Министерство образования и науки Республики Казахстан

Академия логистики и транспорта

Кафедра «Информационно-коммуникационные технологии»

Отчет по преддипломной практике

на тему: «Анализ многоуровневой обработки речевых сигналов при наличии шумов»

Выполнила
студентка гр. РЭТ-19-2 **А.А.Сакенова**

Научный руководитель,
PhD, ассоц.проф.АЛТ **А.М.Достиярова**

Алматы, 2023

Министерство образования и науки Республики Казахстан

Академия логистики и транспорта

Сакенова Алия Алтаевна

Анализ многоуровневой обработки речевых сигналов при наличии шумов

ДИПЛОМНАЯ РАБОТА

Образовательная программа: 6B06209-Радиотехника, электроника и телекоммуникации

Алматы, 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Речевой сигнал можно рассматривать с помощью модели, в которой речевой сигнал является откликом системы с медленно изменяющимися параметрами на периодическое или шумовое возбуждающее колебание. По существу речеобразующий механизм (голосовой тракт) является акустической трубкой, возбуждаемой соответствующим источником при создании желаемого звука. Для звонких звуков источнику возбуждения соответствует квазипериодическая последовательность импульсов, представляющая поток воздуха, протекающий через колеблющиеся голосовые связки. Речевой сигнал можно промоделировать откликом линейной системы с переменными параметрами (голосового тракта) на соответствующий возбуждающий сигнал. При неизменной форме голосового тракта выходной сигнал равен свертке возбуждающего сигнала и импульсного отклика голосового тракта. Однако все разнообразие звуков получается путем изменения формы голосового тракта. Частотная характеристика голосового тракта является гладкой функцией частоты; поскольку голосовой тракт представляет собой полость, то в первую очередь он характеризуется акустическими резонансами, соответствующими резонансным частотам этой полости, которые обычно называются формантными частотами. Спектр речевого сигнала образуется перемножением линейчатого спектра возбуждающего сигнала и спектра, соответствующего голосовому тракту, и, следовательно, тоже является линейчатым, а его огибающая характеризует передаточную функцию голосового тракта. Поскольку при создании различных звуков форма голосового тракта изменяется, огибающая спектра речевого сигнала будет конечно тоже изменяться с течением времени. Аналогично при изменении периода сигнала, возбуждающего звонкие звуки.

Способы представления речевых сигналов: от простейшей периодической дискретизации речевого сигнала до оценок параметров модели. Выбор того или иного способа представления речевого сигнала определяется решаемой задачей, которые разделяются на три класса. К первому классу относят задачи, связанные с анализом речи. Анализ речи является неотъемлемой частью систем распознавания речевых сигналов, а также систем идентификации дикторов по голосу. Ко второму классу относят задачи, связанные с синтезом речи по тексту. Задачи такого типа возникают в многочисленных информационно-справочных системах. В задачах, относящихся к третьему классу, выполняется как анализ системы сжатия речевых сигналов с целью передачи речи по компьютерным сетям или по традиционным линиям связи. Одним из перспективных направлений применения обработки речевых сигналов являются системы распознавания речи в сети Интернет. В этом случае пользователь сети, используя телефон, может соединиться с программой распознавания речи, находящейся на сервере и транслирующей диалог в команды Веб-сервера. Это позволяет получить доступ к распределенным информационным ресурсам сети по телефону.

.

1. Классификация существующих методов обработки применяемых в системах распознавания речи

В настоящее время существуют многочисленные технические средства, могущие воспринимать (распознавать) произносимые речевые сообщения: компьютеры, медицинское электронное оборудование, автомобили, мобильные телефоны и др. Что такое распознавание речи? На первый взгляд, все кажется очень просто: человек произносит слово (фразу), а техническая система адекватно реагирует на него: либо выполняет команду, содержащуюся в слове (фразе), либо набирает диктуемый текст, либо как-то иначе “распоряжается” извлеченной из фразы информацией. Бурное развитие распознавания речи с помощью персонального компьютера началось с 1993 г. Две ключевых задачи распознавания речи – достижение 100 % распознавания на ограниченном наборе команд хотя бы для одного диктора и независимое от диктора распознавание непрерывного речевого потока в реальном масштабе времени произвольного языка с приемлемым качеством – не решены, несмотря на многочисленные попытки решения этих задач в течение последних 50-ти лет. Современные системы распознавания речи уже дают возможность пользователям диктовать слова (фразы) в обычной разговорной манере. Однако процесс непрерывного распознавания речи, дающий до 95 % качества распознавания при оптимальных условиях, все-таки дает на 100 знаков 5 ошибок. Около 200 ошибок на странице формата А4 – слишком много для профессиональной работы. При рассмотрении классификации систем распознавания речи следует отметить, что классификация может осуществляться по различным параметрам. Системы распознавания речи можно классифицировать следующим образом: в зависимости от размера словаря: системы распознавания речи с ограниченным набором слов; системы со словарем большого размера; в зависимости от привязки к диктору: системы, являющиеся дикторо-зависимыми и дикторо-независимыми; в зависимости от типа распознаваемой речи: системы, работающие со слитной речью или раздельной речью.

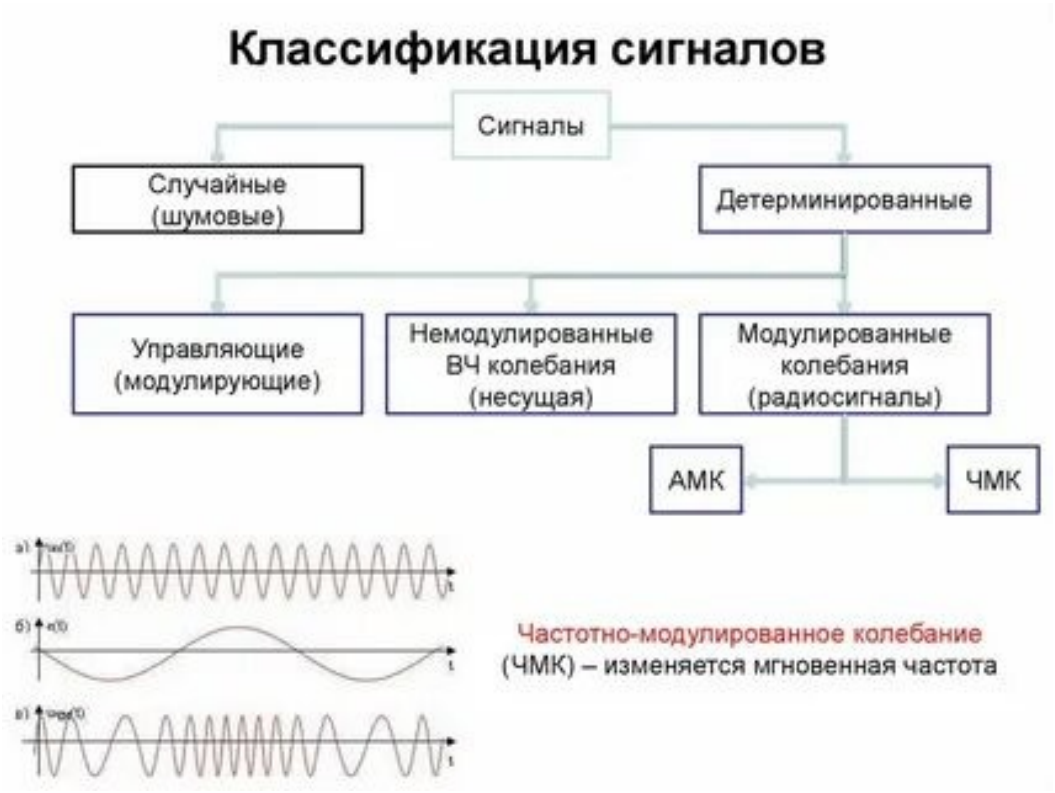


Рисунок 2.1 - Схема ввода речевых сообщений в ЭВМ

Речевой сигнал формируется и передается в пространстве в виде звуковых волн. Источником речевого сигнала служит речеобразующий тракт, который возбуждает звуковые волны в упругой воздушной среде. Приемником сигнала является датчик звуковых колебаний, микрофон - устройство для преобразования звуковых колебаний в электрические. Существует большое количество типов микрофонов (угольные, электродинамические, электростатические, пьезоэлектрические и др.) описанных в специальной литературе. Чувствительным элементом микрофона любого типа является упругая мембрана, которая вовлекается в колебательный процесс под воздействием звуковых волн.

1.1 Детектирование голосовой активности

Детектор голосовой активности - это критически важный компонент системы распознавания речи, который значительно влияет на её точность и производительность. При правильном выделении участков звука, в которых присутствует целевой голосовой сигнал, детектор наличия голоса значительно уменьшает объём данных для обработки системой распознавания речи, что в итоге ускоряет её работу и уменьшает вероятность ложных распознаваний. Особенно это свойство важно при работе систем распознавания речи на персональных устройствах пользователей – смартфонах, ноутбуках, смарт-телевизорах, у которых в отличии от специализированных серверов ограничена процессорная мощность. На базе раннее разработанного авторами линейного 8-ми канального массива MEMS микрофонов с PDM интерфейсом [15] был спроектирован в SolidWorks и изготовлен прототип устройства для аудиовизуального детектирования голосовой активности, представленный на рис. 1. Прототип состоит из трёх печатных плат:

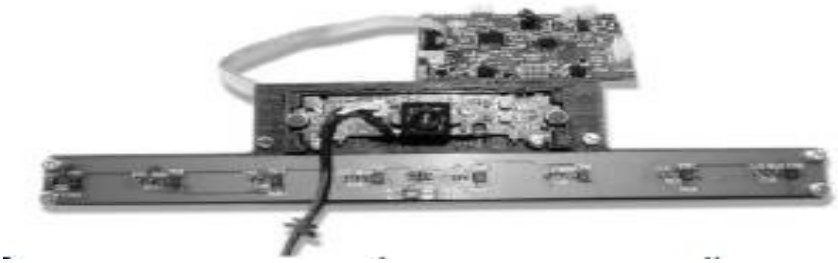


Рис.1. Версия массива микрофонов со встроенной видеокамерой

Прототип по USB шине подключается к компьютеру с видеокартой с поддержкой технологии CUDA, на котором производятся все вычисления. Метод детектирования голосовой активности показал высокое качество разметки голосового сигнала, которое при соблюдении нахождения говорящего человека в поле видимости видеокамеры превышает результаты, показываемые детекторами голоса, использующими только звуковую информацию.

1.2 Оценка качества и разборчивости речи

Применение современных методов компьютерной оценки разборчивости речи является очень полезным при работе звукорежиссера в студиях и театрально-концертных залах (особенно если в них установлена система звукоусиления), а также при оценке качества речевых сигналов при передаче по каналам радиовещания, телефонии, в системах перевода речей и пр. опыт проектирования залов различного назначения (аудиторий, лекционных залов, кинозалов, театральных залов и др.) и результаты многочисленных исследований показали, что разборчивость речи в помещении определяют следующие акустические характеристики:

Уровень прямого речевого сигнала во всех точках зала;

Уровень внешних и внутренних шумов;

Время реверберации;

Структура, уровень и направление прихода отраженных сигналов.
Влияние реверберационного процесса на структуру речевого сигнала можно отчетливо увидеть на примере осциллограмм, записанных в заглушенной камере и в помещении с большим значением реверберации (рис.1). Естественно, что при таком существенном изменении временной структуры речевого сигнала процесс его распознавания существенно ухудшается.

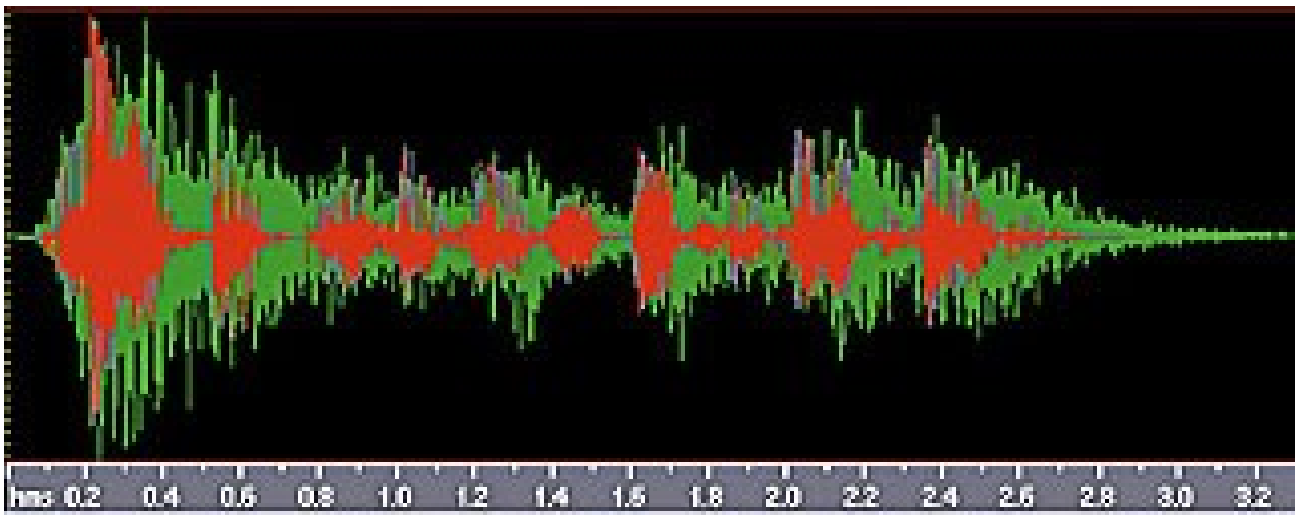


Рис.1 Осциллограмма речевого сигнала в заглушенной камере и в помещении

1.3 Подавление шума

Подавление шума – это то, что подразумевает конструкция и материалы наушников и гарнитур. По сути, это уровень изоляции от внешнего шума, который может обеспечить устройство само по себе, без учета электронных компонентов и алгоритмов. Проще говоря, это то, насколько хорошо наушники справляются с функцией берушей. Шумоподавление достигается путем использования аналоговых или цифровых фильтров и различается по типам реализации — шумоподавление с обратной связью, без обратной связи, а также гибридное. Качественная технология активного подавления шума значительно улучшает акустические характеристики наушников и гарнитур с хорошим пассивным шумоподавлением.

Типичный пример гарнитуры с активным шумоподавлением

1.4 Распознавание речевых сигналов

Точность распознавания зависит от модели распознавания. Также на точность распознавания влияют:

1. Качество исходного звука;

2. Качество кодирования аудио;

3. Разборчивость и темп речи;

4. Сложность фраз и их длина.

Модель распознавания — модель, которая обучена распознавать речь на определенном языке. Для обучения моделей используются массивы данных из сервисов и приложений Яндекса. Это позволяет постоянно улучшать качество распознавания.

Основная поддерживаемая модель для каждого типа распознавания — модель general. Она распознает речь на любую тему на заданном языке: короткие и длинные фразы, а также имена, адреса, даты и числа. Одним из самых больших преимуществ решений распознавания речи Open Source Python является то, что это открытый исходный код. Большинство речевых сигналов представляют собой нестационарные процессы с множеством компонентов, которые могут изменяться по времени и частоте. Классические стационарные методы не в состоянии точно представить эти вариации, в то время как представления позволяют более точно описать нестационарные сигналы. В звуковом речевом сигнале есть две полезные акустические характеристики: основная частота (высота тона) и форманта. Основная частота обычно является самой низкочастотной составляющей сигнала; она представляет собой частоту вибрации голосовых связок во время воспроизведения звука. Форманта - это концентрация акустической энергии вокруг определенной частоты в речевой волне; каждая форманта соответствует резонансу в голосовом тракте. Основные и формантные частоты, представленные основными пиками в спектре, передают важную информацию о речи. На самом деле, большой объем фонетической информации передается соответствующими частями озвученных речевых сигналов. Соответственно, обнаружение и отслеживание формант важны для выделения особенностей речи и распознавания ее эволюционного поведения. Представление озвученного речевого сигнала огибающей амплитуды форманты и мгновенной частотой является богатым, поскольку оно раскрывает как спектральную структуру, так и информацию о времени возбуждения различных формантных диапазонов. Хотя в речевых сигналах в принципе существует бесконечное число формант, четырех формант достаточно для представления характеристик голосового тракта. Следовательно, в большинстве приложений нет необходимости обнаруживать и отслеживать все частотные компоненты речи.

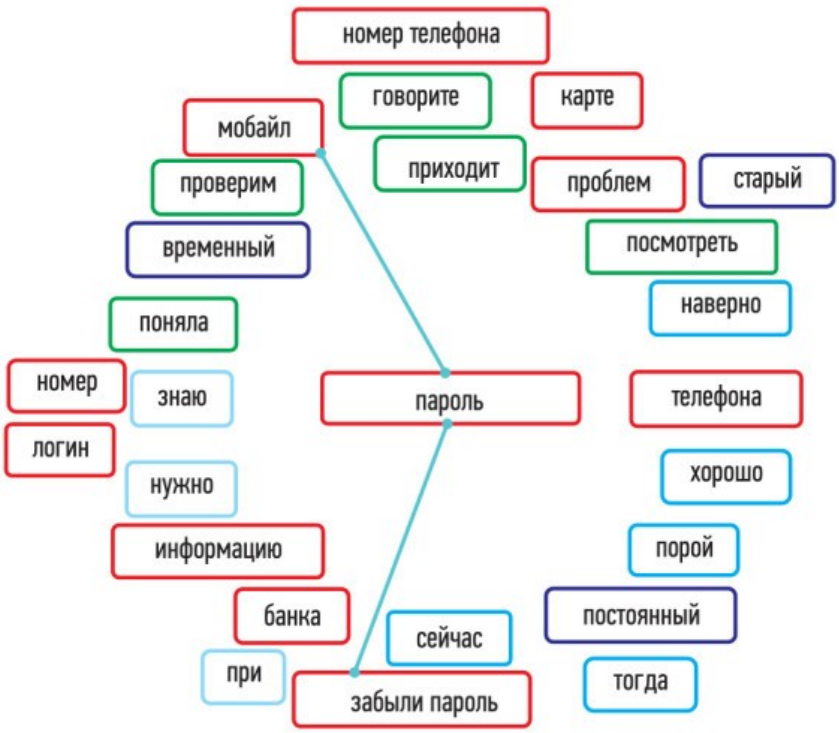
Предварительная обработка речевого сигнала

Преобразование речевого аудио в формат данных, используемый системой ML, является начальным этапом процесса распознавания говорящего. Начните с записи речи с помощью микрофона и преобразования аудиосигнала в цифровые данные с помощью аналого-цифрового преобразователя. Дальнейшая обработка сигнала обычно включает в себя такие процессы, как обнаружение голосовой активности (VAD), подавление шума и выделение признаков. Мы рассмотрим каждый из этих процессов позже. Во-первых, давайте рассмотрим некоторые ключевые методы предварительной обработки речевого сигнала: масштабирование функций и преобразование стерео в моно. Поскольку диапазон значений сигнала сильно варьируется, некоторые алгоритмы машинного обучения не могут должным образом распознавать звук без нормализации. Масштабирование объектов - это метод, используемый для нормализации диапазона независимых переменных или признаков данных. Масштабирование данных устраняет разреженность, приводя все ваши значения к одному масштабу, следуя той же концепции, что и нормализация и стандартизация. Например, вы можете стандартизировать свои аудиоданные с помощью sk. Количество каналов в аудиофайле также может влиять на производительность вашей системы распознавания громкоговорителей. Аудиофайлы могут быть записаны в моно- или стереоформате: моно-аудио имеет только один канал, в то время как стереозвук имеет два или более каналов. Преобразование стереозаписей в моно помогает повысить точность и производительность системы распознавания громкоговорителей. Python предоставляет модуль pydub, который позволяет воспроизводить, разделять, объединять и редактировать аудиофайлы WAV. Вот как вы можете использовать его для преобразования стереофонического WAV-файла в монофонический файл.

Анализ и обработка речи

К технологиям анализа и обработки речи относят быстрый поиск ключевых слов в аудиозаписях, автоматический анализ и оценку телефонных переговоров, интеллектуальный анализ речевой информации. Даннная технология отличается простотой использования и точностью поиска в фонограммах, которая определяется поисковым словарем. Так, для словаря из пяти слов надежность поиска составляет не менее 95%, для словаря из 100 слов — 81%. Интеллектуальный анализ речевой информации позволяет автоматически определять тематику телефонных переговоров. В основе анализа лежат технологии распознавания слитной речи. В результате автоматического распознавания речь дикторов преобразуется в текстовый индексированный файл, пригодный для автоматического лексико-семантического анализа. Решение о принадлежности аудиозаписи к абстрактному тематическому кластеру проводится с учетом частотности и связности слов и словосочетаний, употребляемых дикторами в ходе телефонной беседы (рис. 1).

Рис. 1. Пример семантического облака темы «Восстановление пароля»



Информацию, содержащуюся в речевом сигнале, можно разделить на основную, заключающуюся в передаче смыслового содержания речи, а также дополнительную, которая включает в себя информацию о характеристиках передающей среды. Характеристики передающей среды обычно включают уровень и тип окружающего шума (офисные шумы, уличные шумы, фоновая музыка, голоса других людей и т.д.), шум и искажения в канал передачи (микрофоны, усилители, АЦП, кодеки и т.д. Характеристик передающей среды помогает решать задачи очистки от шума и улучшения качества речевых сигналов, а также оценивать их пригодность для последующего использования в системах автоматического распознавания речи и голоса. Так, например, точность большинства систем автоматического распознавания речи и голоса резко ухудшается при снижении отношения сигнал-шум менее 15 дБ, увеличении уровня реверберации более 0,4 с. Речевые сигналы с «пригодными» параметрами характерны, в основном, для каналов телефонной связи. Речевые сигналы в акустике помещений имеют значительно худшие параметры, что приводит к низкой точности распознавания речи и голоса на таких данных. Кроме того, выполняется оценка качества речевого сигнала для оценки его пригодности для распознавания речи и голоса.

Глава 2. Разработка оптимального алгоритма очистки речевого сигнала, методы, основанные на вычитании амплитудных спектров

2.1 Методы, основанные на оценке спектральных характеристик шума

2.2 Динамическое шумоподавление

2.3 Определения эффективности очистки речевых сигналов разработанным алгоритмом

2.4 Разработка схемы работы программы

2.1 Методы, основанные на оценке спектральных характеристик шума

Звуковой сигнал, записываемый в реальных акустических условиях, часто содержит нежелательные шумы, которые могут порождаться окружающей средой или звукозаписывающей аппаратурой. Один из классов шумов - аддитивные стационарные шумы. Аддитивность означает, что шум суммируется с "чистым" сигналом и не зависит от него, сигнал , в этом случае определяется выражением. Стационарность означает, что свойства шума (мощность, спектральный состав) не меняются во времени. Примерами таких шумов могут являться постоянное шипение микрофона или усилительной аппаратуры, гул электросети. Работа различных приборов, не меняющих звучания по времени (вентиляторы, компьютеры) также может создавать шумы, близкие к стационарным. Не являются стационарными шумами различные щелчки, удары, шелест ветра, шум автомобилей. Для подавления аддитивных стационарных шумов существует алгоритм спектрального вычитания. Он состоит из следующих стадий:

1. Разложение сигнала с помощью быстрого преобразования Фурье или другого преобразования, компактно локализующего энергию сигнала.

2. Оценка спектра шума.

3. "Вычитание" амплитудного спектра шума из амплитудного спектра сигнала.

4. Обратное преобразование - синтез результирующего сигнала.

2.2 Динамическое шумоподавление

Применение данного метода обусловлено тем, что одновременно с подавлением шумов происходит выделение так называемых образцов шума, которые используют на следующем этапе для спектрального вычитания. На втором этапе используются методы, основание на различных модификациях алгоритма вычитания амплитудного спектра. Такой подход оптимален в случае широкополосных непрерывных и импульсно-непрерывных помех, пересекающихся с областью спектра речи. Шумы данного типа не могут быть удалены другими методами (например, адаптивной фильтрацией), поскольку такие помехи являются рассредоточенными по спектру и пересекаются с областью спектра речи. Пусть - спектр зашумленного сигнала на р-м фрейме, - спектр шума, - спектр восстановленного сигнала на р-м фрейме. На практике шум вычисляется на шумовых фреймах сигнала . Это связано с тем, что обычно известен только зашумленный сигнал. Следует отметить, что обрабатывается весь спектр сигнала, а не только речевой диапазон. Выявленные участки относительной тишины сохраняются и используются в дальнейшем шумопонижении. Входной сигнал должен иметь нормированный уровень, для чего в микрофонном усилителе используется автоматическая регулировка усиления, что также позволяет наиболее эффективно использовать АЦП, задействовав всю его разрядность.

2.3 Определения эффективности очистки речевых сигналов разработанным алгоритмом

Для определения эффективности очистки речевых сигналов от шумов помех и искажений, необходимо провести ряд исследований нацеленных на получение качественных результатов, по итогам которых выносится оценка качества и эффективности исследуемого программного или программно-аппаратного средства.Исследования проводились в соответствии с ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости» [10], а также ГОСТ Р 51061-97 «Системы низкоскоростной передачи речи по цифровым каналам. Параметры качества речи и методы измерений» [11]. Стандарты регламентируют получение комплексной оценки качества передачи речи, основанной на методах измерения показателей разборчивости, качества и узнаваемости речи. Разборчивость речи можно определить через относительное количество (в процентах) правильно принятых элементов (слогов, слов, фраз) артикуляционных таблиц. Узнаваемость голоса диктора представляет собой величину, характеризующую степень сохранения субъективно воспринимаемых индивидуальных признаков голоса диктора. В рамках исследований измерения разборчивости речи и узнаваемости голоса диктора проводились методами:

· артикуляционных измерений;

· парных сравнений.