Федеральное агентство по образованию
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет информатики
Кафедра теоретических основ информатики

УДК 681.03

ДОПУСТИТЬ К ЗАЩИТЕ В ГАК
зав. кафедрой ТОИ, д.т.н., проф.
_____ Ю. Л. Костюк
«___» _____ 2005 г.

Рудой Алексей Александрович

ИССЛЕДОВАНИЕ АЛГОРИТМА ДИКТОРОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ РЕЧЕВОГО СИГНАЛА НА ПРИМЕРЕ ФОНЕМ /И/ и /У/

Дипломная работа

Научный руководитель,
аспирант А. Н. Огородников

Исполнитель,
студ. гр. 1401 А. А. Рудой

Электронная версия дипломной работы помещена
в электронную библиотеку. Файл_____
Администратор_____

Томск – 2005

РЕФЕРАТ

Курсовая работа на стр. 44, приложений 3, иллюстраций 17, литературы 11

РАСПОЗНОВАНИЕ РЕЧИ, РЕЧЕВЫЕ СИГНАЛЫ РАЗНЫХ ДИКТОРОВ, VISUAL C++ 6.0, WINDOWS NT/2000/XP

Объект исследования – речевые сигналы, их энергия.

Цель работы – разработка алгоритма дикторонезависимого распознавания фонем /и/ и /у/

Метод и методология проведения работы – аналитический и экспериментальный на ЭВМ

Результат работы:

1) Был разработан алгоритм распознавания фонем /и/ и /у/.
2) Был разработан алгоритм обучения распознаванию.

Область применения – результаты работы будут использоваться в дальнейшем при построении дикторонезависимых систем распознавания речи.

Содержание

ВВЕДЕНИЕ

На сегодняшний день существует немало систем по распознаванию речи. Существует два подхода по распознаванию речи: распознавание ограниченного набора слов и распознавание слитной речи.

Главной опорой распознавания ограниченного набора слов является ограниченный словарь (наиболее популярный набор – список цифр, удобный для многих практических применений). А при распознавании слитной речи обще признано, что обработка слитной речи требует: во-первых, перехода от распознавания слов как целостных звуковых образов к распознаванию звуковых единиц, меньших слова, и, во-вторых, учёта фонетических, синтаксических и семантических ограничений, определяющих возможные языковые структуры речевых сообщений.

В настоящее время уже изучено и создано много систем по распознаванию ограниченного набора слов. Например, числовые последовательности, команды, ключевые слова, собственные имена.

Основное требование к современным системам распознавания речи – многодикторность. Многодикторные системы бывают двух видов: с настройкой на диктора и дикторонезависимые, т. е. не требующие предварительной настройки системы на диктора. Дикторонезависимые системы распознавания слитной речи (например, диктовка, естественный диалог), находятся на стадии лабораторных разработок.

Создание дикторонезависимых систем затруднено тем, что параметры речевых сигналов у разных дикторов сильно варьируют, оставаясь, тем не менее, в определённых границах.

Задачей данной дипломной работы состоит в том, чтобы разработать и исследовать дикторонезависимый алгоритм распознавания гласных фонем на примере фонем /и/ и /у/.

Была выдвинута *гипотеза* о том, что отношение энергии речевого сигнала фонемы /и/ в высоких частотах к энергии в низких частотах будет больше такого же отношения для фонемы /у/ у разных дикторов. Причем существует такой порог, что отношение энергий фонемы /и/ будет больше этого порога, а для /у/ - меньше.

Цель данной дипломной работы:

1) Проверить приведенную выше гипотезу.
2) Разработать алгоритм дикторонезависимого распознавания гласных на примере фонем /и/ и /у/ на основе предложенной гипотезы.

1. АНАЛИТИЧЕСКИЙ ОБЗОР СИСТЕМ РАСПОЗНОВАНИЯ РЕЧИ И ЦИФР

Системы речевого обмена между человеком и ЭВМ можно подразделить на три класса:

1) с речевым ответом;
2) распознавания диктора:
    a) верификация диктора
    b) идентификация диктора
3) распознавания речи.

Системы с речевым ответом предназначаются для выдачи информации пользователю в форме речевого сообщения. Таким образом, системы с речевым ответом — это системы односторонней связи, т. е. от машины к человеку. С другой стороны, системы второго и третьего классов — это системы связи от человека к машине. В системах распознавания диктора задача состоит в верификации диктора (т. е. в решении задачи о принадлежности данного диктора к некоторой группе лиц) или идентификации диктора из некоторого известного множества. Таким образом, класс задач распознавания диктора распадается на два подкласса: верификации и идентификации говорящего.

Последний класс задач распознавания речи также можно разделить на подклассы в зависимости от таких факторов, как размер словаря, количество дикторов, условия произнесения слов и т. д. Основная задача распознающей системы сводится либо к точному распознаванию произнесенной на входе фразы (т. е. система фонетической или орфографической печати произнесенного текста), либо к «пониманию» произнесенной фразы (т. е. к правильной реакции на сказанное диктором). Именно задача понимания, а не распознавания наиболее важна для систем с достаточно большим словарем непрерывных речевых сигналов, в то время как задача точного распознавания более важна для систем с ограниченным словарем, малым количеством дикторов, систем распознавания изолированных слов.

1.1 Системы с речевым ответом

На рис. 1 представлена общая структурная схема системы с речевым ответом на базе ЭВМ. Элементами этой системы являются блоки: памяти для хранения словаря системы с речевым ответом; хранения правил синтеза сообщений по элементам словаря; программ формирования речевого ответа.

На вход системы с речевым ответом поступает сообщение о содержании вопроса, порождаемого либо другой системой обработки информации, либо непосредственно от человека, обратившегося с интересующим его вопросом к информационной системе.

Откликом системы на поставленный вопрос служит выходное сообщение в виде речевой фразы. Простым примером такой системы является автоматическая справочная телефонная служба, которая обнаруживает неправильно набранный номер, определяет причину ошибки (например, телефон отключен или ему присвоен новый номер и т. д.) и посылает на выход системы с речевым ответом сообщение, содержащее необходимые абоненту указания. В таких системах словарь обычно состоит из ограниченного набора изолированных слов (например, цифр с различными окончаниями).

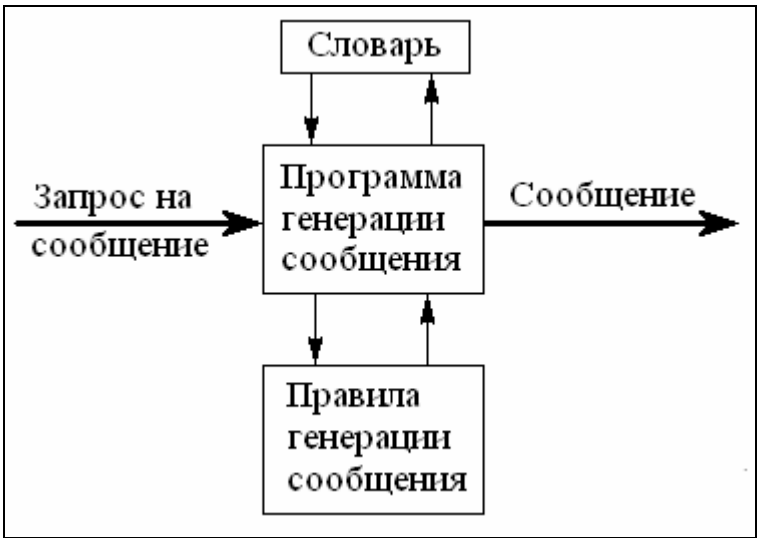

Рисунок 1 - Структурная схема системы с речевым ответом

Существуют два основных подхода к построению систем с речевым ответом. Один из них заключается в попытке построения системы, речевые возможности которой сравнимы с возможностями человека. Такие системы (называемые часто системами синтеза речи по правилам) основаны на модели речеобразования. В этом случае для синтеза достаточно хранить словарь произношений элементов. Сигналы, необходимые для управления речевым синтезатором, в соответствии с моделью речеобразования формируются на основе правил синтеза. Такие системы представляют интерес в том случае, если требуется словарь весьма большого объема. Реализация подобных систем — это проблема, требующая чрезвычайно трудоемких исследований. Однако основная трудность при построении подобных систем состоит в разработке правил управления синтезатором [6].

В системах с речевым ответом второго типа используется ограниченный словарь, и сигнал на выходе таких систем формируется посредством сочленения отдельных элементов реального речевого сигнала, взятых из словаря. Сообщения конструируются в этом случае путем отыскания требуемых слов и фраз в памяти и воспроизведения их в требуемой последовательности. При разработке систем подобного типа следует учитывать три основных соображения:

- Во-первых, способ представления и хранения словаря должен быть выбран таким образом, чтобы в разработанной системе имелась возможность свободного доступа к любому элементу словаря.
- Во-вторых, должен быть выбран способ редактирования речевого материала словаря совместно со способом записи его элементов IB память.
- В-третьих, необходимо обеспечить заданную последовательность выбора и воспроизведения элементов словаря (т. е. способ формирования сообщения).

Поскольку назначение систем с речевым ответом состоит в формировании речевых сообщений, предназначенных для человека, требование к разборчивости становится определяющим. Не менее важное значение, однако, имеют и такие параметры речи, как качество восприятия и натуральность. Таким образом, в разрабатываемой системе необходимо с

предельной полнотой реализовать все три основных условия с тем, чтобы добиться максимально возможной разборчивости и натуральности речевого сигнала.

В настоящее время созданы и исследованы: система речевых команд для выполнения межблочных соединений аппаратуры связи; вспомогательная справочная система; система информации о текущем курсе акций; система информации и контроля банков данных; справочная авиаслужба; система верификации дикторов.

1.2 Системы распознавания дикторов

При распознавании дикторов цифровая обработка речи является тем первым шагом, с которого начинается решение задачи распознавания образов. Как видно из рис. 2, речевой сигнал (представление образа вектором) представлен с использованием таких методов цифровой обработки, которые сохраняют индивидуальные особенности диктора. Полученный образ сравнивается с предварительно подготовленными эталонными образами, а затем применяется соответствующая логика принятия решений для определения голоса заданного диктора среди возможного множества. Системы распознавания дикторов подразделяются на два вида:

1. Идентификация
2. Верификация

При верификации диктора требуется установить его идентичность данному эталону. Устройство верификации принимает одно из двух возможных решений: диктор является тем, за кого он себя выдает, или не является. Для вынесения такого решения используется совокупность параметров, содержащих необходимую информацию об индивидуальности диктора и измеряемых по одной или нескольким фразам. Измеренные значения сравниваются (часто с использованием некоторых существенно нелинейных метрик близости) с аналогичными параметрами эталонных образов подлежащего опознанию диктора.

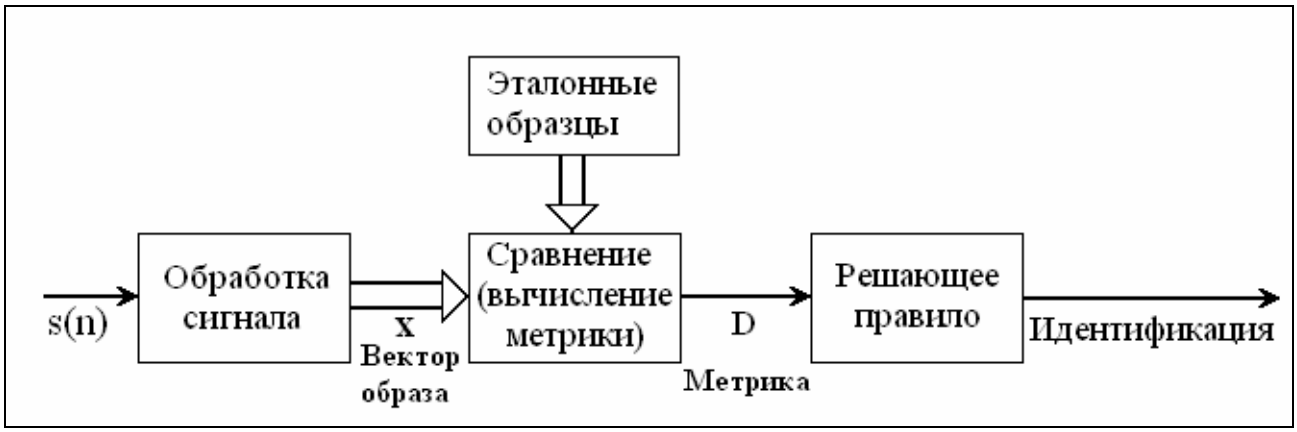

Рисунок 2 - Общее представление задачи распознавания диктора

Таким образом, при верификации диктора требуется однократное сравнение совокупности (совокупностей) измеренных значений со значениями параметров эталонов, на основе которого выносится решение о принятии или отклонении предполагаемой идентичности.

В общем случае вычисляется расстояние между измеренными значениями и распределением эталонов. На основе распределения потерь между возможными типами ошибок (т.е. верификации «самозванца» и отклонении «подлинного» диктора) устанавливается соответствующий порог различимости (расстояния). Вероятность перечисленных выше ошибок практически не зависит от N (числа эталонов, хранимых в системе), поскольку все эталоны голосов других дикторов используются для формирования устойчивого распределения, характеризующего всех дикторов. Записывая сказанное выше в математической форме, обозначим распределение вероятности измеренных значений вектора x для диктора как $p_i(x)$, что приводит к простому решающему правилу вида:

Верифицировать диктора i, если $p_i(x) > c_i * p_{av}(x)$; **(1)**

Отклонить диктора i, если $p_i(x) < c_i * p_{av}(x)$, **(2)**

где c_i — константа для i-го диктора, определяющая вероятности ошибок i-го диктора, а $p_{av}(x)$ — среднее (по всему ансамблю дикторов) распределение вероятности измеренных значений вектора x. Изменяя порог c_i, можно изменять вероятность ошибки, определяемую вероятностями ошибок обоих типов.

Задача идентификации диктора существенно отличается от задачи верификации. В этом случае система должна точно указать одного из дикторов среди N дикторов данного множества. Таким образом, вместо однократного сравнения измеряемых параметров с хранимым в системе эталоном необходимо провести N сравнений. Решающее правило в этом случае сводится к *выбору такого диктора i, для которого*:

$$p_i(x) > p_j(x), \quad j = 1, 2, \ldots N, \quad j \neq i \qquad \textbf{(3)}$$

т. е. выбирается диктор с минимальной абсолютной вероятностью ошибки. С увеличением количества дикторов в ансамбле возрастает и вероятность ошибки, поскольку большое число вероятностных распределений в ограниченном пространстве параметров не может не пересекаться. Все более вероятным становится то, что два или более дикторов в общем ансамбле будут иметь распределения вероятностей, которые близки друг к другу. При таких условиях приемлемая идентификация дикторов становится практически невозможной.

Приведенный выше анализ позволяет сделать вывод, что между задачами идентификации и верификации имеется много общего и много различий. В каждом случае диктор должен произнести одну или несколько тестовых фраз. По этим фразам проводятся некоторые измерения, и затем вычисляются одна или несколько мер различимости («расстояния») между предъявленным и эталонным векторами. Таким образом, с позиции методов цифровой обработки обе эти задачи сходны. Основное различие возникает на этапе вынесения решений.

1.3 Системы распознавания речи

Как и при распознавании диктора, методы цифровой обработки применяются при распознавании речевого сигнала для получения описания распознаваемого образа, которое затем сравнивается с хранимыми в памяти эталонами. Задача распознавания речевого сигнала состоит в определении того, какое слово, фраза или предложение были произнесены.

В отличие от областей машинного речевого ответа и распознавания диктора, где задача в общем случае достаточно определена, область распознавания слов является одной из тех, где, прежде чем поставить задачу, требуется ввести большое число предположений, например:

1) тип речевого сигнала (изолированные слова, непрерывная речь и т. д.);
2) число дикторов (система для одного диктора, нескольких дикторов, неограниченного числа дикторов);
3) тип диктора (определенный, случайный, мужчина, женщина, ребенок);
4) условия произнесения фраз (звукоизолированное помещение, машинный зал, общественное место);
5) система передачи (высококачественный микрофон, узконаправленный микрофон, телефон);
6) тип и число циклов обучения (без обучения, с ограниченным числом циклов обучения, с неограниченным числом циклов обучения);
7) размер словаря (малый объем 80—20 слов, средний объем 20—100 слов и большой объем — более 100 слов).
8) формат произносимых фраз (ограниченный по длительности текст, свободный речевой формат).

Из приведенного перечня условий следует, что при создании систем распознавания речи реализация некоторых из условий может оказаться более предпочтительной. Наиболее распространенными типами систем распознавания речи, в которых широко используются методы цифровой обработки сигналов, являются:

1. Система распознавания изолированных цифр [6]
2. Система распознавания слитной последовательности цифр
3. Система распознавания с большим объемом словаря

Хотя в системах распознавания слитной речи также широко используются цифровые методы обработки [6], однако большая часть усилий при разработке таких систем затрачивается на синтаксический и семантический анализ фраз. Эти области близко примыкают к лингвистической теории речи.

1.3.1 Система распознавания изолированных цифр

Система распознавания изолированных цифр обладает следующими свойствами:

1. Словарь малого объема состоит из изолированных слов, обозначающих десять цифр (0—9).
2. Отсутствуют ограничения на количество дикторов, а такжена их пол и возраст.
3. Условия произнесения: машинный зал, микрофон — узконаправленный или высококачественный.
4. Обучение не предусмотрено.
5. Формат на входе — однословный с паузами между словами.

Основными элементами системы являются: устройство анализа моментов начала и окончания слова (как и в системе распознавания), устройство обработки, формирующее образ или вектор измеренных значений, устройство сегментации фразы на интервалы и блок предварительных и окончательных решений относительно произнесенной цифры.

Хотя существует много способов представления сигнала, которые можно использовать в системах распознавания речи, представления, применяемые в системах, инвариантных к диктору, должны быть достаточно устойчивыми [6]. Измерения параметров должны быть простыми и однозначными, а их измеренные значения должны наиболее полно отражать различия в звуках речи.

Кроме того, измерения должны допускать достаточно простую интерпретацию с позиций систем, инвариантных к диктору. В одной из таких устойчивых систем использованы следующие параметры: среднее число переходов через нуль, энергия, коэффициенты линейного предсказания с использованием двухполюсной модели и погрешность предсказания.

Использование двухполюсной модели было предложено Макоулом и Вульфом [6] как удачный метод описания основных свойств кратковременного спектра.

Самбуром и Рабинером [6] предложен древовидный алгоритм принятия окончательного решения на основе совместной обработки частных решений, полученных по каждому измеренному значению на каждом интервале анализа. При использовании этого алгоритма точность распознавания для 65 дикторов составила от 94,4 до 97,3%.

1.3.2 Система распознавания слитной последовательности цифр

Приведем решение более сложной задачи распознавания слитной последовательности цифр при произнесении их произвольным диктором. Свойства, которыми должна обладать эта система, в основном совпадают со свойствами системы распознавания, рассмотренной при распознавании изолированных цифр, с одним важным исключением. Свойство 5 в данном случае состоит в необходимости распознавания слитной последовательности из трех слов (цифр) без пауз между ними.

Хотя между системами распознавания изолированных цифр и слитной последовательности цифр много общего, реализация этих систем распознавания существенно различается, особенно в блоке анализа или обработки сигнала. Это связано с необходимостью предварительной сегментации слитной последовательности на отдельные цифры перед их распознаванием. Задача сегментации является чрезвычайно сложной, и в настоящее время не найдено простого решения для общего случая.

На рис. 3 представлена структурная схема блока обработки сигнала в системе распознавания слитной последовательности цифр. Записанная последовательность цифр первоначально подвергается анализу с целью определения моментов начала и окончания

фразы. Вслед за определением моментов начала и окончания фразы речевой сигнал подвергается обработке с целью оценивания следующих параметров (100 раз/с): среднего числа переходов через нуль, логарифма энергии, коэффициентов линейного предсказания, логарифма погрешности линейного предсказания и первого коэффициента автокорреляции.

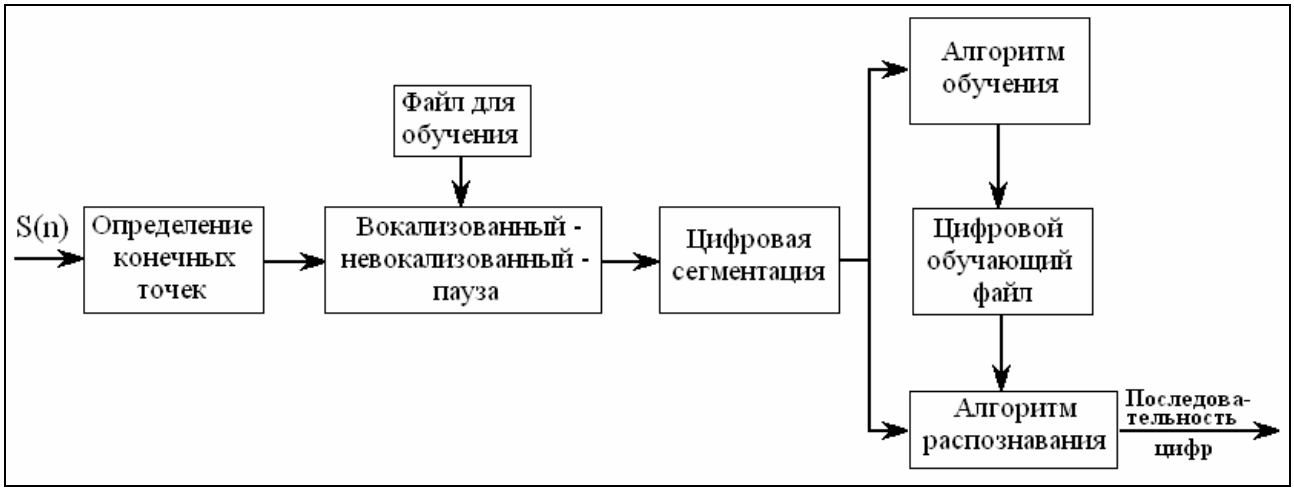

Рисунок 3 - Структурная схема системы распознавания последовательности цифр

Измеренные параметры используются затем в качестве входных сигналов решающего устройства, которое классифицирует каждый 10-миллисекундный интервал как вокализованный, невокализованный или. Для сегментации слитного потока на отдельные цифры используется нелинейно сглаженная траектория признака «вокализованный – невокализованный - пауза» совместно с некоторой статистической информацией о надежности классификации на каждом интервале и измеренными значениями энергии сигнала. Для правильной сегментации в систему необходимо ввести информацию о количестве цифр в фразе.

Сегментация осуществляется на основе использования известных результатов по различным измерениям на каждом 10-милли-секундном интервале. Например, известно, что невокализованный интервал соответствует интервалу, в пределах которого находится искомая граница, поскольку ни одна из цифр не содержит невокализованных звуков внутри слова. Известно также, что глубокие провалы в траектории энергии на вокализованном сегменте почти всегда соответствуют границе между цифрами. Основываясь на этих наблюдениях, можно синтезировать простые и более сложные правила сегментации фразы.

Следующим шагом после сегментации является реализация алгоритма распознавания. Для каждого сегмента цифры вокализованный участок (который определяется по признаку «вокализованный—невокализованный—пауза») подвергается анализу на основе линейного предсказания с использованием десятиполюсной модели. Метод распознавания основан на статистическом решающем правиле, по которому для каждого интервала обрабатываемой фразы (совокупности параметров линейного предсказания) проводится сравнение с соответствующим эталоном и выносится решение о соответствии обрабатываемой цифры той из эталонных, с которой имеется наибольшее сходство (наименьшее расстояние).

Эталонные файлы содержат статистическую информацию о коэффициентах линейного предсказания на каждом интервале анализа и для каждой цифры. В этих файлах содержится информация о среднем и дисперсии по множеству произнесений и множеству дикторов. При распознавании обрабатываемая фраза подвергается преобразованию масштаба времени выравниванием длительностей опознаваемых и эталонных цифр. Поскольку все цифры моно-силлабические, то в большинстве систем распознавания используется линейное

преобразование масштаба. Для каждой эталонной фразы вычисляется среднее расстояние между ее коэффициентами линейного предсказания и соответствующими коэффициентами обрабатываемой фразы, а за произнесенную принимается та цифра, для которой это расстояние оказывается минимальным. Выбор меры различимости, определяющей расстояние между совокупностями параметров линейного предсказания, является одним из наиболее важных факторов, определяющих качество работы подобных систем. Известен ряд мер различимости, используемых при обработке параметров линейного предсказания.

Испытания систем, аналогичных представленной на рис. 3, показали, что, если система настроена на определенного диктора, точность распознавания достигает 98 - 100%, а для произвольного диктора точность составляет около 95% [9.27, 28].

1.3.3 Система распознавания с большим объемом словаря

Третья из описываемых здесь систем распознавания обладает словарем, объем которого значительно превосходит объемы словарей двух первых систем. Однако платой за увеличение объема словаря является то, что система перестает быть не зависимой от диктора, т. е. система должна быть предварительно обучена применительно к каждому предполагаемому пользователю. С учетом обсуждения, проведенного во введении, разработанная Итакурой [6, 11] система с большим словарем обладает следующими свойствами:

1. Словарь состоит из изолированных слов, количество которых составляет 100—500.
2. Система предназначена для одного диктора, но после соответствующего обучения может быть настроена на любого диктора.
3. Отсутствуют ограничения на пол и возраст диктора.
4. Отсутствуют жесткие ограничения -на условия произнесения.
5. Система работает с сигналом телефонного качества.
6. Предусмотрено обучение системы в виде одно- или многократного произнесения каждого слова словаря.
7. Форматом произнесения являются слова, разделенные паузами.

Эта система исследовалась с использованием двух различных словарей. Применяя словарь объемом примерно 120 слов (названия различных городов Япония), Итакура получил частость правильного распознавания, равную 97,3%, а частость отклонения предъявленных слов 1,65%. Для словаря, соответствующего 26 буквам и цифрам от 0 до 9, полученная частость правильного распознавания равна 88,6%. Такое увеличение частости ошибок (11,4% при частости отклонения 0%) обусловлено большим сходством между некоторыми элементами словаря, например: b и d, m и n или i и y.

2. ПРОЦЕСС РЕЧЕОБРАЗОВАНИЯ

2.1 Механизм речеобразования

Речь предназначена для общения. Возможности речи с этой точки зрения можно характеризовать по-разному. Один из количественных подходов основан на теории информации, разработанной Шенноном. В соответствии с этой теорией речь можно описать ее информационным содержанием или *информацией.* Другой способ описания речи заключается в представлении ее в виде си*гнала,* т. е. акустического колебания. Хотя идеи теории информации играют важную роль при построении сложных систем связи, но наиболее полезными на практике являются представления речи в виде колебания или в виде некоторой параметрической модели.

Речевое общение начинается с того, что в мозгу диктора возникает в абстрактной форме некоторое сообщение. В процессе речеобразования это сообщение преобразуется в акустическое речевое колебание. Информация, содержащаяся в сообщении, представлена в акустическом колебании весьма сложным образом. Сообщение сначала преобразуется в последовательности нервных импульсов, управляющих артикуляторным аппаратом (т. е. перемещением языка, губ, голосовых связок и т. д.). В результате воздействия нервных импульсов артикуляторный аппарат приходит в движение, результатом которого является акустическое речевое колебание, несущее информацию об исходном сообщении.

Речь является конечным акустическим продуктом произвольных формализованных движений дыхательных и жевательных органов. Речь развивается, корректируется и поддерживается под воздействием акустической обратной связи органов слуха и кинестетической обратной связи мускулатуры органов речи. Основными органами, участвующими в речеобразовании, являются:

К наиболее важным органам речеобразующей системы человека [6, 8, 10] относятся:

1. *Голосовой тракт* начинается с прохода между голосовыми связками, называемого *голосовой щелью,* и заканчивается у губ. Голосовой тракт, таким образом, состоит из *гортани* (от пищевода до рта) и рта, или *ротовой полости.* У взрослого мужчины общая длина голосового тракта составляет примерно 17 см. Площадь поперечного сечения голосового тракта, которая определяется положением языка, губ, челюстей и небной занавески, может изменяться от нуля (тракт полностью перекрыт) до примерно 20 $см^2$.
2. *Носовая полость* начинается у небной занавески и заканчивается ноздрями. Носовая полость образует вспомогательный путь распространения звуковых колебаний. Он начинается с небной занавески и заканчивается ноздрями. На некотором протяжении носовая полость разделена носовой перегородкой на две полости. Величина акустической связи между носовой и ротовой полостями определяется размерами прохода у небной занавески. В зависимости от величины этой связи звук может излучаться как через рот, так и через ноздри. Связь с носовой полостью существенным образом влияет на характер звука, излучаемого через рот. При образовании не носовых звуков небная занавеска поднята и плотно закрывает вход в носовую полость.
3. К органам речеобразующей системы также относятся такие органы, как легкие, бронхи и трахея, расположенные ниже гортани. Совокупность этих органов служит источником энергии для образования речи. Воздух втягивается в лёгкие при расширении грудной клетки и опускании диафрагмы. Он выталкивается из лёгких при сжимании грудной клетки и увеличении лёгочного давления. Для образования

гласных звуков речи с минимальным возможным уровнем требуется лёгочное давление порядка 4 см водяного столба. Для очень громких высоко тональных звуков обычно развивается давление порядка 20 см водяного столба. В процессе разговора лёгочное давление поддерживается на требуемом уровне благодаря непрерывному и медленному сжиманию грудной клетки.

В настоящее время работа голосовых связок изучена достаточно подробно.

Звуки речи могут быть разделены на три четко выраженные группы по типу возбуждения:

1. *Вокализованные* звуки образуются проталкиванием воздуха через голосовую щель, при котором периодически напрягаются и расслабляются голосовые связки и возникает квазипериодическая последовательность импульсов потока воздуха, возбуждающая голосовой тракт.
2. *Фрикативные* или *невокализованные* звуки генерируются при сужении голосового тракта в каком-либо месте (обычно в конце рта) и проталкивании воздуха через суженное место со скоростью, достаточно высокой для образования турбулентного воздушного потока. Таким образом, формируется источник широкополосного шума, возбуждающего голосовой тракт.
3. При произнесении *взрывных звуков* голосовой тракт полностью закрывается (обычно в начале голосового тракта). За этой смычкой возникает повышенное сжатие воздуха. Затем воздух внезапно высвобождается.

Голосовой тракт и носовую полость можно представить в виде труб с переменной по продольной оси площадью поперечного сечения. При прохождении звуковых волн через эти трубы их частотный спектр изменяется в соответствии с частотной избирательностью трубы. Этот эффект похож на резонансные явления, происходящие в трубах органов и духовых музыкальных инструментов. При описании речеобразования резонансные частоты трубы голосового тракта называют *формантными частотами* или просто *формантами.* Формантные частоты зависят от конфигурации и размеров голосового тракта: произвольная форма тракта может быть описана набором формантных частот. Различные звуки образуются путем изменения формы голосового тракта. Таким образом, спектральные свойства речевого сигнала изменяются во времени в соответствии с изменением формы голосового тракта.

Переменные во времени спектральные характеристики речевого сигнала с помощью звукового спектрографа могут быть высвечены в виде графика [5]. Этот прибор позволяет получить двумерный график, называемый *спектрограммой,* на которой по вертикальной оси отложена частота, а по горизонтальной — время. Плотность зачернения графика пропорциональна энергии сигнала. Таким образом, резонансные частоты голосового тракта имеют вид затемненных областей на спектрограмме. Вокализованным областям сигнала соответствует появление четко выраженной периодичности временной зависимости, в то время как невокализованные интервалы выглядят почти сплошными.

Звуковой спектрограф весьма долго служил основным инструментом исследования речевого сигнала, хотя в настоящее время с помощью цифровой обработки можно получить более гибкие устройства визуального изображения, основные принципы спектрографа используются широко и в настоящее время.

2.2 Акустическая фонетика

Чтобы служить практическим средством передачи информации, язык должен описываться с помощью конечного числа различимых и исключающих друг друга звуков.

Это означает, что язык должен описываться основными лингвистическими единицами, обладающими тем свойством, что если в фразе заменить одну единицу другой, значение фразы изменится. При акустической реализации основная единица может быть подвержена существенным видоизменениям. Подобные видоизменения при восприятии человеком, знающим язык, соотносятся в его сознании с одним и там же лингвистическим элементом. Эти основные лингвистические элементы называются фонемами, а их часто разнообразные, различимые варианты – аллофонами. В каждом языке имеется присущее ему множество фонем, обычно ют 30 до 50.

Например, в английском языке можно выделить 42 фонемы. На рис. 4 приведены различные классы фонем английского языка в его американском произношении. Четыре широких класса звуков образуют гласные, дифтонги, полугласные и согласные. Каждый из классов разбит на подклассы по способу и месту образования звука в голосовом тракте. Каждая фонема рис. 4 может быть отнесена к классу протяжных или кратковременных звуков. Протяжные звуки образуются при фиксированной (инвариантной ко времени) форме голосового тракта, который возбуждается соответствующим источником. К этому классу относятся гласные, фрикативные (вокализованные и невокализованные) носовые согласные. Остальные звуки (дифтонги, полугласные, аффрикаты и взрывные согласные) произносятся при изменяющейся форме голосового тракта. Они образуют класс кратковременных звуков.

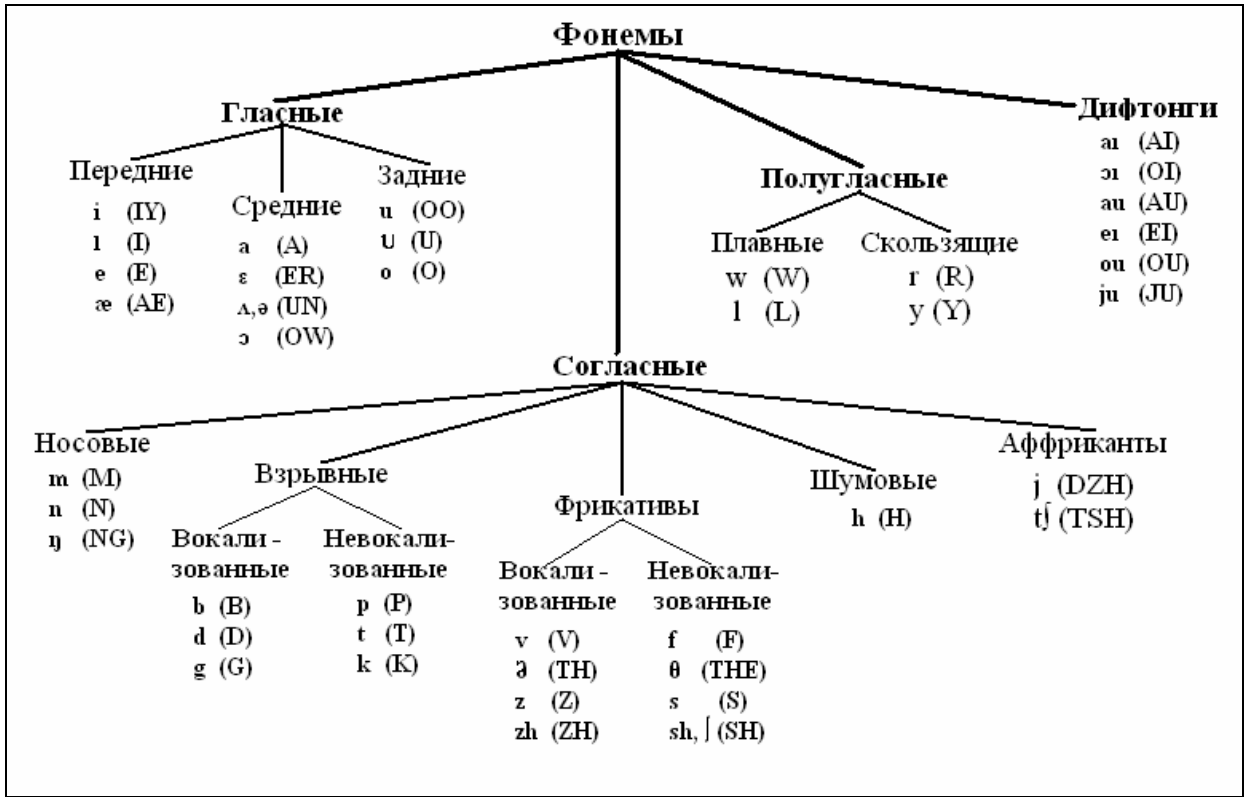

Рисунок 4 - Классификация фонем английского языка в его американском произношении.

Набор фонемных символов и их статические характеристики определяются языком и диалектом, на котором ведётся разговор.

Гласные. Гласные образуются при квазипериодическом возбуждении голосового тракта неизменной формы импульсами воздуха, возникающими вследствие колебания голосовых связок. Зависимость площади поперечного сечения голосового тракта от координаты (расстояния) вдоль его продольной оси определяет резонансные частоты тракта (форманты) и характер произносимого звука. Эта зависимость называется *функцией площади поперечного*

сечения. Функция площади поперечного сечения для каждой гласной зависит в первую очередь от положения языка; вместе с тем на характер звука оказывают влияние положения челюстей, губ и, в меньшей степени, небной занавески. Например, при произнесении звука |a|, как в слове «father», голосовой тракт открыт в начале, а в его конце тело языка образует сужение. Наоборот, при произнесении звука |i|, как в слове «eve», язык образует сужение в начале голосового тракта и оставляет его открытым в конце. Таким образом, каждому гласному звуку может быть поставлена в соответствие форма голосового тракта (функция площади поперечного сечения), характерная для его произношения. Очевидно, что это соответствие неоднозначное, так как у разных дикторов голосовые тракты различны. Другим представлением гласного звука является его описание с помощью набора резонансных частот голосового тракта. Это описание также зависит от диктора. Петерсон и Барней [5] провели измерения формантных (резонансных) частот с помощью звукового спектрографа для гласных, произнесенных различными дикторами.

Дифтонги. Дифтонгом называется участок речи, соответствующий одному слогу, который начинается с одной гласной и затем постепенно переходит в другую. На основе этого определения можно выделить шесть дифтонгов:

1. |*eI*| - как в слове «bay»;
2. |*oU*| - как в слове «boat»;
3. |*aU*| - как в слове «how»;
4. |*oI*| - как в слове «boy»;
5. |*aI*| - как в слове «buy»;
6. |*ju*| - как в слове «you».

Дифтонги образуются путем плавного изменения формы голосового тракта. Дифтонги можно описать изменением во времени функции площади поперечного сечения голосового тракта от значения, соответствующего первой гласной, до значения, соответствующего второй гласной дифтонга.

Полугласные. Группу звуков, содержащих |*w*|, |*l*|, |*r*| и |*y*|, описать довольно трудно. Эти звуки называются *полугласными*, так как по своим свойствам они напоминают гласные звуки. Обычно их характеризуют плавным изменением функции площади поперечного сечения голосового тракта между смежными фонемами. Таким образом, акустические характеристики этих звуков существенно зависят от произносимого текста. Нам удобно рассматривать эти звуки как переходные, сходные с гласными. Их структура близка к структуре гласных и дифтонгов.

Носовые звуки. Носовые согласные |*m*|, |*n*| и |*ŋ*| образуются при голосовом возбуждении. В полости рта при этом возникает полная смычка. Небная занавеска опущена, поэтому поток воздуха проходит через носовую полость и излучается через ноздри. Полость рта, которая вначале закрыта, акустически соединена с гортанью. Таким образом, рот служит резонансной полостью, в которой задерживается часть энергии при определенных частотах воздушного потока. Эти резонансные частоты соответствуют антирезонансам или нулям передаточной функции тракта речеобразования [6, 10]. Более того, для носовых согласных и гласных (т. е. гласных, расположенных перед носовыми согласными) характерны менее выраженные резонансы, чем для гласных. Расширение резонансных областей происходит из-за того, что внутренняя поверхность носового тракта напрягается и при этом носовая полость имеет большое отношение площади поверхности к площади поперечного сечения. Вследствие этого потери за счет теплопроводности и вязкости оказываются большими, чем обычно.

Три носовых согласных различаются местом расположения полной смычки. При произнесении звука |*m*| смычка образуется между губами, |*n*| — у внутренней стороны зубов и

|ŋ| - у небной занавески. Временные колебания согласных |*m*| и |*n*| очень похожи. Это происходит вследствие взаимного влияния резонансов и антирезонансов, образующихся за счет взаимодействия полостей носа и рта [6, 8].

Глухие фрикативные звуки. Глухие фрикативные звуки |*f*|, |*θ*|, |s| и |*sh*| образуются путем возбуждения голосового тракта турбулентным воздушным потоком, возникающим в области смычки голосового тракта. Расположение смычки характеризует тип фрикативного звука. При произнесении звука |*f*| смычка возникает около губ, |*θ*| — около зубов, |s| — в середине полости рта и |*sh*| — в конце полости рта. Таким образом, система образования глухих фрикативных звуков содержит источник шума, расположенный в области смычки, которая разделяет голосовой тракт на две полости. Звуковая волна излучается через губы, т. е. через переднюю полость. Другая полость служит, как и в случае произнесения носовых звуков, для задерживания акустического потока, и таким образом в речеобразующем тракте возникают антирезонансы [6, 10].

Звонкие фрикативные звуки. Звонкие фрикативные звуки |*v*|, |*th*|, |*z*| и |*zh*| являются прототипами глухих звуков |*f*|, |θ|, |*s*| и |*sh*| соответственно. Место расположения смычки для этих пар звуков совпадает. Однако звонкие фрикативные отличаются от своих глухих аналогов тем, что при их образовании участвуют два источника возбуждения. При образовании звонких звуков голосовые связки колеблются и, таким образом, один источник возбуждения находится в гортани. Однако, так как в голосовом тракте образуется смычка, поток воздуха в этой области становится турбулентным.

Звонкие взрывные согласные. Звонкие взрывные согласные |*b*|, |*d*| и |g| являются переходными непротяжными звуками. При их образовании голосовой тракт смыкается в какой-нибудь области полости рта. За смычкой воздух сжимается и затем внезапно высвобождается. При произнесении звука |*b*| смычка образуется между губами, |*d*| - с внутренней стороны зубов, |*g*| — вблизи небной занавески. В течение периода, когда голосовой тракт полностью закрыт, звуковые волны практически не излучаются через губы. Однако слабые низкочастотные колебания излучаются стенками горла (эту область иногда называют голосовым затвором — «voice bar»). Колебания возникают из-за того, что голосовые связки могут вибрировать даже тогда, когда голосовой тракт перекрыт.

Так как структура взрывных звуков изменчива, их свойства существенно зависят от последующего гласного [6, 8]. В этой связи характер временных колебаний несет мало сведений о свойствах этих согласных.

Глухие взрывные согласные. Глухие взрывные согласные |*p*|, |*t*| и |*k*| подобны своим звонким прототипам |*b*|, |*d*| и |*g*|, но имеют одно важное отличие. В течение периода полного смыкания голосового тракта голосовые связки не колеблются. После этого периода, когда воздух за смычкой высвобождается, в течение короткого промежутка времени потери на трение возрастают из-за внезапной турбулентности потока воздуха. Далее следует период придыхания (шумовой воздушный поток из голосовой щели возбуждает голосовой тракт). После этого возникает голосовое возбуждение.

Аффрикаты и звук |h|. Остальными согласными американского произношения являются аффрикаты |*tʃ*| и |*j*| и фонема |*h*|.

Глухая аффриката |*tʃ*| является динамичным звуком, который можно представить как сочетание взрывного |*t*| и фрикативного согласного |ʃ|. Звонкий звук |*j*| можно представить как сочетание взрывного |*d*| и фрикативного звука |*zh*|. Наконец, фонема |*h*| образуется путем возбуждения голосового тракта турбулентным воздушным потоком, т. е. без участия голосовых связок, но при возникновении шумового потока в голосовой щели. (Этот способ возбуждения характерен и для шепота). Структура звука |*h*| не зависит от следующей за ним гласной. Поэтому голосовой тракт может перестраиваться для произнесения следующей гласной в процессе, произнесения звука |*h*|.

2.3 Распространение звуков

Понятие звука почти совпадает с понятием колебаний. Звуковые волны возникают за счет колебаний. Они распространяются в воздухе или другой среде с помощью колебаний частиц этой среды. Следовательно, образование и распространение звуков в голосовом тракте подчиняется законам физики. В частности, основные законы сохранения массы, сохранения энергии, сохранения количества движения вместе с законами термодинамики и механики жидкостей применимы к сжимаемому воздушному потоку с низкой вязкостью, который является средой распространения звуков речи. Используя эти основные физические законы, можно составить систему дифференциальных уравнений в частных производных, описывающую движение воздуха в речеобразующей системе [6]. Составление и решение этих уравнений весьма затруднительны даже для простых предположений относительно формы голосового тракта и потерь энергии в речеобразующей системе. Полная акустическая теория должна учитывать следующие факторы:

1. изменение во времени формы голосового тракта;
2. потери энергии на стенках голосового тракта за счет вязкого трения и теплопроводности;
3. мягкость стенок голосового тракта;
4. излучение звуковых волн через губы;
5. влияние носовой полости;
6. возбуждение голосового тракта.

Построение и создание акустической теории, охватывающей все эти факторы пока еще невозможно.

2.4 Ухо и слух

Конечным приёмником информации в канале речевой связи обычно является человек. Способность человека к восприятию и определяет точность, с какой следует обрабатывать и передавать речевые данные. Эта способность, по существу, задаёт критерий точности при приёме и фактически определяет пропускную способность канала, необходимую для передачи речевых сообщений.

Основным источником знаний об акустико-механических процессах в периферическом отделе слухового анализатора являются эксперименты Бекеши (G. von Bekesy), отмеченные Нобелевской премией в 1961 г.

Ухо – первичный акустический преобразователь, используемый человеком. Акустико-механические компоненты этого органа обычно разделяются на 3 области: наружное, среднее и внутреннее ухо.

2.4.1 Наружное ухо

Обычным термином ухо обозначают ушную раковину, которая окружает вход в слуховой проход. Главное назначение ушной раковины человека состоит в защите слухового прохода, хотя её характеристики направленности на высоких частотах слухового диапазона, вероятно, также облегчают и локализацию источников звука.

Слуховой проход заканчивается мембраной, называемой барабанной перепонкой. Эта мембрана представляет собой относительно жёсткий конус, направленный во внутрь. При грубой аппроксимации слуховой проход можно представить как однородную трубу - открытую на одном конца и закрытую на другом. На резонансных частотах вдоль трубы укладывается нечётное число четвертей длины волны. Таким образом, первый резонанс приходится на частоту около 3000 гц. Можно полагать, что этот резонанс повышает чувствительность слуха в данном диапазоне частот.

2.4.2 Среднее ухо

За барабанной перепонкой находится заполненная воздухом полость среднего уха, содержащая слуховые косточки: молоточек, наковальню, стремечко.

Звуковая волна проходит через наружное ухо и слуховой поход, вызывая колебания барабанной перепонки. Это колебание через три слуховые косточки передаётся во внутреннее ухо. Акустико-механический импеданс внутреннего уха намного превышает импеданс воздуха, поэтому для эффективной передачи энергии звука требуется преобразование (повышение) импеданса. Эту задачу выполняют слуховые косточки.

Среднее ухо выполняет ещё одну важную функцию, а именно защищает от громких звуков более нежное внутреннее ухо.

Одной из важных характеристик среднего уха является частотная характеристика, т. е. Зависимость величины смещения основания стремечка от звукового давления на барабанную перепонку. Ряд исследователей пытались измерить или рассчитать эту характеристику (Бекеше, 1960; Звислоцкий – Zwislocki, 1957, 1959; Мёллер, 1961, 1962). Результаты оказались весьма различными, поскольку частотная характеристика зависит не только от жизненного тонуса человека, но и существенно изменяется от одного индивидуума к другому.

Бекеши, (1960) выполнил ряд измерений передаточной функции среднего уха, непосредственно наблюдая величину смещения круглого окна. Свойства передаточной

функции можно определить на основании знаний о строении среднего уха, входом механическом импедансе внутреннего уха и акустическом импедансе барабанной перепонки. Этот подход использован в работах Звислоцкого (1957, 1958) и Мёллера (1961) для создания схемы – аналога среднего уха. Все эти результаты, согласуясь в общих чертах, свидетельствуют о значительном разбросе характеристик передаточной функции. Совпадает лишь общий вывод о том, что передаточная функция среднего уха имеет характеристики фильтра нижних частот.

2.4.3 Внутреннее ухо

Внутреннее ухо состоит (в нормальном состоянии свёрнутой в плоскую спираль с двумя с половиной оборотами и напоминающей раковину улитки), вестибулярного аппарата и окончаний слухового нерва. В улитке происходит преобразование механических процессов в нервные. Компоненты вестибулярного аппарата (полукружные каналы, мешочек и маточка) служат для ориентации в пространстве и, по-видимому, не используются при анализе слуховых колебаний.

Полость улитки вдоль почти всей её длины разделена перегородкой. Одна половинка, включающая стремечко, называется преддверной лестницей, другая половина – барабанной лестницей. Внутри перегородки улитки имеется канал, называемый улитковым ходом. С одной стороны улитковый ход ограничен костистым выступом со студенистой мембраной, называемой базилярной мембраной, с другой стороны – мембраной, известной как мембраной Рейснера. Перегородка заполнена особой жидкостью – эндолимфмой.

Внутреннее ухо связано со средним ухом подножной пластинкой стремечка. Подножная пластинка поддерживается круговой связкой и располагается в овальном окне. При вибрации стремечко действует как поршень и производит смещение объёма жидкости улитки.

Главным элементом, определяющим основные функциональные динамические свойства перегородки, является базилярная мембрана, на которой покоится орган Корти. Базилярная мембрана у основания уже и значительно жестче и тоньше, чем у вершины, где она более податлива массивна. Поэтому резонансные свойства базилярной мембраной непрерывно изменяются вдоль её длины. На низких частотах базилярная мембрана обычно колеблется синфазно с мембраной Рейснера.

Современные знания акустико-механических свойств базилярной мембраны основываются почти исключительно на результатах исследований Бекеши. Проводя опыты на физиологических препаратах, Бекеши вызывал гармонические колебания основания стремечка и измерял амплитуды и фазы смещений мембраны вдоль улитки.

Однако выполненные Бекеши измерения свидетельствуют, что максимумы частотных характеристик, полученные для резонирующих точек мембраны, увеличиваются с увеличением частоты примерно до 1000 гц со скоростью около 5 дб на октаву; на более высоких частотах амплитуды указанных максимумов примерно одинаковы. Линейным приращениям координаты базилярной соответствуют приращения резонансной частоты почти по логарифмическому закону. Такая закономерность соблюдается, по крайней мере, для частот ниже 1000гц.

2.4.4 Преобразование механических колебаний в нервное возбуждение

Механические движения мембраны превращаются в нервное возбуждение в органе Корти, который состоит из большого числа клеток, среди которых имеются и волосковые клетки. Волоски, выходящие из этих чувствительных клеток, проникают сквозь сетчатую

пластинку и соприкасаются с третьей мембраной перегородки улитки – с покровной мембраной.

Электрофизиологические эксперименты свидетельствуют, что наружные и внутренние волосковые клетки органа Корти различаются по чувствительности к механическим воздействиям (Бекеши, 1953; Дэвис, 1958).

Чувствительные клетки уха соединены с мозгом пучком нервных клеток, или нейронов, образующим слуховой нерв. Общее число нейронов в слуховом нерве доходит, примерно, до 30 000. Нейроны имеют только 2 состояния: активное и заторможенное. При возбуждении входным электрическим сигналом, превышающим некоторый порог, нейроны генерируют стандартный электрический импульс длительностью около 1 мсек., после чего наступает период нечувствительности, длящийся от 1 до 3 мсек. Следовательно, возбуждение нейронов может приводить к появлению разрядов с максимальной частотой до 300-1000 Гц.

Связи между клетками нерва и органом Корти имеют сложную структуру. Функциональное назначение этой сложной системы многократных соединений в настоящее время точно не известно. Высказано предположение, что эта система способствует расширению динамического диапазона слуха (ван Бергейк – van Bergeijk).

Относительно мало известно о механизмах преобразования смещения базилярной мембраны в нервную активность. Ещё меньше известно о способе кодирования информации нервными импульсами и о том, каким образом в мозге возникает слуховое ощущение.

2.4.5 Математическая модель уха

Выше подчёркивалось, что механизм слухового восприятия в целом ещё недостаточно изучен. Тем не менее, современные знания физиологии уха, электрофизиологии нервных клеток и субъективного поведения аудитора при психоакустических испытаниях позволяют установить связь между некоторыми функциями слуха и этими столь различными областями знаний. Установление подобных связей облегчается, если поведение удаётся количественно оценить и аналитически предопределить.

Первым шагом в этом направлении было построение математической модели (рис. 5), описывающей смещение базилярной мембраны под действием произвольного звукового давления у барабанной перепонки (Фланаган, 1962).

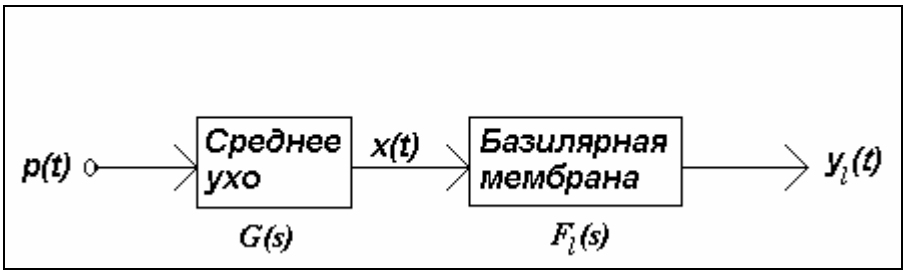

Рисунок 5 - Схематическое изображение уха

$$\frac{Y_l(s)}{P(s)} = \frac{X(s)}{P(s)} * \frac{Y_l(s)}{X(s)} = G(s) * F_l(s) \qquad \textbf{(1)}$$

Обозначения: *p(t)* – давление звука у барабанной перепонки, *x(t)* – эквивалентное линейное смещение основания стремечка, $y_l(t)$ – линейное смещение базилярной мембраны в точке, расположенной на расстоянии l от стремечка. Задача решается в 2 этапа. На I этапе аппроксимируется передаточная функция среднего уха, т. е. устанавливается связь между *x(t)* и *p(t)*. На II этапе аппроксимируется передаточная функция системы на участке от стремечка до указанной точки l на мембране. Аппроксимирующие функции представлены в виде частотных преобразований *G(s)* и $F_l(s)$.

2.5 Речь как процесс фильтрации

Речевая волна представляет собой результат воздействия одного или нескольких источников звука на фильтрующую систему речевого тракта.

Теория речеобразования, основанная на представлениях об источниках звука и фильтрах, характеризуется блок-схемой (рис. 6).

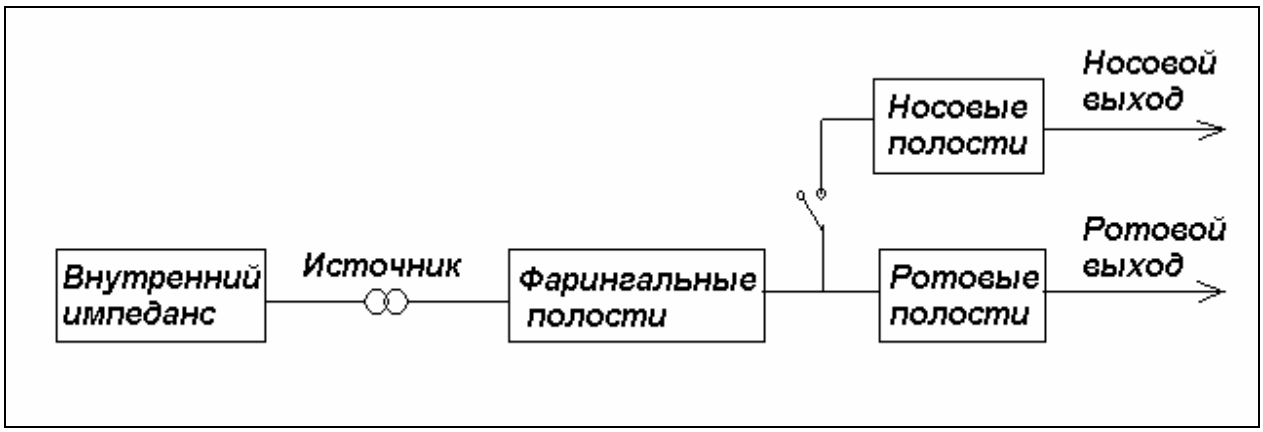

Рисунок 6 - Схематическое представление механизма образования звуков, создаваемых при участии голосового источника

На ней показано несколько соединённых между собой фильтровых звеньев, каждое из которых представляет часть полостей речевого тракта. На рис. 6 в качестве источника звука приняты голосовые связки. Носовая полость присоединена в точке схемы, соответствующей границе между фарингальной и ротовой областями тракта.

Подобные же блок-схемы, но только с двумя фильтровыми звеньями, представляющими соответственно переднюю и заднюю полости, можно считать технической реализацией фонетических представлений. При этом сразу возникают трудности с определением истинных физиологических границ этих полостей.

Полная фильтровая функция, в технике часто называемая функцией передачи, представляет собой частотную зависимость отношения двух величин: звукового давления в звуковом поле на известном расстоянии ото рта говорящего и звукового давления или объёмной скорости источника.

Обозначив через *S* функцию, характеризующую источник, и через *T* – функцию, отображающую свойства фильтра, акустическую характеристику звука речи можно представит равенством:

$$P = S * T \qquad \textbf{(2)}$$

Строго говоря, обе входящие сюда величины в общем случае зависят и от частоты и от времени. Частотную зависимость удобно выразить, используя преобразование Лапласа, как зависимость соответственных функций от комплексной частоты $s = \sigma + j*\omega$:

$$P(s) = S(s) * T(s) \quad \textbf{(3)}$$

т. е. преобразование Лапласа *P(s)* от звукового давления в звуковом поле перед диктором является произведением соответственных преобразований для функции источника *S(s)* и функции передачи *T(s)*.

Говоря о речеобразовании, следует отметить, что источник *S* в формуле (2) представляет собой акустическое возмущение, наложенное на поток выдыхаемого воздуха. Это возмущение вызывается либо препятствием в речевом тракте, обусловливающим наличие трения или внезапное открытие и закрытие прохода, либо, в случае сонорных звуков, квазипериодической модуляцией потока воздуха изменением ширины прохода между голосовыми связками.

Основным свойством голосового источника является периодичность создаваемого звука, которая определяется длительностью T_0 одного цикла работы голосовых связок; обратная ей величина представляет собой основную частоту голоса и равна:

$$F_0 = 1/T_0 \quad \textbf{(4)}$$

«Высота голоса» и «основная частота голоса» не являются синонимами. Строго говоря, высота есть ощущение, связанное с воздействием того или иного тона, а частота – физическое свойство звукового стимула.

Длительность цикла, которым определяется высота голоса, всегда несколько изменяется от периода к периоду. Частью эти изменения имеют систематический характер и связаны с интонационным рисунком речи, частью же представляют собой случайные или, точнее говоря, непреднамеренные колебания; однако эти колебания являются существенным признаком естественной человеческой речи.

Другой характеристикой голосового источника является огибающая спектра создаваемых им колебаний, т. е. зависимость *S(f)* амплитуд составляющих спектра от их частоты. Огибающая определяется регистром голоса, основной частотой и громкостью речи, но отражает также и индивидуальные свойства голоса говорящего.

Одним из основных признаков классификации звуков речи в классической фонетике является характеристика источника. Термин «голос» используется и как характеристика категории источника, и как характеристика специфического вида звуковых колебаний. С точки зрения характера источника возможны следующие случаи:

1) отсутствие источника (пауза);
2) только голосовой источник;
3) одновременно голосовой и шумовой источники;
4) шумовой источник, один или несколько.

Термин «шумовой источник» относится к первичному акустическому возмущению в речевом тракте, возникающему при образовании шёпотных, аспирированных, фрикативных и смычных звуков.

Аналитическое представление речеобразования путём разложения на две компоненты – источник и фильтр – можно продемонстрировать на простом примере, относящемся к сонорному звуку речи. Вследствие большого внутреннего сопротивления голосовой щели характеристикой источника можно считать заданный пульсирующий поток воздуха через голосовую щель. Этот поток, как функция времени, может быть представлен пилообразной кривой.

Использовав преобразование Фурье, можно получить спектр источника в виде ряда гармоник (рис. 7).

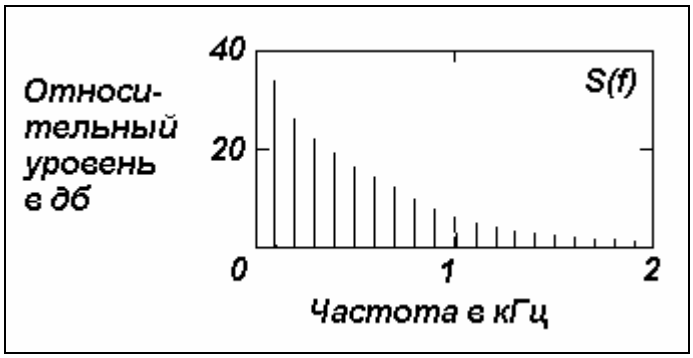

Рисунок 7 - Спектр источника

Для получения амплитудного спектра звука после речевого тракта амплитуды каждой из гармоник спектра источника |S(f)| умножаются на значение фильтровой функции $|T(f)|$ для частоты этой гармоники (5):

$$|P(f)| = |S(f)| * |T(f)| \quad \textbf{(5)}$$

Фаза каждой из гармоник после передачи по речевому тракту может быть получена как сумма фазы данной гармоники в спектре источника и фазы фильтровой функции для частоты этой гармоники (6):

$$\varphi P(f) = \varphi S(f) + \varphi T(f) \quad \textbf{(6)}$$

Это – процесс синтеза, который может быть реализован во всех деталях говорящей машиной.

Технической задачей спектрального анализа речи является получение огибающей функции *P(f)* по временной зависимости *p(t)* звука, воспринимаемого микрофоном.

3. ЦИФРОВАЯ ОБРАБОТКА РЕЧИ

3.1 Задача обработки сигналов

Задача обработки сигналов схематически представлена на рис. 8.

В случае речевых сигналов источником информации является человек. Измерению или наблюдению обычно подвергается акустическое колебание. Обработка сигнала предполагает в первую очередь формирование описания на основе некоторой модели с последующим преобразованием полученного представления требуемую форму. Последним шагом в процессе обработки является выделение и использование информационного содержания сигнала. Этот шаг может осуществляться путем прослушивания сигнала человеком или его автоматической обработки. В качестве примера можно рассмотреть систему идентификации диктора из заданного ансамбля дикторов, в которой используется представление речевого сигнала в виде зависящего времени спектра. Одним из возможных преобразований сигнала в этих условиях является усреднение спектра по всей фразе, сравнение среднего спектра с эталонами, имеющимися для каждого диктора, и затем выбор соответствующего диктора на основе полученных мер сходства спектров. Для данного примера информационным содержанием сигнала являются признаки индивидуальности диктора.

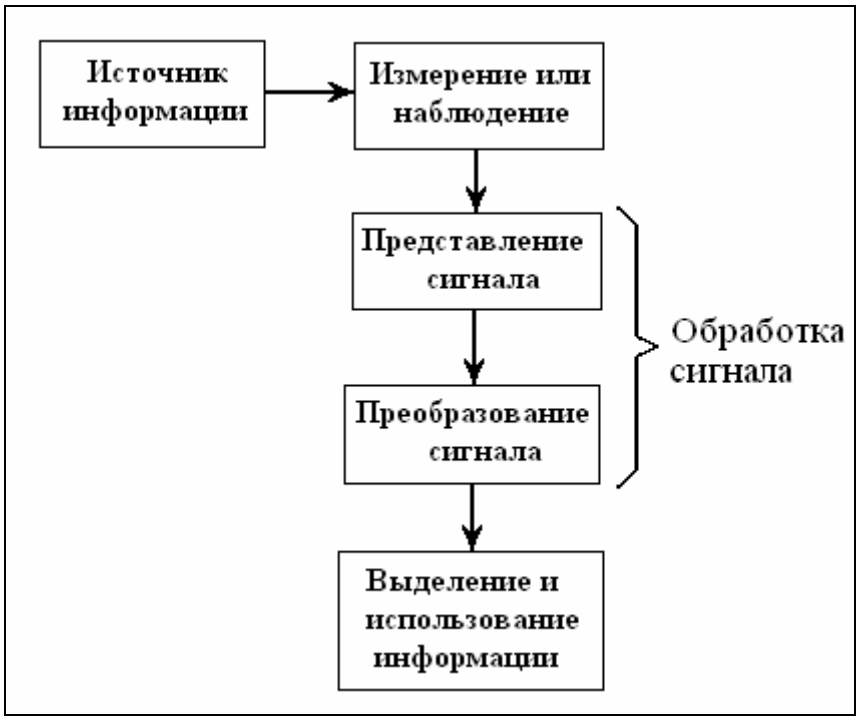

Рисунок 8 - Схема обработки информации

Таким образом, обработка сигнала в общем случае предусматривает решение двух основных задач:

1. Получить общее представление сигнала либо в форме речевого колебания, либо в виде параметров;

2. Преобразовать полученное представление в более удобную для решаемой задачи форму.

Цифровая обработка включает как получение дискретных представлений сигнала, так и теорию, расчет и применение цифровых алгоритмов для преобразования полученных дискретных представлений. Первые методы цифровой обработки речевых сигналов имитировали сложные аналоговые системы [1]. Согласно современной точке зрения система цифровой обработки речевых сигналов, выполненная в виде программы на ЭВМ, реализует точный алгоритм обработки и может быть изготовлена в виде специализированного вычислительного устройства.

Цифровые методы в настоящее время широко применяются при решении задач обработки речевых сигналов [4, 10].

3.2 Способы представления речевых сигналов и их применение

При рассмотрении вопросов применения цифровой обработки речевых сигналов полезно сконцентрировать внимание на трех основных направлениях:

- представление речевых сигналов в цифровой форме
- цифровой реализации аналоговых методов обработки
- методы, основанные исключительно на цифровой обработке.

Представление речевых сигналов в цифровой форме является, конечно, одним из центральных вопросов. Одной из самых основных теорем является *теорема дискретизации* [1, 6] или *теорема Котельникова*, утверждающая, *что всякий ограниченный по полосе частот сигнал может быть представлен в виде последовательности равноотстоящих отсчетов, взятых с достаточно высокой частотой*. Таким образом, процедура дискретизации лежит в основе теории и приложений цифровой обработки. Существует ряд способов дискретного представления речевых сигналов. Как показано на рис. 9, эти способы могут быть разбиты на две большие группы - цифровое и параметрическое представление речевого колебания.

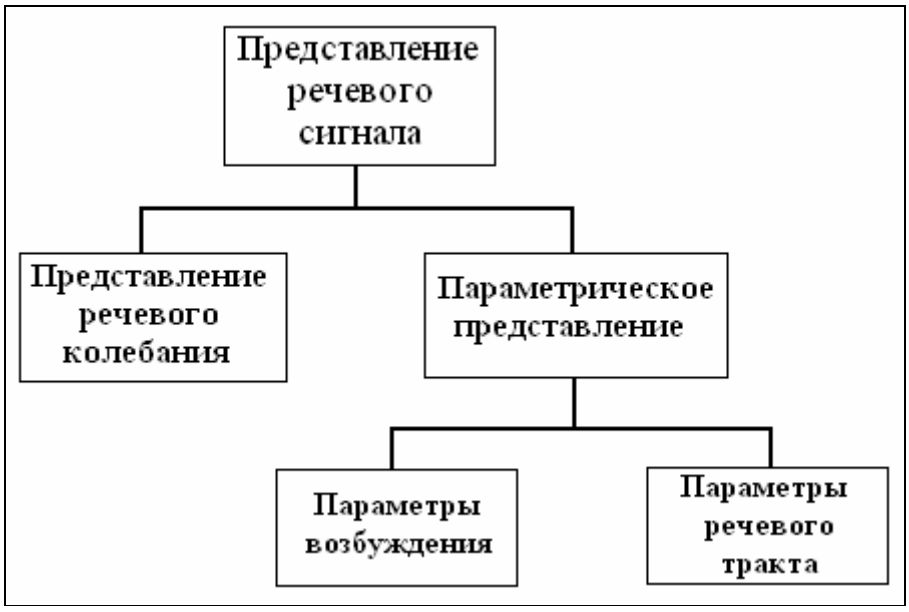

Рисунок 9 - Способы представления речевого сигнала

Цифровое представление речевого колебания, как это следует из названия, основано на сохранении формы колебания в процессе дискретизации и квантования. Параметрическое

представление базируется на описании речевого сигнала, как выходного отклика модели речеобразования. На первом этапе построения параметрического представления речевое колебание подвергается дискретизации и квантованию, а затем обрабатывается для получения параметров модели. Параметры модели обычно разделяются на параметры возбуждения (относящиеся к источнику звуков речи) и параметры голосового тракта (относящиеся непосредственно к отдельным звукам речи).

Наиболее важным фактором, определяющим выбор цифрового представления сигнала и методов цифровой обработки, является специфика решаемой прикладной задачи. На рис. 10 приведено несколько примеров из обширной области передачи и обработки речевых сигналов.

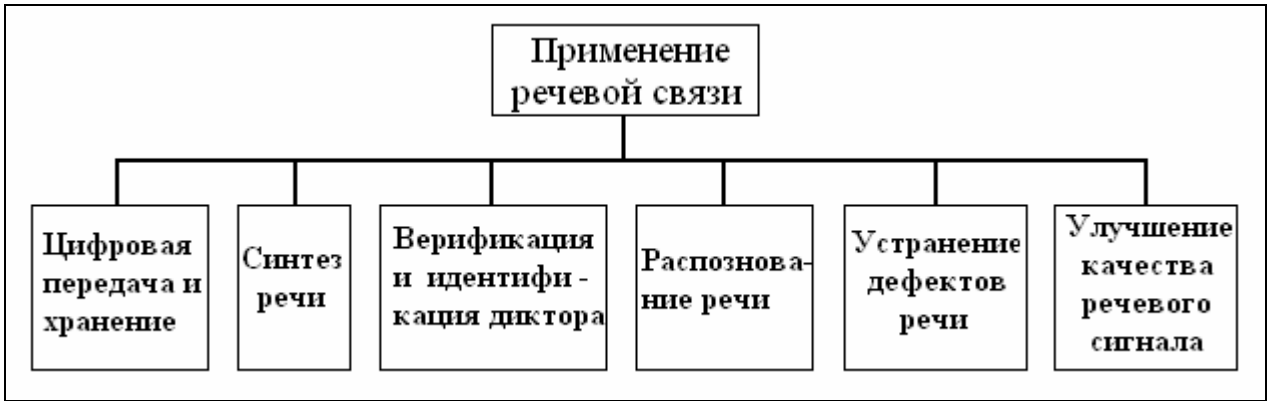

Рисунок 10 - Области применения речевой связи

Цифровая передача и хранение речевого сигнала. Одним из наиболее ранних и наиболее важных примеров применения обработки речевого сигнала является вокодер или кодер голоса (voice-coder), созданный Дадли в 1930-х гг. [6]. Целью разработки вокодера являлось уменьшение полосы частот, необходимой для передачи речи (Более точно, снижение требуемой пропускной способности канала связи при передаче речи). Эта задача актуальна и в настоящее время, несмотря на наличие широкополосных спутниковых, СВЧ ,и оптических систем связи. Кроме того, необходимы дешевые и как можно более низкоскоростные преобразователи речи в цифровую форму для их использования в цифровых телефонных сетях связи. Одной из положительных сторон применения цифровых систем является возможность обеспечения скрытности передачи.

Системы синтеза речи. Большой интерес к системам синтеза речи объясняется необходимостью разработки способа экономического хранения речевого сигнала в системах речевого ответа [6]. Подобная система реализует цифровой алгоритм автоматического сообщения голосом информации, которую запрашивает пользователь с клавиатуры пульта или специального терминала. Поскольку пультом может служить обычный телефонный аппарат с кнопочным набором, система речевого ответа может широко использоваться в коммутируемых телефонных сетях без установки какого-либо дополнительного оборудования [10]. Системы синтеза речи играют большую роль и при обучении правильному произношению речи [6].

Системы верификации и идентификации диктора. Методы верификации и идентификации диктора [6] включают установление подлинности, или идентификации, личности говорящего. Система верификации выносит решение о том, является ли говорящий тем, за кого он себя выдает. Системы такого типа применимы при управлении процессом доступа к информации или ограничении доступа, а также при проведении различного рода автоматических кредитных операций. Системы идентификации диктора должны выдать решение о том, кто из

ограниченного числа дикторов произнес данную фразу. Такие системы могут применяться в области судебной экспертизы.

Системы распознавания речи. В самом общем виде системы распознавания должны преобразовывать речевое сообщение в эквивалентный текст. Сложность задачи распознавания определяется условиями произнесения и контекстом произносимой фразы, а также наличием или отсутствием возможности настройки на диктора. Системы распознавания речи могут применяться в различных устройствах, например, пишущих машинках, управляемых голосом или при речевом общении с ЭВМ. Совместное использование систем распознавания и синтеза речи позволяет получить систему передачи речевого сигнала с минимально возможной скоростью передачи [6].

Устранение дефектов речи. Здесь предполагается обработка речевого сигнала и отображение полученной информации в виде, наиболее приемлемом для обучаемого индивидуума. Например, воспроизведение сигнала, записанного на магнитофонную ленту с различной скоростью, наиболее подходит для слепых, поскольку позволяет им прослушивать текст с любого желаемого места. Разработан также ряд методов цифровой обработки сигнала для сенсорного и визуального отображения информации при обучении глухих речи [6].

Улучшение качества речевого сигнала. В ряде случаев речевой сигнал, поступающий в систему связи, оказывается искаженным, что снижает качество передачи. В этом случае методы цифровой обработки могут быть использованы для улучшения качества восприятия сигнала. Примерами подобных разработок являются устранение реверберации (или эха), устранение шума в речевом сигнале, восстановление/речевого сигнала, записанного в гелиевокислородной среде, которая используется в качестве дыхательной смеси водолазами.

3.3 Сигналы в дискретном времени

С математической точки зрения акустическое колебание, формируемое в речевом тракте человека, можно описать функцией непрерывного времени *t.* Аналоговые (непрерывные во времени) сигналы будут обозначаться через $x_a(t)$. Речевой сигнал можно представить и последовательностью чисел. Последовательности обозначаются далее через x*(n).* Если последовательность чисел представляет собой последовательность мгновенных значений аналогового сигнала, взятых периодически с интервалом *T,* то эта операция дискретизации будет иногда обозначаться через $x_a(nT)$.

3.3.1 Теорема дискретизации

Для применения методов цифровой обработки к такому аналоговому сигналу, как речевое колебание, необходимо представить его в виде последовательности чисел. Обычно это осуществляется путём периодической дискретизации аналогового сигнала для получения последовательности его значений:

$$x(n) = x_a(nT), \qquad -\infty < n < +\infty \tag{1}$$

где n – принимает только целые значения.

Условия, которые должны выполняться для того, чтобы аналоговый сигнал можно было представить последовательностью своих отсчетов единственным образом, хорошо известны и часто формулируются в следующем виде:

Теорема дискретизации **(Котельникова)***:* если сигнал $x_a(t)$ имеет преобразование Фурье $X_a(i\Omega)$ такое, что $X_a(i\Omega)=0$ при $\Omega \geq 2\pi * F_N$, то $x_a(t)$ может быть восстановлен единственным образом по последовательности равноотстоящих отсчетов $x_a(nT)$, $-\infty < n < +\infty$, если $1/T > 2F_N$.

Таким образом, по последовательности отсчетов аналогового сигнала, взятых с частотой, равной, по крайней мере, удвоенной частоте Найквиста F_N, можно восстановить исходный аналоговый сигнал. Применяемые на практике цифровые преобразователи основаны на приближении соотношения:

$$x_a(t)=\sum_{n=-\infty}^{\infty} x_a(nT)\left[\frac{\sin(\pi(t-nT))/T}{\pi(t-nT)/T}\right] \qquad \textbf{(2)}$$

Дискретизация предполагается во многих алгоритмах обработки речевых сигналов, предназначенных для оценки таких важных параметров речи, как частоты формант или период основного тона. В этих случаях аналоговая функция, подвергаемая дискретизации, недоступна наблюдению. Однако параметры изменяются во времени медленно, и поэтому их можно оценивать со скоростью порядка 100 отсч./с (т.е. дискретизировать). Полученные отсчеты параметра являются значениями ограниченной по частоте функции, которую можно восстановить в соответствии с (2).

3.3.2 Прореживание и интерполяция дискретизированного сигнала

Иногда возникает задача изменения частоты дискретизации сигнала, представленного в дискретном времени. Процесс понижения и повышения частоты дискретизации называется *прореживанием* и *интерполяцией* соответственно. В обоих случаях предполагается, что имеется последовательность отсчетов $x(n) = x_a(nT)$, где аналоговая функция $x_a(t)$ имеет ограниченное по частоте преобразования Фурье, такое, что $X_a(i\Omega)=0$, $|\Omega|/2\pi > F_N$.

Прореживание. Пусть требуется понизить частоту дискретизации в M раз, т. е. необходимо построить новую последовательность, соответствующую отсчетам $x_a(t)$, взятым с периодом $T' = M*T$, т.е.:

$$y(n)=x_a(n*T')=x_a(n*T*M) \qquad \textbf{(3)}$$

Заметим, что:

$$y(n)=x(M*n), \qquad -\infty < n < +\infty \qquad \textbf{(4)}$$

Таким образом, $y(n)$ получается путем сохранения только одного из M отсчетов. Из теоремы дискретизации следует, что если $1/T' > 2F_N$, то $y(n)$ также единственным образом описывает исходный аналоговый сигнал. Преобразования Фурье $x(n)$ и $y(n)$ связаны соотношением [7]:

$$Y(e^{i\Omega T})=\frac{1}{M}\sum_{k=0}^{M-1} X\left(e^{i(\Omega T^1-2\pi k)/M}\right) \qquad \textbf{(5)}$$

Структурная схема обобщенной системы прореживания изображена на рис. 11. Фильтр низких частот необходим для того, чтобы не происходило наложение частот.

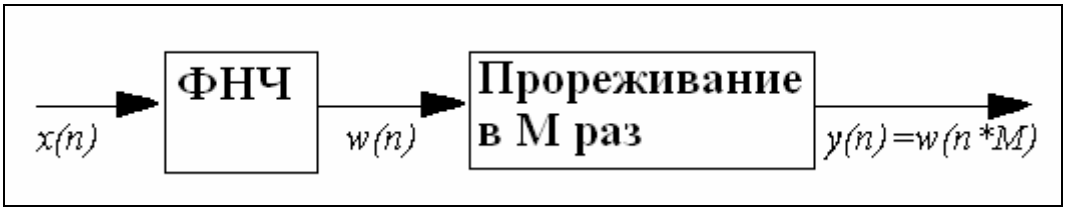

Рисунок 11 - Структурная схема прореживания

Интерполяция. Пусть имеется последовательность отсчетов аналогового сигнала $x(n) = x_a(nT)$. Если необходимо повысить частоту дискретизации в L раз, то следует вычислить новую последовательность, соответствующую отсчетам $x_a(t)$, взятым с периодом $T'=T/L$, т.е.:

$$y(n) = x_a\left(nT'\right) = x_a\left(nT/L\right) \qquad \textbf{(6)}$$

Очевидно, $y(n)=x(n/L)$ для n=0, ±L, ±2L, но для других значений недостающие отсчеты необходимо получить с использованием методов интерполяции [6].

Общая структурная схема процесса интерполяции представлена на рис. 12.

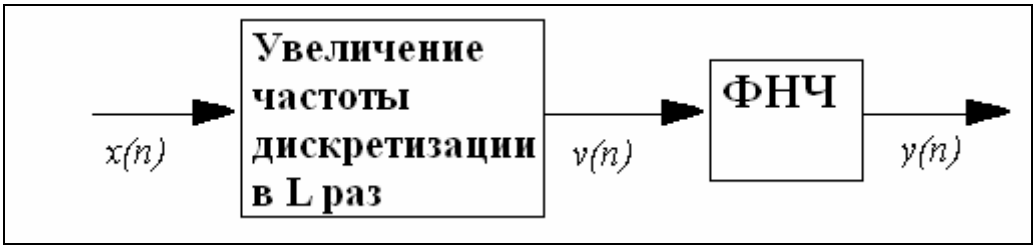

Рисунок 12 - Структурная схема интерполяции

Изменение частоты дискретизации в дробное число раз. Отсчеты, соответствующие периоду дискретизации $T'=MT/L$, можно получить путем комбинаций интерполяции с параметром L и последующей процедуры прореживания с параметром M. Соответствующим подбором целых чисел M и L можно получить любое, необходимое соотношение между частотами дискретизации. Объединив структурные схемы на рис. 11 и 12, легко заметить, что вместо двух достаточно иметь один фильтр нижних частот (рис. 13).

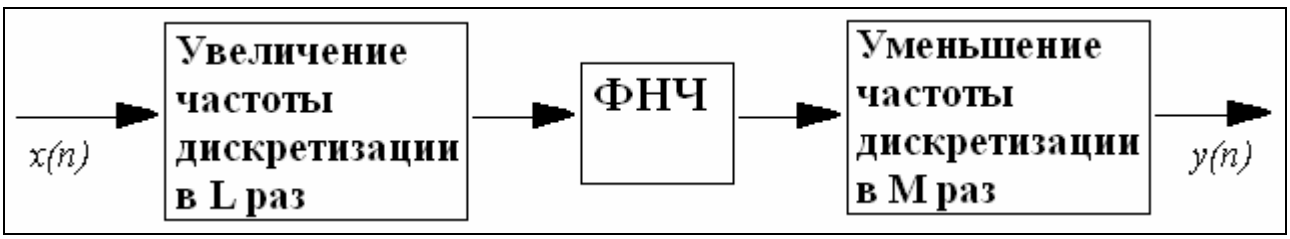

Рисунок 13 Структурная схема повышения частоты дискретизации

Важным аспектом при использовании методов интерполяции и прореживания является выбор фильтра нижних частот. Значительная экономия в объеме вычислений в таких системах достигается использованием фильтров в стандартной прямой форме. Экономия в вычислениях достигается вследствие того, что при прореживании только один из каждых *M* отсчетов подвергается фильтрации, а при интерполяции каждые *L*—1 из L отсчетов равны нулю и потому не влияют на процесс вычисления [6].

Если предположить, что фильтрация будет осуществлена с использованием фильтра нижних частот, то для большого изменения частоты дискретизации (т. е. большого *M* при прореживании и большого L при интерполяции) более целесообразным, оказывается, уменьшать (или увеличивать) частоту дискретизации с помощью серии последовательных прореживаний. В этом случае частота дискретизации уменьшается постепенно и на каждом шаге требуется фильтр нижних частот с менее крутым спадом частотной характеристики [6].

4. РАСПОЗНАВАНИЕ ФОНЕМ /И/ И /У/

Для реализации данной задачи был использован язык программирования высокого уровня VISUAL C++ 6.0. Для записи и редактирования фонем разных дикторов, а также для вычисления частот основного тона разных дикторов, использовалось программное обеспечение SOUND FORGE 5.0.

4.1 Структура спектра фонем /и/ и /у/

Система речеобразования описывается набором резонансов (формант), которые определяются в первую очередь функцией площади поперечного сечения голосового тракта. У фонемы /и/ первая форманта лежит в низкочастотной области спектра, приблизительно в районе 200-300 Гц, а вторая и третья уже за пределами 2000 Гц. А у фонемы /у/ первая и вторая форманты лежат в диапазоне до 600 Гц, а остальные форманты расположены за пределами 2000 Гц, их энергия не значительна.

Передаточная функция речевого тракта не имеет нулей, а имеет полюса, следовательно, амплитуда форманты связана с ее шириной и расположением окружающих ее формант. Амплитуда и ширина – это функции друг друга. Это означает, что, зная амплитуду, можно однозначно определить ширину и наоборот. У разных дикторов ширина основных формант сильно варьируются, но, тем не менее, находятся в определенных границах. Некоторый сдвиг резонансных частот (формант) возникает за счет потерь [6]:

1) Ширина низкочастотных формантных областей (первой и второй) зависит от потерь на стенках голосового тракта.
2) Ширина высокочастотных формантных областей зависит в первую очередь от потерь на вязкое трение, теплопроводность и излучение.

На рис. 14 представлен спектр фонемы /и/. Можно заметить, что первая форманта лежит в районе 230-250 Гц, вторая и третья – 2300 – 3000 Гц, а четвертая – 3900 Гц [5].

Как видно из рис. 14, энергия в низко- и высокочастотной области спектра довольно существенная.

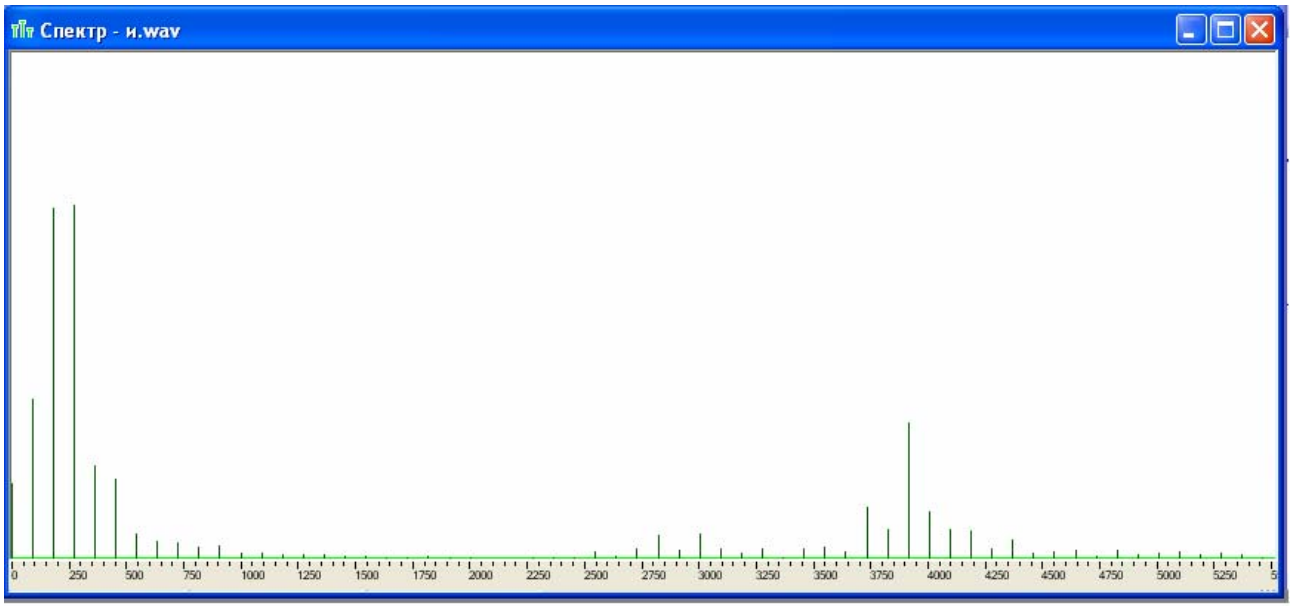

Рисунок 14 – Спектр фонемы /и/

А у фонемы /у/ первая форманта лежит примерно в области 230 Гц, вторая – 600 Гц, третья теоретически должны быть в пределах 2400 Гц, а четвертая – 3900 Гц [5]. Но, как видно из рис. 15, энергия третьей и четвертой формант очень не значительна, т. е. амплитуда спектра в низкочастотной области спектра довольно сильно преобладает над амплитудой в высокочастотной области спектра (рис.15).

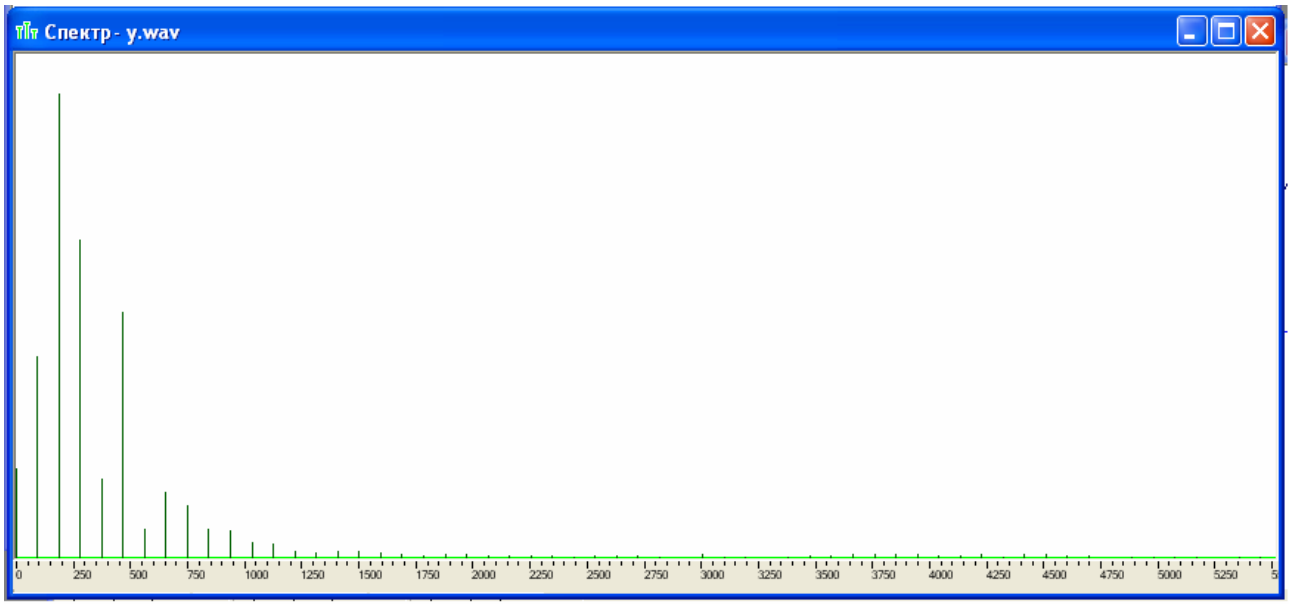

Рисунок 15 – Спектр фонемы /у/

Вследствие всего выше сказанного и того, что амплитуда у фонемы /и/ в высокочастотной области спектра существенно превышает амплитуду фонемы /у/ в той же области, была предложена следующая ***гипотеза*:**
отношение энергии в высокочастотной области спектра к энергии в низкочастотной области спектра фонемы /и/ должно быть больше такого же отношения энергий для фонемы /у/ у разных дикторов. Причем существует такой порог, что отношение энергий фонемы /и/ будет больше этого порога, а для /у/ - меньше.

4.2 Алгоритм распознавания фонем /и/ и /у/

Исходя из выше рассмотренной структуры спектра фонем /и/ и /у/ и предложенной выше гипотезы, предлагается следующий алгоритм по распознаванию рассмотренных фонем /и/ и /у/.

Алгоритм распознавания фонем /и/ и /у/:

1) На вход подается фонема либо /и/, либо /у/ в виде WAV-файла.
2) Загрузка файла в память компьютера – функция `bool MemoryAllotment_Loading(char *FileName)`, `FileName` – имя файла:
    a. Открыть файл на чтение, в случае неудачи выдается сообщение об ошибке открытия файла;
    b. Выделить память под файл;
    c. Загрузить файл в память;
3) Проверка загружаемого файла на принадлежность к классу WAV-файлов – функция `bool TestWaveFile()`. Структура WAV-файла приведена в Приложении 1:
    a. Проверяется заголовок файла, в случае неправильного заголовка выдается сообщение о неправильном формате файла;

b. Указатель (`pWaveFile`) передвигаем на начало области данных.

4) С помощью преобразования Фурье [7] получаем массив спектральных значений: `pSpecSamples[nSpecSamples]`, где `nSpecSamples` – количество спектральных отсчетов в массиве (1):

$$nSpecSamples = \left[nWaveSamples/2\right] + 1 \quad \textbf{(1)}$$

где `nWaveSamples` – количество отсчетов в волновой форме в массиве pWave`Samples`.

Это преобразование Фурье реализовывается с помощью функции, экспортируемой из библиотеки FOURIER.DLL (Приложение 2):

```
UniversalFastFourierTransformID(параметры);
```

5) Находим энергию сигнала в высоко- и низкочастотных областях. Энергия всего сигнала определяется по следующей формуле (2):

$$E_{1,\,nWaveSamples} = \sum_{i=1}^{nWaveSamples} \left(pWaveSamples[i]\right)^2 \quad \textbf{(2)}$$

Из равенства Парсеваля [4], сумма квадратов временных отсчетов равна сумме квадратов спектральных отсчетов, следует формула (3):

$$E_{1,\,nSpecSamples} = \sum_{i=1}^{nSpecSamples} \left(pSpecSamples[i]\right)^2 \quad \textbf{(3)}$$

6) Далее находим соотношение между энергией сигнала в высокочастотной области спектра и в низкочастотной области спектра.

7) Сравниваем получившееся значение с некоторым порогом. В случае если это значение больше порога, то по идее это должна быть фонема /и/, иначе фонема /у/.

4.3 Обучение алгоритма распознавания фонем /и/ и /у/

В представленном выше алгоритме неизвестными являются 2 параметра:

1) Частота разбиения спектра - f_0, относительно которой и будем разделять весь спектр на области высокой и низкой частоты.
2) Порог - D, с которым и будем сравнивать значение получившегося соотношения энергий в высоко- и низкочастотных областях входного сигнала.

Таким образом, задачей обучения алгоритма распознавания фонем /и/ и /у/ является нахождение этих параметров.

Проанализировав структуру спектра фонем /и/ и /у/, можно сделать вывод, что частоту разбиения спектра на область низких и высоких частот должна находиться примерно в диапазоне от 300 до 2000 Гц. Таким образом, зададим множество частот:

$$F=\{f_1, f_2, \ldots, f_K\}, \qquad \text{где } f_1 = 300 \text{ Гц}, \qquad f_K = 2000 \text{ Гц},$$
$$f_{i+1} = f_i + 50 \text{ Гц}, \qquad i = 1, \ldots, K \qquad \textbf{(4)}$$

Так как резонансные частоты (форманты) варьируются у разных дикторов, то для обучения возьмем несколько образцов фонем /и/ и /у/ у разных дикторов.

В результате, для каждой частоты разбиения спектра f_i находим отношение энергий по фонемам /и/ и /у/ для всех дикторов, участвующих в выборке (5):

$$yy[n] = \frac{E^{y}_{f_i, f_{end}}}{E^{y}_{0, f_i}} \qquad \textbf{(5)}$$

где *yy[n]* – массив отношений энергий фонемы /у/ для каждого диктора n = 1, 2, …, N – количество дикторов; $E^{y}_{f_i, f_{end}}$ – энергия сигнала фонемы /у/, вычисляемая по формуле (3);

f_i, f_{end} - частоты, в пределах которых надо вычислить энергию фонемы.

Исходя из анализа спектров различных дикторов с различной длиной речевого тракта от взрослых до детей, можно сделать вывод о том, что значимые форманты, т. е. энергия которых довольно существенна, лежат в пределах 4000 – 4500 Гц. Поскольку энергия остальных формант пренебрежимо мала по сравнению с первыми четырьмя формантами, то было принято решение игнорировать спектр выше 4500 Гц. Тем самым дополнительно перестрахуем себя от искажений сигнала, результатом которых будет появление шума в области частот выше 4500 Гц. Таким образом, в качестве конечной частоты возьмем частоту равную 4500 Гц, т. е. f_{end} = 4500 Гц.

Аналогично формуле (5), вычисляем отношение энергий сигнала для фонемы /и/ и результат помещаем в массив *ii[n]* , n =1,2, …, N.

Таким образом, для каждой частоты разбиения спектра f_i получаем ряд отношений энергий области высоких к области низких частот, являющийся вариационным рядом [3]. Для этого ряда были подсчитаны следующие числовые характеристики, приведенные в табл.1:

1) Максимальные значения соотношений энергий в области высоких и низких частот для фонемы /у/.
2) Минимальные значения соотношений энергий в области высоких и низких частот для фонемы /и/.
3) Среднее значения соотношений энергий для фонем /и/ и /у/ (6):

$$\overline{Mean} = \sum_{n=1}^{N} yy[n] \qquad \textbf{(6)}$$

4) Среднеквадратическое отклонение этих соотношений для фонем /и/ и /у/ (7):

$$s = \sqrt{\frac{\sum_{n=1}^{N} (yy[n] - \overline{Mean})}{N}} \qquad \textbf{(7)}$$

А также были подсчитаны некоторые другие статистики:

5) Разница между минимальным соотношением энергий по фонеме /и/ максимальным соотношением энергий фонемы /у/.
6) Разница между среднеквадратическими отклонениями соотношением энергий по фонеме /и/ и /у/.
7) Порог, вычисленный пропорционально среднеквадратическим отклонениям относительно среднего соотношений энергий фонем /и/ и /у/.
8) Порог, вычисленный пропорционально среднеквадратическим отклонениям относительно минимального соотношения энергий по фонеме /и/ и максимального соотношения энергий по фонеме /у/.

Согласно нашей гипотезе отношение энергий фонемы /и/ одного диктора должно быть больше отношения энергий фонемы /у/ у другого диктора (8):

$$\forall n, j = 1, 2, ..., N \quad ii[n] - yy[j] > 0 \tag{8}$$

Формулу (8) можно записать в следующем виде (9):

$$\underset{n=1,N}{Min}(ii[n]) - \underset{j=1,N}{Max}(yy[j]) > 0 \tag{9}$$

Таким образом, искомая оптимальная частота разбиения спектра f_0 будет равна той частоте f_i, при которой разница между средними значениями соотношения энергий по фонемам /и/ и /у/ будет принимать максимальное значение (10):

$$\underset{i=1,K}{Max}\left(\overline{Mean_u} - \overline{Mean_y}\right) > 0, \quad где\ K = \frac{2000 - 300}{50} = 34 \tag{10}$$

Порог D при частоте f_0:

1) Должен находиться в следующих пределах (11), иначе фонему какого-то диктора из нашей обучающей выборки он не будет распознавать.

$$\underset{j=1,N}{Max}(yy[j]) < D < \underset{n=1,N}{Min}(ii[n]) \tag{11}$$

2) Можно было бы просто взять середину (12). Но экспериментально было определено, что такой порог не подходит, так как не распознает фонемы некоторых дикторов.

$$D = \frac{\underset{n=1,N}{Min}(ii[n]) - \underset{j=1,N}{Max}(yy[j])}{2} \tag{12}$$

Поэтому было принято решение вычислить порог пропорционально среднеквадратическим отклонениям относительно среднего соотношений энергий фонем /и/ и /у/ (13):

$$\frac{S_u}{S_y} = \frac{\overline{Mean_u} - D}{D - \overline{Mean_y}} \quad \textbf{(13)}$$

Откуда порог D есть (14):

$$D = \frac{S_y * \overline{Mean_u} + S_u * \overline{Mean_y}}{S_y + S_u} \quad \textbf{(14)}$$

Результаты вычислений порога по формуле (14) приведены в предпоследнем столбце табл. 1. Как можно увидеть, эти пороги не удовлетворяют условию (11). В таблице об этом свидетельствует знак «-» после величины порога.

Тогда вычислили порог пропорционально среднеквадратическим отклонениям относительно минимального соотношения энергий по фонеме /и/ и максимального соотношения энергий по фонеме /у/ (15):

$$\frac{S_u}{S_y} = \frac{Min_u - D}{D - Max_y} \quad \textbf{(15)}$$

откуда порог D есть (16):

$$D = \frac{S_y * Min_u + S_u * Max_y}{S_y + S_u} \quad \textbf{(16)}$$

В результате получили, как видно из табл. 1, что порог, вычисленный по формуле (16) удовлетворяют условию (11). В таблице об этом свидетельствует знак «+» после величины порога. Практические результаты, полученные при экспериментальных исследованиях на разных дикторах, показали высокую эффективность дикторонезависимого распознавания фонем /и/ и /у/ при этом пороге.

В обучении программы распознавания фонем /и/ и /у/ участвовало 15 дикторов от детей до взрослых.

В результате проведенного обучения была вычислена оптимальная частота разделения спектра f_0 = 1050 Гц, а также был вычислен соответствующий этой частоте D = 0.00152236 (табл. 1). Именно эти значения являются параметрами алгоритма распознавания, относительно которых и будет происходить вычисление для распознавания фонем /и/ и /у/ уже разных дикторов.

Таблица 1 – Числовые характеристики соотношений энергий фонем /и/ и /у/

	Фонема -У-			Фонема -И-			И_min - У_max	И_сред - У_сред	Пороги	
Частоты	Среднее	СКО	MAX	Среднее	СКО	MIN			СКО, И_сред, У_сред (14)	СКО, Имin, Умax (16)
300	10.15086	34.18961	137.9758	234.2316	875.7774	0.011918	-137.9639	224.0807	18.57011 -	132.7922 -
350	0.926962	1.151921	3.975788	0.798116	2.377374	0.011522	-3.964266	-0.12884	0.884908 -	2.681898 -
400	0.515655	0.818282	3.348687	0.173424	0.210219	0.011171	-3.33751	-0.34223	0.243373 -	0.693338 -
450	0.348423	0.722803	3.002811	0.153753	0.205288	0.008635	-2.994176	-0.19467	0.196813 -	0.670929 -
500	0.082605	0.111167	0.469231	0.123496	0.163144	0.002606	-0.466625	0.04089	0.099177 -	0.280127 -
550	0.060861	0.063106	0.240927	0.122952	0.162907	0.002516	-0.238411	0.062091	0.078198 -	0.174360 -
600	0.042403	0.042213	0.152931	0.122776	0.162935	0.002509	-0.150422	0.080373	0.058941 -	0.121979 -
650	0.02697	0.027552	0.109861	0.122449	0.16299	0.002503	-0.107358	0.095479	0.040776 -	0.094337 -
700	0.017994	0.022573	0.080683	0.121977	0.162619	0.002464	-0.078218	0.103983	0.030668 -	0.071149 -
750	0.009642	0.018526	0.064724	0.120742	0.160589	0.001962	-0.062762	0.1111	0.021133 -	0.058232 -
800	0.003098	0.00505	0.01601	0.1206	0.160451	0.001958	-0.014051	0.117502	0.006683 -	0.015581 -
850	0.001497	0.001705	0.006889	0.120553	0.160456	0.001954	-0.004935	0.119056	0.002749 -	0.006837 -
900	0.000915	0.000831	0.003143	0.120492	0.160448	0.001945	-0.001198	0.119577	0.001531 -	0.003137 -
950	0.000664	0.000554	0.002432	0.120429	0.160384	0.001926	-0.000506	0.119764	0.001076 -	0.002430 -
1000	0.000604	0.00054	0.002347	0.120347	0.160255	0.001864	-0.000483	0.119743	0.001006 -	0.002345 -
1050	0.000519	0.000371	0.001522	0.120329	0.160249	0.001863	0.000341	0.119811	0.000795 -	0.001522 +
1100	0.000466	0.000291	0.000986	0.120214	0.160112	0.001862	0.000876	0.119747	0.000683 -	0.000988 +
1150	0.000431	0.000277	0.000876	0.120122	0.160006	0.001862	0.000985	0.119691	0.000638 -	0.000878 +
1200	0.00039	0.000254	0.000873	0.120101	0.159977	0.001859	0.000986	0.119711	0.000579 -	0.000875 +
1250	0.000374	0.000253	0.000856	0.120073	0.159924	0.001853	0.000998	0.119699	0.000563 -	0.000857 +
1300	0.000368	0.000252	0.000851	0.120065	0.159924	0.001853	0.001002	0.119697	0.000557 -	0.000853 +
1350	0.000361	0.00025	0.000847	0.120051	0.15992	0.001853	0.001006	0.11969	0.000547 -	0.000848 +
1400	0.000353	0.000249	0.000835	0.120031	0.15991	0.001852	0.001018	0.119677	0.000540 -	0.000836 +
1450	0.000348	0.000249	0.000833	0.120008	0.15988	0.00185	0.001017	0.11966	0.000534 -	0.000835 +
1500	0.000344	0.000246	0.000824	0.119972	0.159832	0.001847	0.001023	0.119628	0.000528 -	0.000825 +
1550	0.00034	0.000246	0.000822	0.119962	0.159827	0.001846	0.001025	0.119622	0.000524 -	0.000823 +
1600	0.000333	0.000243	0.000811	0.119946	0.159817	0.001846	0.001035	0.119613	0.000515 -	0.000812 +
1650	0.000326	0.00024	0.000777	0.119923	0.159789	0.001846	0.001069	0.119597	0.000506 -	0.000779 +
1700	0.000324	0.00024	0.000776	0.119906	0.159764	0.001845	0.001069	0.119582	0.000503 -	0.000778 +
1750	0.000321	0.000237	0.000755	0.11988	0.159729	0.001845	0.001089	0.119559	0.000498 -	0.000757 +
1800	0.000318	0.000236	0.00075	0.119859	0.159711	0.001844	0.001094	0.119541	0.000495 -	0.000751 +
1850	0.000316	0.000236	0.000749	0.119817	0.159675	0.001844	0.001095	0.119501	0.000493 -	0.000751 +
1900	0.000315	0.000236	0.000748	0.119792	0.159657	0.001844	0.001095	0.119477	0.000491 -	0.000750 +
1950	0.000312	0.000236	0.000748	0.119755	0.159645	0.001842	0.001094	0.119443	0.000489 -	0.000750 +
2000	0.00031	0.000236	0.000748	0.119683	0.159644	0.001841	0.001094	0.119373	0.000486 -	0.000749 +

4.4 Практические результаты

С целью исследования алгоритма был проведен эксперимент по распознаванию фонем /у/ и /и/. В эксперименте участвовало 30 дикторов и ошибок при распознавании не было обнаружено (предполагалось отсутствие искажений и помех в сигнале).

ЗАКЛЮЧЕНИЕ

Был разработан алгоритм распознавания фонем /и/ и /у/.

Был разработан алгоритм обучения распознаванию.

В обучении программы распознавания фонем /и/ и /у/ участвовало 15 дикторов от детей до взрослых.

При практическом распознавании приняло участие 30 дикторов, не участвовавших в обучении программы.

Ошибок при распознавании не было обнаружено (предполагалось отсутствие искажений и помех в сигнале).

.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Винцюк Т. К. Анализ, распознавание и интерпретация речевых сигналов.- Киев: Наукова думка, 1987.- 264с.

2. Гоноровский И. С. Радиотехнические цепи и сигналы.- М: Радио и связь, 1986.- 512с.

3. Кремер Н. Ш. Теория вероятностей и математическая статистика.- М: Юнити-дана, 2001.- 543с.

4. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов.- М: Связь, 1979.- 416с.

5. Потапова Р. К. Тайны современного Кентавра // Речевое взаимодействие «человек-машина».- М., 1992.

6. Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов: пер. с англ. / Под ред. М. В. Назарова, Ю. Н. Прохорова.- М: Радио и связь, 1981.- 496с.

7. Толстов Г.П. Ряды Фурье / 3-е изд.- М.: Наука, 1980.- 384с.

8. Фант Г. Акустическая теория речеобразования: Пер. с англ.- М.: Наука, 1964.- 284с.

9. Фролов А.В, Фролов Г.В Мультимедиа для Windows.- М: Диалог-МИФИ, 1995.- 234с.

10. Фланаган Д. Л. Анализ, синтез и восприятие речи: пер. с англ. / Под ред. А. А. Пирогова.- М.: Связь, 1968.- 396с.

11. Языковые процессоры и распознавание речи.- Тбилиси: Мецниереба,1985.-105с.

ПРИЛОЖЕНИЕ 1 Формат WAV-файлов

Данные, имеющие отношение к мультимедиа (звук, видео и т. п.), хранятся в файлах в так называемом RIFF-формате (Resource Interchange File Format - формат файла для обмена ресурсами). Как wav-файлы, содержащие звук, так и avi-файлы, содержащие видеоинформацию, имеют формат RIFF.

Файл в формате RIFF содержит вложенные фрагменты (chunk's). Внешний фрагмент состоит из заголовка и области данных (Рис.1):

Рисунок 1 – Общий формат WAV-файла

DWORD	DWORD	Целое число байтов
"RIFF"	Размер	Данные

Первое двойное слово заголовка содержит четырехбуквенный код FOURCC, который идентифицирует данные, хранящиеся во фрагменте. Второе двойное слово заголовка - размер области данных в байтах (без учета размера самого заголовка).

Область данных имеет переменную длину, однако она должна быть выровнена на границу слова и при необходимости дополнена в конце нулевым байтом до целого числа слов.

Область, обозначенная как "Данные", может содержать внутри себя другие фрагменты (Рис.2). Для файла, в котором хранятся звуковые данные (wav-файл), эта область содержит идентификатор данных "WAVE", фрагмент формата звуковых данных "fmt " (три символа "fmt" и пробел на конце), а также фрагмент звуковых данных. Файл может дополнительно содержать фрагменты других типов.

Рисунок 2 – Расширенный формат WAV-файла

DWORD	DWORD							
"RIFF"	Размер	Данные						
		"WAVE"	"fmt "	Размер	Формат данных	Фрагмент "data"		
						"data"	Размер	Звуковые данные

Область, обозначенная как "Формат данных", описывает звуковые данные. Формат этой области для файлов PCM (записанных с использованием импульсно-кодовой модуляции) соответствует структуре PCMWAVEFORMAT, определенной в файле mmsystem.h следующим бразом:

```
typedef struct pcinwaveformat_tag {
   WAVEFORMAT wf;
   WORD vBitsPerSample;
} PCMWAVEFORMAT;

typedef PCMWAVEFORMAT *PPCMWAVEFORMAT;
```

Структура WAVEFORMAT также описана в файле mmsystem.h:

```
typedef struct waveformat_tag {
   WORD wFormatTag;          // тип формата
   WORD nChannels;           // количество каналов (моно или стерео)
   DWORD nSamplesPerSec;     // частота дискретизации
   DWORD nAvgBytesPerSec;    // скорость потока данных
   WORD nBlockAlign;         // выравнивание блока данных
} WAVEFORMAT;

typedef WAVEFORMAT *PWAVEFORMAT;
```

Поле wFormatTag описывает тип формата звуковых данных. Для импульсно-кодовой модуляции PCM, которая поддерживается стандартной библиотекой mmsystem.dll, в этом поле должно находиться значение WAVE_FORMAT_PCM, определенное в файле mmsystem.h: #define WAVE_FORMAT_PCM 1

Поле nChannels содержит количество каналов. В нем могут находиться значение 1 (моно) или 2 (стерео).

В поле nSamplesPerSec записана частота дискретизации, то есть количество выборок сигнала в секунду. В этом поле могут находиться стандартные значения (11025 кГц, 22 050 кГц или 44100 кГц) либо нестандартные значения, такие, как 5000 кГц или 4400 кГц. Учтите, что не все драйверы звуковых адаптеров могут работать с нестандартными частотами дискретизации.

Поле nAvgBytesPerSec содержит среднюю скорость потока данных, то есть количество байт в секунду, передаваемых драйверу устройства или получаемых от него. Эта информация может быть использована приложением для оценки размера буфера, необходимого для размещения звуковых данных. Для монофонического сигнала с дискретностью 8 бит численное значение скорости совпадает со значением частоты дискретизации. Для стереофонического сигнала с дискретностью 8 бит она в два раза выше. Точное значение можно подсчитать по формуле (1):

nAvgBytesPerSec = (nChannels * nSamplesPerSec * wBitsPerSanipIe) / 8 (1)

В поле nBlockAlign находится выравнивание блока в байтах, которое подсчитывается по формуле (2):

nBlockAlign = (nChannels * wBitsPerSample) / 8 (2)

Поле wBitsPerSainple находится в структуре PCMWAVEFORMAT и содержит дискретность сигнала, то есть количество бит, используемых для представления одной выборки сигнала. Обычно используются значение 8 или 16.

Что же касается формата самих звуковых данных, то он зависит от количества каналов и от дискретности. Для монофонического сигнала с дискретностью 8 бит звуковые данные представляют собой массив однобайтовых значений, каждое из которых является выборкой сигнала. Для стереофонического сигнала с дискретностью 8 бит звуковые данных имеют формат массива двухбайтовых слов, причем младший байт слова соответствует левому каналу, а старший - правому.

Диапазон изменения значений выборок сигнала определяется дискретизацией. Для 8-битовых данных он составляет от 0 до 255 (Oxff), причем отсутствию сигнала (полной тишине) соответствует значение 128 (0x80). Для 16-битовых данных диапазон изменения составляет от -32768 (-0x8000) до 32767, (Ox7fff), отсутствию сигнала соответствует значение 0.

ПРИЛОЖЕНИЕ 2 Описание функций, экспортируемых библиотекой Fourier.dll

```
int __stdcall UniversalFastFourierTransformID(short* pWaveSamples,
              unsigned nWaveSamples,unsigned nFirstFragSample,
              unsigned nFragSamples, double* pSpecSamples,
              unsigned nSpecSamples, unsigned nInterpolPoints);
```

Указатель на функцию имеет тип PUNIVERSALFASTFOURIERTRANSFORMID.

pWaveSamples – Указатель на массив отсчетов волновой формы.
nWaveSamples - Количество отсчетов волновой формы.
nFirstFragSample - Номер отсчета волновой формы, который является первым отсчетом фрагмента, спектр которого надо найти.
nFragSamples - Количество отсчетов в фрагменте, спектр которого надо найти.
pSpecSamples - Указатель на буфер, в который будет помещены спектральные отсчеты.
nSpecSamples - Количество спектральных отсчетов, которое надо найти.

Функция вычисляет дискретное преобразование Фурье (ДПФ). Если количество отсчетов в фрагменте, ДПФ которого надо найти, равно степени двойки, то ДПФ вычисляется с помощью алгоритма быстрого преобразования Фурье (БПФ). Если количество отсчетов в фрагменте, ДПФ которого надо найти, не равно степени двойки, этот фрагмент передискретизуется так, чтобы число отсчетов в нем стало равно степени двойки, и к передискретизованному фрагменту применяется БПФ. Передискретизация осуществляется с помощью функции ResampleSignal.

На массив отсчетов волновой формы указывает параметр pWaveSamples, количество отсчетов волновой формы задается параметром nWaveSamples. Спектральные отсчеты помещаются в буфер, на который указывает параметр pSpecSamples. Количество спектральных отсчетов, которое надо найти, задается параметром nSpecSamples.

Количество спектральных отсчетов, задаваемое параметром nSpecSamples должно быть меньше либо равно значению параметра nWaveSamples.

Один спектральный отсчет состоит из двух 64-битовых вещественных чисел с двойной точностью:

1) первое число - модуль амплитуды гармоники в диапазоне от 0 до 32768;
2) второе число - начальная фаза гармоники (функция sin).

ВНИМАНИЕ! Если количество отсчетов волновой формы четно, тогда амплитуда гармоники с номером nWaveSamples/2 делится пополам.

Функция возвращает:

- не нуль, если дискретное преобразование Фурье было успешно выполнено;
- нуль, если не хватило памяти.

Последние два символа в имени функций вида *FourierTransform?? обозначают тип входных и выходных отсчетов: I - целочисленные знаковые 16-битовые отсчеты, D - вещественные отсчеты двойной точности. Первый символ из двух описывает тип входных отсчетов, второй - выходных.

ПРИЛОЖЕНИЕ 3 Руководство пользователя

При запуске программы – MFCDialog2.exe возникает диалоговое окно вида (рис. 16):

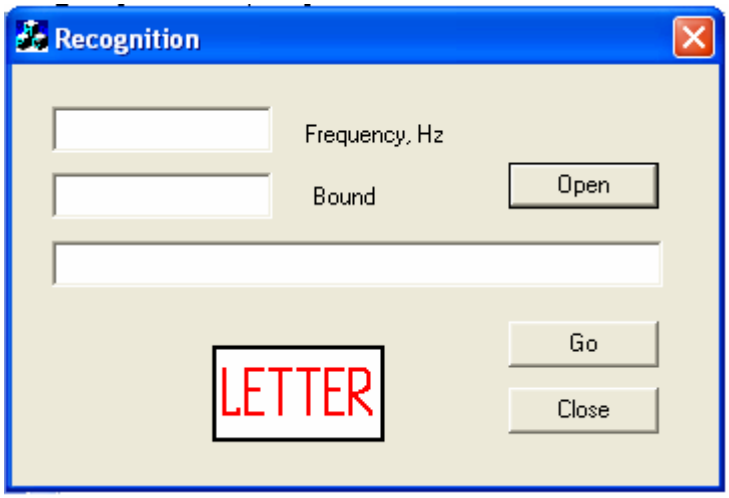

Рисунок 16 – Диалоговое окно Recognition

В поле “Bound” граница, с которой сравнивается отношение энергии в высокочастотной области к энергии в низкочастотной области. Именно эта частота и указывается разбиения спектра в поле “Frequency, Hz”.

При нажатии кнопки “Open” появляется диалоговое окно (Рис. 17), в котором необходимо выбрать WAV-файл для распознавания.

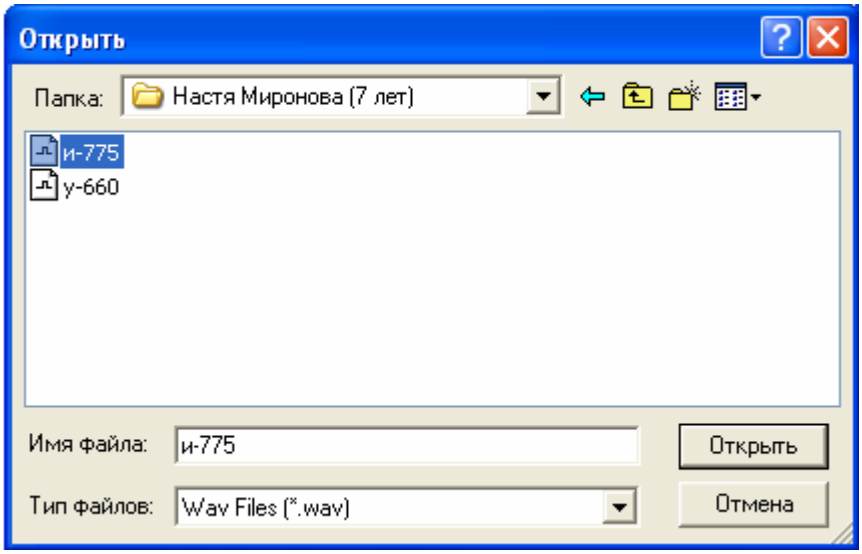

Рисунок 17 – Диалоговое окно Open

При нажатии на кнопку «Go» программа распознавания запускается. И в результате в поле “LETTER” появляется распознанная фонема.