СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Исследования по искусственным нейронным сетям связаны с тем, что способ обработки информации человеческим мозгом в корне отличается от методов, применяемых обычными цифровыми компьютерами. Мозг представляет собой чрезвычайно сложный, нелинейный, параллельный компьютер.

Мозг человека обладает способностью организовывать работу нейронов, так, чтобы они могли выполнять конкретные задачи (такие как распознавание образов, обработку сигналов органов чувств, моторные функции) во много раз быстрее, чем могут позволить самые быстродействующие современные компьютеры. Примером такой задачи обработки информации может служить обычное зрение. В функции зрительной системы входит создание представления окружающего мира в таком виде, который обеспечивает возможность взаимодействия с этим миром. Более точно, мозг последовательно выполняет ряд задач распознавания (например, распознавание знакомого лица в незнакомом окружении). На это у него уходит около 100-200 миллисекунд, в то время как выполнение аналогичных задач даже меньшей сложности на компьютере может занять несколько дней.

Разработка искусственных нейронных сетей началась в начале XX века, но только в 90-х годах, когда вычислительные системы стали достаточно мощными, нейронные сети получили широкое распространение. Создание нейронных сетей было вызвано попытками понять принципы работы человеческого мозга и, без сомнения, это будет влиять и на дальнейшее их развитие. Однако, в сравнении с человеческим мозгом нейронная сеть сегодня представляют собой весьма упрощенную модель, но несмотря на это весьма успешно используются при решении самых различных задач. Хотя решение на основе нейронных сетей может выглядеть и вести себя как обычное программное обеспечение, они различны в принципе, поскольку большинство реализаций на основе нейронных сетей «обучается», а «не программируется»: сеть учиться выполнять задачу, а не программируется непосредственно.

Мозг и цифровой компьютер выполняют совершенно разные задачи и имеют различные свойства. В типичном мозгу человека имеется в 1000 раз больше нейронов, чем логических элементов в процессоре типичного компьютера высокого класса. В соответствии с законом Мура и с учетом того, что по некоторым расчетам, количество нейронов в мозгу должно удваиваться примерно через каждые 2-4 миллиона лет, может быть сделан прогноз, что количество логических элементов в процессоре станет равным количеству нейронов в мозгу примерно к 2020 году. Безусловно, эти прогнозы мало о чем говорят; кроме того, это различие в отношении количества элементов является незначительным по сравнению с различием в скорости переключения и степени распараллеливания. Микросхемы компьютера способны выполнить отдельную команду меньше чем за наносекунду, тогда как нейроны действуют в миллионы раз медленнее. Но мозг сторицей восполняет этот свой недостаток, поскольку все его нейроны действуют одновременно, тогда как большинство современных компьютеров имеет только один процессор (но с несколькими ядрами) или небольшое количество процессоров. Таким образом, даже несмотря на то, что компьютер обладает преимуществом более чем в миллион раз в физической скорости переключения, оказывается, что мозг по сравнению с ним выполняет все свои действия примерно в 100 000 раз быстрее [5].

На сегодняшний день задача распознавания речи является одной из самых приоритетных в направлениях исследования искусственного интеллекта. Данной проблемой занимаются такие гиганты корпоративного мира как Google, Microsoft, IBM, Intel, телефонных компаниях, а так же в ведущих исследовательских центрах.

К примеру, в Intel разрабатываются механизмы синхронизации речевых и сенсорных

команд введенных пользователем ПК. А в IBM технология голосового управления автомобильными навигационными системами.

Но данные технологии решают частные задачи, которые не идут ни в какое сравнение с основной задачей поставленной перед исследователями ИИ – распознавание человеческой речи. Но тут перед разработчиками встает масса проблем:

- Люди в массе своей говорят не разборчиво, и в не самых подходящих шумовых условиях.
- Язык не является застывшим образованием. В большинстве языков имеется несколько диалектов, и даже в рамках одного диалекта существует несколько равноправных произношений одного и того же слова.
- У каждого из нас есть свои речевые особенности, которые могут затруднить распознавание речи
- В естественной речи содержатся звуки-паразиты, которые стоит отфильтровывать
- Для разработок систем распознавания русского языка стоит включит проблему сложности декодирования русской речи. Русский словарь, эквивалентный английскому из 50000 слов, должен содержать около миллиона словоформ. Плюс к этому, в русском языке произвольный порядок слов в предложении. Что усложняет семантический и синтаксический разбор.

Так же сюда стоит включить эффект ко артикуляции (одна и та же фонема может звучать по разному, в зависимости от того какие звуки предшествуют и следуют за ней).

Существующие технологии не могут решить эти проблемы в комплексе. Поэтому от поставленной задачи техники распознавания речи меняются. Общего у них много, но система голосового набора номера в сотовом телефоне на порядок отличается от системы распознавания речи общего назначения.

Долгое время системы распознавания речи требовали, чтобы пользователь выговаривал каждое слово отдельно, однако в самом конце прошлого века появились пакеты, умеющие обрабатывать «слитную» речь. Под данным термином понимается, что уже не обязательно выговаривать каждое слово, но делать паузы в предложениях необходимо. Желательно выдавать ей самостоятельные фрагменты предложений или, целые предложения, если они короткие. Во многих современных пакетах есть синтаксические и семантические модули, и подобная разбивка облегчит распознавание, одновременно улучшив качество. Т.е. под «слитной» речью понимается диктовка текста.

Другой важный критерий – привязка к пользователю. На самом деле все современные система распознавание являются распознаваемыми. А разница лишь в том, что дикторонезависимую систему обучил производитель на нескольких тысячах примеров.

Третий критерий – размер словаря. Чем меньше словарь, тем проще обучить систему и сделать ее дикторонезависимой.

Цель работы состоит в получении общих сведений о нейронных сетях, а так же принципах моделирования нейронных сетей на примере системы распознавания речи. В данной работе мы рассмотрим проблемы, и возможные способы их решения в данной сфере.

1. Общие сведения о нейронных сетях

Разделяют два типа нейронных сетей — биологические нейронные сети и искусственные.

Биологическая нейронная сеть — это так называемая сеть, состоящая из биологических нейронов, связанных или функционально объединёных в нервной системе

Искусственная нейронная сеть — это сеть, состоящая из «искусственных» нейронов, которые в свою очередь моделируются и используются для для имитации свойств биологических нейронов.

Искусственные нейронные сети применяются в различных областях науки: начиная от систем распознавания речи до распознавания вторичной структуры белка, классификации различных видов рака и генной инженерии. Однако, как они работают и чем они хороши

Когда речь идет о задачах, отличных от обработки больших массивов информации, человеческий мозг обладает большим преимуществом по сравнению с компьютером. Человек может распознавать лица, даже если в помещении будет много посторонних объектов и плохое освещение. Мы легко понимаем незнакомцев даже когда находимся в шумном помещении. Но, несмотря на годы исследований, компьютеры все еще далеки от выполнения подобных задач на высоком уровне.

Человеческий мозг удивительно надежный: по сравнению с компьютером он не перестанет работать только потому, что несколько клеток погибнет, в то время как компьютер обычно не выдерживает каких-либо поломок в CPU. Но самой удивительной особенностью человеческого мозга является то, что он может учиться. Не нужно никакого программного обеспечения и никаких обновлений, если мы хотим научиться ездить на велосипеде.

Расчеты головного мозга производятся посредством тесно взаимосвязанных нейронных сетей, которые передают информацию, отсылая электрические импульсы через нейронные проводки, состоящие из аксонов, синапсов и дендритов. В 1943 году, компания McCulloch and Pitts смоделировала искусственный нейрон, как переключатель, который получает информацию от других нейронов и в зависимости от общего взвешенного входа, либо приводится в действие, либо остается неактивным. В узле ИНС пришедшие сигналы умножаются на соответствующие веса синапсов и суммируются. Эти коэффициенты могут быть как положительными (возбуждающими), так и отрицательными (тормозящими). В 1960 годах было доказано, что такие нейронные модели обладают свойствами, сходными с мозгом: они могут выполнять сложные операции распознавания образов, и они могут функционировать, даже если некоторые связи между нейронами разрушены. Демонстрация персептона Розенблатта показала, что простые сети из таких нейронов могут обучаться на примерах, известных в определенных областях. Позже, Минский и Паперт доказали, что простые пресептоны могут решать только очень узкий класс линейно сепарабельных задач (см. ниже), после чего активность изучения ИНС уменьшилась. Тем не менее, метод обратного распространения ошибки обучения, который может облегчить задачу обучения сложных нейронных сетей на примерах, показал, что эти проблемы могут быть и не сепарабельными.

Программа NETtalk применяла искусственные нейронные сети для машинного чтения текста и была первым широкоизвестным приложением. В биологии, точно такой же тип сети был применен для прогнозирования вторичной структуры белка; в самом деле, некоторые из лучших исследователей до сих пор пользуются тем же методом. С этого началась другая волна, вызвавшая интерес к исследованиям ИНС и поднявшая шумиху вокруг магического обучения мыслящих машин.

ИНС могут быть созданы путем имитации модели сетей нейронов на компьютере.

Используя алгоритмы, которые имитируют процессы реальных нейронов, мы можем заставить сеть «учиться», что помогает решить множество различных проблем. Модель нейрона представляется как пороговая величина.

Модель получает данные от ряда других внешних источников, определяет значение каждого входа и добавляет эти значения. Если общий вход выше пороговой величины, то выход блока равен единице, в противном случае – нулю. Таким образом, выход изменяется от 0 до 1, когда общая «взвешенная» сумма входов равна пороговой величине. Точки в исходном пространстве, удовлетворяющие этому условию, определяют, так называемые, гиперплоскости. В двух измерениях, гиперплоскость – линия, в то время как в трех измерениях, гиперплоскость является нормальной (перпендикулярной) плоскостью. Точки с одной стороны от гиперплоскости классифицируются как 0, а точки с другой стороны – 1. Это означает, что задача классификации может быть решена с использованием пороговой величины, если два класса будут разделены гиперплоскостью. Эти проблемы называются линейно сепарабельными.

2. Характеристики и классификация систем распознавания речи

В настоящий момент системы распознавания речи характеризуются следующими признаками:

- дикторозависимость
- раздельность речи
- назначение

Дикторозависимая система предназначена для использования одним диктором, в то время как дикторонезависимая система предназначена для работы с любым диктором.

Дикторонезависимоть — очень ценное качество системы, но в то же время очень труднодостижимое, так как при обучении системы она настраивается на парметры того диктора, на примере которого обучется.

Таким образом, в процессе создания дикторонезависимой системы применяются гораздо более сложные алгоритмы обучения. В таких системах частота ошибок распознавания обычно в 3-5 раз больше, чем в дикторозависимых.

Если в речи слова разделяются интервалами тишины, то говорят, что эта речь — раздельная. Естественная речь, как правило, слитная. Распознавание слитной речи намного труднее в связи с тем, что границы отдельных слов не чётко определены и их произношение сильно искажено смазыванием произносимых звуков.

Назначение системы определяет требуемый уровень абстракции, с которым будет происходить распознавание произносимой речи. Можно выделить 2 типа систем распознавания речи:

- командные системы
- системы диктовки

В командных системах, в общем случае, распознавания слова или фразы происходит как распознавание единого речевого элемента. То есть, при распознавании учитываются исключительно физические характеристики сигнала, а не смысловая нагрузка произносённой речи.

Системы диктовки также анализируют контекст речевого элемента и поэтому требуют большей точности распознавания. Алгоритмы, задействованные в таких системах, например, скрытые сети Макркова, анализируют не только уникальные параметры самого речевого сигнала, но и контекст каждого произнесённого речевого элемента. Также могут применяться алгоритмы динамического программирования. Для анализа контекста в системе необходимо предусмотреть набор грамматических правил, которые должен удовлетворить произносимый и распознаваемый текст. Чем строже эти правила, тем проще реализовать систему распознавания, и тем ограниченней будет набор предложений, которые она сможет распознать.

Существуют следующие подходы к выделению информативных признаков, описывающих речевой сигнал:

- метод линейного предсказания;
- спектральный анализ.

Спектральный анализ отличается от линейного предсказания тем, что оценки среднего значения усредненного шума вычитаются из спектра, вычисленного по зашумленным данным.

Наиболее часто употребляются два подхода к классификации и распознаванию:

- мера близости параметров (такая функция называется метрикой);
- нейронные сети.

Второй подход не использует вспомогательных функций, но моделирует процесс распознавания в биологических системах. Этот подход представляется более перспективным

в настоящее время.

В системах распознавания речи выделяются две основные подсистемы:

- подсистема предварительной обработки речевых сигналов;
- подсистема классификации речевых сигналов.

В настоящей работе представлены модель распознавания речи на основе искусственных нейронных сетей.

2.1 Распознавание по образцу

В качестве примера можно рассмотреть мобильный телефон. Такая система предназначена для ускоренного выбора абонентов из телефонной книги телефона с помощью голоса.

При добавлении нового контакта в телефонную книгу предоставляется возможность привязать голосовую метку, идентифицирующую контакт, например, произнести имя или фамилию абонента. Возможно придётся повторить несколько раз.

Теперь, для вызова абонента достаточно нажать одну из кнопок и произнести голосовую метку. Номер будет вызван из телефонной книги, после чего телефон попытается совершить вызов.

Помимо телефона существуют и другие устройства с подобным голосовым управлением, например компьютерные клавиатуры. Такие клавиатуры оборудуются микрофоном и позволяют привязывать голосовые метки определённым клавишам , комбинациям или последовательностям клавиш. Разработчики утверждают, что так можно ускорить ввод информации, однако качество распознавания зависит от уровня шума.

Это технология работает достаточно хорошо, если устройством пользуется только один человек, а общее количество меток не превышает пары десятков. Если вы «обучите» ваше устройство выполнять ваши команды, то оно будет реагировать только на ваш голос. Таким образом, подобные системы относятся к классу систем, зависимых от диктора.

2.2 Выделение лексических элементов

В данной работе мы сосредоточим своё внимание на подходе к созданию систем, основанном на выделении из речи лексических элементов, таких как фонемы и аллофоны.

Проводя осциологические исследования невозможно выделить из речи фонемы и аллофоны, анализируя только форму огибающе звукового сигнала; нельзя ограничиться составлением баз данных из записей звуковых сигналов всех фонем, аллфонов и других лексических элементов для последующего сравнения формы сигналов в процессе распознавания. Для этого необходимы более сложные методы.

3. Предварительная обработка звуковых сигналов

Перед тем как предпринимать попытки распознавания речи, нужно выполнить предварительную обработку речевого сигнала. В ходе этой обработки следует удалить шумы и посторонние сигналы, частотный спектр которых находится вне спектра человеческой речи. Такую обработку можно выполнить при помощи аналоговых или цифровых полосовых фильтров. Отфильтрованный звуковой сигнал нужно оцифровать, выполнив аналого-цифровое преобразование.

Всю предварительную обработку звукового сигнала можно сделать при помощи стандартного звукового адаптера, установленного в компьютере. Дополнительная цифровая обработка звукового сигнала (например, частотная фильтрация) может выполняться центральным процессором компьютера. Таким образом, при использовании современных персональных компьютеров системы распознавания речи не требуют для своей работы какого-либо специального аппаратного обеспечения.

Важным этапом предварительной обработки входного сигнала является нормализация уровня сигнала. Это позволяет уменьшить погрешности распознавания, связанные с тем, что диктор может произносить слова с разным уровнем громкости.

Заметим, однако, что если входной звуковой сигнал имеет слишком малый уровень громкости, то после нормализации может появиться шум. Поэтому для успешной работы системы распознавания речи необходимо отрегулировать оптимальным образом чувствительность микрофона. Чрезмерная чувствительность может привести к нелинейным искажениям сигнала и, как следствие, к увеличению погрешности распознавания речи.

4. Выделение информативных признаков речевого сигнала

Как уже было сказанно раннее, информации об амплитуде и форме огибающей речевого сигнала не достаточно для выделения из речи лексических элементов. В зависимости от различных обстоятельств форма огибающей речевого сигнала может меняться в широких пределах, что затрудняет задачу распознавания.

Для решения задачи распознавания необходимо выделить первичные признаки речи, которые будут использованы на последующих этапах процесса распознавания. Первичные признаки выделяются посредством анализа спектральных и динамических характеристик речевого сигнала.

4.1 Спектральное представление речи

Для выделения информативных признаков речевого сигнала используется спектральное представление речи. При этом на первом этапе осуществляется получение частотного спектра речевого сигнала с помощью набора программных полосовых фильтров (выполняя так называемое дискретное преобразование Фурье).

На втором этапе выполняются преобразования полученного спектра речевого сигнала:

- логарифмическое изменение масштаба в пространстве амплитуд и частот;
- сглаживание спектра с целью выделения его огибающей;
- кепстральный анализ(cepstral analysis), т.е. обратное преобразование Фурье от логарифма прямого преобразования.

Перечисленные выше преобразования позволяют учитывать такие особенности речевого сигнала, как понижение информативности высокочастотных участков спектра, логарифмическую чувствительность человеческого уха, и т.д.

4.2 Учёт динамики речи

Помимо спектральных характеристик, необходимо учитывать и динамические особенности речи. Для этого используют дельта-параметры, представляющие собой производные по времени от основных параметров.

При этом мы можем отслеживать не только изменение параметров речи, но и скорость их изменения.

5. Выделение фонем и аллофонов

Для выделения фонем и аллфонов применяются нейронные сети и метод формирования нейронных ансамблей.

При этом обучение выделению примитивов речи (фонем и аллофонов) может заключаться в формировании нейронных ансамблей, ядра которых соответствуют наиболее частой форме каждого примитива.

Формирование нейронных ансамблей представляет собой процесс обучения нейронной сети без учителя, при котором происходит статистическая обработка всех сигналов, поступающих на вход нейронной сети. При этом формируются ансамбли, соответствующие наиболее часто встречающимся сигналам. Запоминание редких сигналов происходит позже и требует подключения *механизма внимания* или иного контроля высшего уровня.

6. Уровни распознования слитной речи

Распознавание слитной речи представляет собой многоуровневый процесс. После предварительной обработки речевого сигнала и выделения из него информативных признаков выполняется выделение лексических элементов речи. Это первый уровень распознавания.

На втором уровне выделяются слоги и морфемы, на третьем — слова, предложения и сообщения (рис. 1).

На каждом уровне сигнал кодируется представителями предыдущих уровней. То есть слоги и морфемы составляются из фонем и аллофонов, слова — из слогов и морфем, предложения и сообщения — из слов.

Рис. 1 — Три уровня распознования слитной речи.

При переходе с уровня на уровень помимо представителей сигналов передаются и некоторые дополнительные признаки, временные зависимости и отношения между сигналами. Собирая сигналы с предыдущих уровней, высшие уровни располагают большим объемом информации (или её другим представлением), и могут осуществлять управление процессами на низших уровнях, например, с привлечением *механизма внимания*.

Механизм внимания используется при обучении нейронной сети. В случае использования такого механизма при появлении образца, неизвестного нейронной сети, скорость обучения многократно возрастает. При этом редко встречающийся образец запоминается в нейронной сети.

7. Применение нейронных сетей для распознавания речи

При обучении сети с учителем можно научить сеть распознавать объекты, принадлежащие заранее определенному набору классов. Если же сеть обучается без учителя, то она может группировать объекты по классам в соответствии с их цифровыми параметрами.

Таким образом, на базе нейронных сетей можно создавать обучаемые и самообучающиеся системы. Формулируются следующие требования к самообучающимся системам:

- Разработка системы заключается только в построении архитектуры системы (В процессе создания системы разработчик создает только функциональную часть, но не наполняет (или наполняет в минимальных объемах) систему информацией. Основную часть информации система получает в процессе обучения.)
- Возможность контроля своих действий с последующей коррекцией (Этот принцип говорит о необходимости обратной связи *Действие-Результат-Коррекция* в системе. Такие цепочки очень широко распространены в сложных биологических организмах и используются на всех уровнях — от контроля мышечных сокращений на самом низком уровне до управления сложными механизмами поведения.)
- Возможность накопления знаний об объектах рабочей области (Знание об объекте — это способность манипулировать его образом в памяти.)

Количество знаний об объекте определяется не только набором его свойств, но ещё и информацией о его взаимодействии с другими объектами, поведении при различных воздействиях, нахождении в разных состояниях, и т.д., т.е. его поведении во внешнем окружении.

Например, знание о геометрическом объекте предполагает возможность предсказать вид его перспективной проекции при любом повороте и освещении. Это свойство наделяет систему возможностью абстрагирования от реальных объектов, т.е. возможностью анализировать объект при его отсутствии, открывая тем самым новые возможности в обучении.

- Автономность системы

При интеграции комплекса действий, которые система способна совершать, с комплексом датчиков, позволяющих контролировать свои действия и внешнюю среду, наделенная вышеприведенными свойствами система будет способна взаимодействовать с внешним миром на довольно сложном уровне.

При этом она будет адекватно реагировать на изменение внешнего окружения (естественно, если это будет заложено в систему на этапе обучения). Способность корректировать свое поведение в зависимости от внешних условий позволит частично или полностью устранить необходимость контроля извне, т.е. система станет *автономной.*

Возможность создания на базе искусственных нейронных сетей самообучающихся систем является важной предпосылкой для их применения в системах распознавания (и синтеза) речи.

7. 1 Представление речи в виде набора числовых параметров

После выделения информативных признаков речевого сигнала можно представить эти признаки в виде некоторого набора числовых параметров (т.е. в виде вектора в некотором числовом пространстве). Далее задача распознавания примитивов речи (фонем и аллофонов) сводится к их классификации при помощи обучаемой нейронной сети.

Нейронные сети можно использовать и более высоких уровнях распознавания слитной речи для выделения слогов, морфем и слов.

7.2 Нейронные ансамбли

В качестве модели нейронной сети, пригодной для распознавания речи и обучаемой без

учителя можно выбрать самоорганизующуюся карту признаков Кохонена. В ней для множества входных сигналов формируется нейронные ансамбли, представляющие эти сигналы. Этот алгоритм обладает способностью к статистическому усреднению, что позволяет решить проблему изменчивости речи.

По сравнению с классическим программированием, когда алгоритм решения той или иной задачи задан жестко, нейронные сети позволяют динамически изменять алгоритм простым изменением архитектуры сети.

7.3 Генетические алгоритмы

Возможность изменения алгоритма работы нейронной сети простым изменением ее архитектуры позволяют решать задачи совершенно новым способом, с помощью так называемых *генетических алгоритмов*.

При использовании генетических алгоритмов создаются правила отбора, позволяющие определить, лучше или хуже справляется новая нейронная сеть с решением задачи. Кроме того, определяются правила модификации нейронной сети.

Изменяя достаточно долго архитектуру нейронной сети и отбирая те архитектуры, которые позволяют решить задачу наилучшим образом, рано или поздно можно получить верное решение задачи.

Генетические алгоритмы обязаны своим появлением эволюционной теории (отсюда и характерные термины: популяция, гены, родители-потомки, скрещивание, мутация). Таким образом, существует возможность создания таких нейронных сетей, которые ранее не изучались исследователями (или не поддаются аналитическому изучению), но, тем не менее, успешно решают задачу.

8. Реализация уровня ввода и вывода в системе SAS

SAS (Статистическая Аналитическая Система) — программное обеспечение, разработанное институтом SAS для расширеной аналитики, бизнес сведений, управление данными и прогноза анализов. Это огромнейший интсрумент рыночной расширенной аналитики.

SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, business intelligence, data management, and predictive analytics. It is the largest market-share holder for advanced analytics [6].

Данная сисетма, выполненная с использованием технологии нейронных сетей, предназначена не только для распознавания, но и для синтеза речи.

Блок-схема системы SAS, соответствующая уровню ввода/вывода, показана на рис. 2.

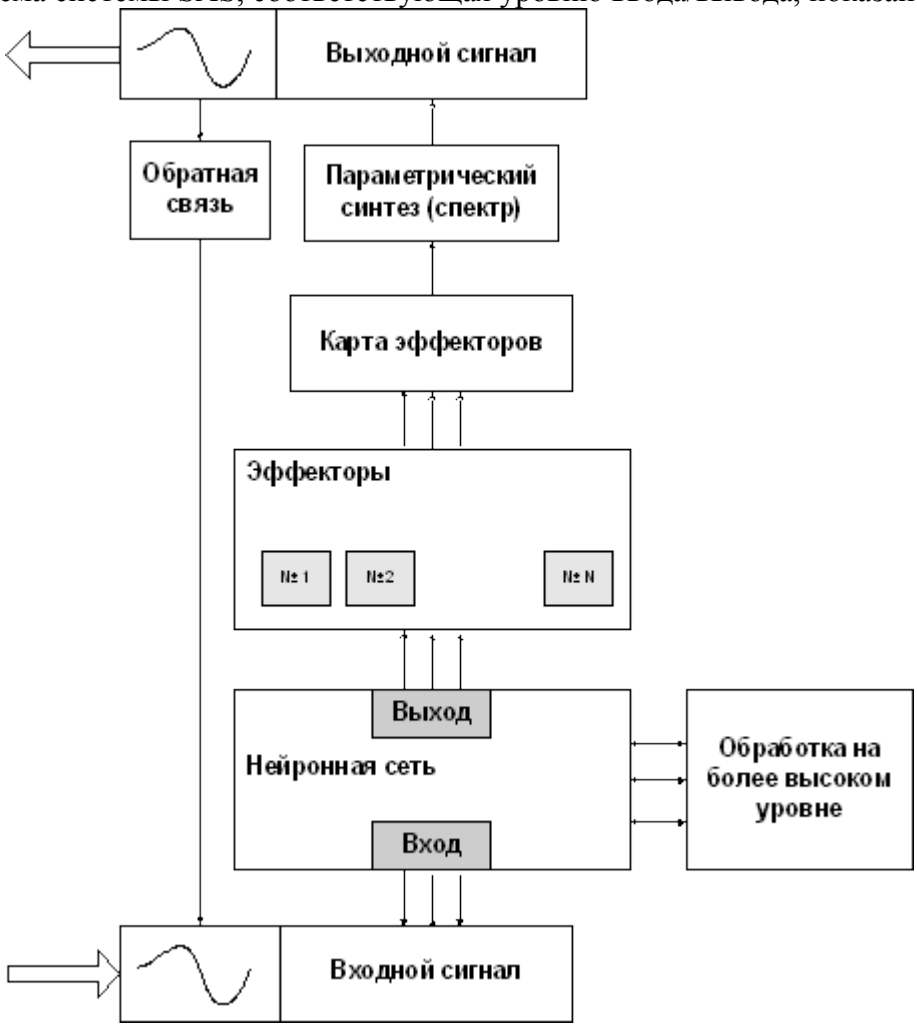

Рис. 2 — Блок схема уровня ввода/вывода

При распознавании речи система SAS осуществляет ввод звуковой информации, предварительную обработку, получение энергетического спектра и выделение примитивов речи.

При синтезе речи осуществляется выделение из нейронной сети запомненного примитива, синтез спектра (частотный параметрический синтез) и преобразование спектра в звуковой сигнал. При обучении последовательным повторением двух вышеописанных процедур осуществляется запоминание примитивов речи в нейронной сети.

8.1 Процесс ввода звука

На рис. 3 изображен процесс ввода звука в системе SAS.

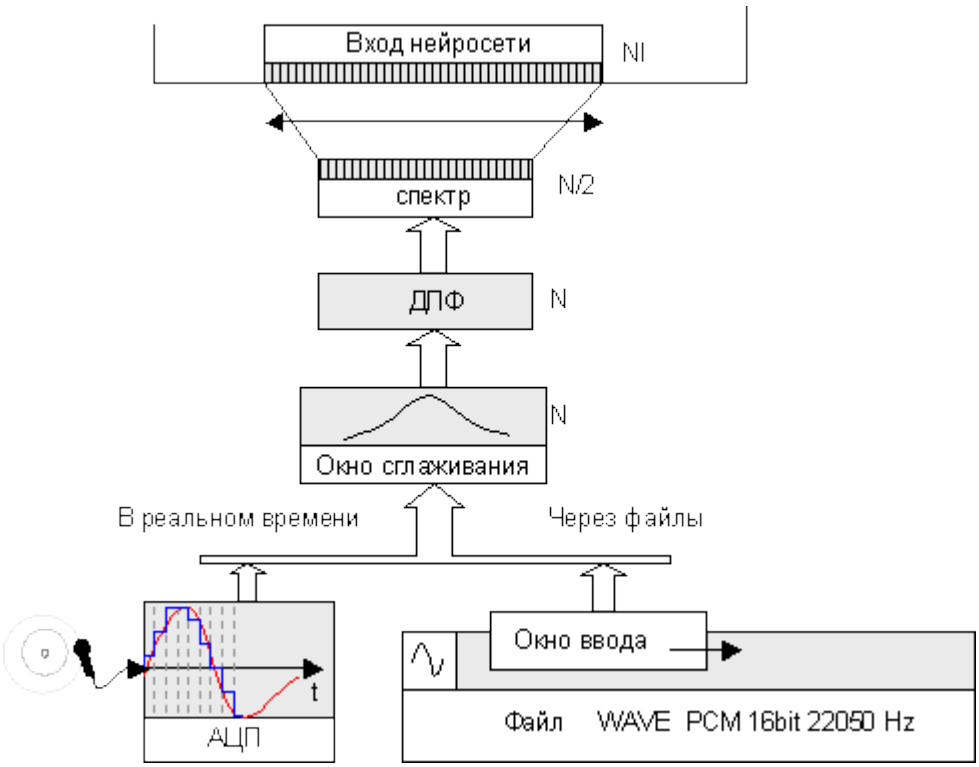

Рис. 3 — Процесс ввода звука в системе SAS

Ввод звука осуществляется в реальном времени через звуковую карту или через файлы формата WAV в кодировке PCM (разрядность 16 бит, частота дискретизации 22 050 Гц). Работа с файлами была предусмотрена, чтобы облегчить многократное повторение обработки нейронной сети, что особенно важно при обучении.

8.2 Предварительная обработка звука

Согласно рис. 3, звуковые сигналы, полученные в реальном времени или введенные из файлов формата WAV, подвергаются в системе SAS предварительной обработке.

При обработке файла по нему перемещается окно ввода, размер которого равен N элементов — размеру окна дискретного преобразования Фурье (ДПФ). Смещение окна относительно предыдущего положения можно регулировать. В каждом положении окна оно заполняется 16-разрядными данными (система работает только с такими звуковыми данными, в которых каждый отсчет кодируется 16 битами).

После ввода данных в окно перед вычислением ДПФ на него накладывается окно сглаживания Хэмминга:

$$newData[i] = Data[i](0.54 - 0.46\cos\frac{2\pi_i}{N-1})$$

Здесь Data — исходный массив данных, newData — массив данных, полученный после наложения окна сглаживания, N — размер ДПФ.

Наложение окна Хэмминга немного понижает контрастность спектра, но позволяет убрать боковые лепестки резких частот, при этом особенно хорошо проявляется гармонический состав речи. Сказанное иллюстрирует рис. 4.

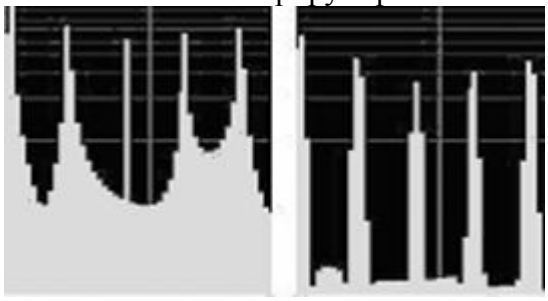

Рис. 4 — Действие окна сглаживания Хэмминга (логарифмический масштаб)

8.3 Выполнение дискретного преобразования Фурье

Результат сглаживания Хэмминга подвергается в системе SAS дискретному преобразованию Фурье по алгоритму быстрого преобразования Фурье. В результате этого преобразования получается амплитудный спектр и информация о фазе сигнала (в реальных и мнимых коэффициентах).

Информация о фазе сигнала отбрасывается и вычисляется энергетический спектр:

$$E[i]=\sqrt{ReC[i]ImC[i]}; i=0...\mathrm{NS}-1, NS=N/2$$

Здесь E[i] – энергии частот.

Так как звуковые данные не содержат мнимой части, то по свойству ДПФ результат получается симметричным, т.е. E[i] = E[N-i]. Таким образом, размер информативной части спектра NS равен N/2.

8.4 Нормирование частотного спектра

Все вычисления в нейронных сетях производятся над числами с плавающей точкой. Поэтому значения параметров объектов, классифицируемых с помощью нейронных сетей, ограничены диапазоном [0.0;1.0].

Для выполнения обработки спектра нейронной сетью в системе SAS полученный спектр нормируется на 1.0. Для этого каждый компонент вектора делится на его длину:

$$newE[i]=\frac{E[i]}{|E|} \qquad |E|=\sum_{i=0}^{NS-1} E[i]$$

8.5 Логарифмическое сжатие спектра

Исследования показали, что информативность различных частей спектра неодинакова: в низкочастотной области спектра содержится больше информации, чем в высокочастотной области спектра.

Поэтому для более экономного использования входов нейронной сети и увеличения необходимо уменьшить число элементов, получающих информацию из высокочастотной области спектра. Это и означает сжатие высокочастотной области спектра в пространстве частот.

В системе SAS применен наиболее распространенный и простой метод — логарифмическое сжатие, или mel-сжатие. Этот метод описан в разделе «Non-linear frequency scales» документа.

Вот формула, по которой выполняется логарифмическое сжатие спектра:

$$m=1125\log(0.0016\mathrm{f}+1)$$

Здесь f —частота в спектре, Гц, m — частота в новом сжатом частотном пространстве.

Рис. 5 иллюстрирует процесс логарифмического сжатия частотного спектра.

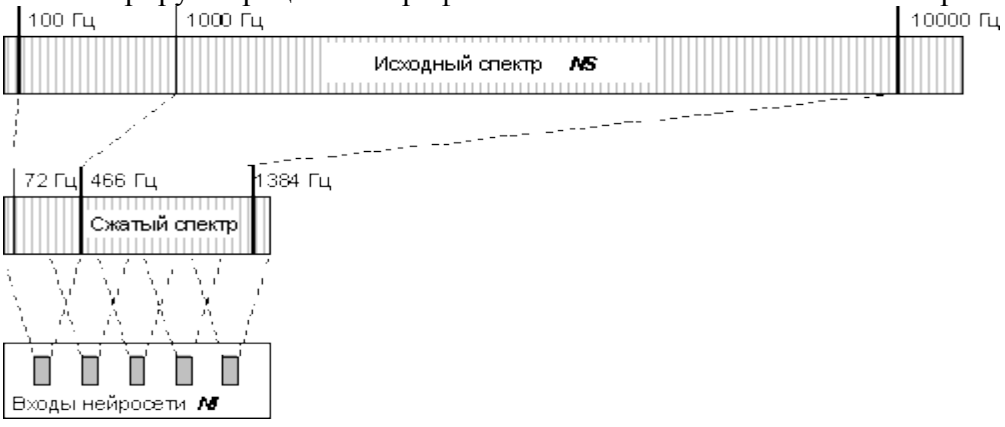

Рис. 5 — Нелинейное преобразование спектра в пространстве частот

9. Применение вейвлет-преобразований

В только что описанной системе SAS для выделения из речи синтаксических элементов применялось быстрое преобразование Фурье.

Однако, как отмечают исследователи, анализ Фурье обладает целым рядом недостатков, в результате которых происходит потеря информации о временных характеристиках обрабатываемых сигналов. Этот анализ подразумевает использование искусственных приемов, с помощью которых осуществляется частотно-временная локализация, например, окон данных (на рис. 3 это окно обозначено как **Окно ввода**).

В современных технологиях обработки и распознавания сигналов применяются так называемые вейвлет-преобразования и вейвлет-анализ.

Термин *вейвлет* (wavelets) можно перевести как «маленькая волна». Вейвлеты представляют собой новый инструмент решения различных задач прикладной математики. Вейвлет-анализ, детальное знакомство с которым требует определенных познаний в математике, лишен недостатков анализа Фурье. Он позволяет достичь неплохих результатов при использовании в системах распознавания речи.

В чем отличие анализа Фурье от вейвлет-анализа?

Фурье-анализ предполагает разложение исходной периодической функции в ряд, в результате чего исходная функция может быть представлена в виде суперпозиции синусоидальных волн различной частоты. Такая суперпозиция и есть спектр сигнала, о котором мы говорили в 3 главе нашей книги.

Что же касается вейвлет-анализа, то здесь входной сигнал раскладывается в базис функций, характеризующих как частоту, так и время. Поэтому с помощью вейвлетов можно анализировать свойства сигнала одновременно и в физическом пространстве (время, координата), и в частотном пространстве. Чтобы подчеркнуть такое обстоятельство, в зарубежной литературе Фурье-анализ называют single spectrum, а спектры, полученные на основе вейвлет-преобразований — time-scale spectrum, или wavelet spectrum.

Функции-базисы для вейвлетных преобразований конструируются на основе производных функций Гаусса. Подробнее об этом Вы сможете прочитать в.

На рис. 6 показаны наиболее часто используемые вейвлеты.

Эти функции имеют свои названия (Табл. 1).

Таблица 1 — Часто используемые вейвлеты.

Обозначение на рисунке 6	**Название**
а	WAWE-вейвлет
б	MHAT-вейвлет. Получил свое название от «мексиканская шляпа, сомбреро» (Mexican Hat)
в	Morlet
г	Paul
д	LMB
е	Daubeshies

При использовании вейвлет-преобразований для распознавания речи разработчик должен выбрать нужную функцию. От правильного выбора зависит успешность распознавания.

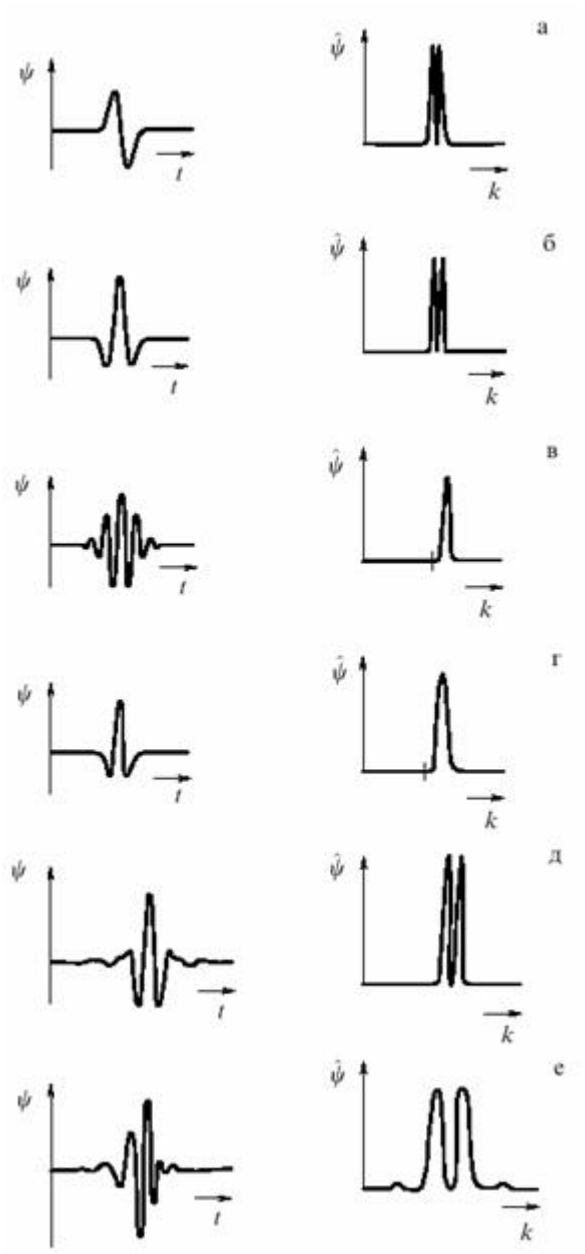

Рис. 6 — Часто используемые вейвлеты

10. Модель распознавания речи на основе искусственных нейронных сетей

Пусть речевой сигнал как входные данные нейронной сети. После обработки звуковых данных получен массив сегментов сигналов. Каждый сегмент соответствует набору чисел, характеризующих амплитудные спектры сигнала. Для подготовки к вычислению для сигнала выхода нейронной сети необходимо записать все наборы чисел в таблицу, строка которой – это набор чисел каждого кадра.

Таблица 2 – Описание набора признаков речевого сигнала

Кадр	**1-ое значение**	**2-ое значение**	**...**	**I-ое значение**
1-ый кадр	x_{11}	x_{21}	...	x_{1I}
2-ой кадр	x_{21}	x_{22}	...	x_{2I}
...	...	...	...	...
N-ый кадр	x_{N1}	x_{N2}	...	x_{NI}

I – Количество значений одного набора чисел

N – Количество наборов чисел (кадр сигнала после нарезки)

Количество входных и выходных нейронов известно. Каждый из входных нейронов соответствует одному набору чисел. А на выходном слое только один нейрон, выход которого соответствует желаемому значению распознавания сигнала.

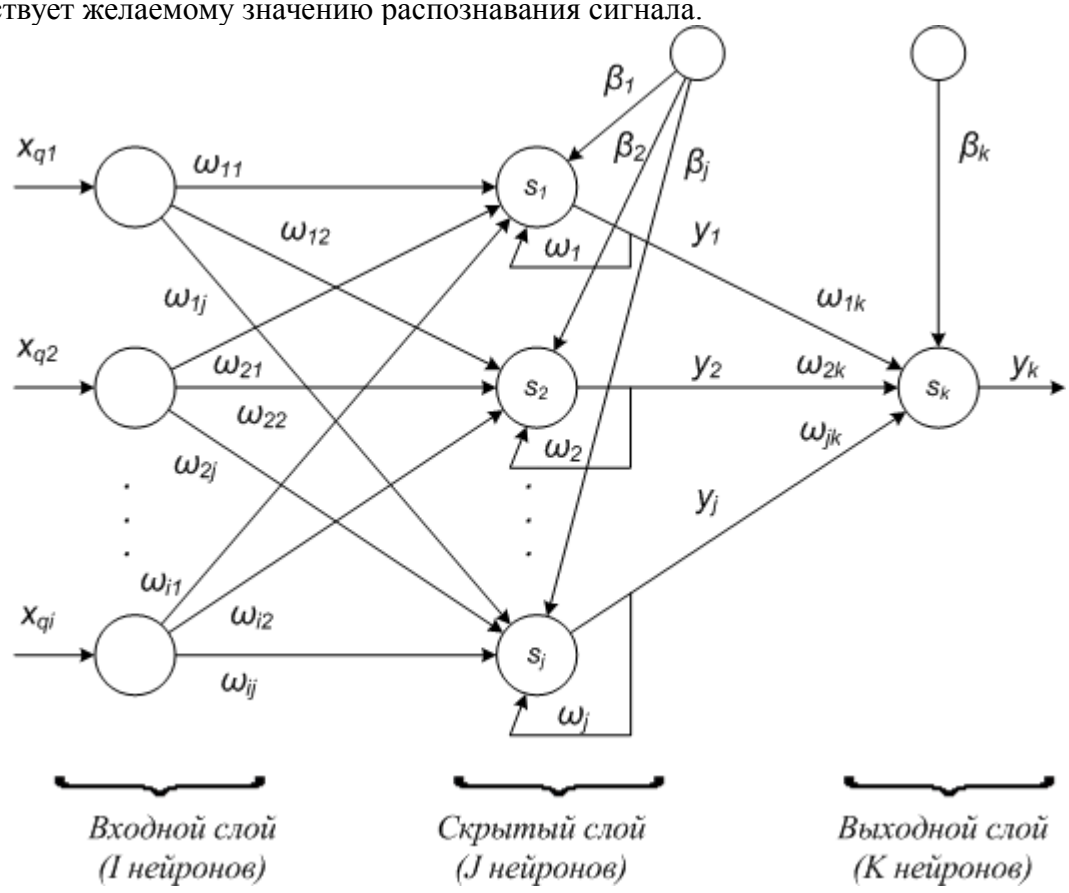

Рис. 7 – Структура нейронной сети с одной обратной связью

Где x_{qi} — i-ое входное значение q-го набора числе;

y_j — выход j-го нейрона слоя;

ω_{ij} — весовой коэффициент связи, соединяющий i-ый нейрон с j-ым нейроном;

ω_j — весовой коэффициент обратной связи j-го нейрона;

β_j — смещение j-го нейрона слоя.

Для вычисления выхода нейронной сети необходимо выполнить следующие последовательные шаги:

Шаг 1: Инициировать все контексты всех нейронов скрытого слоя.

Шаг 2: Подать первый набор чисел на вход нейронной сети. Вычислить для него выходы скрытого слоя.

$$y_j = f\left(\sum_{i=1}^{I} \omega_{ij} x_{1i} + \beta_j + \omega_j x_j\right)$$

где f(x) — нелинейная активиционнная функция $y_i = \frac{1}{1+b^{-aS_j}}$

Шаг 3: Если текущий набор чисел не является последним, то переход на шаг 5, иначе переход на шаг 4.

Шаг 4: Записать выходы нейронов скрытого слоя на контексты $x_j = y_j$. Переход к шагу 2 для следующего набора чисел.

Шаг 5: Вычислить выход нейрона выходного слоя.

$$y_k = f\left(\sum_{j=1}^{J} \omega_{jk} y_j + \beta_k\right)$$

Рассмотрим задачу, которая состоит в распознавании чисел от 0 до 9. Для распознавания одного числа нужно построить собственную нейронную сеть. И так должно построить 10 нейронных сетей. Надиктована база из 250 слов (числа от 0 до 9) с различными вариациями произношения. База случайным образом разделялась на две равные части – обучающую и тестирующую выборки. При обучении нейронной сети распознаванию одного числа, например 5, желаемый выход этой нейронной сети должен быть единицей для обучающей выборки с числом 5, а остальные – нулю.

Обучение нейронной сети осуществляется путем последовательного предъявления обучающей выборки, с одновременной подстройкой весов в соответствии с определенной процедурой, пока ошибка настройки по всему множеству не достигнет приемлемого низкого уровня. Функции ошибки в системе будет вычисляться по следующей формуле:

$$E = \frac{1}{2\mathrm{N}} \sum_{i=1}^{N} (y_{ki} - d_i)^2$$

где N — количество обучающих выборок, обработанных нейронной сетью примеров;

- реальный выход нейронной сети;

- желаемый (идеальный) выход нейронной сети.

Для каждого слова из тестовой выборки реальные выходы вычисляются 10 нейронными сетями распознавания разных чисел. Нейронная сеть, которая имеет максимальное выходное значение, и является нейронной сетью распознавания данного слова. И слово, распознанное нейронной сетью, является результатом распознания.

ЗАКЛЮЧЕНИЕ

В этой работе мы попытались собрать неполную и разрозненную информацию относительно существующих методов распознавания речи.

Прежде всего, мы выделили два подхода к распознаванию речи. Первый подход реализует распознавание элементов речи по образцу и применяется в различного рода системах голосового управления. Второй подход основан на выделении в речи лексических элементов — фонем, аллофонов, морфем и т.д. Этот подход пригоден для создания систем диктовки текста, рассмотренных нами в следующей главе.

Мы также выделили системы распознавания речи, требующие обучения и зависящие от диктора, а также системы, способные работать без предварительного обучения и, следовательно, не зависимые от диктора.

Перед тем как приступить к выделению из речи лексических элементов, необходимо выполнить предварительную обработку речевого сигнала. В ходе этой обработки из сигнала удаляются шумы, выполняется частотная фильтрация и оцифровка, а также нормализация уровня сигнала.

В этой главе мы рассмотрели две методики выделения из речи лексических элементов.

Первая методика предполагает использование дискретного преобразования Фурье. Непосредственно лексические элементы выделяются из оцифрованной речи при помощи нейронной сети, способной к обучению. При этом речь представляется в виде некоторого набора числовых параметров, так как нейронные сети работают именно с наборами таких параметров.

Мы привели несколько выражений, использованных для предварительной обработки сигнала, повышающей контрастность спектра, а также для выполнения дискретного преобразования Фурье и нормирования частотного спектра.

Вторая методика выделения лексических элементов речи основана на применении вейвлет-преобразований. В отличие от дискретного преобразования Фурье, этот метод исключает потерю информации о временных характеристиках обрабатываемых сигналов. Мы отметили, что при использовании вейвлет-преобразований входной сигнал раскладывается не в базисе периодических функций (как в дискретном преобразовании Фурье), а в базисе функций, характеризующих как частоту, так и время.

Техника распознавания речи находится в постоянном развитии. Чтобы всегда быть в курсе событий, следите за публикациями в Интернете и периодической печати. Не исключено, что скоро будут разработаны новые, более совершенные методы распознавания речи.

Так же мы создали не большую модель распознования чисел от 0 до 9, что служит хорошим примером испольщования искусственной нейронной сети.

СПИСОК ЛИТЕРАТУРЫ

1. «Аналитические технологии для прогнозирования и анализа данных» — НейроПроект
2. «Нейронные сети: на пороге будущего» - Даниил Кальченко
3. «Разработка системы распознавания русской речи» - А.А Кибкало, М.М. Лотков, И.Г. Рогожкин, А.А. Туровец
4. «Синтез и распознавание речи. Современные решения» - А.В. Фролов, Г.В. Фролов
5. http://www.aiportal.ru/articles/neural-networks/actuality.html
6. http://en.wikipedia.org/wiki/SAS_(software)
7. http://ru.wikipedia.org/wiki/Нейронная_сеть
8. http://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть
9. http://habrahabr.ru/post/214109/
10. http://habrahabr.ru/post/134998/
11. http://www.moluch.ru/conf/tech/archive/3/712/