672 с.

30. Ниценко, А.В. Алгоритмы пофонемного распознавания слов наперед заданного словаря [Текст] / А.В. Ниценко, В.Ю. Шелепов // Искусственный интеллект. – 2004. – С.633-639.

31. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов: Пер. с англ./Под ред. С. Я. Шаца. — М.: Связь, 1979. 416 с., ил.

32. Рабинер, Л. Теория и применение цифровой обработки сигналов [Текст] / Л.Рабинер, Б.Гоулд – М.: Мир, 1978. – 848с.

33. Рабинер, Л.Р. Цифровая обработка речевых сигналов [Текст] / Л.Р. Рабинер, Р.Ф. Шафер – М.: Радио и связь, 1981. – 496 с.

34. Сергиенко А.Б. Цифровая обработка сигналов. 2 – изд. – СПб.: Питер, 2006. – 608 с.

35. Сорокин, В.Н. Артикуляторно-ориентированная система распознавания речи [текст] / В.Н. Сорокин, А.Н. Ижнин, А.И. Цыплихин, Д.Н. Чепелев // Труды Международного семинара «Диалог - 2003». – 2003. С.657-662.

36. Смоленцев, Н. К. Введение в теорию вейвлетов [Текст] /Ижевск: РХД, 2010. — 292 с.

37. Сорокин, В.Н. Сегментация и распознавание гласных [Текст] / В.Н. Сорокин, А.И. Цыплихин // Информационные процессы. – 2004. – Т.4, №2. – С. 202-220.

38. Сорокин, В.Н. Теория речеобразования [Текст] / В.Н. Сорокин – М.: Радио и связь, 1985. – 312 с.

39. Фланаган, Дж. Л. Анализ, синтез и восприятие речи [Текст]/ пер.с англ. А.А. Пирогова. – М.:Связь, 1968. – 397с.

40. Чистович, Л.А. Физиология речи. Восприятие речи человеком [Текст] /Л.А. Чистович, А.И. Венцов, М.П. Гранстрем и др. – М.: Наука, 1976. – 388 с.

41. Шелепов, В.Ю. К проблеме пофонемного распознавания [Текст] / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – 2005. – №4. – С.662-668.

42. Шелухин, О.И. Цифровая обработка и передача речи [Текст] / О.И. Шелухин, Н.Ф.Лукьянцев; под ред. О.И. Шелухина. – М.: Радио и связь, 2000. – 456с.

43. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004.

44. Allen, J.B., "How Do Humans Process and Recognize Speech?," IEEE Trans. On Speech and Audio Processing, 1994, 2(4), pp. 567-577.

45. Al-Naymat Ghazi, Chawla Sanjay, Taheri Javid “Sparse DTW: A novel approach to speed up Dynamic Time Warping” Proc. of the 8th Australasian Data Mining Conference (AusDM'09) p. 117-127; (2009)

46. Bishop, C. M. Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp

47. Dong Yu Automatic Speech Recognition: a deep learning approach (Signals and Communication Technology) / Springer; 2015 edition (November 11, 2014), p. 321

48. Eamonn J. Keogh, Michael J. Pazzani Derivative Dynamic Time Warping, Section 1Proceedings of the sixth ACM SIGKDD, 2010

49. Giannakopoulos T. Introduction to Audio Analysis: A Matlab Approach 1st Edition / Theodoros Giannakopoulos, Aggelos Pikrakis / Academic Press; 1 edition (April 21, 2014), p. 288

50. Huang X.D. Spoken Language Processing: A Guide to Theory, Algorithm and System Development [Text]/ Xuedong Huang, Alex Acero, Hsiao-Wuen Hon/ Prentice Hall PTC, New Jersey, 2001

51. Mahalanobis, Prasanta Chandra (1936). «On the generalised distance in statistics». Proceedings of the National Institute of Sciences of India 2 (1): 49–55.

52. Stevens, Stanley Smith; Volkmann; John; & Newman, Edwin B. (1937). "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America 8 (3): 185–190.