ЯРОСЛАВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. П.Г. ДЕМИДОВА

На правах рукописи

Сагациян Максим Владимирович

РАЗРАБОТКА И ИССЛЕДОВАНИЕ КОЛЛЕКТИВНЫХ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ДИКТОРОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

Специальность: 05.12.04 Радиотехника, в том числе системы и устройства телевидения

ДИССЕРТАЦИЯ

на соискание учёной степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Брюханов Юрий Александрович

Владимир – 2015

ОГЛАВЛЕНИЕ

СПИСОК СОКРАЩЕНИЙ

БПФ – быстрое преобразование Фурье

ДПФ – дискретное преобразование Фурье

ИНС – искусственная нейронная сеть

ОБПФ – обратное быстрое преобразование Фурье

ОСШ – отношение сигнал/шум

РС – речевой сигнал

СКО – средняя сумма квадратов ошибки

СММ – скрытые марковские модели

ЭВМ – электронная вычислительная машина

ЭОР – эмоционально окрашенная речь

ЭС – эмоциональное состояние

GDX – Gradient Descent Backpropagation with Adaptive Learning Rate

HMM – Hidden Markov Modeling

IBM-PostSNR – Ideal Binary Mask – A Posteriori Signal-to-Noise Ratio

IBM-TSNR – Ideal Binary Mask – Two-Step Noise Reduction

LFPC – Log Frequency Power Coefficients

LMA – Levenberg – Marquardt Algorithm

LOG – logarithm of the spectrum

LPC – Linear Predictive Codes

LPCC – Linear Predictive Cepstral Coefficients

MFCC – Mel Frequency Cepstral Coefficients

PLP – Perceptual Linear Prediction

SCG – Scaled Conjugate Gradient Backpropagation

SNR – Signal-to-Noise Ratio

TEO – Teager Energy Operator

TSNR – Two-Step Noise Reduction

Wiener-PriorSNR – Wiener – A Priori Signal-to-Noise Ratio

ВВЕДЕНИЕ

Актуальность темы и состояние вопроса

В настоящее время вопросы проектирования и создания системы распознавания речевых сигналов, устойчивых к шумам, с низкой частотой появления ошибок, являются актуальной проблемой. Коммерческие программы управления радиотехническими устройствами посредством речевых сигналов появились в начале девяностых годов прошлого века. Они востребованы людьми с ограниченными возможностями, которым из-за травмы руки сложно набирать большое количество текста. Также данные технологии востребованы людьми, у которых по какой либо причине заняты руки. Например, пожарному при чрезвычайной ситуации легче с помощью голоса воспользоваться радиотехническим устройством, чем с помощью рук. Данные программы основаны на обработке сигналов, то есть переводят голос пользователя в текст, таким образом снимая нагрузку с его руки.

Применение технологий распознавания речевых сигналов актуально в области управления радиотехническими устройствами, такими как, например: радиоприемником, рацией, телевизионным устройством, мобильным телефоном, сканером магнитно-резонансной томографии, рентгеновским сканером и др.

В настоящее время растет важность массового внедрения новых интерфейсов взаимодействия человека с радиотехническими системами, поскольку традиционные интерфейсы во многом уже достигли своего совершенства, а вместе с ним и своих пределов [42]. При традиционно высокой значимости информации, поступающей к нам через органы зрения, и ее высокой доли среди всей сенсорной информации, считающейся равной порядка 85% [58], данный канал восприятия человека становится в значительной степени перегружен. И первоочередной

альтернативой здесь видится коммуникация именно по акустическому каналу. Следовательно, в настоящее время технологии распознавания речевых сигналов актуальны не только для людей с ограниченными возможностями, но и для большинства, активно пользующихся техникой, людей. Знания, полученные при исследованиях машинного распознавания речи, в настоящее время являются актуальными и востребованными общественностью.

Интерес к изучению распознавания речевых сигналов нашел свое отражение в многочисленных исследования российских и зарубежных авторов. Для решения данной задачи в настоящее время применяют методы, основанные на искусственных нейронных сетях и скрытых Марковских моделях. Существенный вклад в развитие данных методов внесли труды Дж.К. Бейкера, Л.Е. Баума, Б.Т. Лоуэрра, Л.Р. Липорака, Б. Жуаня, С.Е. Левинсона, Л. Рабинера, Е.К. Левина и др.

Наибольший интерес состоит в создании алгоритма автоматического дикторонезависимого распознавания речевых сигналов ориентированного на большой словарь и дающего высокую точность распознавания даже в присутствии различных шумов. Для отечественного рынка (рынка Российской Федерации) также востребована возможность работы таких систем с русской речью. Такую задачу можно решить, создав интеллектуальный человекоподобный алгоритм аналогичный акустической системе человека. На сегодняшний день наиболее приближены к данной системе искусственные нейронные сети (ИНС) [5]. Проблема создания систем автоматического распознавания речи на основе ИНС изучается с 70-х годов, но из-за низких вычислительных мощностей, больших успехов не достигала. С увеличение вычислительных мощностей ЭВМ возникает среда, в которой можно создавать и тестировать алгоритмы с большой вычислительной сложностью. На сегодняшний день таких мощностей становится достаточно, чтобы с высокой точностью решить поставленную задачу.

Анализируя работы ученных по созданию и исследованию систем дикторонезависимого распознаванию речевых сигналов, можно отметить, что на настоящий момент достигнута вероятность дикторонезависимого распознавания речевых сигналов для малого словаря 93 % и для большого словаря 90,41 % [101]. Также стоит отметить, что данные результаты получены не для русскоязычных речевых сигналов. Следовательно, создание и исследование систем дикторонезависимого распознавания русскоязычных речевых сигналов является весьма актуальной задачей.

Задача распознавания речевых сигналов является частью задачи распознавания слитной речи. Анализируя работы ученых L. Breiman, Lawrence R. Rabiner, Y.T. Chen, S. Furui, W. Siging [62, 63, 65, 89, 98] и спрос современного рынка, можно установить, что для управления радиотехническими устройствами при помощи речевых сигналов система автоматического распознавания речи должна отвечать следующим требованиям:

– возможность работы в режиме реального времени;

– высокое качество распознавания;

– дикторонезависимость;

– возможность работы с русской речью;

– устойчивость к внешним шумам.

Последнее требование связано с тем, что для повышения надежности распознавания речевых сигналов требуется построить систему, не зависящую от внешних шумов для применимости алгоритма в различных условиях.

Существующие методы распознавания речевых сигналов не отвечают абсолютно всем заявленным требованиям. Данное обстоятельство определяет актуальность исследований в данном направлении.

Направление диссертационной работы соответствует области исследований:

1. Разработка методов приема, обработки, отображения и хранения информации. То есть в диссертационной работе исследуется разработка методов приема, обработки, отображения и хранения информации дикторонезависимого распознавания русскоязычных речевых сигналов в радиотехнических устройствах.

2. Разработка перспективных информационных технологий, в том числе цифровых в радиотехнических устройствах. То есть с помощью систем обработки сигналов, выполняющих дикторонезависимое распознавание русскоязычных речевых сигналов, возможно повысить эффективность радиотехнических устройств, таких как, например, радиоприемника, рации, мобильного телефона, телевизионного устройства, сканера магнитно-резонансной томографии, рентгеновского сканера и др.

Целью работы является разработка и исследование результативного алгоритма дикторонезависимого распознавания речевых сигналов для управления радиотехническими системами на базе математического аппарата искусственных нейронных сетей с устойчивостью к внешним шумам.

В соответствии с указанной целью в работе поставлены и решены следующие **задачи**:

1. Анализ существующих моделей, методов и алгоритмов распознавания речевых сигналов с целью выявления степени их соответствия современным требованиям и выбора прототипов для собственных исследований и создания модифицированного алгоритма.

2. Разработка моделей и алгоритмов распознавания речи, обеспечивающих достижение следующих показателей распознавания речевых сигналов:

– скорость работы, достаточная для использования в режиме реального времени;

– высокая вероятность дикторонезависимого распознавания речевых сигналов (для малого словаря не менее 93 % и для большого словаря не менее 90,41 %) [101];

– возможность работы с русской речью;

– устойчивость к шумам без большой потери вероятности распознавания.

3. Программная реализация в среде MatLAB предлагаемых алгоритмов и проведение экспериментальных исследований, подтверждающих их результативность.

Объектом исследования являются системы автоматического дикторонезависимого распознавания речевых сигналов.

Предметом исследования являются модели и алгоритмы распознавания речевых сигналов на основе искусственных нейронных сетей.

При написании работы в методологическом плане применялась следующая **совокупность методов исследования**: теории вероятностей; теории случайных процессов; математического анализа и аналитической геометрии; цифровой обработки сигналов; дискретного преобразования Фурье; теории нейронных сетей и теории программирования.

Научная новизна

Впервые получены следующие научные результаты:

1. Разработан нейросетевой алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением масштабируемых сопряженных градиентов (Scaled Conjugate Gradient Backpropagation, SCG), позволяющий решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов для малого словаря с вероятностью распознавания 97,1 %, что на 4,1 процентных пункта выше существующих результатов.

2. Предложена модификация коллективного нейросетевого алгоритма, позволяющая результативно решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов.

3. Разработан модифицированный коллективный нейросетевой алгоритм на основе персептронов Розенблатта с обучением SCG, позволяющий решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов для большого словаря с вероятностью распознавания 95,7 %, что на 5,29 процентных пункта выше существующих результатов.

4. Разработан коллективный и модифицированный коллективный нейросетевые алгоритмы с блоками шумоподавления для задачи дикторонезависимого распознавания русскоязычных речевых сигналов, работающие в условиях шумов.

Практическая значимость

1. Предложенная модификация коллективного нейросетевого алгоритма расширяет возможности нейросетевых алгоритмов в задаче дикторонезависимого распознавания русскоязычных речевых сигналов.

2. Вероятность распознавания речевых сигналов для разработанного нейросетевого алгоритма bagging-коллектива на основе персептронов Розенблатта с обучением SCG с блоком шумоподавления для малого словаря в интервале от 5 до 20 дБ равняется 93,5 % при использовании алгоритма шумоподавления Скалара на основе винеровской фильтрации.

3. Вероятность распознавания речевых сигналов для разработанного модифицированного нейросетевого алгоритма bagging-коллектива на основе персептронов Розенблатта с обучением SCG с блоком шумоподавления для большого словаря в интервале от 15 до 20 дБ равняется 93,6 % при использовании алгоритма шумоподавления на основе бинарных масок, использующего критерий статистического детектирования на основе апостериорного отношения сигнал/шум.

4. Разработана программа «NN-SCG speech recognition» (свидетельство о государственной регистрации программы для ЭВМ № 2015616920), с помощью которой проведен анализ различных алгоритмов нейросетевого дикторонезависимого распознавания русскоязычных речевых сигналов.

Результаты работы внедрены в соответствующие разработки ООО «ПАНТЕОН» (г. Ярославль) и ООО «А-Вижн» (г. Ярославль). Все результаты внедрения подтверждены соответствующими актами (приложение № 4).

Достоверность материалов диссертационной работы подтверждена согласованностью результатов математического моделирования разработанных алгоритмов и экспериментальной проверки в условиях полунатурного моделирования на реальных речевых сигналах, апробацией в печати и на научно-практических конференциях различного уровня.

Апробация работы. Результаты работы докладывались и обсуждались на следующих конференциях:

– 14-й и 15-й Международной конференции «Цифровая обработка сигналов и её применение», Москва, 2012-2013;

– Международной конференции «Системы синхронизации, формирования и обработки сигналов в инфокоммуникациях», Ярославль, 2013;

– 11-й и 12-й Международных научно-технических конференциях «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации», Курск, 2013, 2015;

– Международной конференции «Перспективные технологии в средствах передачи информации», Владимир, 2013;

– Международной конференции студентов и аспирантов «Путь в науку», Ярославль, 2014-2015;

– 53-й Международной научной студенческой конференции МНСК-2015, Новосибирск, 2015;

– XIII Всероссийской научной конференции «Нейрокомпьютеры и их применение», Москва, 2015.

Публикации. По теме диссертации опубликовано 17 научных работ, из них 3 статьи в журналах, рекомендованных ВАК для публикации результатов кандидатских и докторских диссертаций [31, 38, 40], 14 докладов на научных конференциях [16, 32-37, 39, 48-52]. Получено свидетельство о регистрации программы для ЭВМ [41].

Личный вклад автора. Выносимые на защиту положения предложены и реализованы автором самостоятельно в ходе выполнения научно-исследовательских работ на кафедре динамики электронных систем Ярославского государственного университета им. П.Г. Демидова.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы и четырех приложений. Содержание работы изложено на 134 страницах. Список литературы включает 104 наименований. В работе представлено 28 рисунков и 9 таблиц.

В первой главе освещаются актуальные задачи машинного распознавания речи. Выносится проблема не совершенства существующих алгоритмов дикторонезависимого распознавания русскоязычных речевых сигналов. Приводится описание существующих моделей направленных на решение поставленной задачи. Проведены сравнительный анализ акустических признаков звуков речи и оценка степени их применимости для решения задачи распознавания речи. Рассмотрены основы принципов построения ИНС и алгоритмы коллективного нейросетевого распознавания образов. Ставятся задачи, которые необходимо решить в ходе выполнения работы.

Во второй главе представлена модификация коллективного нейросетевого алгоритма для задачи дикторонезависимого распознавания

русскоязычных речевых сигналов. В качестве алгоритма обучения представлен алгоритм обучения нейронных сетей SCG, который ранее не применялся для коллективных нейронных сетей. Проведено пять серий экспериментов по исследованию: размера bagging-коллектива; количества обучающих дикторов; количества слоев для нейросетевого алгоритма bagging-коллектива; размера словаря для коллективных нейросетевых алгоритмов; работы модифицированного алгоритма нейросетевого распознавания.

В третьей главе проведен анализ работы нейросетевых алгоритмов обучения в задаче дикторонезависимого распознавания русскоязычных речевых сигналов. Исследовано три коллективных нейросетевых алгоритма, основанных на разных алгоритмах обучения: bagging-коллектив 12-слойных персептронов на основе обучения Левенберга-Марквардта; bagging-коллектив 12-слойных сетей Эльмана на основе обучения GDX и bagging-коллектив 12-слойных персептронов на основе обучения SCG.

В четвертой главе проведен анализ работы нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых сигналов в условиях шумов. В данной главе исследованы коллективный и модифицированный коллективный нейросетевые алгоритмы распознавания речевых сигналов с блоками предобработки. Использовалось три алгоритма шумоподавления: IBM-PostSNR; IBM-TSNR и Wiener-PriorSNR.

В заключении подводятся итоги выполнения работы и указываются возможные сферы внедрения полученных результатов.

Основные научные положения и результаты, выносимые на защиту:

1. Алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением SCG для решения задачи дикторонезависимого распознавания русскоязычных речевых сигналов.

2. Модификация коллективного нейросетевого алгоритма, позволяющая решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов для большего размера словаря.

3. Результаты исследования работы коллективных и модифицированных коллективных нейросетевых алгоритмов с блоком шумоподавления для решения задачи дикторонезависимого распознавания русскоязычных речевых сигналов в условиях шумов.

Благодарности. Автор выражает искреннюю признательность своему научному руководителю – д.т.н., профессору Ю.А. Брюханову, а также д.т.н., доценту А.Л. Приорову. Особая благодарность к.т.н. А.И. Топникову за постоянную поддержку в формировании взглядов в научном направлении диссертационной работы. Также автор благодарен коллегам-аспирантам за интересные научные дискуссии и ценные советы.

Отдельная благодарность родным и близким за терпение и предоставленную возможность заниматься научной деятельностью.

ГЛАВА 1. ОБЗОР ИСТОЧНИКОВ И ВЫБОР НАПРАВЛЕНИЯ ИССЛЕДОВАНИЯ

В течение последних 50-55 лет постепенно развилось научное направление создания новых интерфейсов между человеком и электронной вычислительной машиной (ЭВМ). В качестве одного из таких интерфейсов может выступать человеческая речь. Современные исследования в данной области ставят перед собой цель создания речевого интерфейса, позволяющего понимать и воспринимать человеческую речь, причем делать это так, чтобы общение между ЭВМ и человеком было трудно отличимым от общения человека с человеком, то есть, чтобы человек не мог бы даже догадаться, что его собеседник – ЭВМ. Такая система может лишь быть упрощенным функциональным подобием «живого» прототипа, перед ней стоит задача только в воспроизведении и трансформации информации, осуществляемой в «живой» интеллектуальной системе; однако не обязательно интерфейс между человеком и ЭВМ должен повторять конкретную конструкцию «живой» системы [54]. Под понятием «живой» системы подразумевается биологическая система обычного человека, которая умеет воспринимать и понимать человеческую речь.

При процессе разработки модели неизбежно приходится пользоваться рядом упрощений, потому что реальная система восприятия и понимания речи человека является достаточно сложной и трудновоспроизводимой. Некоторые упрощения очевидны, другие являются спорными. В спорных упрощениях желательно обращаться к «живой» системе для их проверки.

В процессе создания модели может возникнуть несколько вариантов решения поставленной задачи. Если разрабатываемая система имеет большую вычислительную сложность, то зачастую проверить, насколько результативна та или иная модель, очень сложно, а иногда практически

нереально до тех пор, пока система не будет полностью спроектирована. В таком случае целесообразнее обратиться к исследованиям «живой» системы с целью понимания, какой из имеющихся вариантов больше согласуется с полученными экспериментальными фактами.

В технологии разработки системы распознавания речи можно провести аналогию между теорией и экспериментальными фактами. Например, люди, профессионально занимающиеся лечением «живых» систем, такие как физиологи и психологи, обычно привыкли считать, что для построения какой-либо теории нужно собрать как можно больше фактов и попытаться дать некоторое обобщение описанным фактам. В данном случае, проектирование системы распознавания речи как будто поставлено наоборот – первостепенным является проектировщик системы, эксперименты в данном случае нужны лишь для ограничения его фантазии. В действительности, конечно, при создании подобных систем экспериментальные факты ограничивают фантазию проектировщика. При создании систем распознавания речи начинать работу нужно с фактов, обобщенных более или менее формализованной теорией. Полученные факты будут больше относиться не к психологии и физиологии, а к другим теориям – акустике и лингвистике.

Теория процессов распознавания и функциональная модель восприятия и понимания речи являются одним и тем же для исследований направленных на создание систем распознавания речи. Следовательно, целью исследований технического, психологического и физиологического изучения является разработка теории или, что то же самое, уточнение структуры и определение параметров этой функциональной модели.

Для того чтобы определить круг вопросов, рассматриваемых в диссертации, нужно коротко остановиться на том, какова, по распространенным сейчас представлениям, общая структура полной модели восприятия и понимания речи [54].

Понимается, что вся система состоит из трех последовательно соединенных моделей. Первая из них, она обычно называется моделью восприятия, производит трансформацию поступающего на вход данной системы акустического речевого сигнала в последовательность фонетических элементов. В данную модель входят блок (блоки) слухового анализа речевого сигнала и блок фонетической интерпретации. Информация о языке, содержащаяся в блоке фонетической интерпретации, еще очень ограничена и касается фонетики языка. То есть, модель может переводить воспринятый ею акустический речевой сигнал в артикуляторные инструкции – указания о том, как нужно произнести то, что модель «услышала». В данном случае модель не знает ни словарного состава языка, ни его грамматики и, тем более, не «понимает» смысла услышанного. Вторая модель производит последовательность фонетических элементов в описание смысла фразы. Она выполняет морфологический анализ и синтаксический анализ, используя для этого словарь (словари) и грамматические правила. Другими словами, это действующая модель анализирующей части данного языка. Описание смысла, получаемое на выходе системы, является описанием тех сведений о «действительности», которые содержались в проанализированной фразе. Третья модель занимается интерпретацией и оценкой полученных сведений о событиях, явлениях и так далее. Она решает, являются эти сведения истинными или ложными, важными или безразличными, что нужно предпринять в результате их получения и т.д. Иначе говоря, модель решает какую-то часть из того, что обозначается как интеллектуальная деятельность. Разработка данной системы в настоящее время добилась серьезных успехов, но имеется ряд не доработанных задач [54, 61, 65].

Исходя из характера задач, решаемых указанными моделями, можно проследить, что их проектированием занимаются специалисты совершенно разного направления, то есть различные модели относятся к компетенции разных направлений науки.

Их хода предыдущих суждений можно сделать предположение, что данные модели можно выполнить последовательно. Но в данном случае возникает ряд проблем. То есть при последовательности действий, при которых первая модель не получит никакой информации с выходов второй и третьей моделей, а вторая модель ничего не знает о том, что производит третья модель, возникает вопрос корректного объединения данных моделей. Решение вопроса объединения моделей заключается в согласовании выхода модели предыдущего уровня с входом модели следующего уровня. Можно пойти путем задания описания последовательности фонетических элементов (какая информация будет в ней содержаться и как она будет представлена) и отображения смысла.

Ясно, что разработка функциональных моделей требует обязательного четкого определения того, что является сигналом на входе и что необходимо получить на выходе. При рассмотрении вопроса о стыковке моделей, естественно, приходится исходить, с одной стороны, из того, какое входное описание необходимо для модели следующего уровня, и, с другой стороны, какое описание реально можно получить на выходе модели предыдущего уровня [46, 54].

В настоящей диссертации рассматриваются экспериментальные данные и теоретические вопросы, касающиеся только третьей из этих трех моделей, определяемой как модель распознавания.

В идеальной ситуации система распознавания речи состоит из двух частей. Данные части можно неявно выделить в самостоятельные блоки или подпрограммы. Какая-нибудь из них может существовать в упрощенном виде, но в любой реализации всегда присутствуют обе части. В литературе можно встретить разные вариации названия данных составных частей. Другими словами можно сказать, что любая система распознавания речи состоит из акустического и лингвистического блоков. Последний блок, впрочем, лингвистическим назван не строго. В общем случае он может включать в себя синтаксическую, фонетическую,

фонологическую, семантическую, морфологическую и лексическую модели языка. Или, другими словами, представят собой упрощенный корреляционный блок. Акустический блок отвечает за представление речевого сигнала. То есть за его трансформацию из временной области в другую форму, в которой в более явном виде присутствует информация о содержании речевого сигнала. Лингвистический блок интерпретирует информацию, которую получает от акустического блока, и отвечает за представление результата распознавания речи потребителю (в роли потребителя может выступать как человек, так и ЭВМ, управляемая речевыми сигналами).

В настоящей диссертации предполагается рассмотреть и исследовать акустическую часть системы распознавания речи. Данные исследования могут дать возможность создать системы распознавания речевых сигналов, которые будут являться в настоящий момент конкурентоспособными по скорости и качеству распознавания относительно существующих аналогичных систем.

1.1. Свойства речевого сигнала

1.1.1. Элементы теории речеобразования

Для возникновения акустического речевого сигнала нужно произвести много сложных координированных телодвижений, которые происходят в ряде органов человека, всю совокупность которых можно назвать речевым аппаратом (рис. 1.1). Легкие, обладающие дыхательной мускулистой анатомией, исполняют роль обеспечения развития давления и возникновения воздушных потоков в речевом тракте. Последний (рис. 1.2, А, Б) можно представить гортанью и рядом воздушных полостей, конфигурация которых существенно изменяется в процессе образования речевого сигнала. Ведущую роль играют движения небной занавески, нижней челюсти, губ и языка [54].

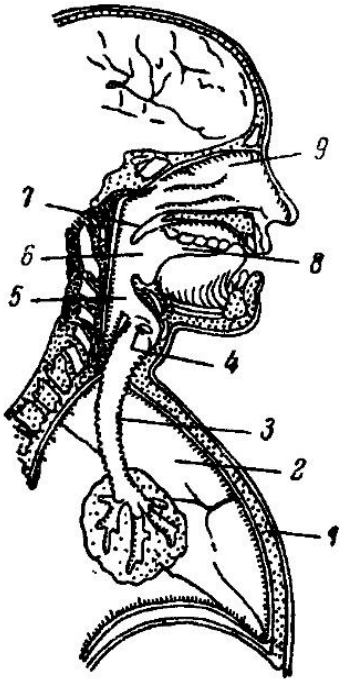

Рис. 1.1. Анатомическая схема речеобразующего аппарата
1 – грудная клетка, 2 – легкие, 3 – трахея, 4 – голосовые связки, 5 – гортанная трубка, 6 – полость глотки, 7 небная занавеска, 8 – полость рта, 9 – полость носа.

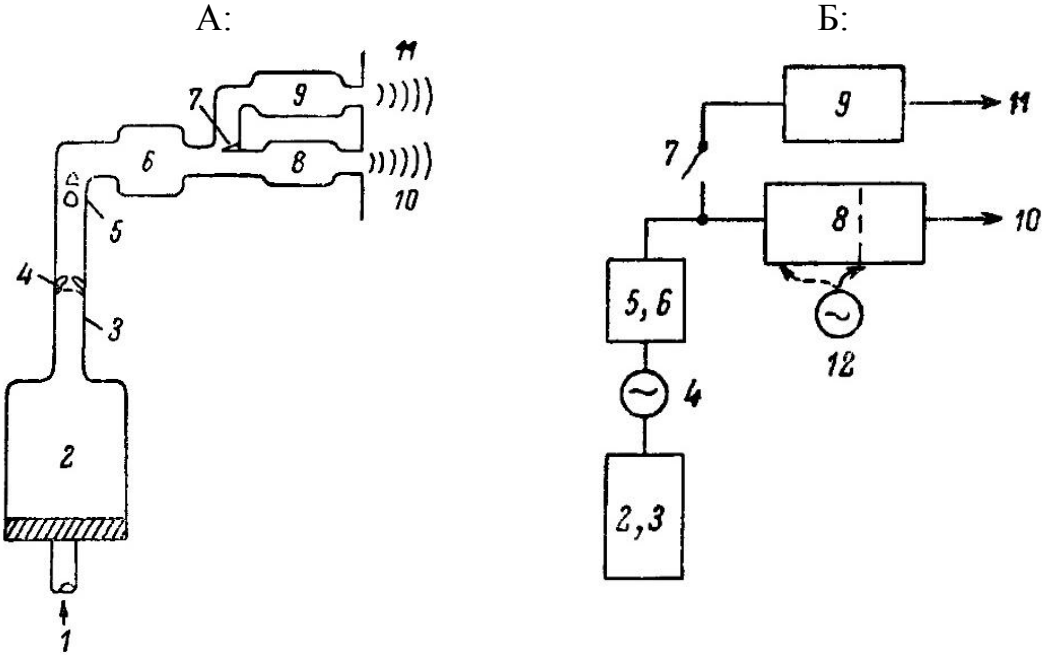

Рис. 1.2. Схема речеобразующего аппарата
А – функциональные элементы; Б – эквивалентная блок-схема.
На А: 1 – сила дыхательных мышц, 2 – объем легких, 3 – трахея, 4 – голосовые связки, 5 – гортанная трубка, 6 – полость глотки, 7 – небная занавеска, 8 – полость рта, 9 – полость носа, 10 – выход воздуха из ротового отверстия, 11 – выход воздуха из носовых отверстий. На Б: 2, 3 – емкость легких и трахеи, 4 – голосовой источник колебаний, 5, 6 – емкость гортани и глотки, 7 – механизм небной занавески, 8 – емкость полости рта, 9 – емкость полостей носа, 10 – выходной сигнал ротового тракта, 11 – выходной сигнал носового тракта, 12 – шумовой источник.

Механизмы возбуждения акустических колебаний связанны либо с работой гортани, либо с возникновением шумовых или импульсных звуков при прохождении воздушного потока через сужения, образующие в определенных местах речевого тракта [54].

Возбуждения акустических колебаний преобразуются с помощью частотной фильтрации в воздушных полостях речевого тракта, действующих по подобию акустических частотных фильтров. Объемы и конфигурация данных полостей в процессе речеобразования определенным образом изменяются. В следствии изменяется и спектр исходных звуковых колебаний, создаваемых акустическими источниками.

Создание воздушных потоков, работа механической гортани, все движения органов, образующих речевой тракт («артикуляторов»), происходят координировано и закономерно. Благодаря данной динамической слаженностью деятельности и возникают сигналы связной речи.

1.1.2. Акустические признаки звуков речи

Основные определения, которыми описывают параметры человеческой речи и связанны с размерами, формой, динамикой трансформации речеобразующего тракта и эмоциональным состоянием человека, можно разделить минимум на четыре группы объективных признаков, которые позволяют различить речевые сигналы: амплитудно-частотные, спектрально-временные, кепстральные и признаки нелинейной динамики [30, 43, 54].

В настоящий момент большое внимание уделяется задачам обработки информации и принятия решений при взаимодействии человека с компьютером. Результативность данного процесса в большинстве случаев зависит от качества информации, полученной от целенаправленности воздействия человека на объекты исследования и пользователя автоматизированной системы. Успешная реализация

диалогового взаимодействия человека и ЭВМ возможно при учете многих признаков, которыми можно описать речевые потоки, возникающие в процессе взаимодействия [43].

Трудовые затраты человека в системах управления ЭВМ (деятельности человека-оператора) связаны с периодическим, иногда довольно продолжительными и изнурительными воздействиями экстремальных показателей социальных, профессиональных и экологических факторов. Данные трудозатраты сопровождаются часто эмоциями, физическим и психическим переутомлением, деструкцией деятельности. Речевой сигнал (РС) является одним из источников эмоций. В русском языке содержится около 40% эмоционально окрашенных слов. Эмоции характеризуются кодированием определенными акустическими параметрами в речевом сигнале, понимание данных особенностей акустического кодирования человеческих эмоций может позволить понять их выражения и сам механизм восприятия эмоций [29].

Исследования речевых сигналов проводились многими учеными с целью описания как технических, так и лингвистических характеристик речи. Большой вклад в развитие науки в области речевой акустики внесли ученые: Г. Фант, М.А. Сапожков, Дж. Фланаган, В.Н. Сорокин, Б.М. Лобанов, В.И. Галунов, Т.К. Винцюк, Л.В. Златоустова, Н.В. Витт, А.В. Аграновский, Р.К. Потапова, Н.Г. Загоруйко, Ю.А. Косарев, М.В. Хитров, А.Л. Ронжин, В.К. Иоффе, В.Г. Михайлов, С.Л. Коваль, В.П. Бондаренко, Е.Л. Чойнзонов, Л.Н. Балацкая и другие [14, 27, 29, 45, 57].

Эмоционально окрашенная речь (ЭОР) находит применение во многих областях жизнедеятельности человека и является актуальной функцией в современных автоматизированных системах управления, реабилитации и протезирования, срочного оповещения и других. В последнее время очень усилился интерес к исследованию РС как объективного показателя эмоционального состояния (ЭС) человека, выполняющего ответственную деятельность таких профессий, как:

космонавта, оператора АЭС, летчика, диспетчера аэропорта и так далее. (Лукьянов, Фролов, 1969; Таубкин, 1977; Williams, Stevens, 1969, 1972; Older, Jenney, 1975; Kuroda, Fujiwara, Okamura, Utsuki, 1976; Congleton, Jones, Shiflett и др., 1997; Rothkrantz, Wiggers и др., 2004; Sigmund, 2004; Хроматиди, 2005; Airas, Alku, 2006; Johannes, Wittels и др., 2007; Соловьева, 2008; Chen, 2008; Siging, 2009; Розалиев, 2009; Калюжный, 2009; Перервенко, 2009; Morist, 2010 [14, 29, 45, 57, 81]). Тем не менее, хоть и проведено множество исследований в данной области, задача автоматического распознавания ЭС говорящего по речевому сигналу на текущий момент не является полностью решенной, также отсутствует модель характеризующая речевые образцы в условиях проявления разных видов эмоций. Модель ЭОР должна отражать взаимосвязь объективных характеристик РС и вида эмоций. В данный момент установление такой взаимосвязи вызывает затруднение у большинства исследователей в данной области [43].

Основная задача получения признаков ЭОР состоит в том, чтобы трансформировать звуковую волну в такое пространство признаков, в котором множество объектов одного класса будет сгруппировано вместе, а множество объектов альтернативных классов максимально разнесено. Соотнесение распознаваемого объекта (в контексте случае под объектом понимается речевой сигнал) с базой объектов, которые нужно идентифицировать, проходит в несколько этапов: 1) выделение того или иного признака объекта; 2) объединение признаков в комплексы или классы; 3) выбор предполагаемого значения из ряда альтернатив. Литературный обзор, который охватывает результаты исследований отечественных и зарубежных авторов [2, 14, 27, 29, 30, 45, 56, 57, 63, 79, 81, 98] показывает, что на данном этапе можно выделить четыре группы объективных признаков, позволяющих различать речевые образцы: спектрально-временные, кепстральные, амплитудно-частотные и признаки

нелинейной динамики (табл. 1.1) [43]. Рассмотрим подробно каждую группу признаков.

Таблица 1.1. Признаки эмоционально окрашенной речи

Название признака	Обозначение	Область		Исследования
		Синтез	Распознование	
1	2	3	4	5
1. Спектрально-временные признаки				
1.1. Спектральные признаки				
1) Среднее значение спектра анализируемого речевого сигнала	$X(i)$	+	+	[29]
2) Нормализованные средние значения спектра	$X_H(i)$	+	+	
3) Относительное время пребывания сигнала в полосах спектра	$t(i)$	+	+	
4) Нормализованное время пребывания сигнала в полосах спектра	$t_H(i)$	+	+	
5) Медианное значение спектра речи в полосах	$m_H(i)$	+	+	
6) Относительная мощность спектра речи в полосах	$P_H(i)$	+	+	
7) Вариация огибающих спектра речи	$V(i)$	+	+	
8) Нормализованные величины вариации огибающих спектра речи	$V_H(i)$	+	+	
9) Коэффициенты кросскорреляции спектральных огибающих между полосами спектра	$R(i,k)$	+	+	
1.2. Временные признаки				
10) Длительность сегмента, фонемы	l	+	+	[2,14,29, 57,81,79]
11) Высота сегмента	h	+	+	[14]
12) Коэффициент формы сегмента	k	+	+	
2. Кепстральные признаки				
13) Мелко частотные кепстральные коэффициенты	$MFCC$	-	+	[63,98]
14) Коэффициенты линейного предсказания с коррекцией на неравномерность чувствительности человеческого уха	PLP	-	+	
15) Коэффициенты мощности частоты регистрации	$LFPC$	-	+	[63]
16) Коэффициенты спектра линейного предсказания	LPC	-	+	
17) Коэффициенты кепстра линейного предсказания	$LPCC$	-	+	
3. Амплитудно-частотные признаки				
18) Интенсивность, амплитуда	i, A	-	+	[2,57,98]

19) Энергия	E	+	+	[79,81]
20) Частота основного тона	$F0$	+	+	[29,45, 56,63,79, 81,98]
21) Формантные частоты	$F1, F2,$ $F3, F4$	+	+	[29,57,63, 79]
22) Джиттер	Ji	-	+	[45,63]
23) Шиммер	Sh	-	+	[45,63,98]
24) Радиальная базисная ядерная функция	$K(x, y)$	-	+	[56]
25) Нелинейный оператор Тигера	TEO	-	+	[45,56,98]
4. Признаки нелинейной динамики				
26) Отображение Пуанкаре	Δt_i	-	+	[27]
27) Рекуррентный график	$R_{i,j}$	-	+	
28) Максимальный характеристический показатель Ляпунова	Yj	-	+	[27,57]
29) Фазовый портрет (аттрактор)	Y_n	-	+	
30) Размерность Каплана-Йорка	D	-	+	[27]

Спектрально-временные признаки [29] характеризуют РС в его физико-математической сущности исходя из наличия компонентов трех видов: 1) периодических (тональных) участков звуковой волны; 2) непериодических участков звуковой волны (шумовых, взрывных); 3) участков, не содержащих РС (речевых пауз). Спектрально-временные признаки позволяют отражать своеобразие формы временного ряда и спектра голосовых импульсов у разных лиц и особенности фильтрующих функций их речевых трактов. А также определяют особенности речевого потока, связанные с динамикой перестройки артикуляционных органов речи говорящего, и являются интегральными характеристиками речевого потока, отражающими своеобразие взаимосвязи или синхронности движения артикуляционных органов говорящего.

В группе спектрально-временных признаков РС рассматривается как некоторый квазистационарный процесс [14]. Среди множества акустических параметров были выделены параметры, инвариантные к действию повышенного уровня сигнала, описывающие статистические

характеристики РС и основного тона, особенности спектральной структуры.

Установлено, что применение только спектральных характеристик невозможно использовать в качестве признаков, позволяющих правильно распознавать и идентифицировать различные эмоции и речевые сигналы [44].

Следует особо выделить такое семейство признаков ЭОР, как кепстральные коэффициенты [63, 98, 96]. Большая часть современных автоматических моделей синтеза и распознавания речи уделяют больше внимания на извлечении частотной характеристики речевого тракта пользователя, убирая при этом параметры сигнала возбуждения. Данный факт можно объяснить тем, что коэффициенты первой системы обеспечивают лучшие показатели делимости звуков. Для решения задачи выделения сигнала возбуждения из сигнала речевого тракта выбирают метод кепстрального анализа. Схематически данный метод можно представить на рисунке 1.3.

Рис. 1.3. Упрощенная схема кепстрального анализа сигнала: БПФ – блок быстрого преобразования Фурье речевого сигнала; LOG – блок логарифмирования спектра сигнала; ОБПФ – блок обратного быстрого преобразования Фурье речевого сигнала

Данное представление является линейным предсказанием и является одним из наиболее результативных методов анализа РС. Данный метод можно назвать доминирующим при оценивании основных параметров РС, таких как, например, период основного тона, форманты, спектр сигнала, функция площади речевого тракта, а также в сжатом представлении речевого сигнала с целью его передачи и хранения на каких-либо носителях информации. Важность данного метода обусловлена высокой

результативностью в получении качественных оценок и простоты вычислений. Основным принципом данного метода линейного предсказания является то, что имеющийся отсчет РС можно представить функцией линейной комбинации предшествующих отсчетов. Коэффициенты предсказания при этом можно определить однозначно минимизацией параметра среднего квадрата разности между отсчетами РС и их предсказанными значениями.

Использование мел-частотных кепстральных коэффициентов MFCC достаточно популярно для представления набора информативных признаков РС. Ключевой идеей данного метода MFCC является реализация модели представления речевой информации «живой» системой, которая будет максимально приближена к информации, поступающей на слуховой анализатор мозга человека. Информативные признаки, построенные на основе MFCC, учитывают особенности психо-акустического восприятия речи человеком, так как используют мел-шкалу. Данная шкала связанна с критическими полосами слуха и вычисляется следующим образом. Речевой сигнал $S(t)$ каждого отдельного слова разбивается на K окон по N отсчетов в каждом, пересекающихся на 1/2 длины:

$$s(t) \to S_n[t], n = 1,\ldots,K.$$

В каждом окне производится дискретное преобразование Фурье (ДПФ):

$$\mathrm{Re}\, X_n[k] = \frac{2}{N}\sum_{i=1}^{N} S_n[i]\cos(\frac{2\pi k(i-1)}{N}),$$

$$\mathrm{Im}\, X_n[k] = -\frac{2}{N}\sum_{i=1}^{N} S_n[i]\sin(\frac{2\pi k(i-1)}{N}),$$

где $k = 1,\ldots,M, M = N/2.$

Находится спектральная плотность мощности получившегося сигнала:

$$P_n[k] = A_n[k]^2,$$

$$A_n[k] = \sqrt{\mathrm{Re}\, X_n[k]^2 + \mathrm{Im}\, X_n[k]^2}.$$

Применяется банк треугольных фильтров:

1. Задается количество фильтров P, а также начальная f_l и конечная f_h частоты (f_h не должна превосходить половины частоты дискретизации).

2. Они переводятся в мелы:

$$f_{mel}(f) = 1127 \ln(1 + \frac{f}{700}),$$

$$f_l^m = f_{mel}(f_l),$$

$$f_h^m = f_{mel}(f_h).$$

3. На мел-шкале отрезок $[f_l^m, f_h^m]$ разбивается на $P+1$ равных непересекающихся подотрезков $[f_j^m, f_{j+1}^m]$, $j = 1,\ldots,P+1$ длины

$$len = \frac{f_h^m - f_l^m}{P+1}.$$

4. Находятся центры данных подотрезков:

$$C^m[i] = f_l^m + i \cdot len,\ i = 1,\ldots,P$$

и переводятся в шкалу Гц:

$$C[i] = f(C^m[i]),\ i = 1,\ldots,P$$

(это центральные частоты треугольных фильтров).

5. Центры треугольных фильтров переводятся из Гц в номера отсчетов массива $P_n[k]$:

$$f_{smp}[i] = \frac{M}{F_s} C[i],\ i = 1,\ldots,P$$

где F_s – частота дискретизации исходного сигнала.

Для каждого фильтра отсчеты спектральной плотности мощности умножаются на соответствующий фильтр:

$$X_n[i] = \sum_{k=1}^{M} P_n[k] H_i[k],\ i = 1,\ldots,P,$$

$$H_i[k] = \begin{cases} 0, & k < f_{smp}[i-1] \\ \dfrac{(k - f_{smp}[i-1])}{f_{smp}[i] - f_{smp}[i-1]}, & f_{smp}[i] \le k \le f_{smp}[i] \\ \dfrac{(f_{smp}[i+1] - k)}{f_{smp}[i+1] - f_{smp}[i]}, & f_{smp}[i] \le k \le f_{smp}[i+1] \\ 0, & k > f_{smp}[i+1] \end{cases}.$$

Вычисляется логарифм $X_n[i] = \ln(X_n[i]),\ i = 1,\ldots,P$ и дискретное косинусное преобразование:

$$C_n[i] = \sum_{k=0}^{P-1} X_n[k] \cos(i(k - \frac{1}{2})\frac{\pi}{P}),\quad i = 1,\ldots,J,$$

где $C_n[i]$ – массив кепстральных коэффициентов, J – желаемое число коэффициентов $(J < P)$.

Все перечисленные кепстральные коэффициенты позволяют уменьшить размерность исходного пространства признаков, что скажется на быстродействии вычисления различных параметров речевых сигналов.

Анализ амплитудно-частотных параметров РС позволяет применять эти характеристики в качестве информативных признаков диагностики ЭС человека и синтеза ЭОР [2, 14, 27, 30, 81]. Амплитудно-частотные параметры позволяют получать оценки, значения которых могут меняться в зависимости от параметров дискретного преобразования Фурье (вида и ширины окна), а также при незначительных сдвигах окна по выборке. РС акустически можно представить как распространяемые в воздушной среде сложные по своему составу звуковые колебания, которые можно охарактеризовать частотой, интенсивностью (амплитудой колебаний) и длительностью. Все указанные характеристики подвержены изменениям

на протяжении одного РС. Они могут быть зафиксированы и описаны посредством специальных электронных приборов (прежде всего осциллографа и спектрографа). Амплитудно-частотные признаки несут необходимую и достаточную информацию для человека по РС при минимальном времени восприятия [43].

В данной диссертационной работе предполагается разработать и создать систему дикторонезависимого распознавания речевых сигналов приближенную к человеческой «живой» модели. Так как данная задача сложная и требует больших вычислительных затрат [36], то желательно для представления речевых сигналов использовать информационно-емкие параметры представления речевых сигналов. Лучшим для такого представления на текущий момент являются MFCC коэффициенты, которые учитывают особенности психо-акустического восприятия речи человеком. Таким образов, в данной работе решено для построения системы дикторонезависимого распознавания речевых сигналов использовать представление РС в виде информативных признаков – MFCC-коэффициентов.

1.2. Классификация систем распознавания речи

Исходя из сложности решаемой задачи, нужно определиться со структурой системы распознавания речи. При выборе возникают сложности в связи с многообразием возможных технических решений на различных уровнях системы. Упростить задачу выбора структуры системы позволяет ее предварительная классификация по ряду признаков, наборам которых могут соответствовать «шаблонные» решения [11].

В качестве признаков можно использовать: тип речи, распознаваемый системой; зависимость системы от распознаваемых голосов дикторов; степень детализации эталонов; количество распознаваемых слов [1, 19]. В [1] вводится понятие полноты словаря и задачи поиска ключевых слов интерпретируются, как распознавание с

неполным словарем. А в [23] предлагаются иные классификационные признаки: назначение системы, ее потребительские свойства и механизмы функционирования.

По **типу речи** различают системы распознавания речевых сигналов и слитной речи (рис. 1.4). В первом случае требуется специальное (дискретное) произнесение слов (речевых команд), при котором паузы между словами значительно превышают внутрисловные паузы. Обычно длительность такой разделительной паузы составляет полсекунды. При распознавании слитной речи пользователь может произносить слова фраз естественно, не делая специальных пауз между словами. Существует и третий вариант работы системы распознавания, при котором система должна обнаруживать произнесение заданных слов в звуковом потоке, независимо от того выделены они паузами или произнесены в окружении других слов. Такой режим распознавания называется режимом поиска ключевых слов.

Рис. 1.4. Признаки классификации систем распознавания речи

По **степени зависимости системы от распознаваемых голосов дикторов** различают дикторозависимые и дикторонезависимые системы, а также системы с автоматической подстройкой. Первые требуют предварительного обучения (адаптации) к голосу пользователя системы, вторые – готовы к работе сразу после установки. Дикторозависимые системы обеспечивают более высокую точность распознавания с голоса основного пользователя системы, чем с любых других голосов. Третий тип систем – системы, автоматически настраивающиеся на голос диктора по мере их использования.

По **степени детализации эталонов** различают системы, использующие в качестве эталонов целые слова и части (монофоны, трифоны, слоги и т.д.) слов. Первые обеспечивают более высокие точность и скорость сравнения, но накладывают значительные ограничения на объем и открытость словаря.

По **количеству распознаваемых слов** (или объему словаря) можно выделить две категории: системы с малыми (обычно, до 100 слов) и большими словарями. В системах с малым словарем есть возможность прямого обучения для каждого слова. В системах с большим словарем такой возможности нет [11].

По **механизму функционирования** можно выделить три категории: простейшие (корреляционные) детекторы [4]; экспертные системы с различным способом формирования и обработки баз данных [10] и вероятностно сетевые модели принятия решения [12].

Подавляющая часть систем распознавания речи относится к системам с вероятностно-сетевыми методами принятия решения. К данной категории можно отнести: скрытое Марковское моделирование [9]; динамическое программирование (алгоритм Витерби) [19] и нейросетевые методы [8]. Например, нейронные сети могут быть использованы для классификации характеристик речевого сигнала и принятия решения о принадлежности к той или иной группе эталонов [8]. Нейронная сеть

обладает способностью к статистическому усреднению, т.е. решается проблема с вариативностью речи. Многие нейросетевые алгоритмы осуществляют параллельную обработку информации, т.е. одновременно работают все нейроны. Таким образом, можно решить проблему со скоростью распознавания – обычно время работы одной нейронной сети составляет несколько итераций. Сейчас многие разработчики используют аппарат нейронных сетей для построения распознавателей [8, 55].

С точки зрения выбора структуры системы представляется целесообразным использование такого классификационного признака, как **тип грамматики**, определяющей структуру распознаваемых высказываний. По типу грамматики системы распознавания речи можно разделить на три класса: командные, с фиксированной грамматикой и системы диктовки.

Командные системы ориентированы на распознавание отдельных слов и/или фраз, включаемых в словарь системы в качестве отдельных элементов. Командные системы не предусматривают возможность распознавания комбинаций элементов словаря.

В системах с фиксированной грамматикой грамматика определяет допустимые комбинации элементов словаря. «Фиксированность» грамматики не означает, что система может работать только с одной, заданной грамматикой – грамматика фиксируется в рамках одной сессии распознавания.

Использование данной классификации упрощает разработку структуры системы распознавания, т.к. для каждого класса имеется набор стандартных решений. Определившись с классом, можно определиться с базовым комплектом методов, моделей и алгоритмов.

1.3. Вероятностно-сетевые методы принятия решений

По характеру функционирования большинство существующих систем распознавания речи можно отнести к классу с вероятностно-

сетевыми методами принятия решений. Для задачи распознавания речевых сигналов применяются такие методы, как CMM (hidden Markov modelling – HMM, скрытая Марковская модель) [28], ИНС (искусственные нейронные сети) [3] или комбинации данных методов [20].

Наиболее популярной и успешно реализованной для распознавания речевых сигналов является CMM. Скрытая Марковская модель выражается как множество переходов и состояний. С каждым переходом из одного состояния I в другое состояние J вычисляется распределение выходных (результирующих) вероятностей P. Данные вероятности определяют возможность того, что при переходе состоится событие X из всего множества наблюдений. Наблюдения состоят из множества начальных и конечных состояний. Любой последовательностью наблюдений в данном методе является результат перехода одного из начальных состояний в одно из конечных. Благодаря тому, что CMM может хорошо описать временные ряды со стохастическими воздействиями, возможно, приближенно обеспечить соответствие данной модели к естественному представлению речи. Метод CMM можно применить для представления любой составляющей речевого сигнала – фраз, сигналов или фонем [28].

Так как направление распознавания речи с помощью модели CMM достаточно изучено и на сегодняшний момент многими ученными по применимости данной модели получено много результатов, то имеет смысл исследовать другие перспективные модели для задачи распознавания речевых сигналов. Одними из таких методов являются различные модификации ИНС.

1.4. Стандартные модели нейронных сетей

ИНС можно представить в виде совокупности процессорных элементов (нейронов). Строение искусственного нейрона представляет собой набор связанных соединений, сумматора и нелинейного оператора. На каждый вход нейрона приходит какое-то значение X_i, которое

умножается на вес данного входа нейрона W_i. Далее подобные значения со всех входов суммируются, и из данной суммы вычитается некоторый порог ε. После этого на получившийся результат действует некоторая активационная функция F, после которой информация подается на выход нейрона Y, как показано на рисунке 1.5 [53].

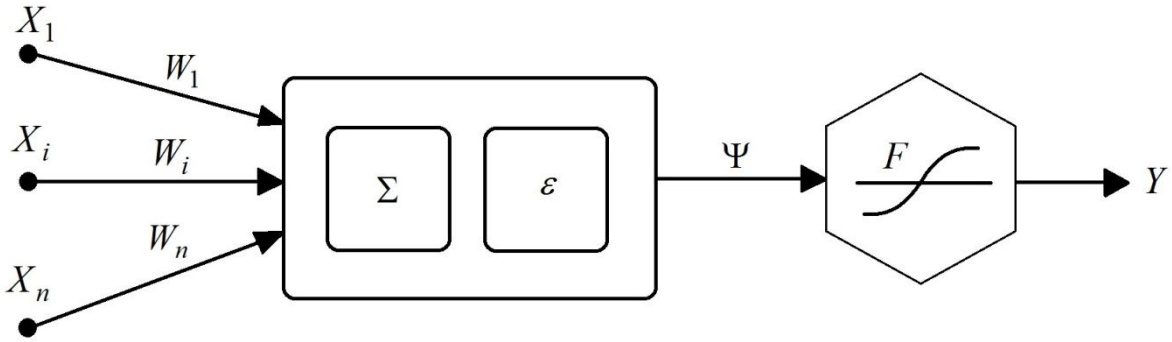

Рис. 1.5. Структурная схема искусственного нейрона

Таким образом, выход каждого нейрона можно вычислить по данной формуле:

$$Y = F\left[\sum_{i=1}^{n}(X_i W_i) - \varepsilon\right]. \qquad (1.1)$$

Введем обозначение произведения входов нейрона на веса данных входов $X_i W_i$ через переменную Ψ.

Выходными (активационными) функциями нейронов F принято использовать нелинейные функции, которые ограничены на множестве значений некоторыми пределами. Наиболее распространенными такими функциями являются: пороговая (случай сигнума), комбинация пороговой и линейной функции, сигмоидальная функция, функция гиперболического тангенса, а также радиально-базисные функции.

Смысл пороговой функции (1.2) заключается в том, что пока суммарный сигнал с входов нейрона не достигнет порогового значения,

нейрон остается неактивным. При достижении данного порога нейрон возбуждается и подает на выход единицу:

$$Y = \begin{cases} 0, & \Psi \le \varepsilon \\ 1, & \Psi > \varepsilon \end{cases}. \qquad (1.2)$$

Пороговая функция может наиболее точно смоделировать нелинейную передаточную характеристику биологического нейрона и благодаря своей простоте дает ИНС большие возможности. Например, для решения задачи классификации образов [18] нейронная сеть с пороговой активационной функцией разбивает пространство на многогранники [68, 100].

Данные активационные функции изображены на рисунке 1.6.

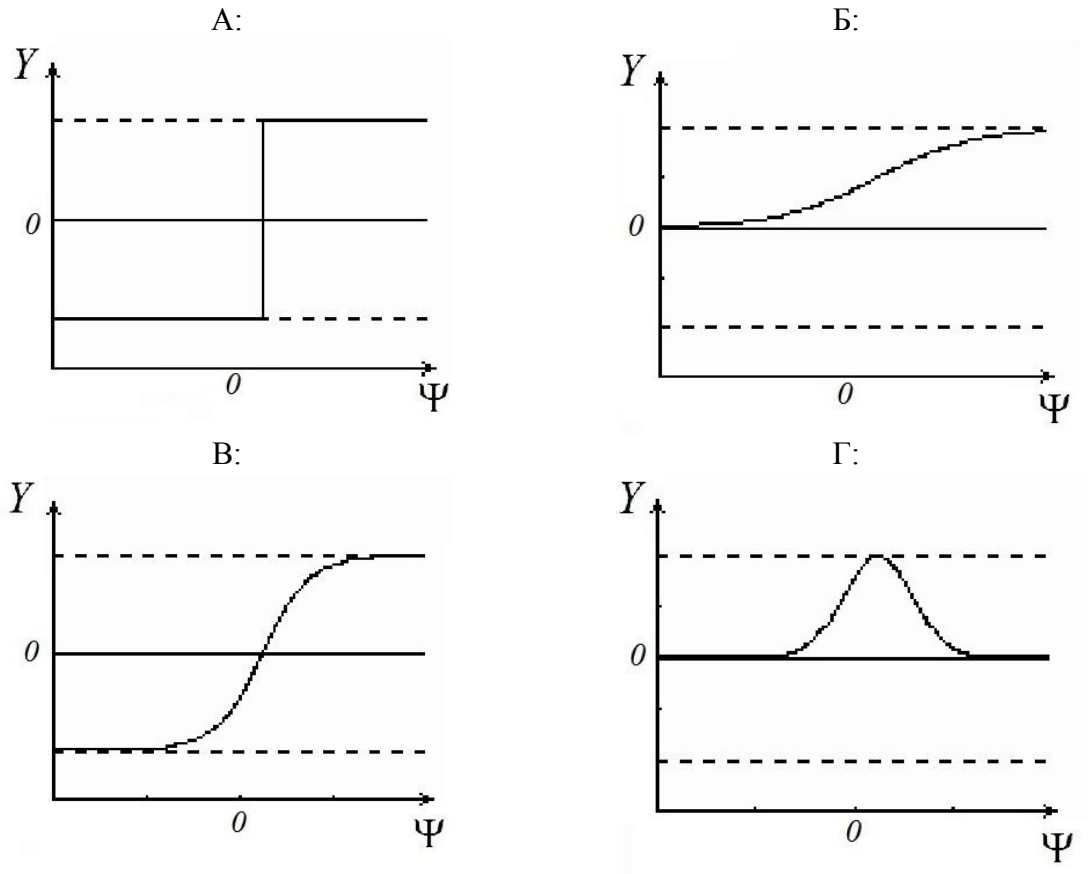

Рис.1.6. Наиболее распространенные виды активационных функций (порог равен 0,5)
А – пороговая, Б – сигмоидальная, В – гиперболический тангенс, Г – радиально-базисная.

При использовании сигмоидальной функции (1.3), диапазон изменений входной величины X уменьшается таким образом, что для любого значения X выходные значения Y расположены в некотором конечном интервале. Также стоит отметить, что данная функция легко дифференцируема. Данное свойство очень полезно в алгоритмах обучения [77, 100].

$$Y = \frac{1}{1 + e^{-(\Psi - \varepsilon)}} \tag{1.3}$$

Функция гиперболического тангенса является модификацией сигмоидальной функции на интервале (–1, 1) и имеет вид:

$$Y = \frac{2}{1 + e^{2(\Psi - \varepsilon)}}.$$

Наиболее интересным и перспективным в настоящее время (особенно в вопросах интерполяции и аппроксимации) является класс – радиальных базисных функций, к которым можно отнести функции, имеющие глобальный экстремум, и которые ведут себя монотонно по мере удаления от него. Например, к радиально базисным функциям можно отнести функцию вида:

$$Y = e^{-\frac{1}{2}(\Psi - \varepsilon)^2}.$$

В данной диссертационной работе использовано несколько разновидностей нейронных сетей, одна из которых называется многослойным персептроном.

Многослойным персептроном (рис. 1.7) называется нейронная сеть, в которой нейроны расположены слоями. Особенность данного построения заключается в связи нейронов одного слоя с нейронами только предыдущего слоя.

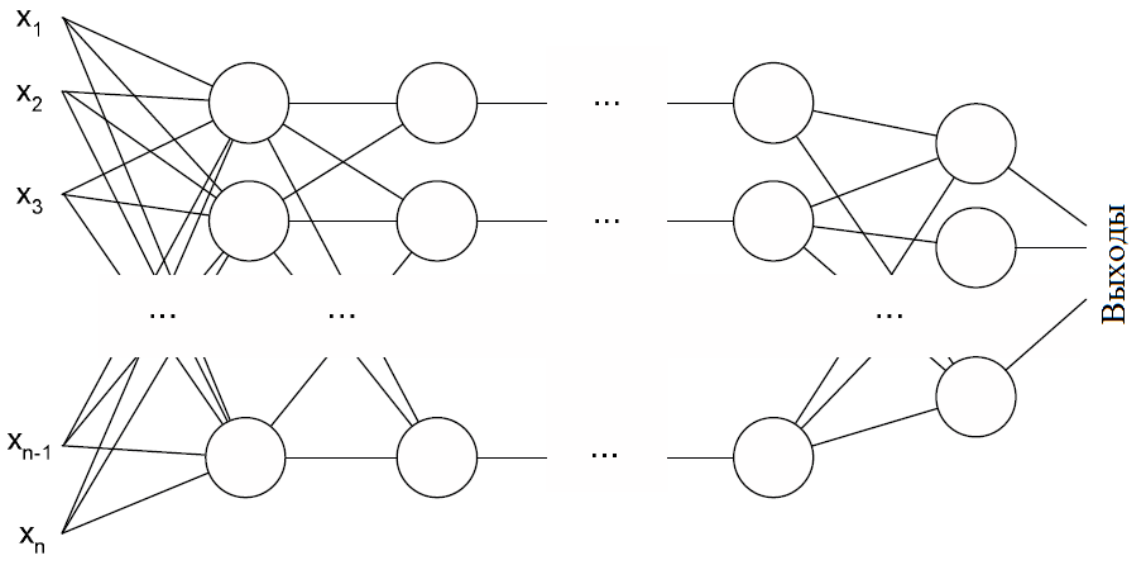

Рис. 1.7. Многослойный персептрон

Для данных нейронных сетей имеется довольно хороший "классический" метод обучения, относящийся к разновидностям алгоритмов градиентного спуска – так называемый метод обратного распространения ошибки (Back Propagation Error). В начале 80-х данный метод сыграл очень важную роль в возникновении интереса к нейронным сетям, а его модификации и до сегодняшнего дня остаются лучшими для рассматриваемого класса нейронных сетей.

Алгоритм обратного распространения позволяет решать задачи, в которых некоторому входному значению вектора $X(x_1, x_2, \ldots, x_n)$ нейронная сеть на выходе выводит требуемый вектор $Y(y_1, y_2, \ldots, y_n)$. В процессе обучения используется ряд пар векторов (X, Y^*), которые называются обучающей парой, где Y^* – целевой вектор. Использование обучающих пар позволяет сети подстраивать свои веса таким образом, чтобы можно было адекватно реагировать на входной вектор X.

Для примера рассмотрим работу алгоритма обратного распространения ошибки в такой модели, в которой активационными функциями являются сигмоидальные функции.

Для процедуры инициализации нейронной сети случайным образом задаются начальные веса и сдвиги сети. В алгоритме обучения нейронной сети обратного распространения можно выделить следующие последовательности действий:

1. Производится выборка очередной обучающей пары (X, Y^*) из обучающего множества и подается входной вектор X на вход данной сети.
2. Вычисляется выход сети Y.
3. Вычисляется разность между реальным (полученным) выходом сети и нужным выходом (целевым вектором обучающей пары).
4. Производится корректировка весов сети так, чтобы ошибка минимизировалась.
5. Пункты с 1 по 4 повторяются для каждого вектора обучающего множества до тех пор, пока ошибка на всем множестве не достигнет минимального значения.

При функционировании обученной сети выполняются пункты 1 и 2.

Вычисления в данной нейронной сети выполняются послойно. На этапе пункта 3 каждый из выходов сети Y вычитается из соответствующей компоненты целевого вектора для вычисления ошибки. Данная ошибка применяется в пункте 4 для правки весов сети, причем размер изменений определяется алгоритмом обучения.

Пункты 1 и 2 рассматриваются как движение вперед, а 3 и 4 – как движение назад. Параметр ошибки распространяется обратно по нейронной сети и применяется для подстройки весов. Данные действия можно описать математически.

Предположим, на входе имеем вектор X, на основе которого вычисляется выходной вектор Y. Далее из вектора Y^* вычитается вектор Y с целью получения вектора ошибки E:

$$E = Y^* - Y.$$

Далее вычисляется величина Ψ каждого нейрона первого слоя как взвешенная сумма входов нейрона. Затем функция активации F сжимает величину Ψ и вычисляет величину OUT для каждого нейрона в данном слое. Полученное множество OUT с одного слоя поступает на вход следующего слоя. Данный процесс повторяется слой за слоем, пока не образуется заключительное множество сети.

Обратный проход нейронной сети отвечает за подстройку весов выходного слоя. Подстройка весов можно легко осуществить с помощью дельта-правила, т.к. для каждого нейрона выходного слоя задано целевое значение. Внутренние слои нейронной сети целевых значений не имеют и называются скрытыми слоями.

Процесс подстройки одного веса от нейрона p в скрытом слое j к нейрону q в выходном слое можно построить следующим образом. Выход OUT слоя k, вычитаясь из целевого значения Y^*, выдает ошибку, которая умножается на производную сжимающей функции (в нашем случае $OUT(1-OUT)$), вычисленную для данного нейрона слоя k, в результате получая величину:

$$\delta = OUT(1-OUT)(Y^* - OUT).$$

Далее δ умножается на функцию OUT нейрона слоя j, от которого образуется рассматриваемый вес. Также данное произведение нужно умножить на коэффициент обучения (скорости обучения) η, для которого определим условие $0{,}01 \le \eta \le 1$. В результате веса выходного слоя после коррекции будут равны:

$$w_{pq,k}^{(n+1)} = w_{pq,k}^{n} + \eta\delta_{q,k} \cdot OUT_{p,q},$$

$$\Delta w_{pq,k} = \eta\delta_{q,k} \cdot OUT_{p,q},$$

где $\delta_{q,k}$ – величина δ для нейрона q в выходном слое k; $OUT_{p,q}$ – величина выхода для нейрона в скрытом слое j; $w_{pq,k}^{n}$ – величина веса от

нейрона в скрытом слое k к нейрону q в выходном слое на шаге n; $w_{pq,k}^{(n+1)}$ – величина веса на шаге $n+1$ после коррекции; $\Delta w_{pq,k}$ – величина изменения веса. Даная процедура выполняется для каждого веса от нейрона скрытого слоя к нейрону в выходном слое.

1.5. Коллективное нейросетевое распознавание

Для повышения точности распознавания речевых сигналов возможно объединить отдельные нейросетевые системы распознавания в единую коллективную систему. Существует ряд методов формирования коллективного распознавания, среди которых можно выделить три основных:

– bagging, или bootstrap aggregation (бутстрап-подмножества) – обучение распознающих нейронных сетей на бутстрап-подмножествах учебного множества [62];

– boosting – это последовательное обучение распознавания членов коллектива, при котором каждое распознавание последующего члена, включенного в коллектив, обучается таким образом, чтобы произошла компенсация недостатков всех предыдущих распознанных членов [97];

– mixture of experts – смесь экспертов, когда в коллектив вводится дополнительное распознавание, которое оценивает компетентность других членов коллектива для каждого входного образа и объединяет их индивидуальные решения с учетом вычисленных оценок [60].

Задача распознавания речевых сигналов характеризуется высокой вычислительной сложностью и большим объемом данных для обучения (например, классический речевой корпус для обучения распознаванию англоязычной речи TIMIT [104] содержит более 500 Мб речевого материала). Для решения такой задачи наиболее целесообразным является использование первого подхода – формирование коллектива нейросетевого распознавания на основе метода bagging, потому что:

– обучение отдельных нейронных сетей осуществляется независимо, что позволяет ускорить формирование коллектива за счет распараллеливания процессов обучения отдельных нейронных сетей;

– данный подход позволяет повысить качество обучения и в последующем распознавания за счет коллективного голосования;

– учебные бутстрап-подмножества могут быть меньше базового учебного множества, что позволяет ускорить процесс обучения каждой нейронной сети.

Коллективный нейросетевой алгоритм bagging строится по принципам равноправного голосования нейронных сетей, входящих в данный алгоритм.

1.6. Алгоритмы шумоподавления

Шумоподавление речевых сигналов очень актуальная задача в областях распознавания речевых сигналов, идентификации диктора и построении различных систем передачи/приема речевой информации, а также в системах, работающих с изображениями и видеосигналами [13, 16, 17].

Задача дикторонезависимого распознавания речевых сигналов может быть реализована достаточно хорошо в не зашумленных средах. Присутствие в речевом тракте фонового шума может значительно ухудшить качество распознавания речевых сигналов. Данное явление происходит по причине несоответствия информативных признаков речевых сигналов, используемых для обучения в чистых акустических условиях, и признаков речевых сигналов, наблюдаемых в зашумленном речевом сигнале [25, 80, 91].

Оценка зашумленности речевого сигнала обычно оценивается отношением сигнал/шум (ОСШ или SNR, Signal-to-Noise Ratio). Данная оценка обычно выражается в децибелах (дБ) и определяется выражением:

$$ОСШ = 10\log_{10}\left(\frac{M_{\text{сигнала}}}{M_{\text{шума}}}\right) = 20\log_{10}\left(\frac{A_{\text{сигнала}}}{A_{\text{шума}}}\right),$$

где M – средняя мощность, A – среднеквадратичное значение амплитуды.

Увеличение значения отношения сигнал/шум способствует ослаблению влияния шума на характеристики системы дикторонезависимого распознавания речевых сигналов.

Процесс шумоподавления в системах дикторонезависимого распознавания речевых сигналов может быть выполнен тремя различными способами: блоком предобработки с помощью различных алгоритмов шумоподавления входных речевых сигналов; на параметрическом этапе, путем представления речевых сигналов с помощью характеристик, устойчивых к шуму; на этапе моделирования, путем адекватного комбинирования моделей шума и чистого сигнала [83]. При модернизации коллективного нейросетевого алгоритма и модифицированного коллективного нейросетевого алгоритма решено использовать первый способ процесса шумоподавления путем внедрения блока предобработки входных речевых сигналов.

В данной работе предлагается исследовать следующие алгоритмы шумоподавления:

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе апостериорного отношения сигнал/шум (Ideal Binary Mask – A Posteriori Signal-to-Noise Ratio, далее IBM-PostSNR) [90];

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе априорного отношения сигнал/шум, для оценки которого используется алгоритм TSNR (Ideal Binary Mask – Two-Step Noise Reduction, далее IBM-TSNR) [6, 87];

– алгоритм шумоподавления Скалара на основе винеровской фильтрации (Wiener – A Priori Signal-to-Noise Ratio, далее Wiener-PriorSNR) [95].

1.6.1. Алгоритмы шумоподавления на основе бинарных масок

Алгоритмы бинарных масок популярны в различных методах анализа речевых сигналов, например, таких как шумоподавления [48, 92, 93], идентификации диктора [52] и автоматического распознавания речи [102].

В настоящее время существует несколько алгоритмов для оценки бинарных масок, которые основаны на следующих принципах: оценки апостериорного отношения сигнал/шум [90], информации о непрерывном тоне [76], Байесовской классификации речевых характеристик [93] и локализации звуковых сигналов [59].

Наиболее популярными алгоритмами являются: оценка бинарных масок, использующая статистический критерий детектирования на основе апостериорного ОСШ и ее модифицированная версия на основе априорного ОСШ [90], оцениваемого с помощью двухступенчатого алгоритма TSNR (Two-Step Noise Reduction) [6, 87].

Статистический критерий детектирования на основе апостериорного ОСШ

Введем обозначение модуля спектра чистого речевого сигнала $|X(\omega,t)|$, зашумленного речевого сигнала как $|Y(\omega,t)|$, а модуль спектра шума как $|N(\omega,t)|$. Используя данные функции можно определить априорное и апостериорное ОСШ:

$$SNR_{prio}(\omega,t)=\frac{|X(\omega,t)|^2}{\big||Y(\omega,t)|-|X(\omega,t)|\big|^2}, \tag{1.4}$$

$$SNR_{post}(\omega,t)=\frac{\left|Y(\omega,t)\right|^2}{\left|N(\omega,t)\right|^2}. \tag{1.5}$$

Если учесть разность фаз между речевым сигналом и шумом, то можно получить следующее выражение:

$$SNR_{prio}(\omega,t)=\frac{\left|X(\omega,t)\right|^2}{\left\|Y(\omega,t)\right|-\left|X(\omega,t)\right\|^2}\geq\frac{\left|X(\omega,t)\right|^2}{\left|N(\omega,t)\right|^2}\geq\left(\frac{\left|X(\omega,t)\right|}{\left|N(\omega,t)\right|}-1\right)^2. \tag{1.6}$$

При постановке выражения (1.5) в выражение (1.6) можно получить связь апостериорного с априорным ОСШ:

$$SNR_{prio}(\omega,t)=\left(\sqrt{SNR_{post}(\omega,t)}-1\right)^2. \tag{1.7}$$

Из выражения (1.7) следует, что апостериорное ОСШ можно использовать как критерий детектирования.

Далее для определения частотных интервалов (в дальнейшем – бинов), в которых присутствует речь, зададим пороговый критерий [90]:

$$SNR_{post}(\omega,t)>\tau. \tag{1.8}$$

При условии нормального распределения модуля спектра шума в каждом бине возможно представление его в виде математического ожидания и дисперсии для каждого интервала частотной полосы [6].

Учитывая нормальное распределение модуля спектра шума, можно определить плотность распределения модуля спектра шума в каждом бине следующим образом:

$$\Xi(\left|N(\omega,t)\right|)=\frac{1}{\sqrt{2\pi}\left|\sigma_N(\omega)\right|}\exp\left(\frac{\left(\left|N(\omega,t)\right|-\mu_N(\omega)\right)^2}{2\sigma_N^2(\omega)}\right), \tag{1.9}$$

где $\mu_N(\omega)$ и $\sigma_N^2(\omega)$ – соответственно, математическое ожидание и дисперсия шума в полосе ω.

Подставив в формулу плотности распределения спектра шума (1.9) выражение для спектра шума (1.5) и используя пороговый критерий (1.8), можно получить уравнение для вероятности того, что апостериорное ОСШ будет выше, чем τ:

$$P\left(|N(\omega,t)| < \frac{|Y(\omega,t)|}{\sqrt{\tau}}\right) = \frac{1}{\sqrt{2\pi}|\sigma_N(\omega)|} \int_{-\infty}^{\frac{|Y(\omega,t)|}{\sqrt{\tau}}} \exp\left(\frac{(n-\mu_N(\omega))^2}{2\sigma_N^2(\omega)}\right) dn$$

В дальнейшем для разделения частотных интервалов на нужные (содержащие речевой сигнал) и не нужные (содержащие шум) нужно ввести пороговое значение вероятности $\rho \in [0;1]$, такое чтобы частотный интервал, содержащий речь, удовлетворял условию [6]:

$$P(SNR_{post}(\omega,t) > \tau) > \rho.$$

Статистический критерий детектирования на основе априорного ОСШ

Описанный выше алгоритм бинарных масок на основе апостериорного ОСШ использовался для оценки априорного ОСШ. Данный алгоритм можно модифицировать [49, 50, 51], изменив пороговый критерий (1.7, 1.8) для априорного ОСШ:

$$SNR_{prio}(\omega,t) > \left(\sqrt{\tau}-1\right)^2. \quad (1.10)$$

Выражение для априорного ОСШ можно записать в следующей форме:

$$SNR_{prio}(\omega,t) = \frac{|X(\omega,t)|^2}{|N(\omega,t)|^2}. \quad (1.11)$$

Подставив в формулу плотности распределения спектра шума (1.9) выражение для спектра шума (1.11) и используя пороговый критерий (1.10), можно получить уравнение для вероятности того, что априорное ОСШ будет выше, чем $\left(\sqrt{\tau}-1\right)^2$:

$$P\left(|N(\omega,t)|<\frac{|X(\omega,t)|}{\sqrt{\tau}}\right)=\frac{1}{\sqrt{2\pi}|\sigma_N(\omega)|}\int_{-\infty}^{\frac{|X(\omega,t)|}{\sqrt{\tau}-1}}\exp\left(\frac{(n-\mu_N(\omega))^2}{2\sigma_N^2(\omega)}\right)dn.$$

В данной формуле неизвестна функция $X(\omega,t)$. Для ее оценки решено использовать двухступенчатый способ TSNR [87].

В дальнейшем для разделения частотных интервалов на нужные (содержащие речевой сигнал) и не нужные (содержащие шум) нужно ввести пороговое значение вероятности $\eta\in[0;1]$, такое чтобы частотный интервал, содержащий речь, удовлетворял условию:

$$P(SNR_{prio}(\omega,t)>\left(\sqrt{\tau}-1\right)^2)>\eta.$$

1.6.2. Алгоритм шумоподавления Скалара на основе винеровской фильтрации

Алгоритм шумоподавления Скалара на основе винеровской фильтрации строится на принципе некоррелированности чистого речевого сигнала и воздействующего аддитивного шума.

При воздействии на речевой сигнал аддитивного шума зашумленную речь можно задать уравнением:

$$y(t)=x(t)+n(t),$$

где $y(t)$ – зашумленная речь, $x(t)$ – чистый речевой сигнал, $n(t)$ – воздействующий аддитивный шум. Данные сигналы можно представить в спектральном виде $Y(\omega,t)$, $X(\omega,t)$ и $N(\omega,t)$.

Целью данного алгоритма является определение модуля спектра чистого речевого сигнала $|X(\omega,t)|$ при условии некоррелированности данного сигнала с шумом (винеровская оценка). Достижение указанной цели возможно с помощью следующего выражения:

$$|X(\omega,t)|^2 = G(\omega,t)|Y(\omega,t)|^2,$$

где $G(\omega,t)$ – функция спектральной коррекции, $|Y(\omega,t)|$ – модуль спектра зашумленного сигнала.

Спектр зашумленного сигнала можно представить в виде:

$$Y(\omega,t) = X(\omega,t) + N(\omega,t),$$

где $N(\omega,t)$ – спектр аддитивного шума.

Для нахождения функции спектральной коррекции необходимо произвести следующие действия:

$$|Y(\omega,t)|^2 = Y(\omega,t)^2 = \big(X(\omega,t) + N(\omega,t)\big)^2,$$

$$|Y(\omega,t)|^2 = |X(\omega,t)|^2 + |N(\omega,t)|^2 + 2|X(\omega,t)\|N(\omega,t)|\cos(\alpha(\omega,t)), \quad (1.12)$$

где $\alpha(\omega,t)$ – разница фаз между $X(\omega,t)$ и $N(\omega,t)$.

Учитывая выражения для априорного и апостериорного ОСШ (1.4, 1.5) выражение (1.11) можно преобразовать к следующему виду:

$$SNR_{post}(\omega,t) = SNR_{prio}(\omega,t) + 1 + 2\sqrt{SNR_{prio}(\omega,t)}\cos(\alpha(\omega,t)). \quad (1.13)$$

Так как чистый сигнал и шум некоррелированны между собой, следовательно $\cos(\alpha(\omega,t)) = 0$. Учитывая данное свойство, выражение (1.13) можно переписать в следующем виде:

$$SNR_{post}(\omega,t) = SNR_{prio}(\omega,t) + 1,$$

$$SNR_{prio}(\omega,t) = SNR_{post}(\omega,t) - 1. \quad (1.14)$$

Поделив выражение (1.14) на $SNR_{post}(\omega,t)$, получим:

$$\frac{SNR_{prio}(\omega,t)}{SNR_{post}(\omega,t)} = \frac{SNR_{post}(\omega,t) - 1}{SNR_{post}(\omega,t)},$$

$$\frac{|X(\omega,t)|^2}{|Y(\omega,t)|^2} = \frac{SNR_{post}(\omega,t) - 1}{SNR_{post}(\omega,t)},$$

$$|X(\omega,t)|^2 = \frac{SNR_{post}(\omega,t) - 1}{SNR_{post}(\omega,t)}|Y(\omega,t)|^2 \quad (1.15)$$

.

Из выражения (1.15) следует, что функция спектральной коррекции определяется выражением:

$$G(\omega,t) = \frac{SNR_{post}(\omega,t) - 1}{SNR_{post}(\omega,t)}.$$

1.7. Выводы по главе

Проведенный в первой главе анализ актуальных задач машинного распознавания речи позволяет говорить о том, что для развития внедрения данной технологии актуальной задачей является дикторонезависимое распознавание речевых сигналов. Для отечественного рынка (рынка Российской Федерации) также востребована возможность работы таких систем с русской речью. Для решения данной задачи в настоящее время целесообразно использовать вероятностно-сетевую модель принятия решений, например, на основе нейросетевых методов. Данный метод актуален тем, что он максимально приближен к человеческой системе восприятия и понимания речевых сигналов, так как нейросетевой подход базируется на принципах работы человеческого мозга. В целом изучение

данного метода очень перспективно, потому что в мире нет искусственно созданных супер-ЭВМ превосходящих хотя бы 5 % производительности человеческого мозга [92].

Проведены сравнительный анализ акустических признаков звуков речи и оценка степени их применимости для решения задачи распознавания речи. Так как задача распознавания речевых сигналов сложная и требует больших вычислительных затрат, то решено было выбрать для представления речевых сигналов наиболее информационно-емкую и приближенную к принципам работы человеческой слуховой системы в виде MFCC коэффициентов.

Рассмотрены основы принципов построения ИНС и алгоритмы коллективного нейросетевого распознавания образов. В результате чего решено для дальнейших исследований взять за основу bagging алгоритм, так как данный алгоритм позволяет ускорить обучение всего нейросетевого алгоритма за счет распараллеливания процессов, а также повысить качество распознавания за счет коллективного голосования.

Для расширения области применения нейросетевых алгоритмов распознавания речевых сигналов в условиях шумов рассмотрено три алгоритма шумоподавления: IBM-PostSNR, IBM-TSNR и Wiener-PriorSNR. В дальнейших исследованиях в условиях шумов предполагается применить данные алгоритмы шумоподавления для повышения качества вероятности распознавания тестовых речевых сигналов.

ГЛАВА 2. РАЗРАБОТКА И ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВОГО АЛГОРИТМА ДИКТОРОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

В настоящее время растет важность массового внедрения новых интерфейсов взаимодействия человека с техническими системами, поскольку традиционные интерфейсы во многом уже достигли своего совершенства, а вместе с ними и своих пределов [42]. При традиционно высокой значимости информации, поступающей к нам через органы зрения, и её высокой доли среди всей сенсорной информации, считающейся равной порядка 85% [58], этот канал восприятия человека становится в значительной степени перегружен. И первоочередной альтернативой здесь видится коммуникация именно по акустическому каналу.

Существует много алгоритмов распознавания слов в речи, но все они могут быть отнесены к одному из двух классов: генеративного и дискриминативного алгоритмов распознавания.

Среди классов генеративных алгоритмов распознавания, наиболее популярными являются скрытые Марковские модели [89] и композиционно-оптимальный подход [7]. Известны более или менее успешные попытки использования этих алгоритмов для словарного распознавания речевых сигналов. Среди таких попыток являются эксперименты по построению систем распознавания речи для английского китайского, русского и тамильского языков [94, 101].

Главным принципом работы, как скрытых Марковских моделей, так и композиционно-оптимального подхода является генерация максимально правдоподобных эталонных сигналов на основе некоторой автоматной грамматики и сопоставление полученных эталонов с речевыми сигналами распознавания. Такой принцип обуславливает как преимущества, так и

недостатки этих алгоритмов. К важным преимуществам генеративных алгоритмов следует отнести результативное моделирование процессов, нелинейно изменяющихся во времени, а к недостаткам можно отнести не очень высокую дискриминантную способность.

К противоположному классу – дискриминативных алгоритмов – относятся алгоритмы, основанные на построении границ между классами распознавания в пространстве признаков. Наиболее распространенным математическим аппаратом для разработки дискриминативных алгоритмов распознавания служат искусственные нейронные сети. Главные достоинства этого математического аппарата в том, что [26]:

– многослойные нейронные сети обладают высокой дискриминантной способностью;

– нейронная сеть во время обучения может найти наилучшую комбинацию ограничений для классификации образов, и при этом нет необходимости в строгих предположениях о распределении входных признаков (что необходимо, например, в скрытых Марковских моделях);

– нейросетевой алгоритм характеризуется хорошими скоростными характеристиками за счет высокой степени параллелизма вычислений.

К недостаткам нейронных сетей можно отнести то, что с помощью этого математического аппарата трудно моделировать высокую временную вариантность распознающих сигналов.

2.1. Алгоритм базового нейросетевого распознавания

Структурная схема базового варианта нейросетевого алгоритма распознавания слов в речевом сигнале, независимого от диктора на примере многослойного персептрона приведена на рисунке 2.1 [39]. Под базовым нейросетевым алгоритмом подразумевается система, в которую входят: один нейросетевой распознаватель; блок вычисления мел-частотных кепстральных коэффициентов (MFCC) с заданными выходными характеристиками и блок селектора слов по степени достоверности. В

качестве нейросетевого распознавателя могут выступать различные многослойные нейронные сети, такие как, например: многослойный персептрон, рекуррентная многослойная сеть Эльмана.

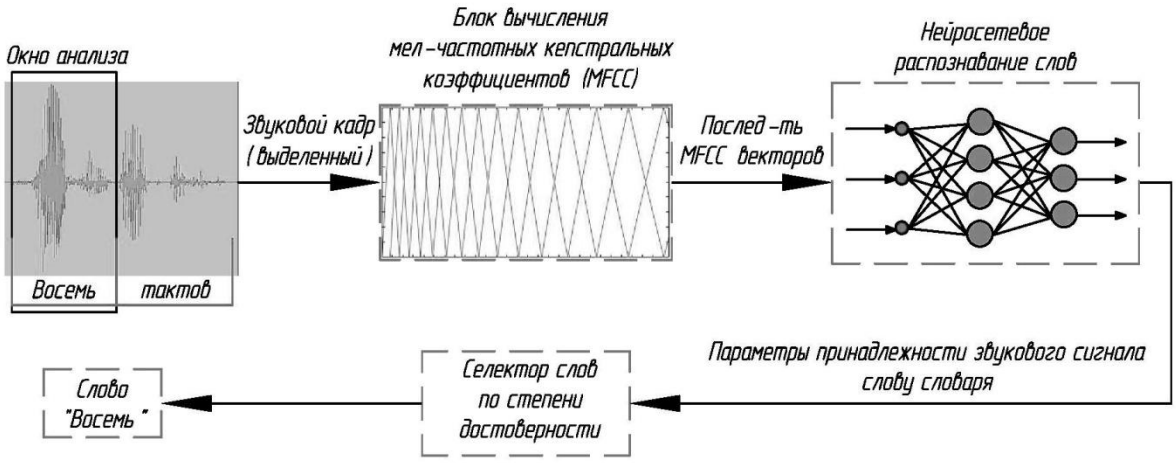

Рис. 2.1. Структурная схема базового нейросетевого алгоритма распознавания слов в речевом сигнале

В качестве параметров речевого сигнала, по которым проводится распознавание, выбран логарифм энергии сигнала по J мел-частотным кепстральным коэффициентам (Mel Frequency Cepstral Coefficients, или MFCC) [74]. MFCC-вектор вычисляется в каждом окне. Компоненты каждого MFCC-вектора нормализуются так, чтобы математическое ожидание каждого компонента стало нулевым, а среднеквадратичное отклонение – единичным. Распознаваемый речевой образ представляет собой последовательность из произведения $J \times K$ нормализованных MFCC-векторов, где J – желаемое число коэффициентов, K – число окон в каждом исследуемом сигнале.

В качестве примера нейронной сети, которая решает задачу распознавания слов, показана на рисунке 2.1 сеть типа "многослойный персептрон" – классическая многослойная сеть с полными последовательными связями и сигмоидальными функциями активации

нейронов. Известно, что двухслойный персептрон может аппроксимировать непрерывную функцию любой сложности, в том числе и функцию, которая описывает нелинейную гиперповерхность, которая разделяет в пространстве признаков отдельные классы образов. Однако более результативным аппроксиматором является трехслойный персептрон, особенно если классы распознавания образуют в пространстве признаков сложные многосвязные участки [86]. Исходя из этого, для распознавания слов был избран многослойный персептрон с R слоями нейронов – один входной, $R-2$ скрытых и один выходной. В дальнейшем в п. 2.3.3. проведено исследование параметра R для нескольких нейросетевых алгоритмов на предмет получения лучших результатов вероятности дикторонезависимого распознавания десяти классов речевых сигналов.

2.2. Алгоритмы коллективного нейросетевого распознавания

2.2.1. Алгоритм коллективного нейросетевого распознавания с обучением SCG

Для повышения точности распознавания слов предложено объединить отдельные нейросетевые распознаватели в единую систему по принципам коллективного равноправного голосования метода bagging [62]. Достоинства выбранного метода были перечислены в п. 1.5.

Вероятность распознавания речевых сигналов алгоритмом bagging P_{bag} можно описать следующим образом:

$$P_{bag} = \frac{\sum_{z=1}^{H} P_{bag}^{z}}{H}, \quad z = 1,\ldots,H,$$

$$P_{bag}^{z} = \frac{\sum_{q=1}^{L} P_{q}^{z}}{L}, \quad q = 1,\ldots,L$$

где P_q^z – вероятность распознавания речевого сигнала z нейронной сетью q, L – количество нейросетевых распознавателей, входящих в алгоритм bagging, H – количество речевых сигналов в речевом корпусе.

В качестве алгоритма обучения сетей, выбран алгоритм масштабируемых сопряженных градиентов (Scaled Conjugate Gradient Backpropagation, SCG) [82], т.к. он стабильный и очень быстрый. В стандартной форме алгоритма сопряженных градиентов требуется использование линейного поиска, что из-за его характера «проб и ошибок» может занять много времени. В модифицированной (данной) версии алгоритма сопряженных градиентов линейный поиск отсутствует. Линейный поиск заменен одномерной формой Левенберга-Марквардта. Основанием для использования именно этого метода было желание обойти сложности, вызываемые не положительностью матрицы Гессе. Формула обновления коэффициентов данного метода:

$$w_{k+1} = w_k + \alpha_k p_k,$$

$$\alpha_k = \frac{\mu_k}{\delta_k},\ \mu_k = -p_k^T E_k',\ \delta_k = p_k^T s_k,$$

$$s_k = E''(w_k)p_k = \frac{E'(w_k + \sigma_k p_k) - E'(w_k)}{\sigma_k} + \lambda_k p_k$$

где $k = 1,2,\ldots,N$, α_k – размер шага, p_k – сопряженный вектор, E_k – вектор ошибки, δ_k – матрица Гессе, $\sigma_k = \frac{\sigma}{|p_k|}$, λ_k – параметр масштабирования матрицы Гессе.

Структурная схема алгоритма коллективного нейросетевого распознавания приведена на рисунке 2.2.

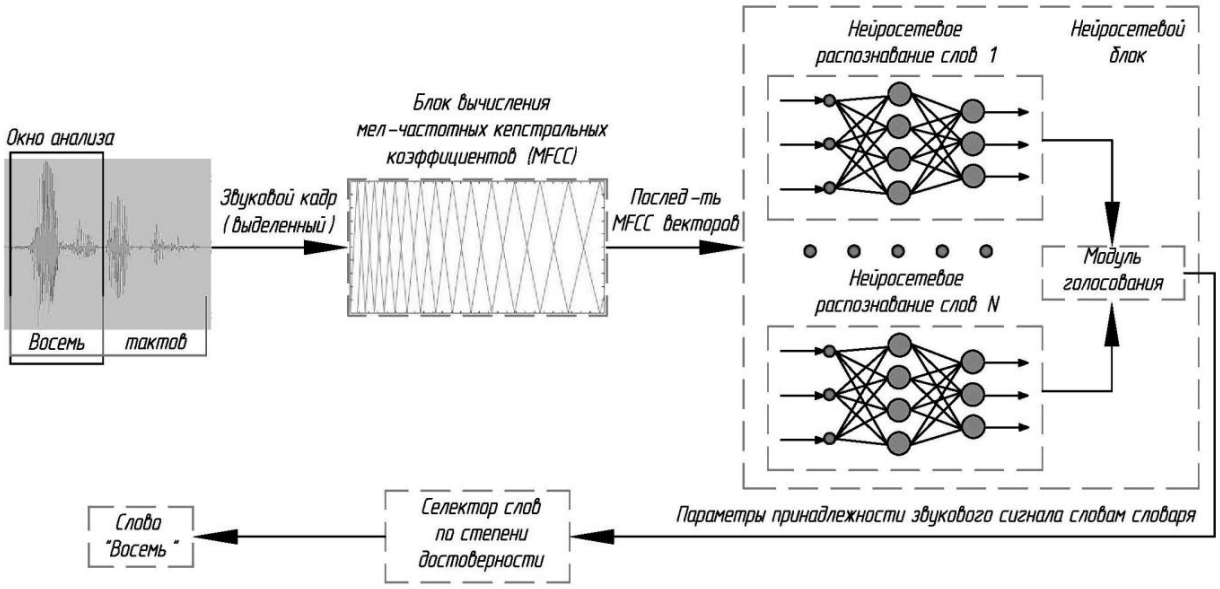

Рис. 2.2. Структурная схема коллективного нейросетевого алгоритма распознавания слов на основе многослойных персептронов

2.2.2. Модифицированный алгоритм коллективного нейросетевого распознавания

Для увеличения технических возможностей распознавания речевых сигналов коллективного нейросетевого алгоритма предложено bagging-алгоритм модифицировать. Данное улучшение алгоритма должно позволить увеличить размер словаря без потери качества дикторонезависимого распознавания речевых сигналов. Данное улучшение позволит расширить сферу применения распознавания речевых сигналов.

Для построения модифицированного bagging-алгоритма предполагается использовать в качестве основного элемента нейросетевой блок коллективного голосования (рис. 2.3). Один нейросетевой блок способен обучиться и распознать речевые сигналы без потери качества распознавания на словаре ограниченной размерности [31]. В данном алгоритме предполагается использовать L нейросетевых блоков.

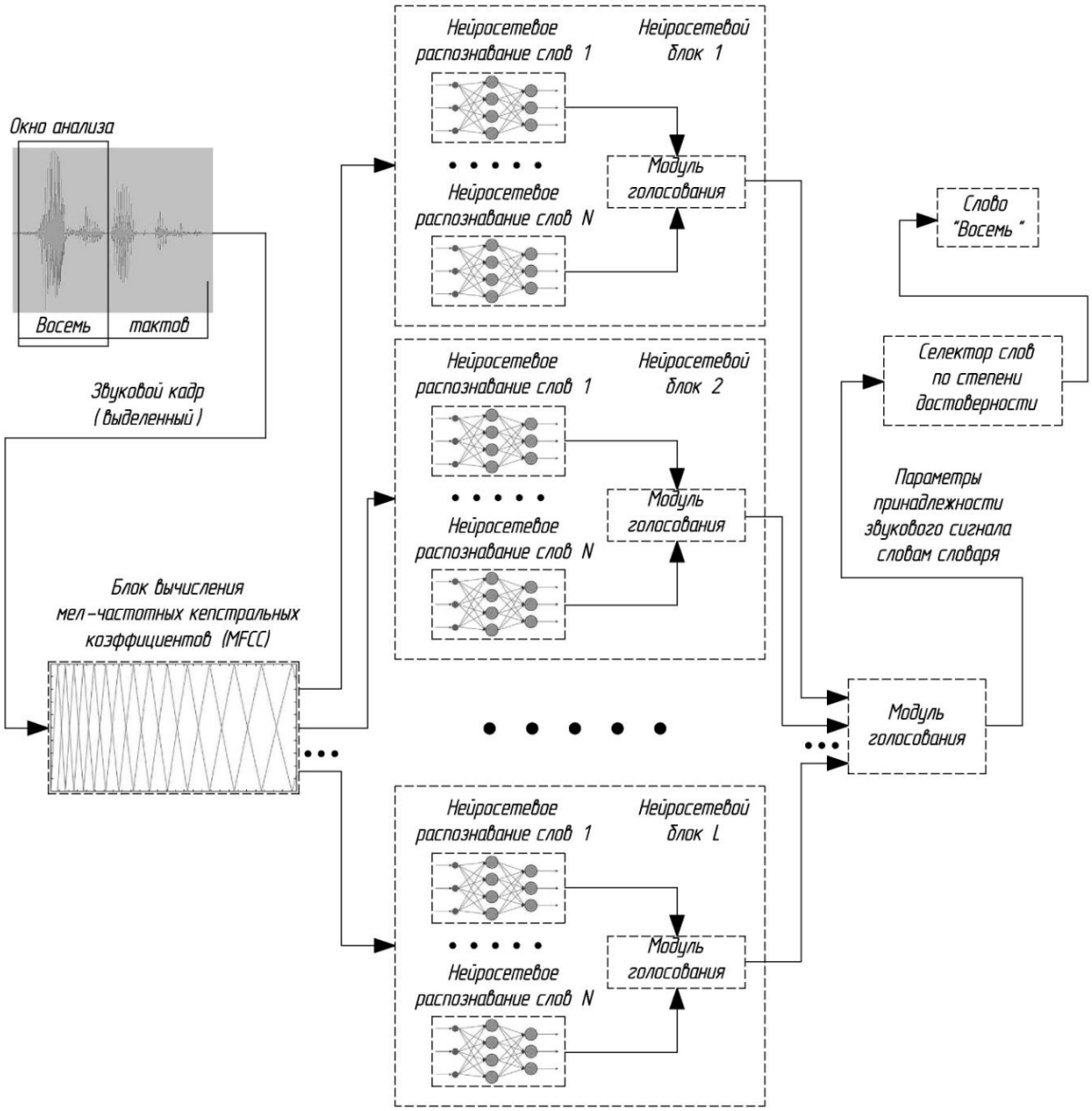

Рис. 2.3. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме тестирования

Количество нейросетевых блоков пропорционально размеру словаря для того, чтобы качество распознавания речевых сигналов оставалось высоким. То есть количество нейросетевых блоков L равняется требуемому размеру словаря Ω, деленному на размер словаря одного нейросетевого блока U:

$$L = \frac{\Omega}{U}.$$

Принцип работы данного алгоритма заключается в следующем. Допустим, требуется построить систему распознавания речевых сигналов размерности словаря Ω. Т.к. один нейросетевой блок не способен обучиться без потери качества распознавания на словаре размерности более определенного размера U для данного нейросетевого блока [31], то предполагается разбить словарь на L словарей с размерностью не более определенного размера словаря для одного нейросетевого блока. То есть предполагается обучить каждый нейросетевой блок на словаре отличном от обучаемых словарей других нейросетевых блоков. И каждый нейросетевой блок обучается на словаре с размерностью, не превышающей определенного размера словаря для данного нейросетевого блока. В качестве параметров речевого сигнала, по которым проводится распознавание, выбран логарифм энергии сигнала по J мел-частотным кепстральным коэффициентам [74]. Преобразование входного речевого сигнала в массив MFCC-коэффициентов производится так же, как и в базовом нейросетевом алгоритме распознавания. Данный способ описан в п. 2.1.

Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме обучения представлена на рисунке 2.4. В состав схемы (рис. 2.4) исследуемого алгоритма входит L нейросетевых блоков, каждый из которых состоит из N нейросетевых распознавателей. В качестве нейросетевого распознавателя слов берется стандартная нейросетевая многослойная сеть, такая как, например: многослойный персептрон, рекуррентная многослойная сеть Эльмана. Количество нейросетевых блоков берется исходя из размера обучающего множества классов (речевых сигналов) по правилу, описанному выше.

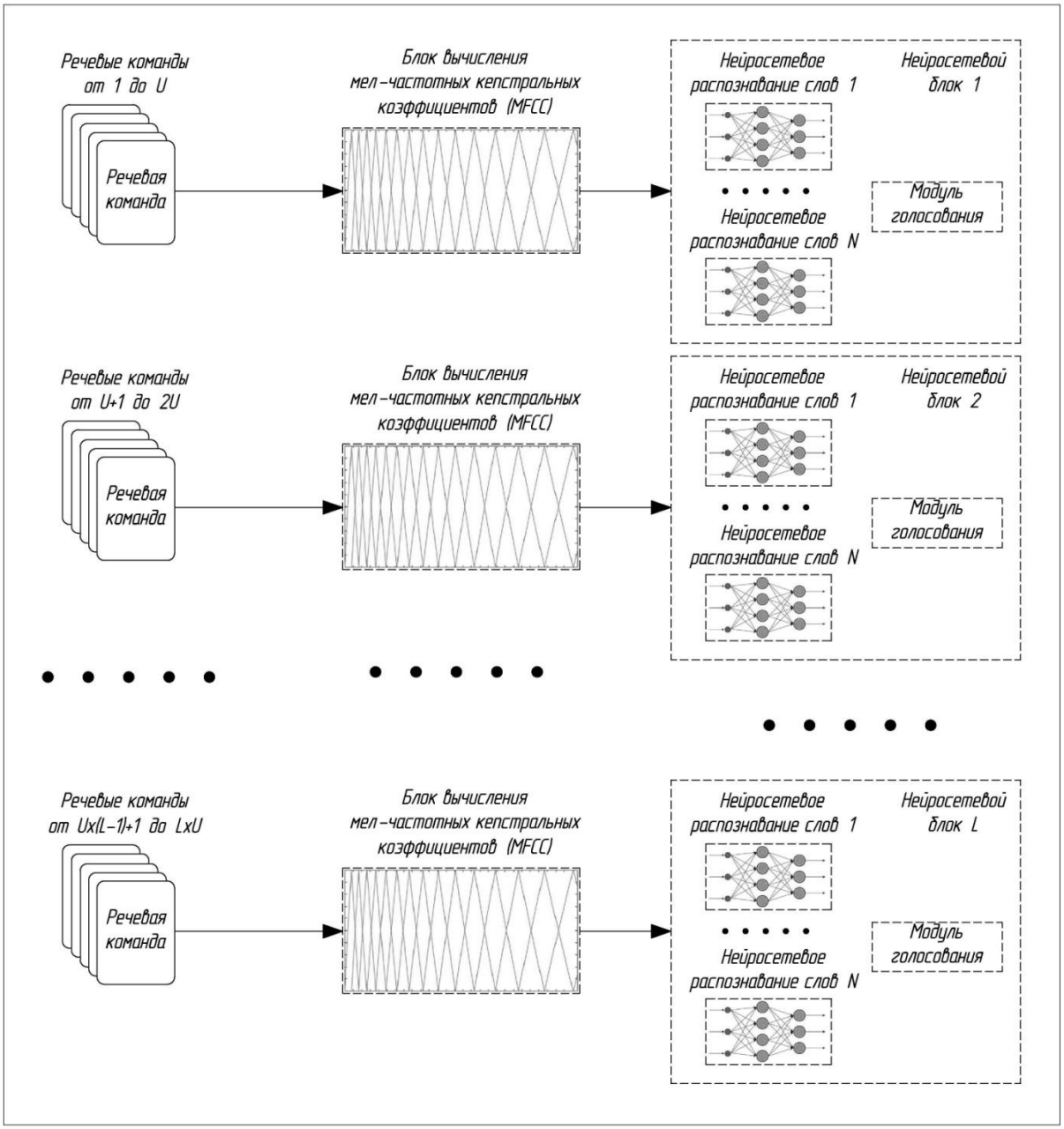

Рис. 2.4. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме обучения

После обучения каждого нейросетевого блока на соответствующем словаре возможно тестирование всего алгоритма. При тестировании на вход системы, изображенной на рисунке 2.3, подается любой речевой сигнал, имеющийся в исследуемом словаре с размерностью Ω. В блоке вычисления мел-частотных кепстральных коэффициентов из данного речевого сигнала вычисляется массив $J \times K$ MFCC-коэффициентов

данного речевого сигнала. Затем полученные MFCC-коэффициенты поступают на входы всех имеющихся в данном алгоритме нейросетевых блоков. Если нейросетевой блок обучался на словаре, содержащем данный речевой сигнал, то данный нейросетевой блок с определенной вероятностью ее распознает. А если нейросетевой блок не обучался на словаре, содержащем данный речевой сигнал, то данный нейросетевой блок с достаточной вероятностью ее не распознает. Далее информация о распознании речевого сигнала с нейросетевых блоков поступает на модуль голосования. От модуля голосования информация в виде параметров принадлежности звукового сигнала слову словаря поступает на блок селектора слов по степени достоверности. Селектор слов по степени достоверности определяет итоговый результат тестирования данного алгоритма.

2.3. Исследование нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов

Исследование состоит из пяти серий экспериментов, состоящих в определении:

– размера нейросетевого bagging-коллектива;

– количества обучающих дикторов и сравнении одиночного нейросетевого распознавания и bagging-коллектива нейросетевого распознавания;

– количества слоев для нейросетевого алгоритма bagging-коллектива;

– количества речевых сигналов для нейросетевого алгоритма bagging-коллектива.

Также проводится исследование работы модифицированного нейросетевого алгоритма bagging-коллектива.

Для исследования каждого алгоритма произведена выборка значений, показывающих результат распознавания исследуемым алгоритмом тестируемые значения, равная *V* измерений:

$$V = Z \times A \times B,$$

где Z – число исследуемых сигналов, A– количество исследуемых тестируемых записей одного сигнала, B – число проведенных экспериментов над всеми тестируемыми речевыми сигналами. Во всех экспериментах $A = 50,\ B = 3$ и $Z = \overline{5,102},$ следовательно $V = \overline{750,15300}.$ При произведенной выборке частота распознавания приближенно равна вероятности распознавания.

2.3.1. Выбор размера нейросетевого bagging-коллектива в задаче дикторонезависимого распознавания речевых сигналов

Предполагается определить размер нейросетевого bagging-коллектива (рис. 2.2), при котором вероятность распознавания речевых сигналов достигает порогового значения, после которого с ростом размера bagging-коллектива данная вероятность растет медленно.

В качестве материала для данных экспериментов использовался речевой корпус «Г» речевой базы «КРИПТОН-01» на основе собственных записей [34, 40] (приложение № 1), содержащий более получаса звукозаписей различных русскоязычных фраз, которые были записаны 8 дикторами. Речевой корпус разбит разработчиками на два непересекающихся множества: учебное и тестовое. В качестве обучающих дикторов взяты люди разного пола (50 % мужчины – 2 человека, 50 % женщины – 2 человека), разного возраста (17-30 лет) и разного эмоционального состояния. В качестве тестирующих дикторов взяты люди разного пола (75 % мужчины – 3 человека, 25 % женщины – 1 человек), разного возраста (18-28 лет) и разного эмоционального состояния. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (4 диктора), а оценка точности распознавания – на тестовом подмножестве (остальные 4 диктора). Запись сигналов производилась на микрофон BBK dm-150 в условиях малого «повседневного» белого шума. В качестве сигналов были взяты

произношения цифр от «0» до «9», которые каждый обучающий диктор произнес по 12 раз и каждый тестирующий диктор также произнес по 12 раз в разном эмоциональном состоянии.

Параметрами речевого сигнала, по которым проводится обучение и тестирование нейронных сетей является логарифм энергии сигнала по 13 мел-частотным кепстральным коэффициентам [74]. Распознаваемый речевой образ представляет собой последовательность из $J \cdot K = 13 \cdot 29 = 377$ нормализованных MFCC-векторов, где J – желаемое число коэффициентов, K – число окон в каждом исследуемом сигнале. Мел-кепстральное представление обучающего «Г.1» и тестирующего «Г.2» разделов речевой базы «Г» (приложение № 1), записанных восьмью разными дикторами. Вид мел-кепстрального представления обучающего раздела «Г.1» речевого корпуса «Г» речевой базы «КРИПТОН-01» показано на рисунке 2.5.

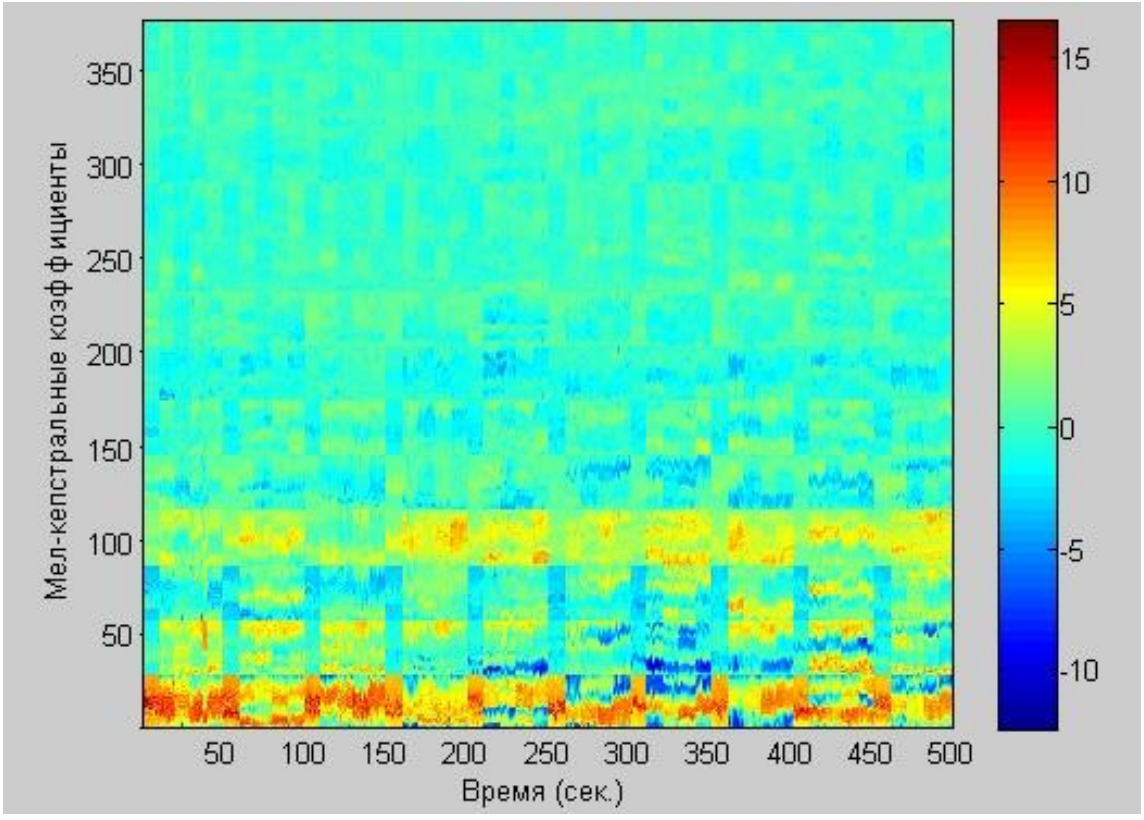

Рис. 2.5. Мел-кепстральное представление обучающего раздела «Г.1» речевого корпуса «Г» речевой базы «КРИПТОН-01»

Вид мел-кепстрального представления тестирующего раздела «Г.2» речевого корпуса «Г» речевой базы «КРИПТОН-01» показано на рисунке 2.6.

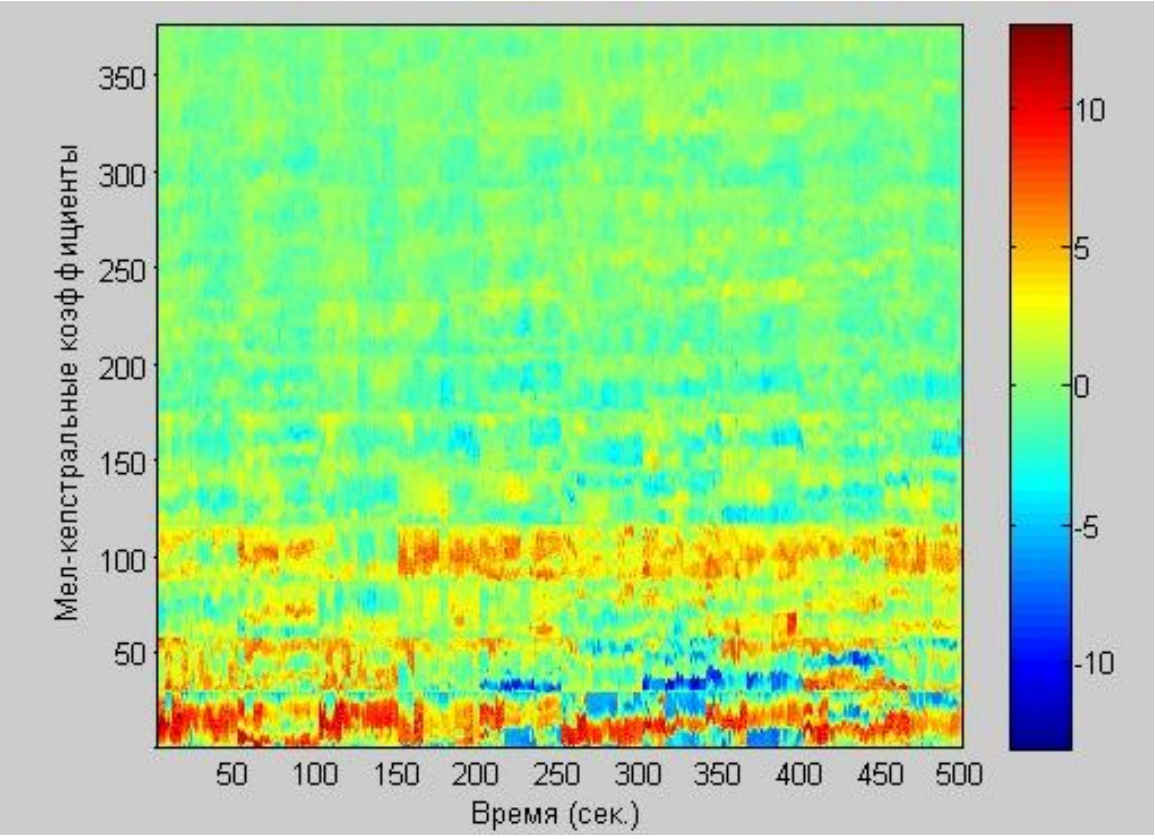

Рис. 2.6. Мел-кепстральное представление тестирующего раздела «Г.2» речевого корпуса «Г» речевой базы «КРИПТОН-01»

В экспериментах исследуются восемь bagging-алгоритмов коллективного дикторонезависимого нейросетевого распознавания с размерностью: 1, 2, 5, 10, 20, 30, 40, 50. При размерности $L=1$ алгоритм коллективного нейросетевого распознавания становится базовым алгоритмом. В bagging-алгоритме в качестве распознавателя выбран десятислойный персептрон Розенблатта [86]. Алгоритмом обучения сетей выбран алгоритм масштабируемых сопряженных градиентов (Scaled Conjugate Gradient Backpropagation, SCG) [82].

Результаты экспериментов приведены на рисунке 2.7, на котором представлена зависимость вероятности распознавания речевых сигналов от размера bagging-коллектива нейросетевых распознавателей.

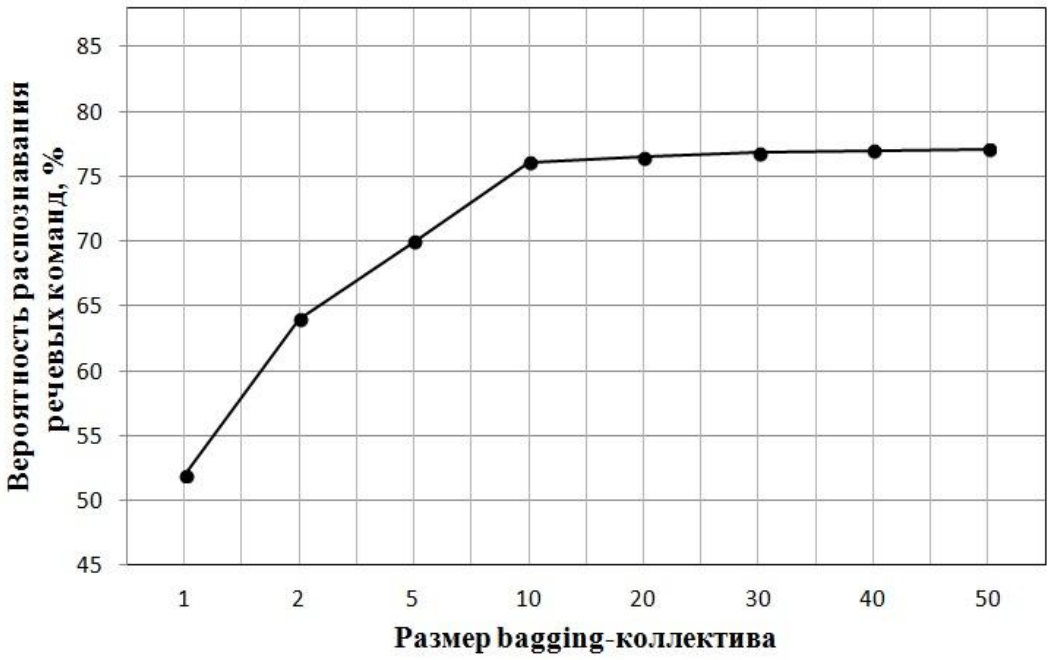

Рис. 2.7. Зависимость вероятности распознавания речевых сигналов от размера bagging-коллектива нейросетевых распознавателей

Из данных результатов следует, что при десяти нейросетевых распознавателях происходит насыщение вероятности распознавания речевых сигналов. Данная вероятность равна 76,1 % правильно распознавания тестируемого множества. При количестве нейросетевых распознавателей более десяти вероятность распознавания речевых сигналов растет медленно относительно до пороговых значений. С увеличением размерности коллективного распознавания увеличивается в арифметической прогрессии вычислительная сложность распознавания речевых сигналов. Следовательно, из полученных результатов можно сделать вывод, что для дальнейших исследований лучше выбирать bagging-коллектив с размерностью десять.

2.3.2. Выбор количества обучающих дикторов в задаче дикторонезависимого распознавания речевых сигналов

Предполагается определить количество обучающих дикторов, при котором вероятность распознавания речевых сигналов достигает порогового значения, после которого с ростом числа обучающих дикторов данная вероятность растет медленно [32]. Также предполагается сравнение базового нейросетевого алгоритма и нейросетевого алгоритма bagging из 10 нейросетевых распознавателей. В bagging-алгоритме в качестве распознавателя выбран десятислойный персептрон Розенблатта [86]. Для обучения сетей персептрона Розенблатта, выбран алгоритм SCG [33, 82].

В качестве материала для данных экспериментов использовалось 12 речевых корпусов на основе собственных записей [34, 40], содержащие около восьми часов звукозаписей различных русскоязычных фраз, которые были записаны 24 дикторами. Речевой корпус разбит разработчиками на два непересекающихся множества: учебное и тестовое. В качестве обучающих и тестирующих дикторов взяты люди разного пола, разного возраста и разного эмоционального состояния. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (12 дикторов), а оценка точности распознавания – на тестовом подмножестве (остальные 12 дикторов). Запись речевых сигналов производилась на микрофон BBK dm-150 в условиях малого «повседневного» белого шума. В качестве речевых сигналов были взяты произношения цифр от «0» до «9». В исследованиях введено обозначение речевых корпусов заглавной буквой и обозначение раздела корпуса заглавной буквой и цифрой через точку. Цифра в данном случае обозначает учебное или тестовое множество речевых сигналов, где «1» обозначает учебное множество, а «2» – тестовое множество. То есть каждый корпус состоит из двух разделов. Речевые характеристики дикторов и информация о речевых корпусах представлены в приложении № 1.

С помощью данных речевых корпусов произведено сравнение базового нейросетевого распознавания и нейросетевого алгоритма bagging-коллектива из десяти нейросетевых распознавателей (рис. 2.8).

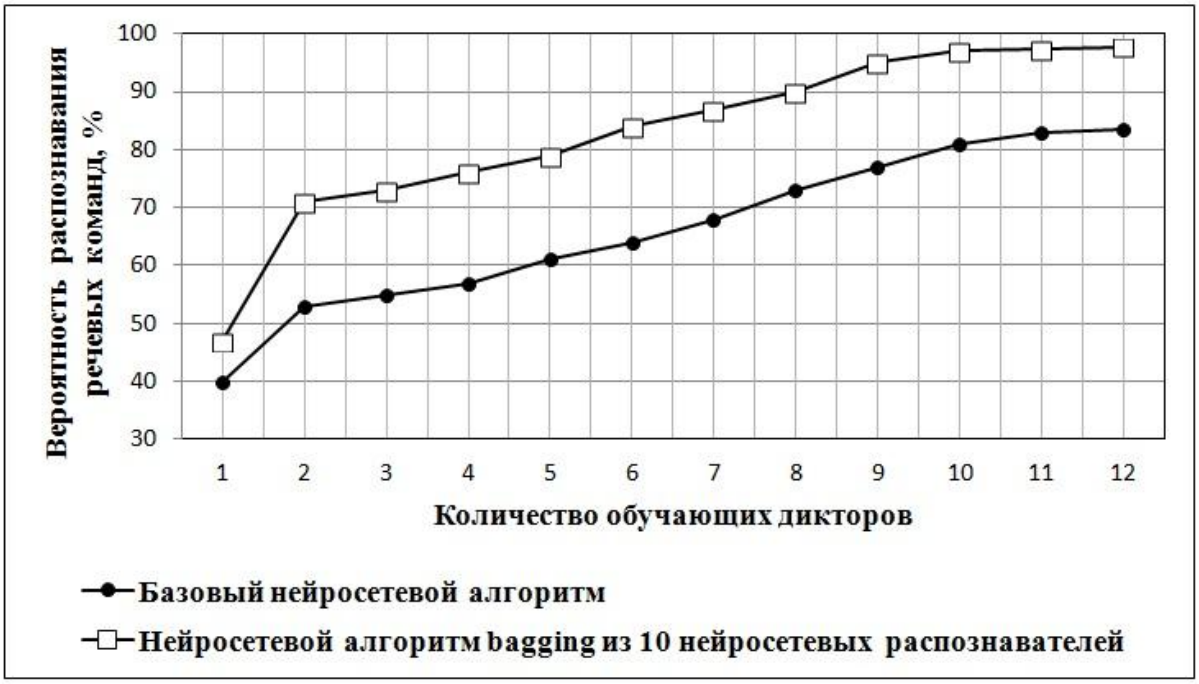

Рис. 2.8. Зависимость вероятности распознавания речевых сигналов от количества обучающих дикторов

Из данных результатов следует, что при 10 обучающих дикторах происходит насыщение вероятности распознавания речевых сигналов коллективного нейросетевого алгоритма bagging. Данная вероятность равна 97,1 % правильно распознанного тестируемого множества. Также исследования показывают насыщение вероятности распознавания речевых сигналов базового нейросетевого алгоритма при 11 дикторах. Вероятность распознавания речевых сигналов базового нейросетевого алгоритма при 11 дикторах равна 83 % правильного распознанного тестируемого множества. При числе обучающих дикторов более 11 вероятность распознавания речевых сигналов растет медленно относительно допороговых значений.

Также из данных экспериментов следует, что нейросетевой алгоритм bagging из 10 нейросетевых распознавателей гораздо результативнее базового нейросетевого алгоритма.

2.3.3. Выбор количества слоев нейросетевого алгоритма bagging-коллектива

Известно, что многослойная нейронная сеть может смоделировать функцию практически любой степени сложности [15, 24], причем число слоев и число нейронов в каждом слое определяют сложность функции. Предполагается определить количество слоев нейронной сети, при котором вероятность распознавания речевых сигналов достигает порогового значения, после которого с ростом числа слоев данная вероятность растет не значительно. В данных экспериментах исследуется количество слоев для персептрона Розенблатта [86] и сети Эльмана [69, 84] в задаче дикторонезависимого распознавания речевых сигналов.

В качестве материала для данных экспериментов использовался собственный речевой корпус «К» речевой базы «КРИПТОН-01» на основе собственных записей [34, 40] (приложение № 1). Данный корпус содержит более получаса звукозаписей различных речевых сигналов (на русском языке), которые были записаны 20 дикторами. Речевой корпус разбит разработчиками на два непересекающихся множества: учебное и тестовое. В качестве обучающих дикторов взяты люди разного пола (70 % мужчины – 7 человек, 30 % женщины – 3 человека), разного возраста (17-38 лет) и разного эмоционального состояния. В качестве тестирующих дикторов взяты люди разного пола (80 % мужчины – 8 человек, 20 % женщин – 2 человека), разного возраста (18-35 лет) и разного эмоционального состояния. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (10 дикторов), а оценка точности распознавания – на тестовом подмножестве (остальные 10 дикторов). Запись сигналов производилась на микрофон BBK dm-150 в

условиях малого «повседневного» белого шума. В качестве речевых сигналов были взяты произношения цифр от «0» до «9», которые каждый обучающий диктор произнес по 5 раз, и каждый тестирующий диктор также произнес по 5 раз.

С помощью данного речевого корпуса оценено количество слоев персептрона Розенблатта и сети Эльмана в bagging-алгоритме, состоящем из 10 одинаковых подобных сетей, в задаче дикторонезависимого распознавания речевых сигналов. В экспериментах исследуются 10 персептронов и 10 сетей Эльмана с числом слоев от 1 до 17. Для обучения сетей персептрона Розенблатта, выбран алгоритм SCG [82]. Для обучения сетей Эльмана, выбран алгоритм градиентного спуска с учетом моментов и с адаптивным обучением (Gradient Descent Backpropagation with Adaptive Learning Rate, GDX) [26].

Результаты экспериментов приведены на рисунке 2.9.

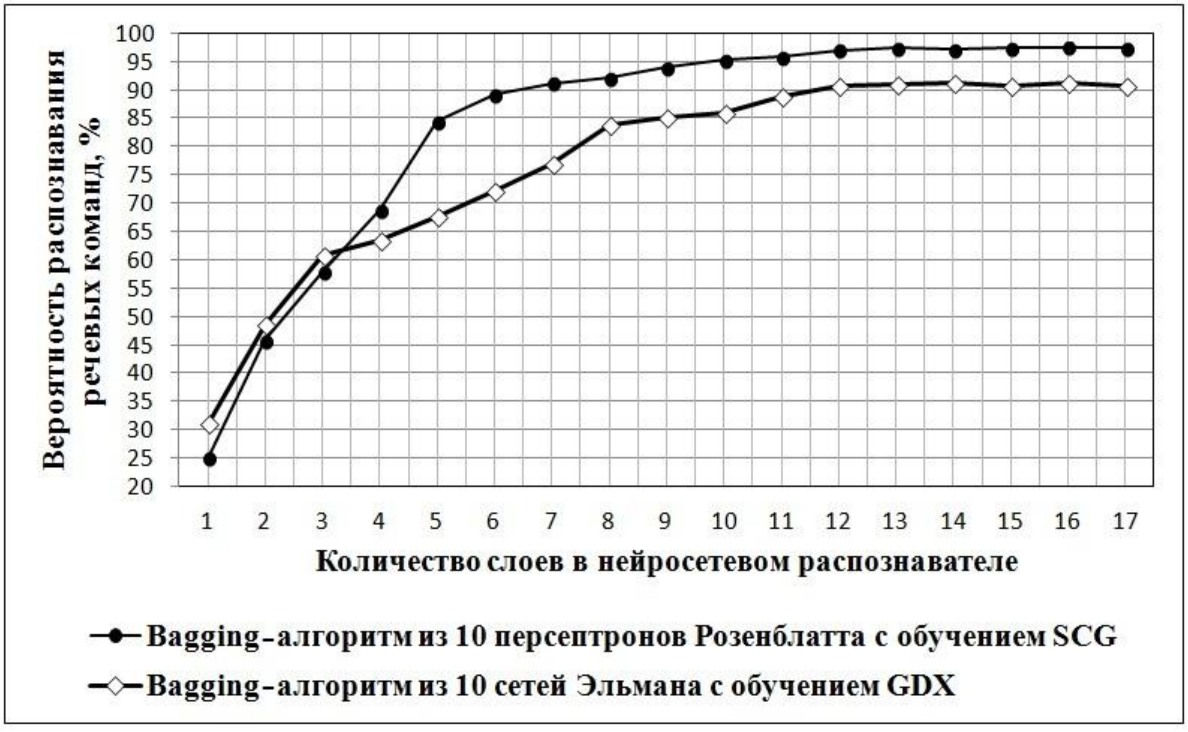

Рис. 2.9. Зависимость вероятности распознавания речевых сигналов от количества слоев нейросетевых распознавателей

Из данных результатов следует, что при 12 слоях нейросетевых распознавателей происходит насыщение вероятности распознавания речевых сигналов исследуемых коллективных нейросетевых алгоритмов bagging. У bagging-алгоритма из 10 персептронов Розенблатта с обучение SCG данная вероятность равна 97 % правильно распознанного тестируемого множества. У bagging-алгоритма из 10 сетей Эльмана с обучение GDX данная вероятность равна 90,7 % правильно распознанного тестируемого множества. При числе слоев нейросетевых распознавателей более 12 вероятность распознавания речевых сигналов растет медленно относительно допороговых значений. Из рисунка 2.7 следует, что при небольшом количестве слоев (от 1 до 3) нейросетевых распознавателей алгоритм bagging из 10 сетей Эльмана с обучением GDX лучше алгоритма bagging из 10 персептронов Розенблатта с обучение SCG. Также из данных экспериментов следует, что при большом количестве слоев нейросетевых распознавателей (от 4 до 17) нейросетевой алгоритм bagging из 10 персептронов Розенблатта с обучение SCG гораздо результативнее bagging алгоритма из 10 сетей Эльмана с обучением GDX.

2.3.4. Выбор размера словаря коллективных нейросетевых алгоритмов

Предполагается определить количество речевых сигналов, при которых вероятность распознавания речевых сигналов достигает порогового значения, после которого с ростом числа речевых сигналов данная вероятность резко падает. В данных экспериментах исследуется количество речевых сигналов для задачи дикторонезависимого распознавания речевых сигналов в различных коллективных нейросетевых алгоритмах. В качестве коллективных нейросетевых алгоритмов выбраны bagging-коллектив из 10 персептронов Розенблатта [86] с обучением SCG и bagging-коллектив из 10 сетей Эльмана [69, 84] с обучением GDX.

В качестве материала для данных экспериментов использовался собственный речевой корпус «С» речевой базы «КРИПТОН- 02» на основе собственных записей [34, 40] (приложение № 2). Данный корпус содержит более 8 часов звукозаписей различных речевых сигналов (на русском языке), которые были записаны 20 дикторами. Речевой корпус «С» разбит разработчиками на два непересекающихся множества: учебное и тестовое. В качестве обучающих дикторов взяты люди разного пола (70 % мужчины – 7 человек, 30 % женщины – 3 человека), разного возраста (17-38 лет) и разного эмоционального состояния. В качестве тестирующих дикторов взяты люди разного пола (80 % мужчины – 8 человек, 20 % женщин – 2 человека), разного возраста (18-35 лет) и разного эмоционального состояния. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (10 дикторов), а оценка точности распознавания – на тестовом подмножестве (остальные 10 дикторов). Запись речевых сигналов производилась на микрофон BBK dm-150 в условиях малого «повседневного» белого шума. В качестве речевых сигналов были взяты произношения 102 распространенных речевых сигналов (приложение № 2). Данные речевые сигналы каждый обучающий диктор произнес по 5 раз и каждый тестирующий диктор также произнес по 5 раз.

В экспериментах исследуются 10 вариаций (частей) речевой базы «КРИПТОН-02». Данные вариации содержат различное количество различных речевых сигналов: 5, 7, 10, 12, 14, 15, 20, 30, 50, 102. С помощью данной речевой базы оценен максимальный размер словаря для нейросетевого алгоритма bagging-коллектива, при котором вероятность дикторонезависимого распознавания речевых сигналов сохраняет высокое значение.

Результаты экспериментов приведены на рисунке 2.10, на котором представлена зависимость вероятности распознавания речевых сигналов от размера словаря.

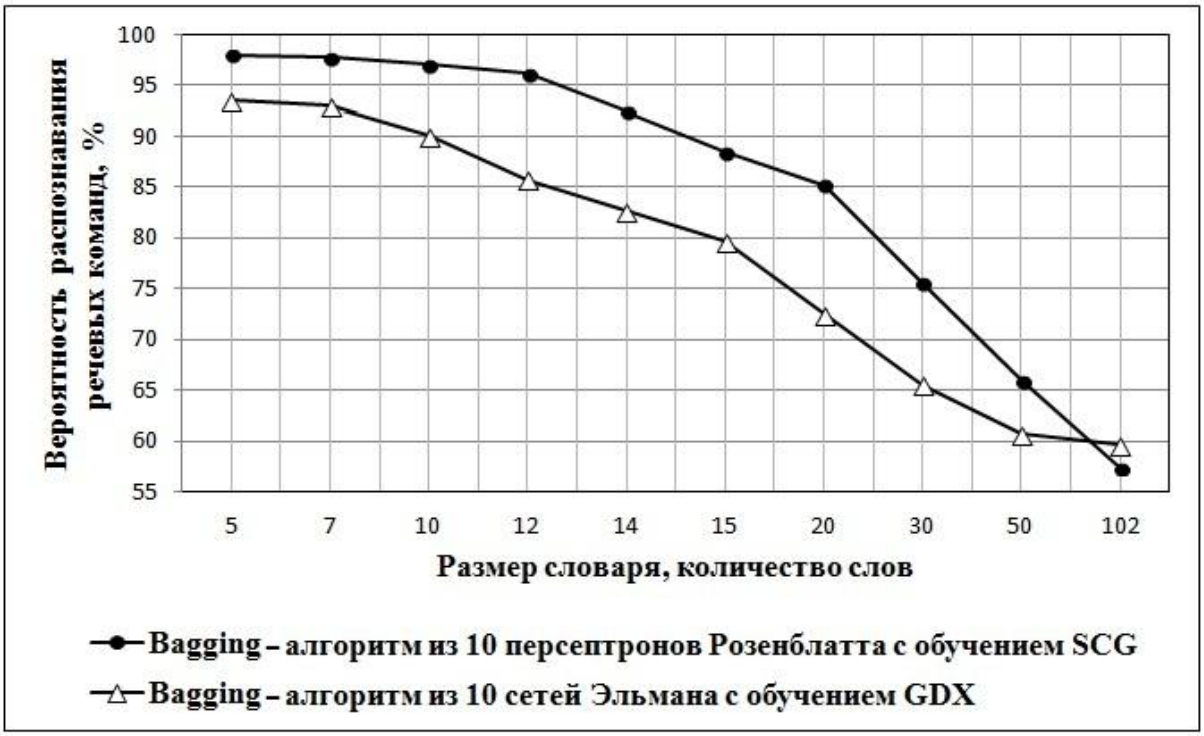

Рис. 2.10. Зависимость вероятности распознавания речевых сигналов от размера словаря

Из данных результатов следует, что при размере словаря более определенного количества классов (слов) у исследуемых нейросетевых алгоритмов происходит быстрый спад вероятности распознавания речевых сигналов. У алгоритма bagging-коллектива из 10 персептронов Розенблатта с обучением SCG данный спад наступает при размере словаря более 12 классов (видов речевых сигналов), а у алгоритма bagging-коллектива из 10 сетей Эльмана с обучением GDX данный спад наступает при размере словаря более 7 классов. Данные исследования показали, что вероятность распознавания речевых сигналов у алгоритма bagging-коллектива из 10 персептронов Розенблатта с обучением SCG при размере словаря 5, 7, 10 и 12 классов равна соответственно 98 %, 97,8 %, 97,1 % и 96,1 %. Вероятность распознавания речевых сигналов у bagging-коллектива из 10 сетей Эльмана с обучением GDX при 7 классах равна 93%. Из данных показателей видно, что первый алгоритм существенно результативней

второго. Также из экспериментов следует, что данные нейросетевые алгоритмы применимы для качественного распознавания речевых сигналов только для не большого размера словаря. При большом размере словаря данные алгоритмы справляются с задачей распознавания речи с малым показателем вероятности распознавания речевых сигналов, что существенно влияет на качество решаемой данной задачи.

2.3.5. Исследование работы модифицированных алгоритмов коллективного нейросетевого распознавания

В данном исследовании предполагается оценить работу модифицированных алгоритмов коллективного нейросетевого распознавания на большом словаре. В качестве основного элемента взят bagging-коллектив. Предполагается исследовать модифицированные алгоритмы на основе двух разновидностей нейронных сетей: персептрона Розенблатта с обучением SCG [82] и рекуррентной сети Эльмана с обучением GDX [26].

В качестве материала для данных экспериментов использовался собственный речевой корпус «С» речевой базы «КРИПТОН-02» на основе собственных записей [34, 40] (приложение № 2). Речевой корпус «С» разбит на два непересекающихся множества: учебное и тестовое. Данный корпус записан 20 дикторами. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (10 дикторов), а оценка точности распознавания – на тестовом подмножестве (остальные 10 дикторов). В качестве речевых сигналов были взяты произношения 102 речевых сигналов. В исследуемой речевой базе каждый речевой сигнал промаркирован обозначением – «Sθ», где θ – означает номер того или иного речевого сигнала. То есть исследуемый речевой корпус состоит из речевых сигналов S1, S2, S3, S4, …, S102.

Далее нужно определиться с размерностью модифицированного коллективного нейросетевого алгоритма. Количество нейросетевых блоков

пропорционально размеру словаря для того, чтобы качество распознавания речевых сигналов оставалось высоким. В предыдущих исследованиях было установлено, что bagging-коллектив из 10 персептронов Розенблатта с обучением SCG распознает до 10 слов без существенного падения вероятности распознавания соответствующих речевых сигналов. Также установлено, что алгоритм bagging-коллектива из 10 сетей Эльмана с обучением GDX распознает до 7 слов без существенного падения вероятности распознавания соответствующих речевых сигналов. Из данных результатов следует, что для исследования данных алгоритмов нужно построить два модифицированных алгоритма с разным количеством нейросетевых блоков и разной размерностью обучающих словарей. То есть модифицированный bagging-алгоритм на основе персептронов Розенблатта с обучением SCG должен содержать 11 нейросетевых блоков и 11 обучающих словарей с максимальной размерностью 10 (рис. 2.9). А модифицированный bagging-алгоритм на основе сетей Эльмана с обучением GDX должен содержать 15 нейросетевых блоков и 15 обучающих словарей с максимальной размерностью 7.

Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме обучения на тестируемой части речевой базы «КРИПТОН-02» представлена на рисунке 2.11. В состав схемы (рис. 2.11) исследуемого алгоритма входит 11 нейросетевых блоков, каждый из которых состоит из 10 нейросетевых распознавателей. В качестве нейросетевого распознавателя слов берется многослойный персептрон с обучением SCG. Количество нейросетевых блоков берется исходя из размера обучающего множества классов (речевых сигналов) по правилу, описанному выше.

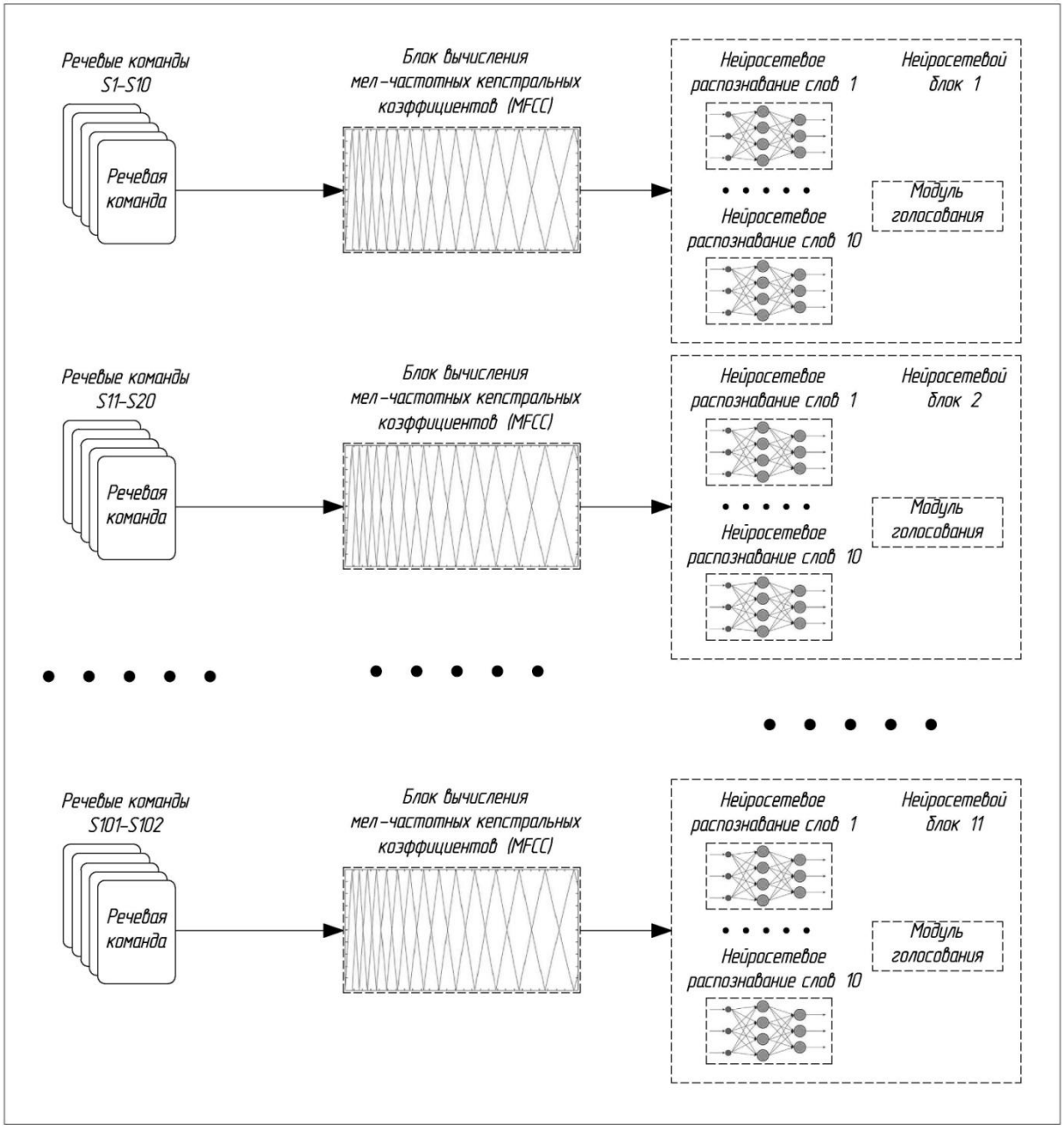

Рис. 2.11. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме обучения на основе персептронов Розенблатта с обучением SCG

Тестирование данных модифицированных коллективных нейросетевых алгоритмов на основе персептронов Розенблатта с обучением SCG (рис. 2.12) и сетей Эльмана с обучением GDX производилось на тестовом подмножестве речевой базы «КРИПТОН-02».

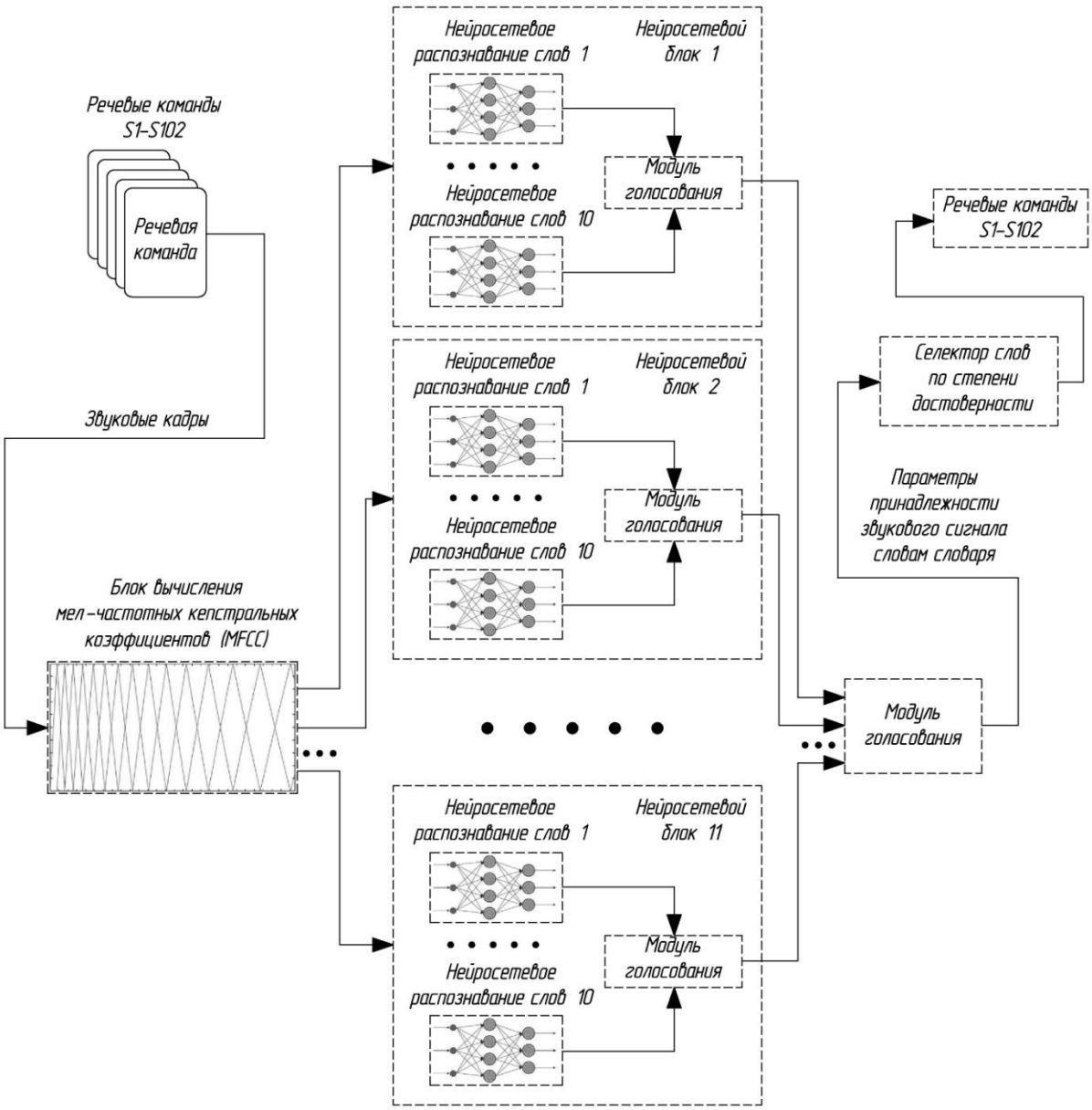

Рис. 2.12. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме обучения на основе персептронов Розенблатта с обучением SCG

Для исследования каждого алгоритма произведена выборка значений, показывающих результат распознавания исследуемым алгоритмом тестируемое значение, равная 15300 измерений. При произведенной выборке частота распознавания приближенно равна вероятности распознавания. Результаты тестирования исследуемых алгоритмов представлены в таблице 2.1. Все операции выполнялись

последовательно. Время обучения и тестирования можно существенно сократить за счет распараллеливания процессов обучения и тестирования нейросетевых блоков и увеличения вычислительной мощности компьютера. Для вероятности попадания всех значений точности распознавания из генеральной выборки 0,95 оценен доверительный интервал определенной вероятности распознавания речевых сигналов.

Таблица 2.1

Алгоритм	**Точность распознавания, %**	**Доверительный интервал, п.п.**	**Время выполнения, сек.**	
			Обучения сетей	**Тестирования сетей**
Модифицированный bagging-коллектив на основе сетей Эльмана с обучением GDX	91,5	± 5,5	3030	380
Модифицированный bagging-коллектив на основе многослойных персептронов с обучением SCG	**95,7**	± 3,2	2688	381

Из данных результатов следует, что модифицированный алгоритм нейросетевого bagging-коллектива хорошо справляется с задачей дикторонезависимого распознавания речевых сигналов на словаре, состоящем из 102 классов речевых сигналов. Также из данных результатов следует, что исследуемый алгоритм на основе многослойных персептронов с обучением SCG распознает речевые сигналы из речевой базы «КРИПТОН-02» с вероятностью 95,7 %, а исследуемый алгоритм на основе сетей Эльмана с обучением GDX распознает речевые сигналы из той же речевой базы с вероятностью 91,5 %. Из данных показателей следует, что модифицированный bagging-коллектив на основе многослойных персептронов с обучением SCG результативнее алгоритма на основе сетей Эльмана с обучением GDX.

Модифицированный коллективный нейросетевой алгоритм на основе персептронов Розенблатта с обучением SCG позволяет решать задачу

дикторонезависимого распознавания русскоязычных речевых сигналов для большого словаря с вероятностью распознавания 95,7 %, что на 5,29 процентных пункта выше существующих результатов [101]. Учитывая доверительный интервал полученных значений, следует, что с вероятностью 0,95 точность распознавания речевых сигналов также лучше существующих результатов [101].

Данные исследования проводились в программе MATLAB® R2009a на персональном компьютере с техническими параметрами: процессор – Intel® Core™ 2 Duo, тактовая частота процессора – 2 ГГц, оперативная память – 3 ГБ. Установлено, что для обучения нейросетевых алгоритмов было задействовано около 75% производительности центрального процессора и около 10% оперативной памяти. Так как при тестировании произведено распознавание исследуемыми алгоритмами 5100 речевых сигналов, то распознавание одного сигнала модифицированным bagging-коллективом на основе многослойных персептронов с обучением SCG в среднем заняло 0,08 секунд. Данного времени более чем достаточно для распознавания в реальном времени. Для режима реального времени достаточно, чтобы алгоритм распознавал один тестовый сигнал за одну секунду. Учитывая количество загрузки процессора и задействованной оперативной памяти персонального компьютера в данном эксперименте можно приближенно задать требования к радиотехническим устройствам, в которые возможно интегрировать исследуемые алгоритмы. Для реализации в радиотехническом устройстве модифицированного bagging-коллектива на основе многослойных персептронов с обучением SCG для задачи дикторонезависимого распознавания русскоязычных речевых сигналов для размера словаря в 102 сигнала нужно примерно выполнять 250×10^6 операций в секунду и требуется 125 МБ оперативной памяти.

2.4. Выводы по главе

Во второй главе представлены алгоритмы базового и коллективного нейросетевого распознавания. Также в данной главе разработан модифицированный коллективный нейросетевой алгоритм для задачи дикторонезависимого распознавания речевых сигналов. В качестве основных многослойных нейронных сетей выбраны: персептрон Розенблатта и сеть Эльмана. Для обучения нейросетевых алгоритмов выбрано два алгоритма обучения: масштабируемых сопряженных градиентов (SCG) и градиентного спуска с учетом моментов и с адаптивным обучением (GDX). Для исследований использовались собственные речевые базы «КРИПТОН-01» и «КРИПТОН-02». Все исследования проводились с условием дикторонезависимого распознавания речевых сигналов.

Для исследования данных нейросетевых алгоритмов было проведено пять серий экспериментов:

– В результате проведения первой серии экспериментов по исследованию размера нейросетевого bagging-коллектива в задаче дикторонезависимого распознавания речевых сигналов определено пороговое значение размера bagging-коллектива, после которого вероятность распознавания с ростом размера bagging-коллектива растет медленно. Данный порог равен 10 нейросетевым распознавателям в коллективном нейросетевом алгоритме, при которых вероятность дикторонезависимого распознавания 10 речевых сигналов равняется 76,1 % распознавания для четырех обучающих дикторов. В качестве нейросетевого распознавателя использовался многослойный персептрон Розенблатта с обучением SCG.

– В результате проведения второй серии экспериментов по исследованию количества обучающих дикторов в задаче дикторонезависимого распознавания речевых сигналов исследованы

базовый и коллективный нейросетевые алгоритмы. Определено пороговое значение количества обучающих дикторов, после которого вероятность распознавания с ростом количества дикторов растет медленно. Для базового нейросетевого алгоритма данный порог определился при 11 обучающих дикторах и равен 83 % вероятности распознавания 10 классов речевых сигналов. Для коллективного нейросетевого алгоритма данный порог определился при 10 обучающих дикторах и равен 97,1 % вероятности распознавания 10 классов речевых сигналов. В качестве нейросетевого распознавателя использовался многослойный персептрон Розенблатта с обучением SCG.

– В результате проведения третьей серии экспериментов по исследованию количества слоев для нейросетевого алгоритма bagging-коллектива исследованы два нейросетевых алгоритма bagging-коллектива из: 10 персептронов Розенблатта с обучением SCG и 10 сетей Эльмана с обучением GDX. Определено пороговое значение количества слоев для каждого алгоритма, после которого вероятность распознавания с ростом количества слоев растет медленно. Для исследованных алгоритмов данный порог определился при 12 слоях нейросетевых распознавателей и равен для алгоритмов bagging-коллектива из: 10 персептронов Розенблатта с обучением SCG – 97 %, а для 10 сетей Эльмана с обучением GDX – 90,7 %.

– В результате проведения четвертой серии экспериментов по исследованию размера словаря для коллективных нейросетевых алгоритмов исследованы два нейросетевых алгоритма bagging-коллектива из: 10 персептронов Розенблатта с обучением SCG и 10 сетей Эльмана с обучением GDX. Определено пороговое значение размера словаря, после которого вероятность распознавания с ростом количества речевых сигналов резко падает. Для bagging-коллектива из 10 персептронов Розенблатта с обучением SCG данный порог определился на 12 речевых сигналах и равен 96,1 % распознавания. Для bagging-коллектива из 10

сетей Эльмана с обучением GDX данный порог определился на 7 речевых сигналах и равен 93 % распознавания.

– В результате проведения пятой серии экспериментов исследована работа модифицированных алгоритмов нейросетевого распознавания. Были исследованы модифицированные алгоритмы на основе двух разновидностях нейронных сетей: 10 персептронов Розенблатта с обучением SCG и 10 сетей Эльмана с обучением GDX. Модифицированный bagging-коллектив на основе 10 сетей Эльмана распознал 102 речевых сигнала с вероятность распознавания 91,5 % при времени обучения данного алгоритма 3030 секунд и времени тестирования 380 секунд. Модифицированный bagging-коллектив на основе 10 персептронов Розенблатта распознал 102 речевых сигнала с вероятность распознавания 95,7 % при времени обучения данного алгоритма 2688 секунд и времени тестирования 381 секунд. Данные результаты показали конкурентоспособность модифицированного коллективного нейросетевого алгоритма для распознавания речевых сигналов словаря, содержащего около 100 речевых сигналов.

ГЛАВА 3. ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ОБУЧЕНИЯ В ЗАДАЧЕ ДИКТОРОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ

Постоянное совершенствование инструмента нейронных сетей в данной задаче является актуальной задачей на сегодняшний день. Одним из методов улучшения работы коллективных нейронных сетей в задаче дикторонезависимого распознавания речи является правильный подбор алгоритма обучения нейронных сетей [71].

3.1. Алгоритмы обучения коллективных нейронных сетей дикторонезависимого распознавания речевых сигналов

3.1.1. Алгоритм bagging-коллектива многослойных персептронов с обучением Левенберга-Марквардта

В настоящее время одним из самых популярных алгоритмов обучения нейронных сетей является алгоритм Левенберга-Марквардта (Levenberg - Marquardt Algorithm, LMA) [103], так как он стабильный и достаточно быстрый. Структурная схема нейросетевого распознавания речевых сигналов, на которой применим данный алгоритм, изображена на рисунке 2.2. Моделирование данных нейронных сетей осуществляется с помощью функции «newff» пакета расширения MatLab – Neural Network Toolbox, содержащего средства для проектирования, моделирования, разработки и визуализации нейронных сетей [40, 47]. Формула обновления коэффициентов данного метода:

$$w_{k+1} = w_k - (J_k^T J_k + \mu I)^{-1} J_k E_k$$

где $k = 1,2,\ldots,N$, E_k – вектор ошибки, μ – комбинационный коэффициент, J_k – якобиан, I – единичная матрица.

Алгоритм Левенберга-Марквардта имеет вычислительную сложность $O(N^3)$ и задействует $O(2N^3)$ памяти на каждой итерации [73, 78].

С выходными сигналами нейронных сетей осуществляется процедура SOFTMAX-нормализации, после чего данные выходные сигналы можно трактовать как векторы распределения вероятности распознавания [39]. Далее векторы распределения вероятности распознавания от T нейронных сетей (T – количество нейронных сетей в данном bagging-коллективе) попадают на модуль голосования, на котором вычисляется средний вектор распределения вероятности распознавания.

Окончательное решение о распознавании принимается в селекторе слов по уровню достоверности (рис. 2.2) с учетом апостериорной информации, то есть распределения вероятности распознавания слов в речевых кадрах (данные вероятности вычисляются нейронными сетями).

3.1.2. Алгоритм bagging-коллектива сетей Эльмана с обучением GDX

Исследуемой сетью является сеть Эльмана. Данная сеть характеризуется частичной рекуррентностью в форме обратной связи между скрытым и входным слоем, реализуемой с помощью однократных элементов запаздывания [69, 84, 85]. Каждый скрытый нейрон имеет свой аналог в контекстном слое, образующем совместно с внешними входами сети входной слой. Данный алгоритм приведен на рисунке 3.1. Моделирование данных нейронных сетей осуществляется с помощью функции «newelm» пакета расширения MatLab – Neural Network Toolbox [47]. В качестве алгоритма обучения сетей выбран алгоритм GDX [26], так как он стабильный и очень быстрый. Формула обновления коэффициентов данного метода:

$$w_{k+1} = w_k - \eta J_k E_k + \alpha_k \Delta w_{k-1},$$

$$\Delta w_{k-1} = -\eta J_{k-1} E_{k-1},$$

где $k = 1,2,\ldots,N$, E_k – вектор ошибки, η – коэффициент обучения, J_k –

якобиан, α_k – коэффициент момента, $\alpha_k \in (0,1)$.

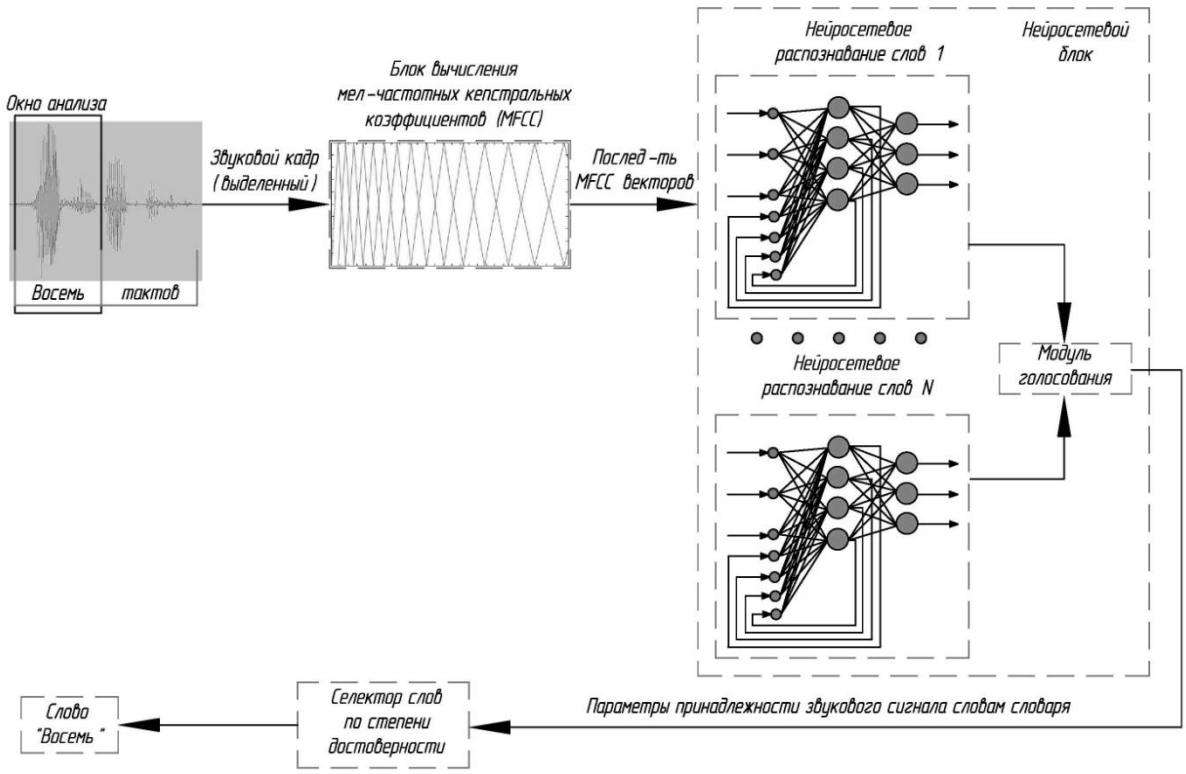

Рис. 3.1. Структурная схема коллективного нейросетевого алгоритма распознавания слов на основе сетей Эльмана

Так как исследуемой сетью является рекуррентная сеть Эльмана, то сложно дать оценку вычислительной сложности данного алгоритма. Данную оценку можно дать по итогам эксперимента с данной сетью.

С выходными сигналами данных нейронных сетей осуществляется процедура SOFTMAX-нормализации [39]. Далее векторы распределения вероятности распознавания от P нейронных сетей (P – количество нейронных сетей в данном bagging-коллективе) попадают на модуль голосования, на котором вычисляется средний вектор распределения вероятности распознавания.

Окончательное решение о распознавании принимается в селекторе слов по уровню достоверности (рис. 3.1).

3.1.3. Алгоритм bagging-коллектива многослойных персептронов с обучением SCG

Исследуемой сетью является многослойный персептрон Розенблатта [21]. В качестве алгоритма обучения сетей, выбран алгоритм SCG [82], так как он стабильный и очень быстрый. В стандартной форме алгоритма сопряженных градиентов требуется использование линейного поиска, что из-за его характера «проб и ошибок» может занять много времени. В модифицированной (данной) версии алгоритма сопряженных градиентов линейный поиск отсутствует. Линейный поиск заменен одномерной формой Левенберга-Марквардта. Идея заключается во вводе скаляра λ_k, который, как предполагается, будет регулировать неопределенность матрицы Гессе $E''(w_k)$. Основанием для использования именно этого метода было желание обойти сложности, вызываемые не положительностью матрицы Гессе [82]. Это делается путем установки:

$$s_k = \frac{E'(w_k + \sigma_k p_k) - E'(w_k)}{\sigma_k} + \lambda_k p_k .$$

Формула обновления коэффициентов данного метода:

$$w_{k+1} = w_k + \alpha_k p_k ,$$

$$\alpha_k = \frac{\mu_k}{\delta_k},\ \mu_k = -p_k^T E_k',\ \delta_k = p_k^T s_k ,$$

где $k = 1,2,\ldots,N$, α_k – размер шага, p_k – сопряженный вектор, E_k – вектор ошибки, δ_k – матрица Гессе, $\sigma_k = \frac{\sigma}{|p_k|}$, λ_k – параметр масштабирования матрицы Гессе.

Если матрица Гессе $\delta_k \le 0$ является отрицательно определенной или равна нулю, тогда нужно увеличить значение параметра масштабирования

λ_k до того момента, пока матрица Гессе не станет положительно определенной.

Для каждой итерации происходит один вызов функции $E(w)$ и два вызова $E'(w)$, которые дают сложность расчетов на одной итерации $O(3N^2)$. Когда алгоритм реализован, данная сложность может быть уменьшена до $O(2N^2)$, так как два вызова $E'(w)$ могут быть посчитаны со сложностью $O(N^2)$ и, соответственно, потребуется задействовать $O(N^2)$ вычислительной памяти на каждую операцию [82].

Данный алгоритм приведен на рисунке 2.2. Моделирование данных нейронных сетей осуществляется с помощью функции «newpr» пакета расширения MatLab – Neural Network Toolbox [47].

С выходными сигналами данных нейронных сетей осуществляется процедура SOFTMAX-нормализации [39]. Далее векторы распределения вероятности распознавания от X нейронных сетей (X – количество нейронных сетей в данном bagging-коллективе) попадают на модуль голосования, на котором вычисляется средний вектор распределения вероятности распознавания.

Окончательное решение о распознавании принимается в селекторе слов по уровню достоверности (рис. 2.2).

3.2. Сравнение работы алгоритмов обучения коллективных нейронных сетей

В данном исследовании предполагается сравнить работу нескольких алгоритмов обучения. В экспериментах исследуются три вида нейросетевых блоков, основанных на различных алгоритмах обучения:

– bagging-коллектив десяти 12-слойных персептронов на основе обучения Левенберга-Марквардта;

– bagging-коллектив десяти 12-слойных сетей Эльмана на основе обучения GDX;

– bagging-коллектив десяти 12-слойных персептронов на основе обучения SCG.

Работу алгоритмов обучения предполагается сравнить по следующим критериям: вероятности распознавания при тестировании алгоритмов обучения, количеству циклов обучения, значению времени обучения и показателям среднеквадратичной ошибки [99].

В качестве материала для данных экспериментов использовался собственный речевой корпус «С» речевой базы «КРИПТОН-02» на основе собственных записей [34, 40] (приложение № 2). Речевой корпус «С» разбит разработчиками на два непересекающихся множества: учебное и тестовое. Данный корпус записан 20 дикторами. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (10 дикторов), а оценка точности распознавания – на тестовом подмножестве (остальные 10 дикторов). В качестве сигналов были взяты произношения 102 распространенных речевых сигналов.

Данные исследования проводились на персональном компьютере с техническими параметрами: процессор – Intel® Core™ 2 Duo, тактовая частота процессора – 2 ГГц, оперативная память – 3 ГБ. Установлено, что для обучения нейросетевых алгоритмов было задействовано около 55 % производительности центрального процессора и около 8 % оперативной памяти. Все операции выполнялись последовательно. Время обучения и тестирования можно существенно сократить за счет распараллеливания процессов обучения и тестирования нейросетевых блоков и увеличения вычислительной мощности компьютера.

При обучении нейронных сетей возникает проблема переобучения. Она заключается в том, что при условии, когда на элементах обучающего множества ошибка обучения достигла малого значения, погрешность существенно возрастает при предоставлении новых данных. Для исключения явления переобучения в экспериментах использовался метод формирования представительной выборки [22]. Смысл метода связан с

организацией целенаправленной процедуры прерывания обучения. Для этого из исходных данных выделяется 3 подмножества. Первое – обучающее подмножество, второе – контрольное (проверочное) подмножество и третье тестовое подмножество. Обучающее подмножество используется для настройки параметров сети; контрольное подмножество используется для настройки параметров сети; контрольное подмножество используется в течение всего процесса обучения для того, чтобы контролировать представительность используемой выборки. Как правило, ошибка для контрольного подмножества на начальной фазе обучения уменьшается, также как и ошибка для обучающего подмножества. Однако когда ошибка для контрольного подмножества начинает увеличиваться, это означает, что в сети начал проявляться эффект переобучения. В этом случае фиксируется итерация, на которой ошибка для контрольного подмножества была минимальной, и восстанавливаются соответствующие значения настраиваемых параметров сети. Соответствующая длина выборки признается представительной.

Ошибка для тестового подмножества обычно не используется в процессе обучения, а применяется для сравнения различных моделей. Однако полезно рассчитывать погрешность для тестового подмножества в течение всего процесса обучения. Если соответствующая ошибка достигает минимума при ином числе итераций, чем для контрольного подмножества, то это может указывать на неудачное выделение подмножеств из набора данных.

Для оценки качества обучения нейронных сетей выбран стандартный критерий ошибок – средняя сумма квадратов ошибки (СКО) обучения:

$$СКО = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(d_i - y_i)^2 ,$$

где N – объем выборки (число примеров в обучающем множестве), e_i – ошибка сети, d_i – желаемая величина выхода, y_i – реально полученные на сети значения для каждого примера i.

Оцениваемыми нейронными сетями выбраны: 12-слойный персептрон на основе обучения Левенберга-Марквардта; 12-слойная сеть Эльмана на основе обучения GDX; 12-слойный персептрон на основе обучения SCG. В качестве обучающей выборки выбрано несколько вариантов 10 речевых сигналов из речевой базы «КРИПТОН-02».

Результаты экспериментов оценки СКО от количества циклов обучения с учетом метода формирования представительной выборки для исключения эффекта переобучения представлены на рисунках 3.2 – 3.4. В данных экспериментах была поставлена цель достичь показателя СКО равного 0,01 и посмотреть на возможность появления эффекта переобучения. При достижении поставленной цели СКО обучение прекращается и считается, что сеть обучена.

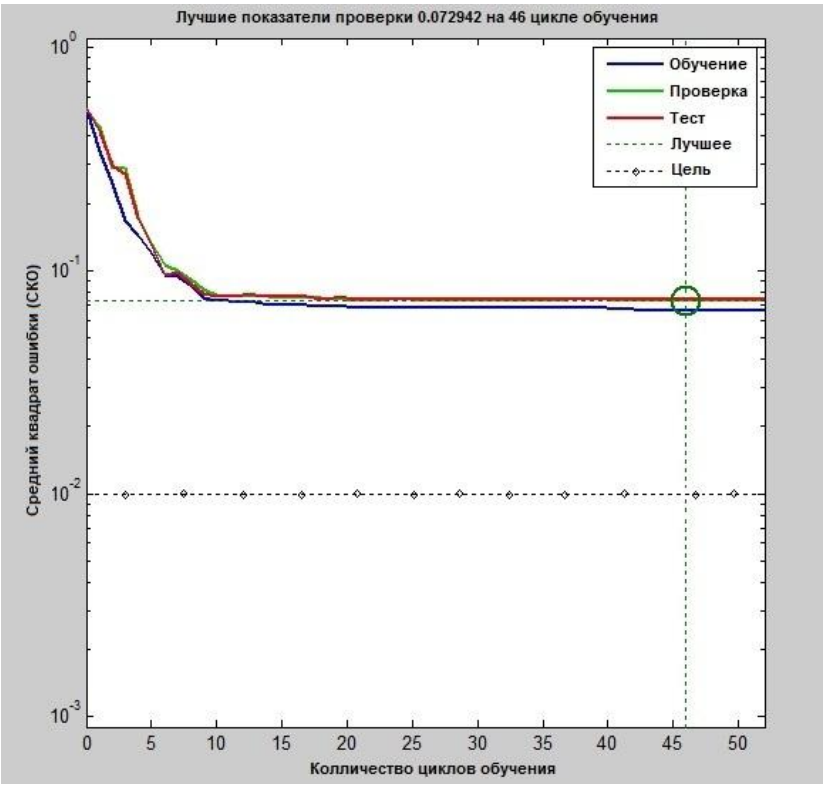

Рис. 3.2. Зависимость СКО от количества циклов обучения для 12-слойного персептрона на основе обучения Левенберга-Марквардта

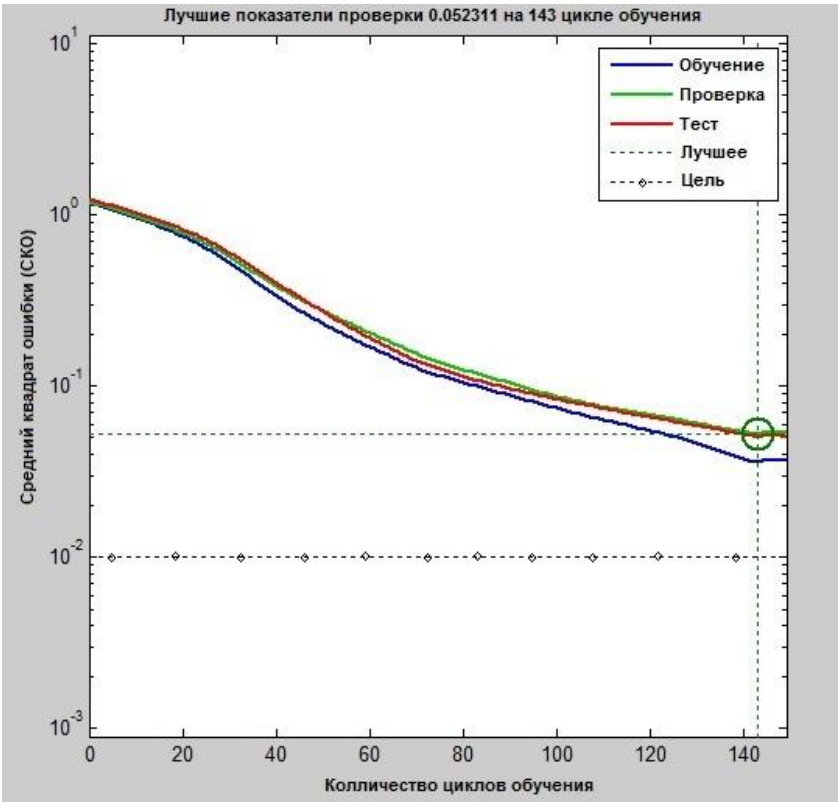

Рис. 3.3. Зависимость СКО от количества циклов обучения для 12-слойной сети Эльмана на основе обучения GDX

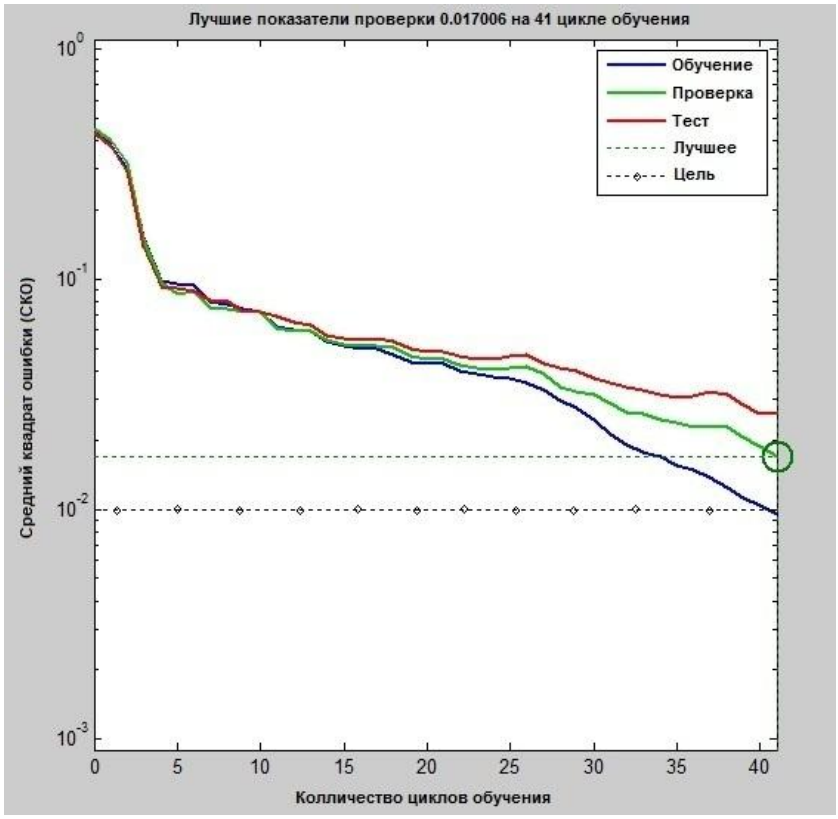

Рис. 3.4. Зависимость СКО от количества циклов обучения для 12-слойного персептрона на основе обучения SCG

Из рисунков 3.2 – 3.4 следует, что эффект переобучения при цели СКО в 0,01 замечен только у двух алгоритмов: 12-слойного персептрона на основе обучения Левенберга-Марквардта и 12-слойной сети Эльмана на основе обучения GDX. У первого алгоритма прерывание произошло на 46 цикле обучения при показателе проверки СКО=0,073, у второго алгоритма прерывание произошло на 143 цикле обучения при показателе проверки СКО=0,052. У третьего алгоритма 12-слойного персептрона на основе обучения SCG эффекта переобучения при тех же условиях не замечено. Третий алгоритм достиг значения СКО=0,017 на 41 цикле обучения. Более детальные показатели данного эксперимента представлены в таблице 3.1.

Таблица 3.1

Сравнение алгоритмов обучения нейронных сетей

Алгоритм	**Целевое СКО**	**Достижимое СКО**	**Количество циклов обучения**	**Время обучения, сек.**
12-слойный персептрон на основе обучения Левенберга-Марквардта	0,01	0,073	46	1784
12-слойная сеть Эльмана на основе обучения GDX		0,052	143	31
12-слойный персептрон на основе обучения SCG		**0,017**	**41**	**27**

Из полученных результатов следует, что при сравнении трех алгоритмов по всем показателям оказался лучшим алгоритм 12-слойного персептрона на основе обучения SCG. Стоит отметить, что обучение алгоритмом Левенберга-Марквардта нейронной сети 12-слойного персептрона оказалось очень долгим. Данное время обучения на несколько

порядков больше времени обучения других исследуемых алгоритмов. Такое поведение алгоритма обучения можно объяснить большой вычислительной сложностью (табл. 3.2) относительно других исследованных алгоритмов [73, 78].

Таблица 3.2

Вычислительная сложность алгоритмов обучения нейронных сетей

Алгоритм обучения	Вычислительная сложность	Количество задействованной памяти
Левенберга-Марквардта	$O(N^3)$	$O(2N^3)$
GDX*	–	–
SCG	$O(2N^2)$	$O(N^2)$
* – вычислительную сложность алгоритма GDX оценить сложно, т.к. данный алгоритм адаптивен и предназначен для рекуррентных нейронных сетей		

Для исследования каждого алгоритма произведена выборка значений, показывающих результат распознавания исследуемым алгоритмом тестируемое значение, равная 1500 измерений. При произведенной выборке частота распознавания приближенно равна вероятности распознавания. Для оценки вероятности дикторонезависимого распознавания речевых сигналов решено провести эксперименты над bagging-коллективами исследуемых нейросетевых блоков. То есть предполагается исследовать три алгоритма: bagging-коллектива 10 многослойных персептронов на основе обучения Левенберга-Марквардта; bagging-коллектива 10 сетей Эльмана на основе обучения GDX; bagging-коллектива 10 многослойных персептронов на основе обучения SCG. В качестве материала для экспериментов предполагается использовать речевой корпус «К» речевой базы «КРИПТОН-01». В ходе серии экспериментов (табл. 3.3): bagging-коллектив из 10 многослойных персептронов на основе обучения SCG показал лучшие результаты 97,1 % точности распознавания; bagging-коллектив из 10 многослойных

персептронов на основе обучения Левенберга-Марквардта показал худшие результаты; bagging-коллектив из 10 сетей Эльмана на основе обучения GDX показал средние результаты. В качестве результатов взята средняя точность распознавания всех речевых сигналов. Для вероятности попадания всех значений точности распознавания из генеральной выборки 0,95 оценен доверительный интервал определенной вероятности распознавания речевых сигналов.

Таблица 3.3
Сравнение алгоритмов распознавания

Алгоритм	**Точность распознавания, %**	**Доверительный интервал, п.п.**	**Время обучения сетей, сек.**
Bagging-коллектив из 10 многослойных персептронов на основе обучения Левенберга-Марквардта	84	± 5,3	18 000
Bagging-коллектив из 10 сетей Эльмана на основе обучения GDX	90,5	± 4,1	300
Bagging-коллектив из 10 многослойных персептронов на основе обучения SCG	**97,1**	± 2,8	262

Нейросетевой алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением SCG позволяет решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов для малого словаря с вероятностью распознавания 97,1 %, что на 4,1 процентных пункта выше существующих результатов [101]. Учитывая доверительный интервал полученных значений, следует, что с вероятностью 0,95 точность распознавания речевых сигналов также лучше существующих результатов [101].

3.3. Выводы по главе

В третьей главе проведен анализ работы нейросетевых алгоритмов обучения в задаче дикторонезависимого распознавания речевых сигналов. В данной главе рассмотрено три коллективных нейросетевых алгоритма, основанных на разных алгоритмах обучения: bagging-коллектив 12-слойных персептронов на основе обучения Левенберга-Марквардта; bagging-коллектив 12-слойных сетей Эльмана на основе обучения GDX и bagging-коллектив 12-слойных персептронов на основе обучения SCG. Для исследований использовалась собственная речевая база «КРИПТОН-02». Все исследования проводились с условием дикторонезависимого распознавания речевых сигналов. Сравнение работы алгоритмов обучения проведено по следующим критериям: вероятности распознавания при тестировании алгоритмов обучения, количеству циклов обучения, значению времени обучения и показателям среднеквадратичной ошибки.

При сравнении работы алгоритмов обучения нейронных сетей по критерию СКО при целевом СКО=0,01: лучшие результаты, равные 0,017 при 41 цикле обучения и 27 секундах, затраченных на данное обучение распознавать 10 речевых сигналов, показал 12-слойный персептрон на основе обучения SCG; средние результаты, равные 0,052 при 143 циклах обучения и 31 секундах, затраченных на обучение, показала 12-слойная сеть Эльмана; худшие результаты, равные 0,073 при 46 цикле обучения и 1784 секундах, затраченных на данное обучение, показал 12-слойный персептрон на основе обучения Левенберга-Марквардта. Большое время обучения 12-слойного персептрона на основе обучения Левенберга-Марквардта можно объяснить большой вычислительной сложностью алгоритма и трудностью алгоритма решить сложную задачу обучения нейронной сети распознавания 10 речевых сигналов.

При сравнении работы исследуемых коллективных нейросетевых алгоритмов по критерию вероятности распознавания при тестировании

алгоритмов обучения: лучшие результаты, равные 97,1 % вероятности распознавания 10 речевых сигналов при затраченном времени обучения 262 секунд, показал bagging-коллектив из 10 многослойных персептронов на основе обучения SCG; средние результаты, равные 90,5 % вероятности распознавания 10 речевых сигналов при затраченном времени обучения 300 секунд, показал bagging-коллектив из 10 многослойных сетей Эльмана на основе обучения GDX; очень плохие результаты, равные 84 % вероятности распознавания 10 речевых сигналов при затраченном времени обучения 18 000 секунд, показал bagging-коллектив из 10 многослойных персептронов на основе обучения Левенберга-Марквардта.

ГЛАВА 4. АНАЛИЗ РАБОТЫ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ДИКТОРОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ В УСЛОВИЯХ ШУМОВ

4.1. Алгоритм коллективного нейросетевого распознавания с встроенным блоком шумоподавления

Алгоритм коллективного нейросетевого распознавания с определенным успехом работает в условии отсутствия или малого присутствия в тестируемых речевых сигналах различных шумов [35]. При практическом использовании данного алгоритма дикторонезависимое распознавание тестируемых речевых сигналов может быть не результативным вследствие влияния посторонних шумов, таких как: шум ветра; шелест листьев; стук дождя; скрип дверей; шум двигателя машины; шум со строительной площадки и так далее. Данное явление объясняется тем, что при обучении коллективного нейросетевого алгоритма используется обучающая речевая база без шума или с незначительной примесью. То есть коллективный нейросетевой алгоритм обучается на речевой базе с большим числом информативных признаков, с помощью которых каждый речевой сигнал может быть с определенной вероятностью отличной от всех остальных. При тестировании коллективного нейросетевого алгоритма на зашумленной речевой базе вероятность распознавания может уменьшаться вследствие слияния информативных признаков тестируемых речевых сигналов с присутствующими признаками шумов. Следовательно, уникальность тестируемых речевых сигналов будет уменьшаться с увеличением мощности шума, присутствующего в тестируемой речевой базе [37, 38].

Для решения проблемы возможной зашумленности тестовых речевых сигналов в задаче дикторонезависимого распознавания речевых

сигналов, решено в исследуемый коллективный нейросетевой алгоритм добавить блок шумоподавления (рис. 4.1).

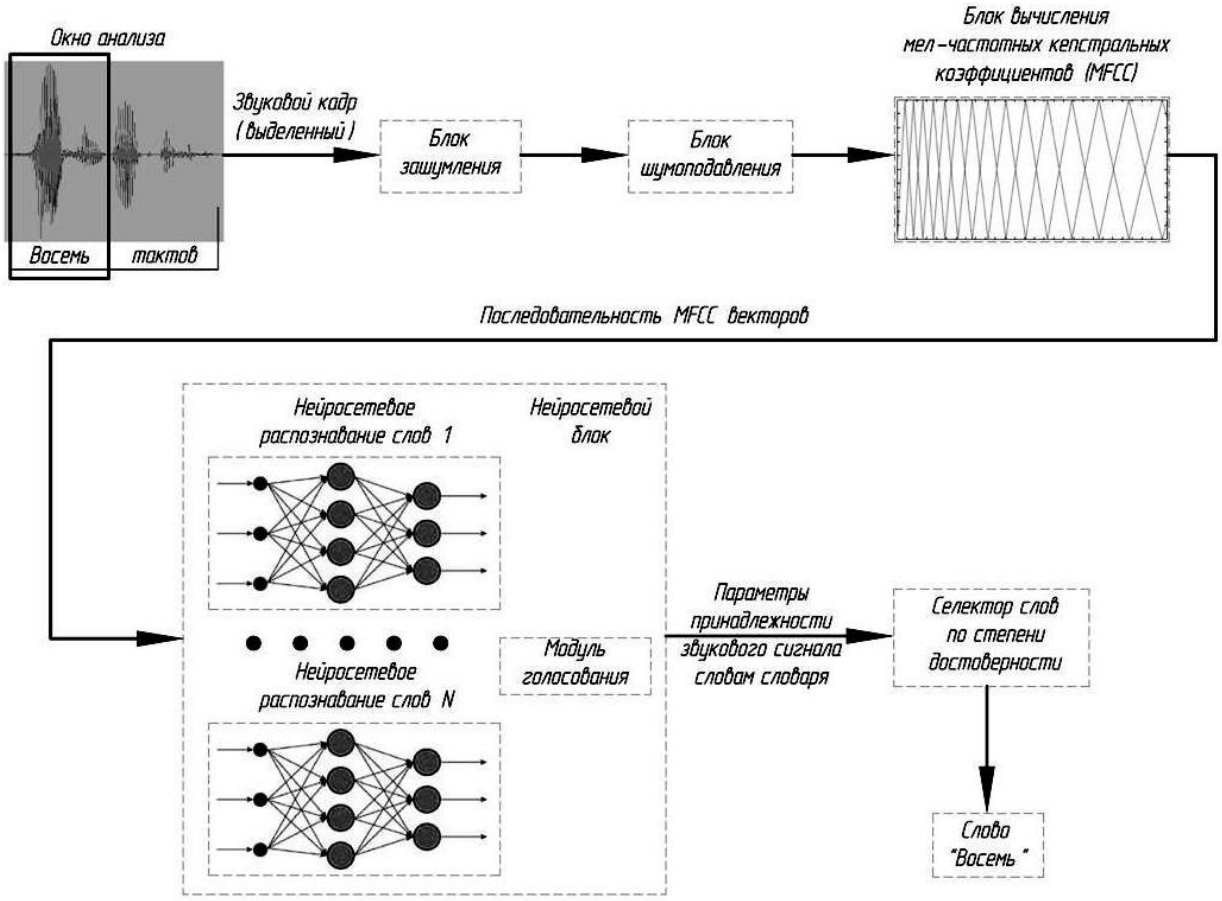

Рис. 4.1. Структурная схема коллективного нейросетевого алгоритма распознавания речевых сигналов с блоками предобработки

Для моделирования зашумленности тестовых речевых сигналов решено входные тестовые речевые сигналы дополнить блоком зашумления (рис. 4.1). То есть в модель, созданную для тестирования коллективного нейросетевого алгоритма, встраивается регулируемый блок зашумления и блок предобработки (шумоподавления) на основе различных алгоритмов шумоподавления.

Обучение данного коллективного нейросетевого алгоритма осуществляется на не зашумленной речевой базе, то есть без блоков зашумления и шумоподавления.

В данном коллективном нейросетевом алгоритме в качестве параметров речевого сигнала, по которым проводится обучение и распознавание, используется логарифм энергии сигнала по J MFCC-коэффициентам [74]. Методика преобразования входного речевого сигнала в массив MFCC-коэффициентов описана в п. 2.1.

4.2. Алгоритм модифицированного коллективного нейросетевого распознавания с встроенным блоком шумоподавления

В целях увеличения технических возможностей распознавания речевых сигналов коллективного нейросетевого алгоритма предложено bagging-алгоритм модифицировать. Данное улучшение алгоритма должно позволить увеличить размер словаря без потери качества дикторонезависимого распознавания речевых сигналов. Соответственно данное улучшение позволит расширить сферу применения распознавания речевых сигналов.

При построении модифицированного bagging-алгоритма предполагается использовать в качестве основного элемента нейросетевой блок коллективного голосования (рис. 2.3). Один нейросетевой блок способен обучиться и распознать речевые сигналы без существенной потери качества распознавания речевых сигналов на словаре с ограниченной размерностью [38]. В данном алгоритме предполагается использовать L нейросетевых блоков.

Практическое использование данного модифицированного коллективного нейросетевого алгоритма может быть не результативным вследствие влияния упомянутых посторонних шумов. Для решения проблемы возможной зашумленности тестовых речевых сигналов в задаче дикторонезависимого распознавания речевых сигналов, решено исследуемый модифицированный коллективный нейросетевой алгоритм модернизировать блоком шумоподавления (рис. 4.2).

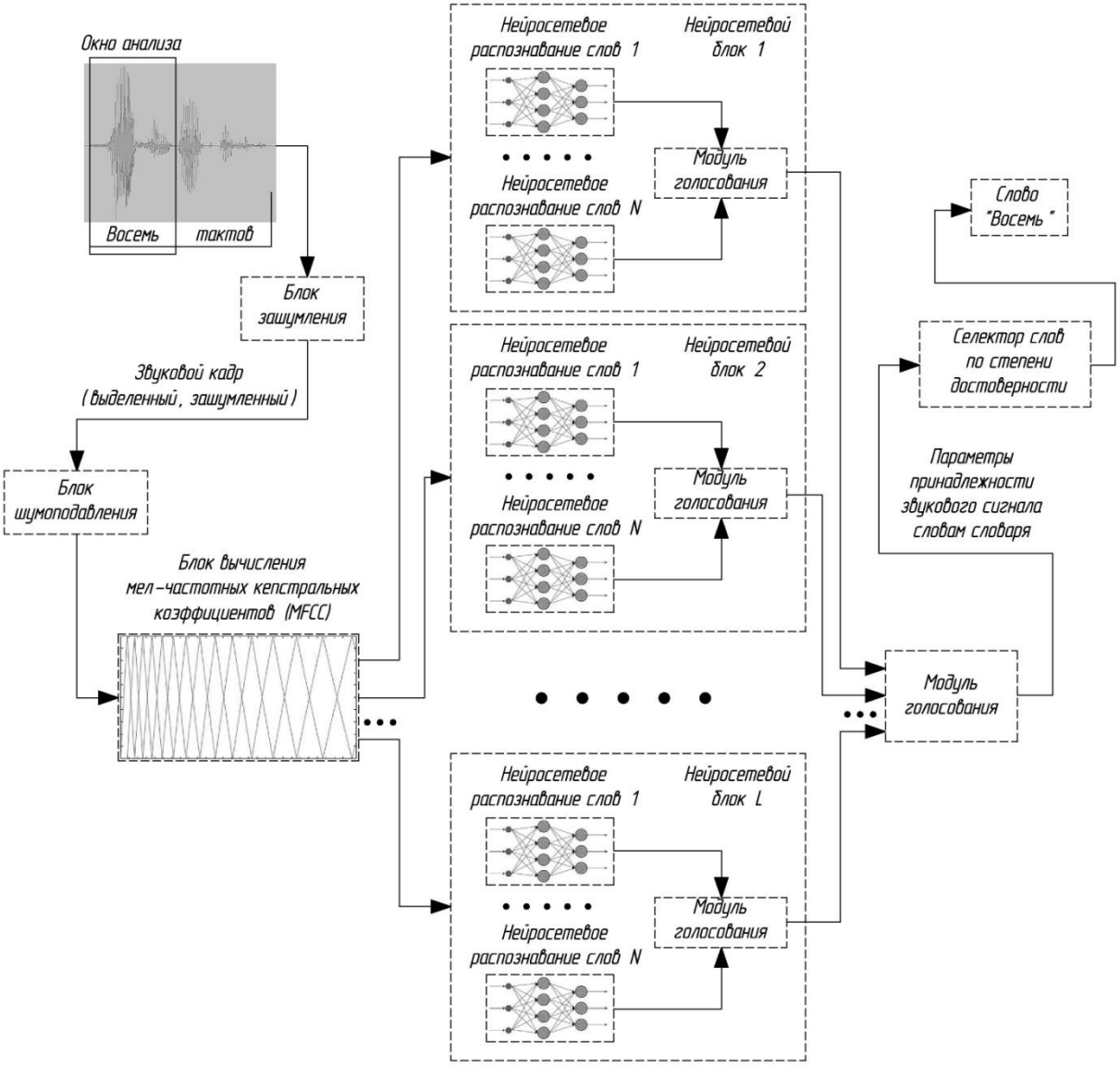

Рис. 4.2. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов в режиме тестирования с блоками предобработки

Обучение данного модифицированного коллективного нейросетевого алгоритма осуществляется на не зашумленной речевой базе, то есть без блоков зашумления и шумоподавления в данном алгоритме.

4.3. Исследование коллективного нейросетевого алгоритма с встроенным блоком шумоподавления

Целью исследования коллективного нейросетевого алгоритма с встроенным блоком шумоподавления (рис. 4.1) является оценка работы его в условиях шумов. В качестве алгоритмов шумоподавления выбрано три алгоритма:

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе апостериорного отношения сигнал/шум [90];

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе априорного отношения сигнал/шум, для оценки которого используется алгоритм TSNR [6, 87];

– алгоритм шумоподавления Скалара на основе винеровской фильтрации [95].

В экспериментах исследуется bagging-коллектив 10 многослойных персептронов на основе обучения SCG [82].

В качестве материала для данных экспериментов использовался собственный речевой корпус «К» речевой базы «КРИПТОН-01» (приложение № 1) на основе собственных записей [40], содержащий около двух с половиной часов звукозаписей различных речевых сигналов (на русском языке), которые были записаны двадцатью дикторами. Речевой корпус разбит на два непересекающихся множества: учебное и тестовое. В качестве обучающих дикторов взяты люди разного пола (70 % мужчины – 7 человек, 30 % женщины – 3 человека), разного возраста (17-38 лет) и разного эмоционального состояния. В качестве тестирующих дикторов взяты люди разного пола (80 % мужчины – 8 человек, 20 % женщин – 2 человека), разного возраста (18-35 лет) и разного эмоционального состояния. Обучение всех алгоритмов распознавания проводилось, соответственно, на учебном подмножестве (10 дикторов), а оценка

точности распознавания – на тестовом подмножестве (другие 10 дикторов). Запись речевых сигналов производилась на микрофон BBKdm-150 в условиях малого «повседневного» белого шума. В качестве речевых сигналов были взяты произношения цифр от «0» до «9», которые каждый обучающий диктор произнес по 50 раз и каждый тестирующий диктор также произнес по 50 раз.

Далее из имеющегося речевого корпуса было получено 8 речевых корпусов путем различного зашумления аддитивным белым Гауссовым шумом в отношении сигнал/шум (ОСШ): -15, -10, -5, 0, 5, 10, 15, 20 дБ. В дальнейшем из каждого речевого корпуса получено 4 речевых корпуса путем использования разных блоков шумоподавления, из которых: 3 речевых корпуса обработаны тремя алгоритмами шумоподавления (IBM-PostSNR, IBM-TSNR, Wiener-PriorSNR); 1 речевой корпус не обработан от шумов. В итоге для исследования было получено 32 речевых корпуса.

В ходе серии экспериментов (рис. 4.3) на алгоритме bagging-коллектива из 10 многослойных персептронов на основе обучения SCG в условиях шумов (рис. 4.1) произведено сравнение различных алгоритмов шумоподавления. Полученные результаты сравнены с результатом работы коллективного нейросетевого алгоритма с чистой речевой базой. Также получены результаты работы коллективного нейросетевого алгоритма без блока шумоподавления в условиях шумов. Из рисунка 4.3 видно, что результаты для трех исследуемых алгоритмов шумоподавления (IBM-PostSNR, IBM-TSNR, Wiener-PriorSNR) в целом оказались примерно одинаковыми. При более детальном рассмотрении поведения коллективного нейросетевого алгоритма распознавания речевых сигналов с блоком шумоподавления, основанном на исследуемых алгоритмах шумоподавления: на участке от -15 до -10 дБ показатели распознавания речевых сигналов оказались примерно одинаковыми; на участке от -5 до 0 дБ показатели распознавания речевых сигналов оказались примерно одинаковыми для IBM-PostSNR, Wiener-PriorSNR, а для алгоритма IBM-

TSNR данные показатели оказались ниже примерно на 5 % вероятности распознавания; при ОСШ 5 дБ показатели распознавания речевых у трех исследуемых алгоритмов снова примерно выровнялись; на участке от 5 до 15 дБ показатели распознавания речевых сигналов оказались примерно одинаковыми для IBM-PostSNR и IBM-TSNR, а для алгоритма Wiener-PriorSNR данные показатели оказались выше примерно на 5 % вероятности распознавания; при показателе ОСШ в 20 дБ показатели распознавания речевых у трех исследуемых алгоритмов снова примерно выровнялись.

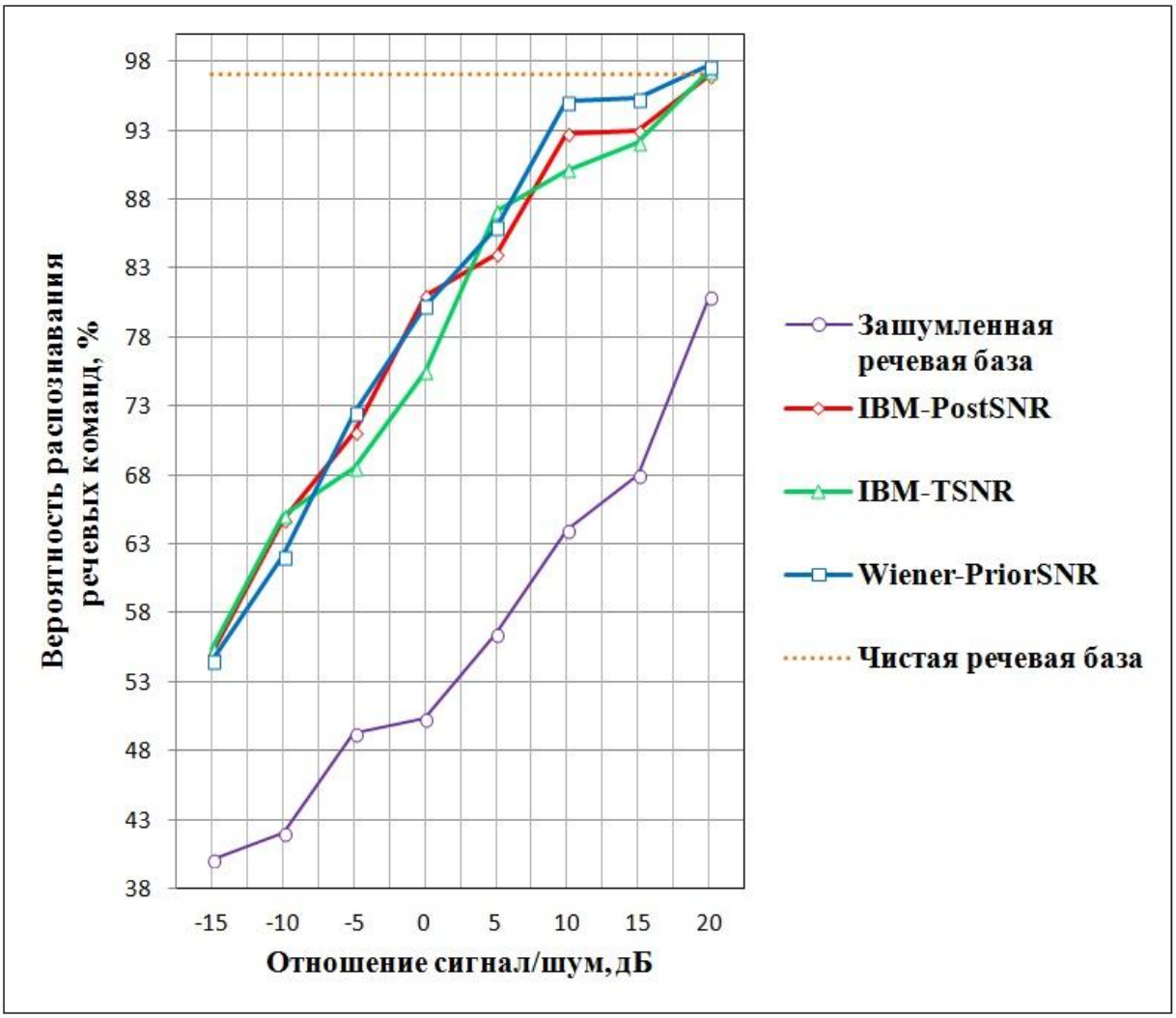

Рис. 4.3. Оценка работы исследуемых алгоритмов шумоподавления в задаче распознавания речевых сигналов. Коллективный нейросетевой алгоритм

При сравнении работы коллективного нейросетевого алгоритма без блока шумоподавления в условиях с чистой тестируемой речевой базой и работой с блоком шумоподавления с зашумленной речевой базой следует:

– вероятность распознавания с блоком шумоподавления при зашумлении тестируемой речевой базы со значениями ОСШ от -15 до 20 дБ уменьшается с уменьшением показателя ОСШ и меньше вероятности распознавания без блока шумоподавления при использовании чистой тестируемой речевой базы;

– вероятность распознавания с блоком шумоподавления при зашумлении тестируемой речевой базы со значением ОСШ 20 дБ больше вероятности распознавания при использовании чистой тестируемой речевой базы без блока шумоподавления в целом на долю процента. Данное поведение можно объяснить не идеально записанной (созданной) чистой речевой базой, так как данная база была записана в условиях малого «повседневного» шума.

Из полученных результатов (рис. 4.3) также следует, что коллективный нейросетевой алгоритм в условиях шумов и без блока шумоподавления распознает речевые сигналы малоэффективно.

При количественной оценке шумоподавления (табл. 4.1) данных алгоритмов средняя вероятность распознавания речевых сигналов на интервале от -15 дБ до 20 дБ оказалась лучшей у алгоритма Скалара – 80,4%.

Таблица 4.1

Сравнение работы шумоподавления в задаче распознавания речевых сигналов

Коллективный нейросетевой алгоритм с блоками предобработки				
ОСШ, дБ	**Средняя вероятность распознавания речевых сигналов, %**			
	Алгоритмы шумоподавления			
	Без шумоподавления	IBM-PostSNR	IBM-TSNR	Wiener-PriorSNR
-15	40,1	55,0	55,3	54,5
-10	42,0	64,7	65,0	62,1

-5	49,3	71,1	68,5	72,5
0	50,3	81,0	75,5	80,3
5	56,5	84,0	87,1	86,0
10	64,0	92,7	90,1	95,1
15	68,0	93,0	92,1	95,3
20	81,0	97,0	97,3	97,7
[-15, 20]	56,3	79,8	78,9	80,4
[5, 20]	67,3	91,6	91,7	93,5

На практике наиболее часто встречается зашумление речевого сигнала от 5 дБ до 20 дБ [25]. При данных показателях зашумления использование исследуемых алгоритмов шумоподавления дает высокие показатели вероятности распознавания речевых сигналов (табл. 4.1).

4.4. Исследование модифицированного коллективного нейросетевого алгоритма с встроенным блоком шумоподавления

Целью исследования модифицированного коллективного нейросетевого алгоритма с встроенным блоком шумоподавления является оценка его работы в условиях шумов. В качестве алгоритмов шумоподавления выбрано три алгоритма:

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе апостериорного отношения сигнал/шум [90];

– алгоритм на основе бинарных масок, использующий критерий статистического детектирования на основе априорного отношения сигнал/шум, для оценки которого используется алгоритм TSNR [6, 87];

– алгоритм шумоподавления Скалара на основе винеровской фильтрации [95].

В экспериментах исследуется модифицированный bagging-коллектив 10 многослойных персептронов на основе обучения SCG [82].

Структурная схема тестирование модифицированного коллективного нейросетевого алгоритма на основе персептронов Розенблатта с обучением SCG (рис. 4.4) с блоками предобработки представлена на рисунке 4.4.

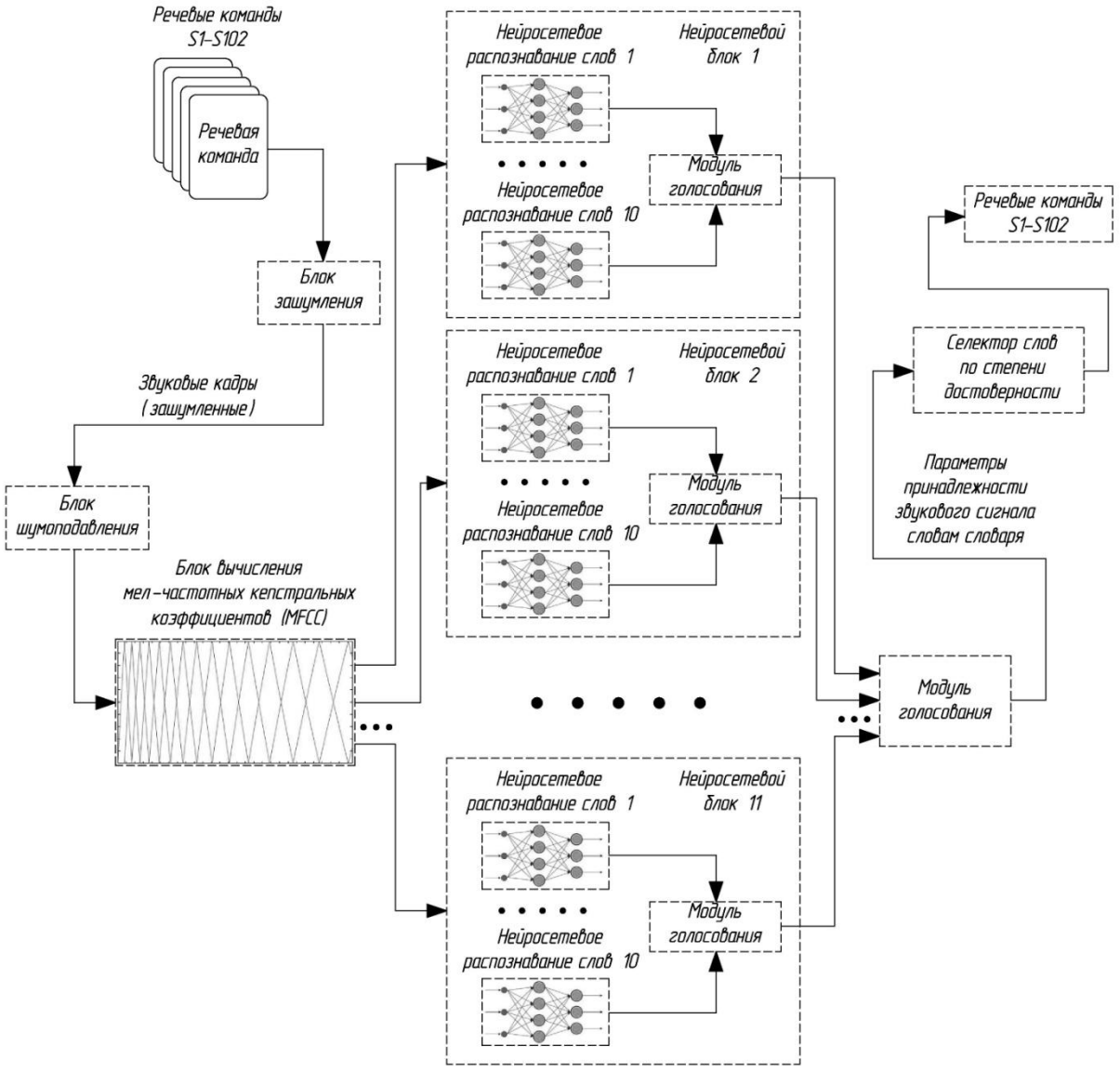

Рис. 4.4. Структурная схема модифицированного коллективного нейросетевого алгоритма распознавания речевых сигналов с блоками предобработки

В качестве материала для данных экспериментов использовался собственный речевой корпус «С» речевой базы «КРИПТОН-02» на основе собственных записей [34, 40] (приложение № 2). Методика создания речевого корпуса «С» речевой базы «КРИПТОН-02» описана в п. 2.3.5.

Обучение модифицированного алгоритма производилось на чистом обучающем подмножестве речевой базы «КРИПТОН-02». В предыдущих исследованиях было установлено, что bagging-коллектив из десяти персептронов Розенблатта с обучением SCG распознает до 10 слов без существенного падения вероятности распознавания соответствующих речевых сигналов. То есть модифицированный bagging-алгоритм на основе персептронов Розенблатта с обучением SCG должен содержать 11 нейросетевых блоков и 11 обучающих словарей с максимальной размерностью 10 (рис. 2.9).

Для тестирования исследуемых алгоритмов из имеющегося речевого корпуса было получено 8 речевых корпусов путем различного зашумления аддитивным белым Гауссовым шумом в отношении сигнал/шум (ОСШ): -10, -5, 0, 5, 10, 15, 20 дБ. В дальнейшем из каждого речевого корпуса получено 4 речевых корпуса путем использования разных блоков шумоподавления, из которых: 3 речевых корпуса обработаны тремя алгоритмами шумоподавления (IBM-PostSNR, IBM-TSNR, Wiener-PriorSNR); 1 речевой корпус не обработан от шумов. В итоге для исследования было получено 28 речевых корпуса.

В ходе серии экспериментов по тестированию модифицированного Bagging-коллектива из 10 многослойных персептронов на основе обучения SCG в условиях шумов (рис. 4.4) было произведено сравнение различных алгоритмов шумоподавления.

Полученные результаты (рис. 4.5) сравнены с результатом работы модифицированного коллективного нейросетевого алгоритма с чистой речевой базой. Также получены результаты работы модифицированного коллективного нейросетевого алгоритма без блока шумоподавления в условиях шумов. Из рисунка 4.5 видно, что результаты для двух исследуемых алгоритмов шумоподавления (IBM-TSNR, Wiener-PriorSNR) в целом оказались примерно одинаковыми, а для алгоритма IBM-PostSNR данные результаты несколько ниже.

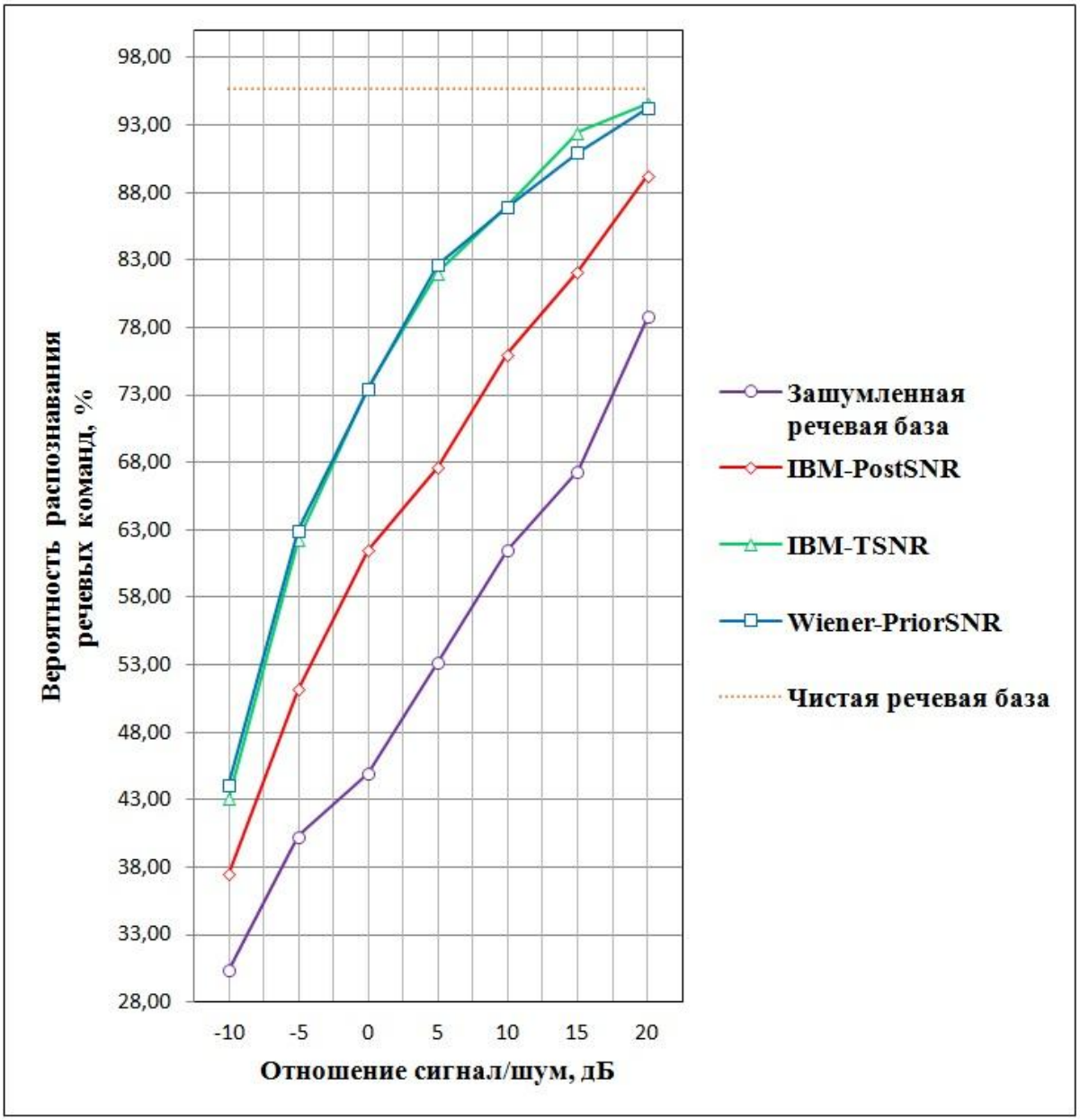

Рис. 4.5. Оценка работы исследуемых алгоритмов шумоподавления в задаче распознавания речевых сигналов. Модифицированный коллективный нейросетевой алгоритм с блоками предобработки

При сравнении работы модифицированного коллективного нейросетевого алгоритма без блока шумоподавления в условиях с чистой тестируемой речевой базой и работой с блоком шумоподавления с зашумленной речевой базой следует:

– вероятность распознавания с блоком шумоподавления при зашумлении тестируемой речевой базы при значениях ОСШ от -10 до 20

дБ уменьшается с уменьшением ОСШ и меньше вероятности распознавания без блока шумоподавления при использовании чистой тестируемой речевой базы;

– вероятность распознавания с блоком шумоподавления при зашумлении тестируемой речевой базы при ОСШ равном 20 дБ больше вероятности распознавания при использовании чистой тестируемой речевой базы без блока шумоподавления в целом на долю процента. Данное поведение можно объяснить не идеально записанной (созданной) чистой речевой базой, так данная база была записана в условиях малого «повседневного» шума.

Из полученных результатов (рис. 4.5) также следует, что коллективный нейросетевой алгоритм в условиях шумов и без блока шумоподавления распознает речевые сигналы малоэффективно.

При количественной оценке шумоподавления (табл. 4.2) данных алгоритмов средняя вероятность распознавания речевых сигналов на интервале от -10 дБ до 20 дБ оказалась лучшей у алгоритмов IBM-TSNR и Wiener-PriorSNR.

Таблица 4.2

Сравнение работы шумоподавления в задаче распознавания речевых сигналов

Модифицированный коллективный нейросетевой алгоритм с блоками предобработки				
ОСШ, дБ	**Средняя вероятность распознавания речевых сигналов, %**			
	Алгоритмы шумоподавления			
	Без шумоподавления	IBM-PostSNR	IBM-TSNR	Wiener-PriorSNR
-10	30,4	37,5	43,1	44,1
-5	40,3	51,2	62,3	63,0
0	44,9	61,5	73,6	73,5
5	53,1	67,7	82,1	82,7
10	61,5	76,0	87,0	86,9
15	67,4	82,2	92,5	91,0
20	78,8	89,3	94,6	94,3
[-10, 20]	53,77	66,49	76,45	76,50
[5, 20]	65,20	78,80	89,05	88,73
[15, 20]	73,10	85,75	93,55	92,65

На практике наиболее часто встречается зашумление речевого сигнала от 5 дБ до 20 дБ [25]. При данных показателях зашумления использование исследуемых алгоритмов шумоподавления дает удовлетворительные показатели вероятности распознавания речевых сигналов (табл. 4.2). Высокие показатели распознавания дают алгоритмы шумоподавления IBM-TSNR и Wiener-PriorSNR в условиях слабой зашумленности (при 15 – 20 дБ).

4.5. Выводы по главе

В четвертой главе проведен анализ работы нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых сигналов в условиях шумов. В данной главе исследованы коллективный и модифицированный коллективный нейросетевые алгоритмы распознавания речевых сигналов с блоками предобработки. Представлено три алгоритма шумоподавления: IBM-PostSNR; IBM-TSNR и Wiener-PriorSNR. Для исследований использовалась собственные речевые базы «КРИПТОН-01» и «КРИПТОН-02». Все исследования проводились с условием дикторонезависимого распознавания речевых сигналов. Обучение исследуемых нейросетевых алгоритмов производилось на чистой речевой базе. Тестирование производилось на речевых базах с различной зашумленностью (- 15, -10, -5, 0, 5, 10, 15, 20 дБ). В качестве шума выбран аддитивный белый Гауссовый шум. Для обучения нейросетевых алгоритмов выбран алгоритм обучения SCG. Каждый нейросетевой блок модифицированного bagging-коллектива состоит из 10 многослойных персептронов на основе обучения SCG.

Проведен сравнительный анализ работы коллективного и модифицированного коллективного нейросетевых алгоритмов с различными блоками шумоподавления и различно зашумленными тестируемыми речевыми базами.

Для алгоритма bagging-коллектива из 10 многослойных персептронов на основе обучения SCG с блоками предобработки результаты для трех алгоритмов шумоподавления в целом оказали одинаковыми. При количественной оценке шумоподавления данных алгоритмов средняя вероятность распознавания речевых сигналов на интервале от -15 дБ до 20 дБ оказалась лучшей у алгоритма Скалара – 80,4%. При показателях зашумления от 5 дБ до 20 дБ алгоритмы шумоподавления дают высокие показатели вероятности распознавания речевых сигналов, такие как 93,5 %, 91,7 %, 91,6 % вероятности распознавания, соответственно, для алгоритмов шумоподавления Wiener-PriorSNR, IBM-TSNR и IBM-PostSNR. Из полученных результатов также следует, что коллективный нейросетевой алгоритм в условиях шумов и без блока шумоподавления распознает речевые сигналы малоэффективно.

Для алгоритма модифицированного bagging-коллектива, состоящего из 11 нейросетевых блоков и блоков предобработки, результаты для двух исследуемых алгоритмов шумоподавления (IBM-TSNR, Wiener-PriorSNR) в целом оказались примерно одинаковыми, а для алгоритма IBM-PostSNR данные результаты несколько ниже. При количественной оценке шумоподавления данных алгоритмов средняя вероятность распознавания речевых сигналов на интервале от -10 дБ до 20 дБ оказалась лучше у алгоритмов IBM-TSNR и Wiener-PriorSNR. При показателях зашумления от 5 дБ до 20 дБ использование исследуемых алгоритмов шумоподавления дает удовлетворительные показатели вероятности распознавания речевых сигналов. Высокие показатели распознавания, равные, соответственно, 93,55 % и 92,65 %, дают алгоритмы шумоподавления IBM-TSNR и Wiener-PriorSNR в условиях слабой зашумленности от 15 до 20 дБ, что также имеет практическую значимость исследуемых алгоритмов. Из полученных результатов также следует, что модифицированный коллективный нейросетевой алгоритм в условиях шумов и без блока шумоподавления распознает речевые сигналы малоэффективно.

Дополнение коллективного и модифицированного коллективного нейросетевых алгоритмов блоками шумоподавления существенно расширяет возможности применения данных нейросетевых алгоритмов для решения задачи дикторонезависимого распознавания речевых сигналов.

ЗАКЛЮЧЕНИЕ

Проведенный анализ актуальных задач машинного распознавания речи позволяет говорить о том, что дикторонезависимое распознавание речевых сигналов является актуальной задачей. Для отечественного рынка (рынка Российской Федерации) также востребована возможность работы таких систем с русской речью. Решение данной задачи в настоящее время актуально использовать вероятностно-сетевую модель принятия решений, такую как нейросетевой метод.

Применение технологий распознавания речевых сигналов актуально в области управления радиотехническими устройствами, такими как, например: радиоприемником, рацией, телевизионным устройством, мобильным телефоном, сканером магнитно-резонансной томографии, рентгеновским сканером и др.

Созданы речевые базы «КРИПТОН-01» с размерностью 10 сигналов и «КРИПТОН-02» с размерностью 102 сигнала для тестирования нейросетевых алгоритмов.

Разработан алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением SCG для решения задачи дикторонезависимого распознавания русскоязычных речевых сигналов.

Разработана модификация коллективного нейросетевого алгоритма, позволяющая решать задачу дикторонезависимого распознавания русскоязычных речевых сигналов для большего размера словаря.

Разработана научно-исследовательская программа, с помощью которой можно проводить анализ алгоритмов дикторонезависимого распознавания русскоязычных речевых сигналов, путем математического моделирования данных алгоритмов обучающих и тестирующих на речевых базах «КРИПТОН-01» и «КРИПТОН-02». Авторские права защищены свидетельством о государственной регистрации программы для ЭВМ.

Проведен анализ параметров bagging-коллектива многослойных персептронов Розенблатта с обучением SCG, в результате чего было определенно, что рациональнее: выбирать размер bagging-коллектива 10; использовать 10 обучающих дикторов; устанавливать по 12 слоев в каждом нейросетевом распознавателе и использовать размер словаря не больше 10. При данных параметрах получена вероятность дикторонезависимого распознавания русскоязычных речевых сигналов 97,1 %, что на 4,1 процентных пункта выше существующих результатов. Учитывая доверительный интервал полученного значения $\pm 2{,}8$ процентных пункта, следует, что с вероятностью 0,95 точность распознавания речевых сигналов также лучше существующих результатов.

Проведены исследования модифицированных алгоритмов на основе двух разновидностях нейронных сетей: 10 персептронов Розенблатта с обучением SCG и 10 сетей Эльмана с обучением GDX. Модифицированный bagging-коллектив на основе 10 сетей Эльмана распознал 102 речевых сигналов с вероятность распознавания 91,5 % при времени обучения данного алгоритма 3030 секунд и времени тестирования 380 секунд. Модифицированный bagging-коллектив на основе 10 персептронов Розенблатта распознал 102 речевых сигналов с вероятность распознавания 95,7 % при времени обучения данного алгоритма 2688 секунд и времени тестирования 381 секунд, что на 5,29 процентных пункта выше существующих результатов. Учитывая доверительный интервал $\pm 3{,}2$ процентных пункта для полученного значения вероятности распознавания модифицированного bagging-коллектив на основе персептронов Розенблатта с обучением SCG, следует, что с вероятностью 0,95 точность распознавания речевых сигналов также лучше существующих результатов.

Проведен анализ работы нейросетевых алгоритмов обучения в задаче дикторонезависимого распознавания русскоязычных речевых сигналов. Рассмотрено три коллективных нейросетевых алгоритмов основанных на

разных алгоритмах обучения: bagging-коллектив 12-слойных персептронов на основе обучения Левенберга-Марквардта; bagging-коллектив 12-слойных сетей Эльмана на основе обучения GDX и bagging-коллектив 12-слойных персептронов на основе обучения SCG. В результате было показано преимущество алгоритма обучения SCG.

Проведен анализ работы нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых сигналов в условиях шумов. Исследованы коллективный и модифицированный коллективный нейросетевые алгоритмы распознавания речевых сигналов с блоками предобработки. Представлено три алгоритма шумоподавления: IBM-PostSNR; IBM-TSNR и Wiener-PriorSNR. Обучение исследуемых нейросетевых алгоритмов производилось на чистой речевой базе. Тестирование производилось на речевых базах с различной зашумленностью (- 15, -10, -5, 0, 5, 10, 15, 20 дБ). В качестве шума выбран аддитивный белый гауссовский шум. Каждый нейросетевой блок модифицированного bagging-коллектива состоит из 10 многослойных персептронов на основе обучения SCG.

Для алгоритма bagging-коллектива из 10 многослойных персептронов на основе обучения SCG с блоками предобработки результаты для трех алгоритмов шумоподавления в целом оказали близкими. При количественной оценке шумоподавления данных алгоритмов средняя вероятность распознавания речевых сигналов на интервале от 5 дБ до 20 дБ алгоритмы шумоподавления дают высокие показатели вероятности распознавания речевых сигналов, такие как 93,5 %, 91,7 %, 91,6 % вероятности распознавания соответственно для алгоритмов шумоподавления Wiener-PriorSNR, IBM-TSNR и IBM-PostSNR.

Модифицированный алгоритм bagging-коллектива, состоящий из 11 нейросетевых блоков и блоков предобработки показал высокие результаты распознавания равные соответственно 93,55 % и 92,65 % для алгоритмов

шумоподавления IBM-TSNR и Wiener-PriorSNR в условиях слабой зашумленности от 15 до 20 дБ.

Дополнение коллективного и модифицированного коллективного нейросетевых алгоритмов блоками шумоподавления расширяет возможности применения данных нейросетевых алгоритмов для решения задачи дикторонезависимого распознавания речевых сигналов.

СПИСОК ЛИТЕРАТУРЫ

1. Аграновский, А.В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов / А.В. Аграновский, Д.А. Леднов. – М.: Издательство «Радио и связь», 2004. – 164 с.
2. Алдошина, И.А. Связь акустических параметров с эмоциональной выразительностью речи и пения / И.А. Алдошина, А. Иринина // Звукорежиссер. – Санкт-Петербург: 2003. – № 2(33).
3. Бондарева, О.В Состязательные искусственные нейронные сети в системе распознавания речи / О.В. Бондарева, В.Н. Бондарев // Системы автоматики и автоматическое управление. Материалы студенч. науч.-техн. конф. г. Севастополь, 14-15 мая 2001г. – Севастополь: Изд-во СевНТУ,2002. – С. 29-33.
4. Бочаров, И.В. Распознавание речевых сигналов на основе корреляционного метода / И.В. Бочаров, Д.Ю. Акатьев // Электронный журнал «Исследовано в России». – М.: МФТИ, 2003. – № 6. – С.1547-1557.
5. Бураков, М.В. Нейронные сети и нейроконтроллеры: учебное пособие / М.В. Бураков. – СПб.: ГУАП, 2013. – 284 с.
6. Веселов, И.А. Использование априорного отношения сигнал/шум для построения бинарных масок в задаче подавления шума в речевых сигналах / И.А. Веселов, А.В. Куликов, Я.М. Скопинцев, Г.С. Тупицин // доклад 15-ой международной конференции «Цифровая обработка сигналов и её применение». – Москва, 2013. – Т.1. – С. 246-249.
7. Винцюк, Т.К. Анализ, распознавание и интерпретация речевых сигналов / Т.К. Винцюк. – Киев: Наукова думка, 1987. – 264 с.
8. Гапочкин, А.В. Нейронные сети в системах распознавания речи / А.В. Гапочкин // Science Time. – Казань: 2014. – № 1(1). – С. 29-36.
9. Гребнов, С.В. Аналитический обзор методов распознавания речи в системах голосового управления / С.В. Гребнов // Вестник ИГЭУ. – 2009. –

№ 3. – С. 83-85.

10. Громов, Ю.Ю. Интеллектуальные информационные системы и технологии : учебное пособие / Ю.Ю. Громов, О.Г. Иванова, В.В. Алексеев и др. – Тамбов : Изд-во ФГБОУ ВПО «ТГТУ», 2013. – 244 с.

11. Гусев, М.Н. Методы и модели распознавания русской речи в информационных системах / М.Н. Гусев. – Санкт-Петербург: Диссертация на соискание уч. ст. д.т.н., 2014. – 378 с.

12. Доррер, Г.А. Теория принятия решений: Учебное пособие для студентов направления 23010.62 – Информатика и вычислительная техника / Г.А. Доррер. – Красноярск: ФГАОУ ВПО «Сибирский федеральный университет», 2013. – 180 с.

13. Калинкина, Д. Проблема подавления шума на изображениях и видео и различные подходы к ее решении. / Д. Калинкина, Д. Ватолин // Компьютерная графика и мультимедиа. – 2005. – № 3(2).

14. Калюжный, М.В. Система реабилитации слабовидящих на основе настраиваемой сегментарной модели синтезируемой речи / М.В. Калюжный. – Санкт-Петербург: Диссертация на соискание уч. ст. к.т.н., 2009. – 171 с.

15. Колмогоров, А.Н. О представлении непрерывных функций нескольких переменных суперпозициями непрерывных функций меньшего числа переменных / А.Н. Колмогоров // ДАН СССР. – 1956. – Т. 108. – № 2. – С. 179-182.

16. Кравцов, С.А. Алгоритм неэталонной оценки степени зашумленности речевых сигналов / С.А. Кравцов, А.В. Куликов, М.В. Сагациян, Г.С. Тупицин // Докл. 14-й междунар. конф. «Цифровая обработка сигналов и её применение». – М.: 2012. – Т.1. – С. 177-179.

17. Левин, Е.К. Разработка средств исследования и повышения помехоустойчивости систем автоматического распознавания голосовых команд в телефонии / Е.К. Левин. – Владимир: Диссертация на соискание уч. ст. д.т.н., 2014. – 257 с.

18. Лепский, А.Е. Математические методы распознавания образов: Курс лекций. / А.Е. Лепский, А.Г. Броневич // – Таганрог: Изд-во ТТИ ЮФУ, 2009. – 155 с.

19. Мазуренко, И.Л. Компьютерные системы распознавания речи / И.Л. Мазуренко // Интеллектуальные системы. – М.: 1998. – № 3(1-2). – С. 117-134.

20. Маковкин, К.А. Гибридные модели: скрытые Марковские модели и нейронные сети, их применение в системах распознавания речи / К.А. Маковкин // Модели, методы, алгоритмы и архитектуры систем распознавания речи. Вычислительный центр им А.А. Дородницына. – М.: 2006. – С. 40-95.

21. Марьина, О.А. Методы обучения многослойного персептрона. Попытки оптимизации задачи поиска глобального минимума функции энергии / О.А. Марьина, Д. А. Ладяев // Электронное научное издание «Электроника и информационные технологии». – 2009. – № 1(5).

22. Медведев, В.С. Нейронные сети. MATLAB 6. / В.С. Медведев, В.Г. Потемкин // Под общ. ред. В.Г. Потемкина. – М.: ДИАЛОГ-МИФИ, 2001. – 630 с.

23. Морозов, М.Н. Курс лекций по дисциплине "Системы искусственного интеллекта" [Электронный ресурс] / М.Н. Морозов // Режим доступа: http://khpi-iip.mipk.kharkiv.edu/library/ai/conspai/10.html #part_9 (дата обращения: 22.03.2015).

24. Назаров, А.В. Нейросетевые алгоритмы прогнозирования и оптимизации систем / А.В. Назаров, А.И. Лоскутов // – СПб.: Наука и Техника, 2003. – 384 с.

25. Новоселов, С.А. Подавление шума в речевых сигналах на основе метода нелокального усреднения / С.А.Новоселов, А.И.Топников, А.И.Савватин, А.Л.Приоров // Цифровая обработка сигналов. – 2011. – №4. – С. 23-28.

26. Осовский, С. Нейронные сети для обработки информации / С.

Осовский. Перевод с польского И.Д. Рудинского. – М.: Финансы и статистика, 2002. – С. 22-24.

27. Перервенко, Ю.С. Исследование инвариантов нелинейной динамики речи и принципы построения системы аудио анализа психофизиологического состояния / Ю.С. Перервенко. – Таганрог: Диссертация на соискание уч. ст. к.т.н., 2009. – 175 с.

28. Рабинер, Л.Р. Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи / Л.Р. Рабинер // ТИИЭР. – 1989. – № 2. – С. 86-120.

29. Розалиев, В.Л. Моделирование эмоциональных реакций пользователя при речевом взаимодействии с автоматизированной системой / В.Л. Розалиев // Известия ВолгГТУ. – Волгоград: ВГТУ, 2009. – № 8(6). – С. 76-79.

30. Романенко, В.О. Эмоциональные характеристики речи и их связь с акустическими параметрами / В.О. Романенко // Общество. Среда. Развитие. – 2010. – № 4. – С. 124-128.

31. Сагациян, М.В. Анализ эффективности нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых команд / М.В. Сагациян, Г.С. Тупицин // Журнал «Информационные системы и технологии». – Орел: 2015. – № 3. – С. 16-26.

32. Сагациян, М.В. Зависимость точности дикторонезависимого распознавания речевых команд нейросетевыми алгоритмом от количества обучающих дикторов / М.В. Сагациян, Г.С. Тупицин // Докл. 11-й междунар. научно-технической конф. «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации». – Курск: 2013. – С.189-191.

33. Сагациян, М.В. Коллективное нейросетевое распознавание речи с алгоритмом обучения масштабируемых сопряженных градиентов / М.В. Сагациян // XIII Всероссийская научная конференция «Нейрокомпьютеры и их применение». – М.: ГБОУ ВПО МГППУ, 2015. – С. 45.

34. Сагациян, М.В. Метод обучения и тестирования нейронных сетей для выполнения задачи дикторонезависимого распознавания речевых команд / М.В. Сагациян // Докл. 66-й Всероссийской научно-технической конф. студентов, магистрантов и аспирантов с международным участием. – Ярославль: Издательство ЯГТУ, 2013. – 119-121 с.

35. Сагациян, М.В. Нейросетевое распознавание речевых команд в условиях шумов / М.В. Сагациян // Международная молодежная научно-практическая конференция «Путь в науку», секция «цифровая обработка сигналов и изображений». – Ярославль: 23-30 апреля 2015.

36. Сагациян, М.В. Обучение нейронной сети алгоритмом SCG в задаче дикторонезависимого распознавания речи / М.В. Сагациян, Г.С. Тупицин // Докл. 12-й междунар. научно-технической конф. «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации». – Курск: 2015.

37. Сагациян, М.В. Повышение эффективности коллективного нейросетевого алгоритма в задаче дикторонезависимого распознавания речевых команд в условиях шумов с помощью бинарных масок / М.В. Сагациян, С.А. Кравцов // Докл. 53-й Международная научная студенческая конференция МНСК-2015, секция «Радиотехника и связь». – Новосибирск: 2015. – С. 39.

38. Сагациян, М.В. Повышение эффективности коллективного нейросетевого алгоритма на основе обучения SCG в задаче дикторонезависимого распознавания речевых команд в условиях шумов / М.В. Сагациян, Г.С. Тупицин, С.А. Кравцов, А.Л. Приоров // Журнал «Информационные системы и технологии». – Орел: 2015. – № 4.

39. Сагациян, М.В. Разработка и исследование нейросетевого алгоритма дикторонезависимого распознавания слов в устной речи / М.В. Сагациян, С.А. Кравцов, Г.С. Тупицин // Докл. 15-й междунар. конф. «Цифровая обработка сигналов и её применение». – М.: 2013. – Т.1. – С. 252-255.

40. Сагациян, М.В. Разработка и исследование нейросетевого алгоритма

дикторонезависимого распознавания речевых команд / М.В. Сагациян, А.В. Куликов, Г.С. Тупицин // Вестник Поволжского государственного технологического университета. Сер.: Радиотехнические и инфокоммуникационные системы. – Йошкар-Ола: 2014. – № 1(20). – С. 62-68.

41. Сагациян, М.В. NN-SCG speech recognition – научно-исследовательская программа по изучению алгоритмов нейросетевого дикторонезависимого распознавания речевых команд / М.В. Сагациян, Г.С. Тупицин // Свидетельство о государственной регистрации программы для ЭВМ № 2015616920 от 30 апреля 2015г.

42. Сапунов, Г.В. Система автоматического распознавания речевых команд для параллельных архитектур / Г.В. Сапунов. – М.: Диссертация на соискание уч. ст. к.т.н., 2005. – 129 с.

43. Сидоров, К.В. Анализ признаков эмоционально окрашенной речи / К.В. Сидоров, Н.Н. Филатова // Вестник Тверского государственного технического университета. – 2012. – № 20. – С. 26-32.

44. Сидоров, К.В. К вопросу оценки эмоциональности естественной и синтезированной речи по объективным признакам / К.В. Сидоров, М.В. Калюжный // Вестник Тверского государственного технического университета. – Тверь: 2011. – № 18. – С. 81-85.

45. Соловьева, Е.С. Методы и алгоритмы обработки, анализа речевого сигнала для решения задач голосовой биометрии / Е.С. Соловьева. – М.: Диссертация на соискание уч. ст. к.т.н., 2008. – 149 с.

46. Сорокин, В.Н. Распознавание личности по голосу: аналитический обзор / В.Н.Сорокин, В.В.Вьюгин, А.А. Тананыкин // Информационные процессы. – 2012. – Т. 12. – № 1. – С. 1-30.

47. Список функций Neural Network Toolbox: Функции создания новой сети. – 2012 [электронный ресурс]. Дата обновления: 21.04.2012. – URL: http://matlab.exponenta.ru/neuralnetwork/book2/11/newff.php (дата обращения: 08.01.2013).

48. Тупицин, Г.С Использование бинарных масок для повышения качества идентификации диктора / Г.С Тупицин, М.В. Сагациян // Международная конференция студентов и аспирантов «Путь в науку». – Ярославль: 2014.

49. Тупицин, Г.С. Использование априорного отношения сигнал/шум для построения бинарных масок в задаче идентификации диктора / Г.С. Тупицин, А.В. Куликов, М.В. Сагациян // Докл. междунар. конф. «Системы синхронизации, формирования и обработки сигналов в инфокоммуникациях». – Ярославль: 2013. – Т.1. – С. 168-170.

50. Тупицин, Г.С. Повышение качества закрытой текстонезависимой идентификации диктора с помощью бинарных масок / Г.С. Тупицин, М.В. Сагациян // Международная молодежная научно-практическая конференция «Путь в науку», секция «цифровая обработка сигналов и изображений». – Ярославль: 23-30 апреля 2015.

51. Тупицин, Г.С. Повышение качества закрытой текстонезависимой идентификации диктора в условиях шумов с помощью бинарных масок / Г.С. Тупицин, М.В. Сагациян // Докл. 12-й междунар. научно-технической конф. «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации». – Курск: 2015.

52. Тупицин, Г.С. Повышение качества идентификации диктора в условиях шумов с помощью бинарных масок / Г.С. Тупицин, А.В. Куликов, М.В. Сагациян // Доклад международной конференции «Перспективные технологии в средствах передачи информации». – Владимир: 2013.

53. Уоссермен, Ф. Нейрокомпьютерная техника: Теория и практика / Ф. Уоссермен // Пер. с англ. Ю.А. Зуева и В.А. Точенова. – М.: Мир, 1992. – 184 с.

54. Физиология речи. Восприятие речи человеком. / Л.А. Чистович [и др.]; под ред. Н.П. Бехтерева [и др.]. – Ленинград: издательство «Наука»,

1976. – 388 с.

55. Хайкин, С. Нейронные сети: полный курс. / С. Хайкин. – М.: Вильямс, 2005. – 1104 с.

56. Хейдоров, И.Э. Классификация эмоционально окрашенной речи с использованием метода опорных векторов / И.Э. Хейдоров, Я. Цзинбинь, [и др.] // Речевые технологии. – Санкт-Петербург, 2008. – № 3. – С. 63-71.

57. Хроматиди, А.Ф. Исследование психофизиологического состояния человека на основе эмоциональных признаков речи / А.Ф. Хроматиди. – Таганрог: Диссертация на соискание уч. ст. к.т.н., 2005. – 154 с.

58. Что такое психология / Ж. Годрфруа. – М.: «Мир», 1999. – 496 с.

59. Anzalone, M. Determination of the potential benefit of time-frequency gain manipulation / M. Anzalone, L. Calandruccio, K. Doherty, L. Carney // Ear and Hearing. – 2006. – Vol. 27. – № 5. – P. 480-492.

60. Avnimelech, R. Boosted Mixture of Experts: An Ensemble Learning Scheme / R. Avnimelech, N. Intrator // Neural Computation. – 1999. – Vol. 11(2). – P. 483-497.

61. Bansal, S. Speaker identification system using close set / S. Bansal, A. Hooda, Anima // International journal of research in Engineering and Technology. – 2012. – Vol. 1(3). – P. 411-414.

62. Breiman, L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – Vol. 24(2). – P. 123-140.

63. Chen, Y.T. A study of emotion recognition on mandarin speech and its performance evaluation: Ph. D. dissertation / Y.T. Chen. – Tatung, 2008.

64. Fletcher, R. Practical Methods of Optimization / R. Fletcher // John Wiley & Sons. – 1975.

65. Furui, S. An overview of speaker recognition technology / S. Furui // ESCA Workshop on Automatic Speaker Recognition, Identification and Verification. – 1994. – P. 1-9.

66. Gibak, K. Why do speech-enhancement algorithms not improve speech intelligibility? / K. Gibak, C Loizou Phillips // Processing of ICASSP-2010. –

2010. – Vol. 1. – P. 397-400.

67. Gill, P.E. Practical Optimization / P.E. Gill, W. Murray, M.H. Wright // Academic Press. Inc. – 1980.

68. Hangartner, R.D. Probabilistic computation by Neuromine Networks / R.D. Hangartner , P. Cull // BioSystems. – 2000. – Vol. 14. – P. 167-176.

69. Haykin, S. Neural networks, a comprehensive foundation / S. Haykin. – New York: Macmillan College Publishing Company, 1994.

70. Hestenes, M. Conjugate Direction Methods in Optimization / M. Hestenes // Springer Verlag. – New York: 1980.

71. Hinton, O. E. Learning and relearning in Boltzmann machines / O. E. Hinton, T. J. Sejnowski // In Parallel distributed processing. – Cambridge, MA: MIT Press. 1986. – Vol. 1. – P. 282-317.

72. Johansson, F.M. Backpropagation Learning for Multi-Layer Feed-Forward Neural Networks Using the Conjugate Gradient Method / F.M. Johansson, F.U. Dowla, D.M. Goodman // Lawrence Livermore National Laboratory, Preprint UCRL-JC-104850. – 1990.

73. Kauchik, Mitra. A Scalable Projective Bundle Adjustment Algorithm using the L_∞ Norm / Mitra Kauchik, Chellappa Rama // Dept. of Electrical and Computer Engineering University of Maryland, College Park, MD. – USA: 2008. – P. 79-86.

74. Kotomin, A.V. Voice Commands Recognition Using Convolutional Neural Networks/ A.V. Kotomin // Proceedings of Junior research and development conference of Ailamazyan Pereslavl university. – Pereslavl, 2012. – P. 1-10.

75. Largest neuronal network simulation achieved using K computer. – 2013. [электронный ресурс]. Дата обновления: 02.08.2013. – URL: http://www.riken.jp/en/pr/press/2013/20130802_1/ (дата обращения: 08.05.2015).

76. Li, N. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction / N. Li, P.C. Loizou // JASA. – 2008. – Vol.

123. – № 3. – P. 1673-1682.

77. Loura, L.M. Fluid-fluid membrane microheterogenety: a fluorescence resonance energy transfer study / L.M. Loura // Biophysical Journal. – 2001. – № 80. – P. 776-788.

78. Mago, Vijay Kumar. Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition: Advancing Technologies: Advancing Technologies / Vijay Kumar Mago, Nitin Bhatia. – USA: IGI Global. 2011. – 786 p.

79. Makarova, V. RUSLANA: a database of russian emotional utterances / V. Makarova, V.A. Petrushin // ICSLP. – 2002. – P. 2041-2044.

80. May, T. Noise-robust speaker recognition combining missing data techniques and universal background modeling / T. May, S. van de Par, A. Kohlrausch // IEEE Trans. Audio, Speech, Lang. Process. – 2012. – Vol. 20, – №1. – P. 108-121.

81. Morist, M.U. Emotional speech synthesis for a radio dj: corpus design and expression modeling: master thesis MTG-UPF dissertation / M.U. Morist. – Barcelona, 2010.

82. Müller, M.F. A scaled conjugate gradient algorithm for fast supervised learning / M.F. Müller // Neural Networks. – 1993. – Vol. 1. – P.525-534.

83. Ortega-Garcia, J. Overview of speech enhancement techniques for automatic speaker recognition / J. Ortega-Garcia, J. Gonzalez-Rodriquez // Proc. Int. Conf. Spoken Lang. Process. – 1996. – Vol. 2. – P. 929-932.

84. Osowski, S. Sieci neuronowe w ujeciu algorytmicznym / S. Osowski. – Warszawa: WNT, 1996.

85. Pham, D.T. Training of Elman networks and dynamic system modeling / D.T. Pham, X. Liu // International Journal of Systems Science. – 1996. – Vol. 27. – № 2. – P. 221-226.

86. Pinkus, A. Approximation theory of the MLP model in neural networks / A. Pinkus // Acta Numerica. – 1999. – Vol. 8. – P. 143-195.

87. Plapous, C. Improved signal-to-noise ratio estimation for speech

enhancement / C. Plapous, C. Marro, P. Scalart // IEEE Transactions on Audio, Speech, and Language Processing. – 2006. – Vol. 14(6). – P. 2098-2108.

88. Powell, M. Restart Procedures for the Conjugate Gradient Method / M. Powell // Mathematical Programming. – 1977. – P.241-254.

89. Rabiner, L.R. A tutorial on Hidden Markov models and selected application in speech recognition / L.R. Rabiner // Proceesding of the IEEE. – 1989. – Vol. 77(2). – P. 257-286.

90. Renevey, P. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion / P. Renevey, A. Drygajlo // in Proc. Consistent and Reliable Acoustic Cues for Sound Analysis Workshop. – 2001. – P. 71-74.

91. Rodriguez, R. Noisy Spiking Neurons and Networks / R. Rodriguez // BioSystems. – 1998. – Vol. 48. – P. 187–194.

92. Roman, N. Pitch-based monaural segregation of reverberant speech / N. Roman, D. Wang // The Journal of the Acoustical Society of America. – 2006. – Vol. 120. – P. 458-469.

93. Roman, N. Speech segregation based on sound localization / N. Roman, D. Wang, G. Brown // The Journal of the Acoustical Society of America. – 2003. – Vol. 114. – P. 2236-2252.

94. Ronzhin, A.L. Survey of Russian Speech Recognition Systems / A.L. Ronzhin, R.M. Yusupov, I.V. Li, A.B. Leontieva // In Proc. Of 11-th International Conference SPECOM 2006. – St. Petersburg: «Anatoliya», 2006. – P. 54-60.

95. Scalart, P. Speech enhancement based on a priori signal to noise estimation / P. Scalart, J.V. Filho // IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96). – 1996. – Vol. 2. – P. 629-632.

96. Seltzer, M. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition / M. Seltzer, B. Raj, R. Stern // Speech Communication. – 2004. – Vol. 43. – P. 379–393.

97. Shrestha, D. L. Experiments with Ada Boost. RT, an Improved Boosting

Scheme for Regression / D. L. Shrestha, D. P. Solomatine // Neural Computation. – 2006. – Vol. 18(7). – P.1678-1710.

98. Siging, W. Recognition of human emotion in speech using modulation spectral features and support vector machines: Master of Science dissertation / W. Siging. – Kingston, 2009.

99. Thimm, G. Neural network initialization / G. Thimm, E. Fiesler, J. Mira, F. Sandoval // From Natural to Artificial Neural Computation. – Malaga: IWANN, 1995. – P. 533-542.

100. Van Schaik, A. Building blocks for electronic spiking neural networks / A. Van Schaik // Neural Networks. – 2001. – Vol. 14. – P. 617-628.

101. Vimala, C. A Review on Speech Recognition Challenges and Approaches / C. Vimala, Dr.V. Radha // World of Computer Science and Information Technology Journal (WCSIT). – 2012. – Vol. 2(1). – P. 1-7.

102. Wang, D. Eds. Computational Auditory Scene Analysis / D. Wang, G. J. Brown // Wiley & IEEE Press, Hoboken. – New Jersey: 2006.

103. Wilamowski, B.M. Improved Computation for Levenberg–Marquardt Training / B. M. Wilamowski, H. Yu // Neural Networks, IEEE Transactions on Neural Networks. – 2010. – Vol.21, №6. – P.930-937.

104. Xuedong, H. Spoken language processing: a guide to theory, algorithm, and system development/ H. Xuedong, A. Acero, Hsiao-Wuen Hon. – New Jersey: Prentice-Hall PTR Upper Saddle River, 2001. – P. 19-68.

ПРИЛОЖЕНИЕ 1. ИНФОРМАЦИЯ О РЕЧЕВОЙ БАЗЕ «КРИПТОН-01»

Обучающие дикторы												
Номер раздела корпуса	**А.1**	**Б.1**	**В.1**	**Г.1**	**Д.1**	**Е.1**	**Ж.1**	**З.1**	**И.1**	**К.1**	**Л.1**	**М.1**
Количество дикторов	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Количество мужчин	1	1	2	2	3	3	4	5	6	7	8	9
Количество женщин	0	1	1	2	2	3	3	3	3	3	3	3
Возрастной интервал, лет	25	25/30	19/30	17/30	17/32	17/33	17/35	17/35	17/37	17/38	17/38	17/38
Тестирующие дикторы												
Номер раздела корпуса	**А.2**	**Б.2**	**В.2**	**Г.2**	**Д.2**	**Е.2**	**Ж.2**	**З.2**	**И.2**	**К.2**	**Л.2**	**М.2**
Количество дикторов	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Количество мужчин	1	1	2	3	3	4	5	6	7	8	8	9
Количество женщин	0	1	1	1	2	2	2	2	2	2	3	3
Возрастной интервал, лет	21	18/21	18/28	18/28	18/29	18/31	18/32	18/32	18/34	18/35	18/35	18/35

Обучающие и тестирующие дикторы												
Номер корпуса	**А**	**Б**	**В**	**Г**	**Д**	**Е**	**Ж**	**З**	**И**	**К**	**Л**	**М**
Количество дикторов	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**
Количество мужчин	2	2	4	5	6	7	9	11	13	15	16	18
Количество женщин	0	2	2	3	4	5	5	5	5	5	6	6
Возрастной интервал, лет	21/25	18/30	18/30	13/30	17/32	17/33	17/35	17/35	17/37	17/38	17/38	17/38

Речевые сигналы:

№	**Значение сигнала**	**№**	**Значение сигнала**	**№**	**Значение сигнала**	**№**	**Значение сигнала**	**№**	**Значение сигнала**
S1	Один	S3	Три	S5	Пять	S7	Семь	S9	Девять
S2	Два	S4	Четыре	S6	Шесть	S8	Восемь	S10	Ноль

Речевая база «КРИПТОН – 01» разработана автором Сагациян М.В. специально для проведения исследований по данной диссертационной работе

ПРИЛОЖЕНИЕ 2. ИНФОРМАЦИЯ О РЕЧЕВОЙ БАЗЕ «КРИПТОН-02»

Обучающие дикторы	
Номер раздела корпуса	**C.1**
Количество дикторов	**10**
Количество мужчин	7
Количество женщин	3
Возрастной интервал, лет	17/38
Тестирующие дикторы	
Номер раздела корпуса	**C.2**
Количество дикторов	**10**
Количество мужчин	8
Количество женщин	2
Возрастной интервал, лет	18/35
Обучающие и тестирующие дикторы	
Номер корпуса	**C**
Количество дикторов	**20**
Количество мужчин	15
Количество женщин	5
Возрастной интервал, лет	17/38

Речевые сигналы:

№	Значение сигнала	№	Значение сигнала	№	Значение сигнала
S1	Здравствуйте	S35	Будет	S69	Шесть
S2	Досвидания	S36	Было	S70	Семь
S3	Включить	S37	Есть	S71	Восемь
S4	Выключить	S38	Погода	S72	Девять
S5	Свет	S39	Солнечно	S73	Десять
S6	Отопление	S40	Дождь	S74	Одиннадцать
S7	Вентиляция	S41	Снег	S75	Двенадцать
S8	Температура	S42	Град	S76	Тринадцать
S9	Показания	S43	Пасмурно	S77	Четырнадцать
S10	Поставить	S44	День	S78	Пятнадцать
S11	На охрану	S45	Неделя	S79	Шестнадцать
S12	Почта	S46	Месяц	S80	Семнадцать
S13	Написать	S47	Год	S81	Восемнадцать
S14	Письмо	S48	Утро	S82	Девятнадцать
S15	Адрес	S49	День	S83	Двадцать
S16	Название	S50	Вечер	S84	Тридцать
S17	События	S51	Ночь	S85	Сорок
S18	Новости	S52	Сегодня	S86	Пятьдесят
S19	Перевод	S53	Завтра	S87	Шестьдесят

S20	Спорт	S54	Вчера	S88	Семьдесят
S21	Финансы	S55	Аптека	S89	Восемьдесят
S22	Курс валют	S56	Магазин	S90	Девяносто
S23	Показать	S57	Автосервис	S91	Сто
S24	Видео	S58	Вокзал	S92	Тысяча
S25	Фотографии	S59	Аэропорт	S93	Миллион
S26	Отчет	S60	Электростанция	S94	Рубль
S27	Ближайшее	S61	Завод	S95	Доллар
S28	Автоответчик	S62	Номер	S96	Юань
S29	Громкость	S63	Ноль	S97	Евро
S30	Тише	S64	Один	S98	Россия
S31	Громче	S65	Два	S99	Европа
S32	Время	S66	Три	S100	Китай
S33	Будильник	S67	Четыре	S101	Ярославль
S34	Радио	S68	Пять	S102	Москва

Речевая база «КРИПТОН – 02» разработана автором Сагациян М.В. специально для проведения исследований по данной диссертационной работе

ПРИЛОЖЕНИЕ З. СВИДЕТЕЛЬСТВО О РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭЛЕКТРОННОЙ ВЫЧИСЛИТЕЛЬНОЙ МАШИНЫ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2015616920

«NN-SCG speech recognition - научно-исследовательская программа по изучению алгоритмов нейросетевого дикторонезависимого распознавания речевых команд»

Правообладатели: ***Сагациян Максим Владимирович (RU), Тупицин Геннадий Сергеевич (RU)***

Авторы: ***Сагациян Максим Владимирович (RU), Тупицин Геннадий Сергеевич (RU)***

Заявка № **2015614146**

Дата поступления **30 апреля 2015 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***25 июня 2015 г.***

Врио руководителя Федеральной службы по интеллектуальной собственности

Л.Л. Кирий

ПРИЛОЖЕНИЕ 4. АКТЫ ВНЕДРЕНИЯ РЕЗУЛЬТАТОВ РАБОТЫ

УТВЕРЖДАЮ
Ген. Директор ООО «ПАНТЕОН»
_____ С.И. Тигин
«05» ___марта_____ 2015 г.

АКТ

о внедрении результатов диссертационной работы Сагациян Максима Владимировича, выполненной в Ярославском государственном университете им. П.Г. Демидова (ЯрГУ), на тему «Разработка и исследование коллективных нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов»

Результаты диссертационной работы Сагациян М.В. «Разработка и исследование коллективных нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов» нашли применение в разработках многоотраслевого ООО «ПАНТЕОН». Особый практический интерес представляют следующие результаты диссертации:

1. Научно-исследовательская программа «NN-SCG speech recognition» для исследования коллективных и модифицированных коллективных нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых сигналов.
2. Речевые базы «КРИПТОН-01» и «КРИПТОН-02», для строительства системы речевого интерфейса вывода с ЭВМ для различных приложений и для анализа коллективных и модифицированных коллективных нейросетевых алгоритмов в задаче дикторонезависимого распознавания речевых сигналов.

Главный инженер _____ С.И. Вовчек

УТВЕРЖДАЮ

Ген. директор ООО «А-Вижн»

_____ И.В. Апальков

«16» ___марта____ 2015 г.

АКТ

о внедрении результатов диссертационной работы Сагациян Максима Владимировича, выполненной в Ярославском государственном университете им. П.Г. Демидова (ЯрГУ), на тему «Разработка и исследование коллективных нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов»

Результаты диссертационной работы Сагациян М.В. «Разработка и исследование коллективных нейросетевых алгоритмов дикторонезависимого распознавания речевых сигналов» нашли применение в разработках ООО «А-Вижн». Особый практический интерес представляют следующие результаты диссертации:

1. Нейросетевой алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением масштабируемых сопряженных градиентов (Scaled Conjugate Gradient Backpropagation, SCG) с блоком шумоподавления дикторонезависимого распознавания русскоязычных речевых сигналов работающий в условиях шумов.

2. Модифицированный нейросетевой алгоритм bagging-коллектива на основе персептронов Розенблатта с обучением SCG с блоком шумоподавления дикторонезависимого распознавания русскоязычных речевых сигналов работающий в условиях шумов.

Комиссия в составе:

Технический директор _____ А.С. Конюхов

Инженер-программист _____ Н.Б. Герасимов

Инженер _____ Е.А. Жемчугова