Министерство образования и науки Украины
Национальный Технический Университет Украины
«Киевский Политехнический Институт»

На правах рукописи

Ладошко Ольга Николаевна

УДК 534.78, 004.934

ПОВЫШЕНИЕ РОБАСТНОСТИ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ МЕТОДАМИ ОБРАБОТКИ СИГНАЛОВ

05.09.08 – прикладная акустика и звукотехника

Диссертация на соискание научной стрепети
кандидата технических наук

Научный руководитель:
Продеус Аркадий Николаевич
доктор технических наук,
профессор кафедры акустики и
акустоэлектроники НТУУ «КПИ»

Київ – 2016

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Стремительное развитие современных достижений в области цифровой обработки сигналов способствует широкому распространению аппаратно-программных систем автоматического стенографирования, портативных устройств, управляемых голосом, таких как мобильные телефоны и планшеты, системы автоматического распознавания речи в телефонных и сетевых линиях связи. Неустойчивая и ненадежная работа систем автоматического распознавания речи (АРР) в этих системах и устройствах в условиях шума и реверберации приводит к вынужденному ограничению их использования.

Несмотря на многолетнюю историю и значительный вклад в достижения в области распознавания речи совершенными такими украинскими учеными как Т. К. Винцюк [1], Н. Н. Сажок [2], Т. В. Людовик [3], В. В. Пилипенко [4], О. Н. Карпов [5], Т. В. Ермоленко [6], а также рядом зарубежных ученых Л. Рабинер [7], С. Янг [8], Р. Стерн, , Г. Хемански [9], Д. Элис [10] и многими другими, решение задачи автоматического распознавания речи в условиях помех остается актуальным по сей день.

Для повышения точности распознавания речи в условиях шума и реверберации необходимо проводить корректировку искаженных сигналов. Решение большинства задач коррекции речевых сигналов для повышения точности распознавания систем АРР базируется на методах предварительной обработки сигналов и методах параметрического представления речевых сигналов в системе АРР [11-13].

Центральное место среди методов предварительной коррекции речевых сигналов занимают спектральные методы ослабления шумовой помехи и поздней реверберации. Наиболее популярными в настоящее время являются такие методы ослабления шумовой и реверберационной помех как спектральное вычитание, винеровская фильтрация, а также предложены

Я. Ефраимом и Д. Малахом метод минимизации среднеквадратичной ошибки оценки кратковременного амплитудного спектра (MMSE) и метод минимизации среднеквадратичной ошибки оценки логарифма кратковременного амплитудного спектра (logMMSE) [14].

Между тем, следует отметить, что задача оптимизации такого ослабления в системах АРР недостаточно рассмотрена в литературе. В частности, не исследована в полной мере оценка степени снижения точности распознавания речи при ослаблении шума и поздней реверберации, а также недостаточно рассмотрен вопрос согласованности точности распознавания с критериями качества речевых сигналов.

Решение большинства задач параметрического представления речевых сигналов в системах АРР базируется на ограниченном использовании классических моделей слухового восприятия, как это изложено в трудах С. Снефа, А. Гитза, Р. Лайона [12], что обусловлено желанием уменьшить вычислительную сложность систем АРР. Кроме того, существующие методы параметрического представления речевого сигнала, такие как мел-кепстральные частотные коэффициенты (MFCC), разработанные П. Мермелстейн, С. Дэвис [15] и перцепционные коэффициенты линейного предсказания (PLP) [16], предложенные Г. Хемански на данный момент морально устарели, поскольку конструировались без учета необходимости обеспечения робастности систем АРР в сложных помеховых условиях. Такое положение дел привело к необходимости разработки более сложных методов параметрического представления речевых сигналов, одним из которых является метод нормализованных по мощности кепстральных коэффициентов (PNCC) [17, 18 Kim]. При исчислении таких коэффициентов пытаются удалить медленно изменяющуюся шумовую составляющую сигнала, а также учесть временное маскирование речевого сигнала шумом. К сожалению, следует отметить неэффективность применения этого метода в условиях нестационарного шума.

Таким образом, разработка новых, более действенных в системах автоматического распознавания речи, методов обработки речевых сигналов является **актуальной научно-технической задачей** коррекции речевых сигналов, которая имеет важное прикладное значение.

Связь работы с научными программами, планами, темами

Диссертационную работу было выполнена на кафедре акустики и акустоэлектроники Национального технического университета Украины «Киевский политехнический институт». Исследования, результаты которых изложены в диссертации, полученные при выполнении госбюджетной темы "Разработка нового поколения медицинских приборов - широкополосных ультразвуковых физиотерапевтических излучателей с возможностью одновременного ионофореза" (№ ДР 0114U002485) в виде математического обеспечения коррекции сигналов. Полученные в диссертационной работе результаты внедрены в практику создания цифровой гидролокационной системы обнаружения и классификации акустических сигналов в рамках опытно-конструкторской работы «Зарница-58250» № 0107U000073Т в виде математического обеспечения систем обнаружения и классификации сигналов.

Внедрение результатов диссертационной работы подтверждено соответствующими актами.

Цель и задачи исследования. Целью работы является разработка новых методов обеспечения помехоустойчивости систем АРР.

Для достижения поставленной цели решены следующие задачи:

1. Повышение точности АРР при ослаблении поздней реверберации методом предварительной коррекции сигнала и при слепой оценке времени реверберации.

2. Определение целесообразности использования параметризации речевых сигналов в пространстве признаков PNCC.

3. Обоснование структуры детектора голосовой активности, обеспечивающего робастность системы АРР при использовании PNCC признаков.

4. Выявление возможности повышения робастности системы АРР за счет введения нового признака «траектория частоты основного тона» в детекторе голосовой активности.

Объектом исследования являются процессы обработки акустических речевых сигналов, направленные на повышение робастности систем АРР.

Предметом исследования являются методы обработки искаженных сигналов в системе АРР.

Методы исследования. В работе использовались методы частотной коррекции с использованием алгоритма logMMSE для вычисления коэффициента передачи корректирующего фильтра для ослабления поздней реверберации. Для слепого оценивания времени реверберации использовался метод максимального правдоподобия. Методы спектрального анализа речевых сигналов, положения физиологической акустики на основе психоакустического сглаживания и сжатия спектральной характеристики речевого сигнала, методы теории фильтрации использовались для определения целесообразности использования признаков PNCC. Моделирование систем АРР выполнялось с использованием аппарата скрытых марковских моделей для построения акустических моделей речи и методов статистического моделирования речи. Обоснование структуры детектора голосовой активности выполнялось методами построения нейронных сетей и цифровой обработки сигналов. Оценка нового признака «траектория частоты основного тона» в детекторе голосовой активности выполнялась методами автокорреляционной обработки сигналов и методом динамического программирования для получения наиболее вероятного траектории частоты основного тона. Для экспериментальной проверки разработанных алгоритмов использованы реальные и искусственные речевые сигналы.

Научная новизна полученных результатов:

1. Впервые решена задача построения нейросетевого детектора голосовой активности системы APP с использованием признаков нормализованные по мощности кепстральные коэффициенты PNCC для эксплуатации в условиях нестационарных помех.

2. Впервые предложен метод повышения помехоустойчивости детектора голосовой активности за счет оценивания признака «траектория основного тона», включенного в перечень классификационных признаков нейросетевого детектора голосовой активности системы автоматического распознавания речи.

3. Усовершенствован метод обучения нейросетевого детектора голосовой активности на основе адаптивной коррекции параметров, что позволяет ускорить процедуру обучения.

4. Усовершенствован метод ослабления поздней реверберации, что позволяет повысить точность систем автоматического распознавания речи даже в условиях недостаточности априорной информации о параметрах реверберации.

Практическая значимость полученных результатов.

1. Предложенный способ определения новых, шумовых и паузных участков речевого сигнала, на котором базируется разработан нейросетевой детектор речевой активности, позволяет повысить точность систем автоматического распознавания речи.

2. Предложено использование дополнительной классификационного признака «траектория частоты основного тона», а также адаптивная коррекция весовых коэффициентов в разработанном нейросетевом детекторе голосовой активности позволяют уменьшить вычислительные затраты и достичь желаемой точности определения тоновых, шумовых и паузных участков речевого сигнала в условиях действия шумовых помех.

3. Усовершенствование метода ослабления поздней реверберации позволяет существенно повысить точность систем автоматического

распознавания речи даже в условиях недостаточности априорных данных о времени реверберации и его зависимости от частоты.

Получены и описаны в данной работе результаты компьютерного моделирования, экспериментальных исследований и расчетов нашли отражение в патенте Украины на полезную модель, которая описывает трансформацию образцов или операции, направленные на повышение устойчивости системы АРР к действию шумов в канале или к изменению условий эксплуатации.

Полученные результаты также могут использоваться в учебном процессе высших учебных заведений Украины, в т. ч. при подготовке инженеров-акустиков в курсах лекций кафедры акустики и акустоэлектроники НТУУ "КПИ", а именно: в курсе лекций по дисциплине «Устройства регистрации и отображения информации» в разделе «Кодирование акустических сигналов», а также в курсе лекций по дисциплине «Компьютерные акустические системы», в разделе «Коррекция речевых сигналов».

Личный вклад соискателя отражен публикациями [19-23]. В научных работах, выполненных в соавторстве соискателю принадлежит: в работе [24] моделирование системы автоматического распознавания речи, проведения экспериментальных исследований и анализ полученных результатов, в работах [25-26] - разработка компьютерных средств анализа и проведения статистического анализа полученных результатов, создание разметки спонтанного украинской речи, в работе [27] - разработка алгоритма для компьютерного моделирования траектории основного тона, проведения экспериментальных исследований и анализ полученных результатов; в работе [28] соискатель сформулировал задачу, исследовал влияние адаптивной коррекции весовых коэффициентов нейросети, провел необходимые экспериментальные исследования и проанализировал полученные результаты; в работе [29] личный вклад соискателя заключается в исследовании и моделировании методов ослабления реверберационной

помехи, проведении экспериментальной проверки правильности теоретических расчетов, в работе [30] - в анализе полученных результатов по критерию точности распознавания речи.

В работе [31] соискатель разработал способ определения тоновых, шумовых и паузных участков речевого сигнала, провел моделирование детектора выделения голосовой активности на основе предложенного способа, провел экспериментальную проверку предложенной полезной модели. В работе [32] соискатель провел экспериментальные исследования системы автоматизированного стенографирования и выполнил статистический анализ полученных результатов, в работе [33] - соискатель разработал способ выделения тональных, шумовых и паузных участков устной речи, в [34] - провел моделирование методов ослабления реверберационной помехи и проверил экспериментально правильность теоретических расчетов.

Апробация результатов диссертации. Основные положения и результаты диссертации обсуждались на 8-ми научно-технических конференциях и 2-х школах-семинарах МННЦИТиС НАН и МОН Украины: Международная научно-техническая конференция «Искусственный интеллект. Интеллектуальные системы ИИ», АР Крым, Кацивели, 2010, 2011; Международная научно-техническая конференция Акустический симпозиум «Консонанс», г. Киев 2011 2013; Международная научно-техническая конференция «Моделирование и компьютерная графика», г. Донецк, 2011; III Международная научно-техническая конференция студентов, аспирантов и молодых ученых. "Информационные управляющие системы и компьютерный мониторинг", м. Донецк, 2011; Одиннадцатая всеукраинская международная конференция «Обработка сигналов и изображений и распознавания образов», г. Киев, 2012; Тридцать четвертая международная научная конференция IEEE «Электроника и нанотехнологии», Киев, 2014. Школа-семинар Украинской ассоциации по обработке информации и распознавания образов совместно с МННЦИТиС НАН и МОН Украины

«Устноязыковые технологии и проблемы создания корпусов», г. Киев, 2010; «Распознавание и синтез спонтанной речи», г. Киев, 2011.

Публикации. По результатам исследований опубликовано 16 научных работ, в том числе 7 статей в научных изданиях Украины, из них 1 статья в изданиях Украины, включенных в международные наукометрические базы данных, 1 патент на полезную модель, 8 тезисов докладов в сборниках материалов конференций.

Структура и объем диссертационной работы. Диссертация состоит из введения, четырех глав, заключения, списка использованных литературных источников, приложений. Общий объем составляет 185 страниц, в том числе 150 страниц основного текста. Работа содержит 89 рисунков и 38 таблиц, 8 приложений и список использованных источников из 122 наименований.

ГЛАВА 1.
ОБЗОР МЕТОДОВ ПОВЫШЕНИЯ РОБАСТНОСТИ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Принципиально постановка задачи повышения робастности систем автоматического распознавания речи (АРР) изображена на рис 1.1. Речевой сигнал $x(t)$ подвергается воздействию помех $n(t)$ (фоновый шум и реверберация) и мешающих факторов (различие частотных характеристик микрофонов и фильтров, ошибки кодирования).

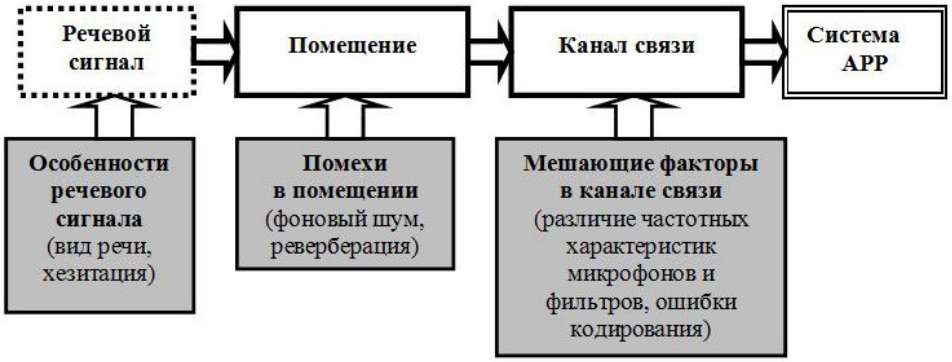


Рис. 1.1. Влияние помех на робастность системы АРР

Аналитическая модель создания речевого сигнала шумом и реверберацией определяется выражением:

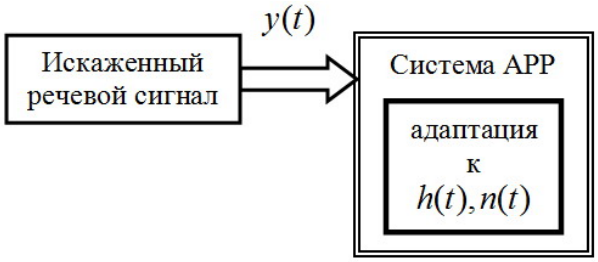
$$y(t) = x(t) \otimes h(t) + n(t), \qquad (1.1)$$

где $x(t)$ - неискаженный речевой сигнал; $h(t)$ - импульсная характеристика (ИХ) помещения ; $n(t)$ - случайный шумовой процесс (шум окружения).

Существует различное количество подходов к повышению робастности систем АРР (рис. 1.2), однако проблема повышения помехоустойчивости (робастности) систем АРР должна решаться по двум основным направлениям: подходы реализующие коррекцию речевых сигналов до

поступления на вход. предварительная коррекция речевого сигнала (рис. 1.3,а) и внутри системы АРР (рис. 1.3,б) (см. обзоры [12 Virtatnen, 13 Benesty, 11 Jinyu]).

а - предобработка



б - постобработка

Рис. 1.2. Два подхода к подавлению влияния шума и реверберации

Как правило к методам предварительной коррекции относят подавление помех в препроцессоре коррекции (рис. 1.3), ориентированное на амплитудные критерии качества откорректированного сигнала, такие как соотношение сигнал шум [12 Virtatnen].

Ко вторым относятся подходы, работающие внутри системы АРР: робастное к шумовой помехе параметрическое представление речевого сигнала, адаптация акустических моделей к шумовым условиям, учет языковых (лингвистических) особенностей речевого сигнала (обработка спонтанной речи, построение грамматик и моделей языка для систем АРР) [11 Jinyu].

В частности, среди помех можно выделить два основных вида: аддитивный фоновый шум от нескольких источников (суммарный сигнал от различных источников звука, поступивший на вход записывающего устройства, шумы записывающего устройства) и реверберация (суммарный сигнал, поступивший на вход записывающего устройства после

многократного переотражения от стен помещения, в котором происходила запись).

Для достижения подавления аддитивной помехи традиционно используют несколько методов, а именно: винеровская фильтрация или восстановление во временной области [35-38], метод подпространства [39-40], восстановление в спектральной области [41-42], методы, основанные на моделях [43-46].

Несмотря на развитость математического аппарата для подавления аддитивной помехи в речевом сигнале методами предварительной коррекции, задача подавления поздней реверберации недостаточно рассмотрена в литературе [13 Benesty, 11 Jinyu]. Кроме того в современных системах АРР до сих пор используется представление речевых сигналов в виде мел-частотных кепстральных коэффициентов MFCC [15] и перцепционных коэффициентов линейного предсказания PLP [16], которые не содержат в себе механизма обеспечения помехоустойчивости систем АРР.

В связи с перечисленными выше недостатками методов повышения робастности в системах АРР предлагается задачу повышения робастности систем АРР, разбить на два этапа: предварительная коррекция искаженного сигнала до поступления в систему АРР и предварительная обработка сигнала внутри системы АРР (параметрическое представление) (рис. 1.3).

1.1. Модель автоматического распознавания речи в условиях помех

В общем случае цель системы АРР состоит в получении оптимальной, в смысле критерия максимального правдоподобия, последовательности слов W, распознанных в искаженном речевом сигнале Y:

$$\hat{W} = \arg\max_{W} P_{\Gamma}(W) \cdot \sum_{\theta} \prod_{t=1}^{T} p_{\Lambda}(\overline{x_t}(Y) \mid \theta_t) \cdot P_{\Lambda}(\theta_t \mid \theta_{t-1}), \qquad (1.2)$$

где $P_{\Gamma}(W)$ - вероятностная модель соответствующего языка; $\overline{x_t}(Y)$ - вектор признаков, полученный из откорректированного сигнала для методов предварительной коррекции или робастный вектор признаков, полученный из искаженного сигнала Y методов параметрического представления сигнала в системах АРР; $\{x_t\}$ - совокупность векторов, которая описывается скрытыми Марковский моделями с последовательностью всех возможных состояний θ_t для транскрипции W; t - момент наблюдения. Λ и Γ - соответственно параметры акустической и модели языка.

Структура работы системы АРР в условиях помех отображена на рис. 1.4.

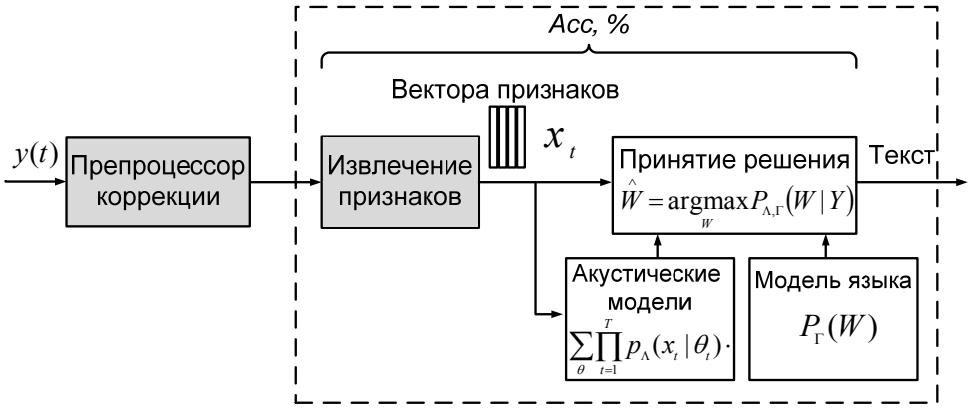


Рис. 1.4. Структура системы АРР в условиях помех

Критерием качество полученной последовательности слов является точность распознавания на выходе системы АРР, определяемая из выражения:

$$Acc\% = \frac{N - D - S - I}{N}, \tag{1.3}$$

Применение методов предварительной коррекции или робастного параметрического представления сигнала в системе АРР позволяет исключить необходимость адаптировать параметры акустической модели

$p_{\Lambda}(\overline{x_t}(Y)|\theta_t)$ (рис. 1.4) к искаженному сигналу Y системы АРР, не усложнять вычисления и не менять структуру и параметры уже существующих систем АРР (модуль принятия решения), характерно всем методам, основанным на адаптации моделей [11, 13]. Поэтому в данной диссертации такому направлению, как коррекция речевых сигналов путем их предварительной обработки, решено уделить значительное внимание.

1.2. Повышение устойчивости систем АРР к влиянию реверберации

Задачу повышения устойчивости (робастности) систем АРР к влиянию шума и реверберации можно сформулировать следующим образом (рис. 1.5). Обычно системы АРР обучают на образцах неискаженной речи $x(t)$ [47, 48]. Между тем, при эксплуатации таких систем в реальной обстановке микрофон часто расположен на расстоянии от источника речи, поэтому на вход АРР поступает речевой сигнал, искаженный шумом и реверберацией (1.1).

Из-за различия условий обучения и эксплуатации качество работы систем АРР существенно снижается. Это явление именуют неустойчивостью систем АРР к действию шума и реверберации на речевой сигнал [49, 50].

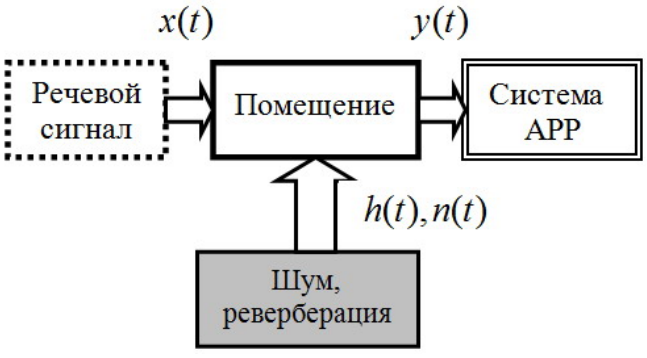


Рис. 1.5. Влияние помех на речевой сигнал

Выше были указаны два принципиально разных подхода к повышению робастности систем АРР: предобработка сигнала $y(t)$ в так называемом

препроцессоре (рис. 1.3,а) с целью восстановить сигнал $x(t)$ и постобработка сигнала $y(t)$ путем адаптации параметров системы АРР к свойствам шума $n(t)$ и ИХ $h(t)$ (рис. 1.3,б).

Каждый из указанных подходов обладает достоинствами и недостатками. Чрезвычайно важным и решающим достоинством предобработки является возможность не изменять структуру и параметры уже существующих систем АРР (что является весьма непростой задачей ввиду принципиальной сложности алгоритмов АРР). Ввиду этого в данной работе избран подход, базирующийся на предварительной коррекции речевого сигнала. Учитывая значительный опыт, накопленный к настоящему времени в области коррекции речевых сигналов [42, 47-48], можно надеяться на получение заметных результатов в плане повышения качества работы системы АРР.

В отличие от задачи подавления шума, задача подавления реверберации, и в первую очередь - поздней реверберации, в системах АРР все еще далека от завершения. Ее актуальность особенно возросла в последние годы, в связи с быстрым ростом средств мобильной связи (мобильная телефония, IP-телефония).

1.2.1. Спектральный подход к подавлению поздней реверберации.

Реверберация – это процесс многолучевого распространения акустического звука от его источника к микрофону. Если излучаемый сигнал представляет собой короткий (несколько десятков микросекунд) импульс, тогда полученный микрофоном сигнал можно трактовать как «отклик» помещения на воздействие в виде δ-функции. Иными словами, такой сигнал является импульсной характеристикой (ИХ) помещения (рис. 1.6), состоящей из прямого звука, отражений, которые прибывают вскоре после прямого звука («ранняя реверберация») и отражений, которые прибывают после ранней реверберации («поздняя реверберация»).

Протяженность области ранних отражений невелика – около 50 мс – и структура ИХ в этой области имеет дискретный характер. Напротив, область

поздних протяжений может быть весьма обширной – до нескольких секунд – и имеет стохастический характер.

Рис.1.6. Импульсная характеристика помещения

Как следствие, поздняя реверберация оказывает значительно более разрушающее действие на речевой сигнал. Поэтому подавляющее большинство современных работ посвящено поиску методов эффективного подавления именно поздней реверберации [48, 49-51].

Аналитически модель речевого сигнала $y(t)$, искаженного реверберацией, можно представить в виде свертки «чистого» речевого сигнала $x(t)$ с импульсной характеристикой $h(t)$ помещения:

$$y(t) = \int_0^\infty h(v)x(t-v)dv = x(t) \otimes h(t). \qquad (1.4,a)$$

Используя модель Полэка [50] для ИХ помещения

$$h(t) = \xi(t)e^{-\delta t},$$

где $\xi(t)$ - стационарный белый шум, $\delta = 3\ln 10 / T_{60}$ - показатель скорости затухания уровня звука в помещении (T_{60} - время реверберации), а также выделяя в ИХ $h(t)$ (рис. 1.6) области, соответствующие ранним и поздним отражениям

$$h_i(t) = \begin{cases} h(t), & 0 \le t \le T_l; \\ 0, & \partial p.\, t \end{cases} \quad h_l(t) = \begin{cases} h(t+T_l), & t \ge 0; \\ 0, & \partial p.\, t \end{cases},$$

искаженный реверберацией сигнал можно представить в виде:

$$y(t) = h_i(t) \otimes x(t) + r(t), \qquad (1.4,б)$$

где $r(t) = h_l(t) \otimes x(t - T_l)$ - компонент, обусловленный действием поздней реверберации; T_l - момент начала поздней реверберации.

Полагая, что слагаемые соотношения (1.4,б) статистически независимы, приходим к трактовке поздней реверберации как разновидности аддитивного шума. К сожалению, нестационарный характер этого шума делает неэффективным непосредственное применение традиционных методов подавления шума, рассчитанных на стационарный или медленный нестационарный шум [42].

Классифицируя методы подавления реверберации, различают полное и частичное подавление реверберации [51]. Очевидно, подавление поздней реверберации является частичным, поскольку действие ранней реверберации не компенсируется. Кроме того, из соотношения (1.4,а) следует, что задача полного подавления реверберации относится к обратным задачам, относящимся к классу так называемых «некорректных задач» [52], и потому должна решаться путем деконволюции с учетом сложностей, обусловленных некорректностью решаемой задачи – такой способ решения именуют «методом обратной фильтрации» [49].

Действие поздней реверберации на речевой сигнал является значительно более разрушающим, по сравнению с действием ранней реверберации. Кроме того, действие ранней реверберации сравнительно легко компенсируется на этапе постобработки речевого сигнала, при

вычислении кепстральных признаков [49]. Поэтому в данной диссертации внимание уделено задаче подавления поздней реверберации.

В работе [50] предложено подавлять позднюю реверберацию методом спектрального вычитания, предварительно оценивая спектр мощности поздней реверберации. В работе [51] показана возможность подавления поздней реверберации с использованием метода частотной коррекции, осуществляемой в соответствии с выражением:

$$\hat{\lambda}_x^{1/2}(l,k) = G(l,k)\lambda_y^{1/2}(l,k),$$

где $\lambda_y(l,k)$ - спектр мощности l-го сегмента сигнала $y(t)$ на частоте $f_k = kF_s / N_{fft}$; F_s - частота дискретизации; N_{fft} - параметр быстрого преобразования Фурье; k - номер частотной выборки; $\hat{\lambda}_x(l,k)$ - оценка спектра мощности l-го сегмента сигнала $x(t)$ для k-й частотной выборки; $G(l,k)$ - коэффициент передачи корректирующего фильтра для l-го сегмента сигнала $y(t)$ на частоте f_k.

Преимущество метода частотной коррекции, по сравнению с методом обратной фильтрации, состоит в устойчивости результатов к перемещениям диктора относительно микрофона. Еще одно важное преимущество метода частотной коррекции состоит в его тесной связи с методом спектрального вычитания и винеровской фильтрацией [47]. Действительно, представляя коэффициент передачи корректирующего фильтра в виде:

$$G(l,k) = \left(1 - \left(\frac{1}{\gamma(l,k)}\right)^{\beta_1}\right)^{\beta_2},$$

где $\gamma(l,k) = \lambda_y(l,k) / \lambda_n(l,k)$ - апостериорное отношение сигнал-шум, $\lambda_y(l,k)$ - спектр мощности l-го сегмента сигнала $y(t)$ на частоте $f_k = kF_s / N_{fft}$, $\lambda_n(l,k)$

- спектр мощности l-го сегмента помехи, можно показать, что ситуации вычитания амплитудных спектров соответствуют $\beta_1 = 1/2$ и $\beta_2 = 1$, ситуации вычитания спектров мощности соответствуют $\beta_1 = 1$ и $\beta_2 = 1/2$, а для метода винеровской фильтрации следует принимать $\beta_1 = 1$ и $\beta_2 = 1$.

Для расстояний между источником звука и микрофоном, больших критической дистанции (в данной работе ограничимся рассмотрением именно этой ситуации, при которой поздняя реверберация оказывает наиболее деструктивное действие на принимаемый микрофоном сигнал [48]), спектр мощности поздней реверберации $\lambda_r(l,k)$ может быть вычислен через спектр $\lambda_y(l,k)$ наблюдаемого сигнала $y(t)$ [50,51]:

$$\lambda_r(l,k) = e^{-2\delta(k)T_l} \cdot \lambda_y(l - N_l, k), \tag{1.5}$$

где $N_l = T_l F_s / R$; R - сдвиг фреймов, выраженный в выборках; $\delta(k) = 2\ln 10 / T_{60}(k)$; $T_{60}(k)$ - время реверберации.

Смысл соотношения (1.5) достаточно прост: текущие звуки речи маскируются «реверберационным хвостом», тянущимся за предыдущими звуками.

Для повышения точности измерений спектра $\lambda_y(l,k)$ К. Леберт [50] предложил производить скользящее усреднение:

$$\hat{\lambda}_y(l,k) = \eta_z \hat{\lambda}_y(l-1,k) + (1-\eta_z)\left|Y(l,k)\right|^2, \tag{1.6}$$

где $\eta_z \approx 0{,}9$ - параметр, регулирующий степень усреднения.

В работе Э. Хабетса [51] параметр усреднения η_z предложено считать частотно-зависимым:

$$\hat{\lambda}_y(l,k) = \eta_z(k) \hat{\lambda}_y(l-1,k) + (1-\eta_z(k))\left|Y(l,k)\right|^2, \tag{1.7}$$

где $Y(l,k)$ - дискретное преобразование Фурье l-го сегмента сигнала $y(t)$;

$$\eta_z(k) = \begin{cases} \eta_z^d(k), & |Y(l,k)|^2 \le \hat{\lambda}_y(l-1,k); \\ \eta_z^a(k), & \textit{в остальных случаях}. \end{cases} \tag{1.8}$$

При этом верхнее значение параметра $\eta_z^d(k)$ ($0 \le \eta_z^d(k) < 1$) предложено ограничивать величиной

$$\eta_z^d(k) = \frac{1}{1 + 2\delta(k)R/F_s}, \tag{1.9}$$

а значение параметра $\eta_z^a(k)$ предложено выбирать, исходя из условия

$$0 \le \eta_z^a(k) < \eta_z^d(k). \tag{1.10}$$

Вместе с тем, к соотношениям (1.5)-(1.10) можно предъявить ряд претензий:

- выбор значения параметра усреднения η_z в (1.6) не обоснован;
- не обоснована необходимость зависимости параметра усреднения η_z от частоты в соотношении (1.7);
- не обоснованы соотношения (1.9) и (1.10) по выбору значений параметров $\eta_z^d(k)$ и $\eta_z^a(k)$;
- выбор значения параметра T_l (граница между ранними отражениями и поздней реверберацией) не рассматривался с позиций максимизации таких критериев как качество речевого сигнала и качество распознавания речи.

Таким образом, выбор оптимальной структуры оценки спектра поздней реверберации, а также оптимизация параметров такой оценки составляет содержание одной из задач, решаемых в данной диссертации.

1.2.2. Оценка параметров реверберации при отсутствии знаний о импульсной характеристике помещения. Переходя к формулировке второй задачи, отметим, что поскольку в соотношении (1.5) фигурирует время реверберации $T_{60}(k)$, это означает, что время реверберации необходимо предварительно оценить по имеющейся импульсной характеристике помещения [53]. Между тем, в ряде случаев такая оценка невозможна (отсутствует образец записи ИХ помещения, отсутствует доступ в помещение и т.п.). В этом случае необходимо произвести слепое измерение времени реверберации $T_{60}(k)$ по речевому сигналу $y(t)$, с последующим восстановлением сигнала $x(t)$ [54, 55].

Существует многообразие методов слепого измерения времени реверберации. В работах [58] и [59] изложен нейросетевой подход. Подход, базирующийся на таком сегментировании речи, при котором выделяются паузы, предложен в [50]. В работах [60] и [61] представлен подход, при котором предполагается, что частотная характеристика фильтра, обратного ИХ помещения, является минимально-фазовой, что на практике обычно не выполняется.

В работе [54 Ratnam] впервые предложен метод, обладающий хорошей помехоустойчивостью и не уступающий (и даже несколько превосходит) конкурентным методам в точности измерений [62 Gaubitch], именуемый методом максимального правдоподобия (МП).

Сущность измерений времени реверберации T_{60} методом МП состоит в следующем. Информацию о параметре T_{60} извлекают преимущественно в паузах речевого сигнала, где действие реверберации проявляется в виде звуковых «шлейфов», тянущихся за последними звуками слов. Структура этих шлейфов подобна структуре ИХ канала передачи:

$$y(n) = \xi(n)a(n), \qquad a(n) = e^{-\delta n/F_s},$$

где $\xi(n),\ n \geq 0; n = tF_s$ - дискретный гауссовый белый шум с параметрами $[0, \sigma]$.

Если наблюдается N-мерный вектор $\mathbf{y}$, тогда в силу статистической независимости выборок процесса $\xi(n)$, многомерная плотность вероятностей этого вектора, именуемая также функцией правдоподобия, имеет вид произведения N одномерных гауссовских распределений:

$$L(\mathbf{y}; \mathbf{a}, \sigma) = \frac{1}{a(0)\cdots a(N-1)} \left(\frac{1}{2\pi\sigma^2} \right)^{N/2} \exp\left(- \frac{\sum_{n=0}^{N-1} (y(n)/a(n))^2}{2\sigma^2} \right). \qquad (1.11)$$

Используя соотношение (1.11), необходимо оценить параметры σ и $\mathbf{a}$, где $\mathbf{a}$ - N-мерный вектор, по имеющимся N значениям вектора $\mathbf{y}$. Учитывая экспоненциальный характер функции $a(n)$:

$$a(n) = a^n, \quad a = \exp(-\delta / F_s), \qquad (1.12)$$

из (1.11) с учетом (1.12) следует

$$L(\mathbf{y}; a, \sigma) = \left(\frac{1}{2\pi a^{(N-1)}\sigma^2} \right)^{N/2} \exp\left(- \frac{\sum_{n=0}^{N-1} a^{-2n} y(n)^2}{2\sigma^2} \right). \qquad (1.13)$$

Логарифмируя (1.13), получают

$$\ln L(\mathbf{y}; a, \sigma) = -\frac{N(N-1)}{2} \ln(a) - \frac{N}{2} \ln(2\pi\sigma^2) - \frac{\sum_{n=0}^{N-1} a^{-2n} y(n)^2}{2\sigma^2}. \qquad (1.14)$$

Приравнивая нулю частные производные от выражения (1.14) по параметрам σ и a, получают систему уравнений для неизвестных σ и a:

$$\frac{\sum_{n=0}^{N-1} n a^{-2n} y(n)^2}{a\sigma^2} - \frac{N(N-1)}{2a} = 0, \tag{1.15}$$

$$\sigma^2 = \frac{1}{N}\sum_{n=0}^{N-1} a^{-2n} y(n)^2 . \tag{1.16}$$

Подставляя (1.15) в (1.16), получаем соотношение, определяющее параметр a:

$$\frac{\sum_{n=0}^{N-1} n a^{-2n} y(n)^2}{\sum_{n=0}^{N-1} a^{-2n} y(n)^2} - \frac{(N-1)}{2} = 0 . \tag{1.17}$$

В силу нелинейности уравнения (1.17) алгоритм нахождения неизвестного параметра a должен быть итерационным. Найденный, таким образом, параметр a позволяет найти необходимое время реверберации T_{60}, используя соотношения:

$$T_{60} = 6{,}91/\delta, \qquad \delta = -F_s \ln a . \tag{1.18}$$

Не смотря на теоретическую обоснованность, слепые измерения времени реверберации параметра $T_{60}(k)$ неизбежно сопровождаются погрешностью измерений, снижающей эффективность подавления поздней реверберации. Другим недостатком слепых измерений является приближенный характер оценивания частотной зависимости параметра T_{60} [56Lollmann, 57Jeub]. Оба упомянутых фактора негативно сказываются на качестве функционирования системы АРР, снижая точность распознавания речи.

Выработка рекомендаций по минимизации снижения эффективности процедуры деверберации, обусловленной отсутствием информации о времени реверберации, составляет содержание второй задачи, решаемой в данной диссертации.

1.3. Методы робастного параметрического представления речевого сигнала в системе АРР

Как было отмечено выше методы предварительного ослабление помех сопровождаются искажением речевых сигналов, что в свою очередь отрицательно сказывается на точности АРР [13Benesty, 63Gunawan].

Большинство современных систем АРР используют информацию о речевом сигнале (блок извлечения признаков рис. 1.4) в виде спектрального представления на выходе гребенки фильтров:

$$MF[m,l] = \sum_{k=0}^{(K/2)-1} \left| X[m,e^{j\omega_k}] H_l[e^{j\omega_k}] \right|^2 , \qquad (1.19)$$

где $X[\cdot]$ - кратковременный спектр сигнала; $H_l[\cdot]$ - оконная функция сглаживания; $\omega_k = 2\pi k / F_s$, F_s - частота дискретизации; m и l - индекс фрейма и частотного канала анализа.

Следует подчеркнуть, что главным преимуществом такого представления речевого сигнала в системе АРР в отличие от методов предварительной коррекции сигнала заключается в его эффективности как для чистого так и для искаженного сигнала одновременно [11Jinyu, 12Virtanen]. В этом случае эффект шума не подавляется в сигнале, а уменьшается в процессе извлечения признаков (параметризации), который максимизирует критерий (1.2) получение наиболее вероятной последовательности слов.

Можно предположить, что включение компенсации остаточного шума, имеет место в восстановленном сигнале $\bar{x}(t)$, в процедуру (1.19), позволит избежать чрезмерного искажения сигнала и таким образом повысить точность АРР.

Наиболее распространенным спектральным представлением речевого сигнала в системах АРР является мел-частотные кепстральные коэффициенты [15 Davis] и перцепционные коэффициенты линейного предсказания [16Hermansky], основные этапы и суть вычисления которых изображены на рис. 1.7.

Рис. 1.7. Стадии вычисления MFCC и PLP коэффициентов

Эти коэффициенты получают путем выделения в спектре наиболее информативных, с точки зрения восприятия речевых сигналов. Для каждого фрейма на выходе гребенки фильтров (1.19) вычисляются n MFCC коэффициентов:

$$mfcc[n] = \frac{1}{L}\sum_{l=1}^{L}\log(MF[m,l])\cos\left[\frac{2\pi}{L}\left(l+\frac{1}{2}\right)n\right], \qquad (1.20)$$

где $l = 1,2,...,L$ - l-й номер фильтра,

или n PLP кепстральных перцепционных коэффициентов на основе коэффициентов линейного предсказания $b[i]$ согласно рекурсивному соотношению:

$$plp[n] = -b[n] + \tfrac{1}{n}\sum_{i=1}^{n-1}(n-i)\cdot b[i]\cdot plp[n-i]. \qquad (1.21)$$

Главным недостатком методов MFCC и PLP является ограничение для получения спектральных характеристик речевого сигнала на выходе гребенки фильтров (1.19) в виде усредненных значений.

Для учета влияния канала связи в процедуры получения MFCC и PLP представления речевого сигнала сочетают с фильтром RASTA [65Hermansky], что позволяет подавить медленно и быстро изменяемую частоты, которые нехарактерны для речевого сигнала на выходе гребенки фильтров (1.19). Тем не менее фильтр RASTA повышает устойчивость представлений речевого сигнала в системах АРР лишь в отношении линейных спектральных искажений, реализуя эффект полосовой фильтрации временных траекторий кепстральных коэффициентов.

Кроме того методы MFCC и PLP часто сочетают с нормализацией кепстральных среднего (CMN), что позволяет подавить аддитивный эффект влияния канала связи в логарифмически-мел-кепстральной области. Однако применение метода CMN, как и RASTA-фильтрации, в реальном режиме работы имеет существенный недостаток, а именно необходимость проведения расчетов на длительном отрезке сигнала. Помимо усреднения кепстральных значений, расхождение между данными тренировки и тестирования в системах АРР учитывают путём выравнивания закона распределения гистограммным методом (HEQ) [66Hilger]. Главная проблема этого подхода заключается в получении надежной оценки кумулятивной

функции распределения тестовых и тренировочных данных, которую трудно оценить для коротких предложений, т.е. недостаточной статистической выборки.

Помимо того существенный недостаток методов MFCC и PLP заключается в том что они используют лишь базовые свойства моделей слухового восприятия С. Снефа, О. Гитза, Р Лайона [12Virtanen], такие как: анализ в критических полосах слуха, нелинейность слухового восприятия и зависимости его от частоты, предыскажения кривой равной громкости, нелинейная зависимость между интенсивностью и громкостью, воспринимаемой человеком. В то же время методы MFCC и PLP не позволяют осуществлять коррекцию спектральных характеристик речевого сигнала на частотах отличных от частот критических полос слухового восприятия [64SternHearingIsBelieving]

В связи с указанными трудностями повышение робастности представления сигнала в системах АРР (рис. 1.4) за счет извлечения устойчивых к искажениям признаков целесообразно пересмотреть с позиций извлечения из речевого сигнала наиболее информативных, с точки зрения восприятия речевых сигналов и обработки его органами слуха, признаков для поиска возможных путей учета маскировки речевого сигнала шумом.

Исследователями Х. Занг, М. Хэитц, И. Брюс и Л. Кани [12Virtanen, 64Stern], на основе трёх классических моделей слухового восприятия С. Снефа, О. Гитза, Р Лайона [12Virtanen] была разработана более сложная, в сравнении с традиционными MFCC и PLP представлениями речевого сигнала, «физиологическая» модель активности слухового нерва для систем АРР [67Zhang], которая позволила описать отзыв активности слухового нерва на речь, маскированную шумом(рис. 1.8,а).

Как видно из рис. 1.8,а. модель Х. Занг, М. Хэитц, И. Брюс и Л. Кани включает в себя два пути: прохождения сигнала τ_c и контрольный путь τ_k, регулирующий временную константу τ_c нелинейных фильтров пути прохождения сигнала и ответственен за сжатие и подавление эффектов,

возникающих в отклике слухового нерва [67Zhang]. Путь прохождения сигнала моделирует настройку базилярной мембраны и состоит из каскада нелинейных, с изменяющимися во времени параметрами, узкополосных и линейных фильтров .

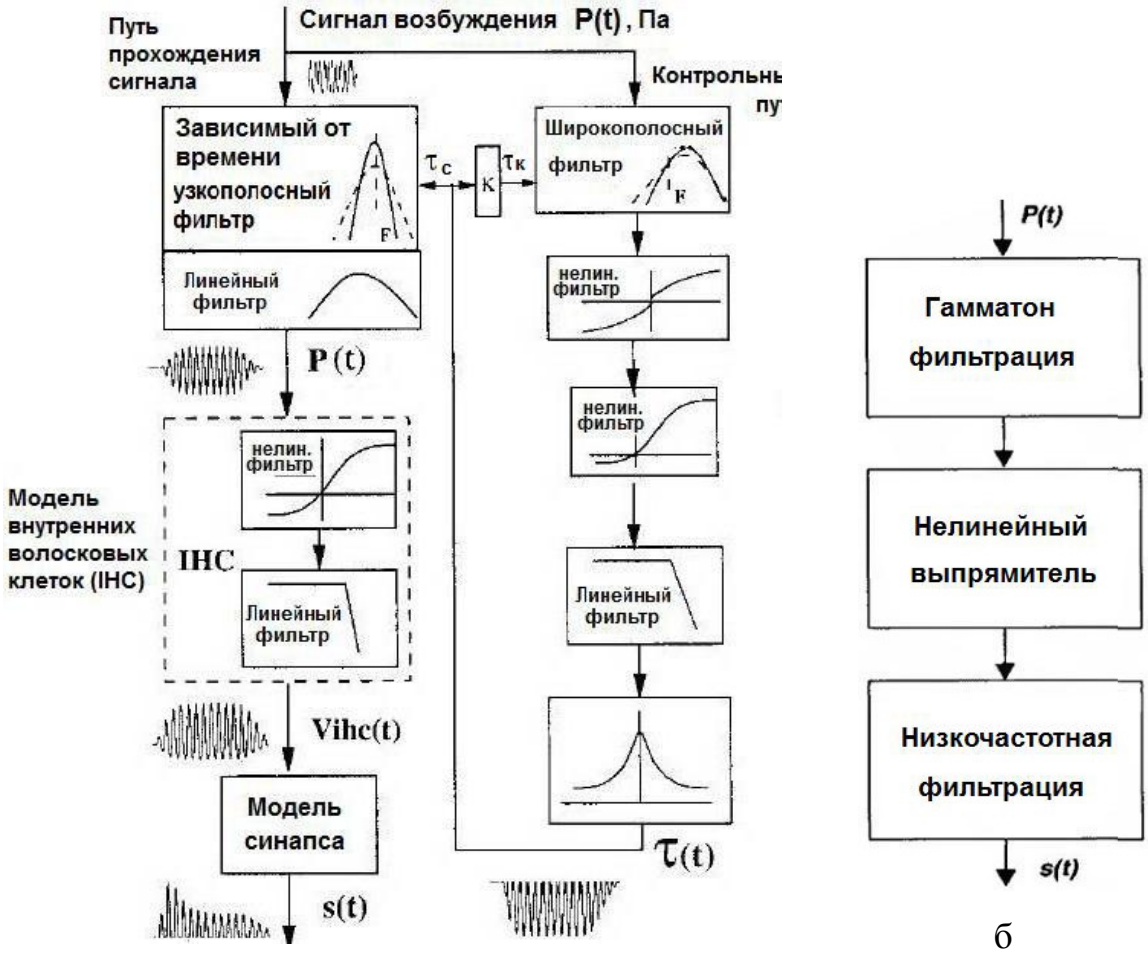


Рис. 1.8. Физиологические модели представления речевого сигнала

В работе С. Кима и Р. Стерна [68Kim] были проведены исследования модели Х. Занга, представления речевого сигнала MFCC и более упрощенной модели Х. Занга (состоящей из полосового фильтра, нелинейного выпрямителя, низкочастотного фильтра в каждом канале), представленной на рис. 1.8,б. Результаты этих исследований в виде точности автоматического распознавания речи показали, что полная слуховая модель Х. Занга обеспечивает 15 дБ выигрыша в соотношении сигнал-шум в сравнении с MFCC представлением сигнала. В тоже время упрощенная модель рис. 1.8,а

обеспечила выигрыш на 10дБ. Однако объем вычислений, которая потребовала модель Х. Занга (рис. 1.8,а), в 250 раз превышала аналогичный для модели MFCC.

Это обстоятельство стало причиной разработки более эффективного представления речевого сигнала на основе нормализованных по мощности кепстральных коэффициентов PNCC (рис. 1.9).

Компенсация шума, по методу PNCC [17, 18Kim], проводится по оценкам средней по времени мощности, получаемой путем усреднения в течение нескольких фреймов кратковременной оценки спектра мощности на выходе гребенки фильтров (1.19). Оценивая меняющийся во времени порог шума и отнимая его от кратковременной оценки спектра мощности, получают откорректированный речевую составляющую сигнала (1.22).

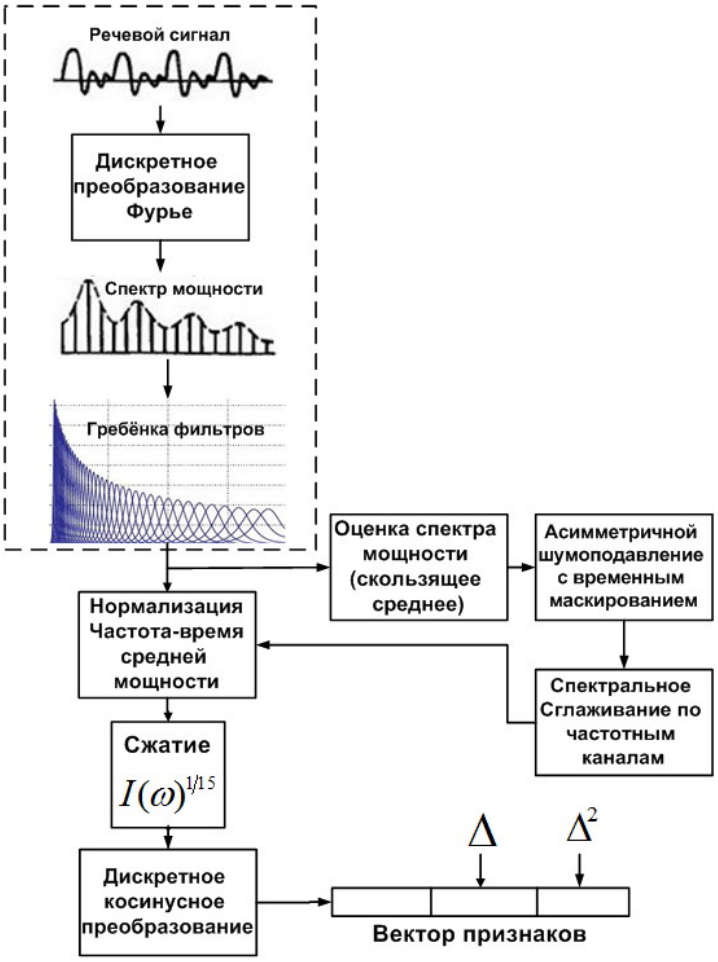


Рис. 1.7. Стадии вычисления PNCC коэффициентов

$$Q_{вих}[m,l]=\begin{cases}\lambda_a Q_{вих}[m-1,l]+(1-\lambda_a)Q_{вх}[m,l], \\ \quad if \quad Q_{вх}[m,l]\geq Q_{вих}[m-1,l] \\ \lambda_b Q_{вих}[m-1,l]+(1-\lambda_b)Q_{вх}[m-1,l], \\ \quad if \quad Q_{вх}[m,l]\geq Q_{вих}[m-1,l]\end{cases}, \tag{1.22}$$

где $Q_{вх}[m,l]$ - результат скользящего усреднения $MF[m,l]$ на протяжении M фреймов; $Q_{вих}[m,l]$ - усреднённая по времени мощность, при коэффициентах забывания $1>\lambda_a>\lambda_b>0$.

Проблема вычисления коэффициентов PNCC заключается в несовершенстве процедуры выявления речевого сигнала на фоне шума. Так, после получения откорректированного речевого сигнала к нему применяют идеальный линейный полупериодный выпрямитель. После этого речевой сигнал должен обрабатываться в первом канале, путем повторного применения фильтра (1.22) для определения порогового уровня мощности только для фреймов, где отсутствует речи. В то же время речевой сигнал должен обрабатываться во втором канале путем применения процедуры временного маскировки:

$$Q_{tm}[m,l]=\begin{cases}Q_0[m,l], & Q_0[m,l]\geq\lambda_t\cdot Q_P[m-1,l] \\ \mu_t\cdot Q_P[m-1,l], & Q_0[m,l]<\lambda_t\cdot Q_P[m-1,l]\end{cases}, \tag{1.23}$$

где $Q_0[m,l]$ - сигнал на выходе полупериодного выпрямителя; λ_t - коэффициент забывания; $Q_P[m,l]=\max\left(\lambda_t\cdot Q_P[m-1,l],Q_0[m,l]\right)$ - для каждого канала l и фрейма m.

Эта процедура позволяет подчеркнуть компоненты речевого сигнала, поступающие к слуховому аппарату первыми и подавляя возможное влияние запоздавших компонент, пришедших из других направлений распространения сигнала.

Таким образом формирования вектора коэффициентов PNCC полностью зависит от правильного определения речевого сигнала в данном случае определяется на основе порогового энергетического детектора

речевой активности, выбирая значения 1-го (соотношение (1.22)) или 2-го (соотношение (1.23)) каналов обработки следующим образом:

$$PNCC_n[m,l] = \begin{cases} \max\left(Q_{tm}[m,l], Q_f[m,l]\right), & речь \\ Q_f[m,l], & речь \quad отсутствует \end{cases}, \tag{1.24}$$

где n - количество полученных коэффициентов; $Q_{tm}[m,l]$ - сигнал на выходе процедуры временного маскирования; $Q_f[m,l]$ - значение нижней огибающей, полученное при помощи (1.22) на выходе полупериодного выпрямителя.

Выявленный существенный недостаток в вычислении PNCC коэффициентов может привести к некорректной обработки речевого сигнала в случае нестационарного шума окружения. Указанные выше недостатки обусловили постановку задач, решаемых в данной диссертации.

Вместе с тем к приведенным выше результатам работ [17, 18Kim] можно предъявить ряд претензий:

1. не обоснована эффективность алгоритмов MFCC, PLP и PNCC для искажений и шумов, связанных с наличием телефонного канала связи;

2. выбор значений коэффициентов забывания $\lambda_a \approx 0{,}999$ и $\lambda_b \approx 0{,}5$ блока ассиметричной фильтрации (1.22) было введено экспериментальным путем лишь для гауссового белого шума 5дБ, музыкального шума 5дБ и времени реверберации с $T_{60} = 0.5\text{с}$;

3. выбор значений коэффициента забывания $\lambda_t = 0{,}85$ и коэффициента подавления $\mu_t \le 0{,}2$ для блока временного маскирования (1.23) не рассматривался с позиций максимизации такого критерия как качество распознавания речи в телефонном канале связи.

Поиск ответов на данные вопросы составляет содержание поставленной задачи, решаемой в разделах – проверка целесообразности использования параметризации речевого сигнала мощностно-

нормализованными кепстральными коэффициентами PNCC и оптимизация параметров метода PNCC в задаче АРР в телефонном канале связи.

Переходя к формулировке второй задачи, отметим, что для метода PNCC принципиально важной процедурой обработки является эффективность разделения процедуры ассиметричной фильтрации и временного маскирования на основании определения, к какому типу, речь или пауза, относится обрабатываемый кадр анализа, (соотношения (1.22) и (1.23)). Поэтому в последующих главах данной диссертационной работы будет рассмотрена задача построения процедуры выявления голосовой активности во входящем сигнале.

1.4. Обзор методов выделения голосовой активности в речевом сигнале

Усовершенствование процедуры раздельной обработки речевого сигнала при определении голосовой активности и при определении параметров шума методом PNCC, упомянутое в предыдущем разделе, требует проведения анализа существующих подходов выявления голосовой активности. Общеизвестным является тот факт, что проектирования детектора голосовой активности, который в дальнейшем будем обозначать аббревиатурой VAD (Voice Activity Detector), является нетривиальной задачей, если шум (шумовая помеха) носит нестационарный характер или подобный речевому сигналу. VAD в свою очередь кроме применения в системах АРР используются также в мобильной телефонии и для разработки алгоритмов коррекции речевых сигналов. По этой причине существует огромное число различных подходов, зачастую разрабатываемых для конкретного вида его применения и поэтому могут не подходить для решения задачи повышения робастности системы АРР [12Virtanen].

1.4.1. Общие особенности традиционных подходов к детектированию голосовой активности.

Среди существующих традиционных подходов построения VAD наибольшее распространение получили методы, которые строятся по схеме, представленной на рис. 1.8. [12Virtanen, 1Rabiner, 70Benyassine, 69Atal, 71Архипов]

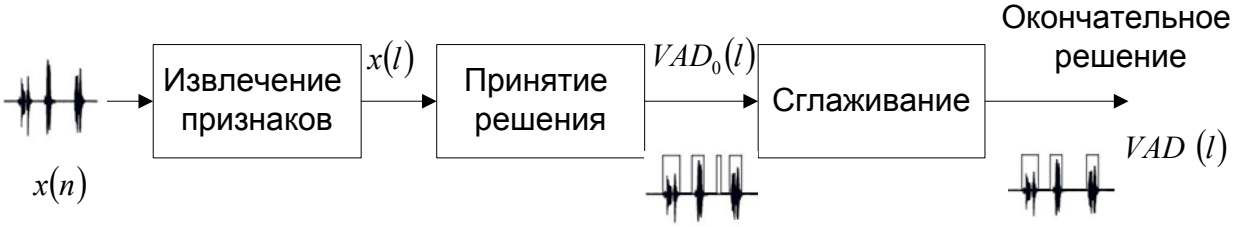


Рис. 1.8. Типичная схема всех методов VAD

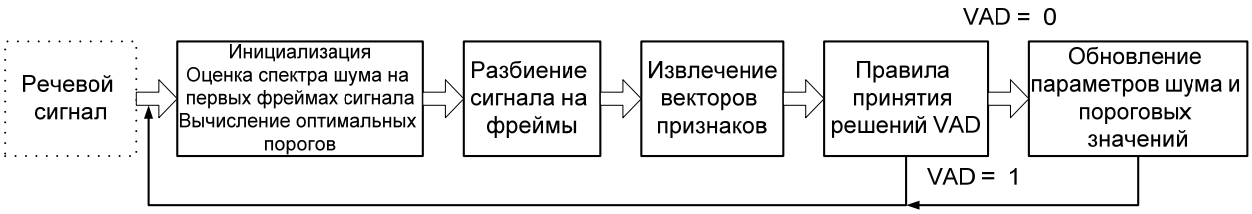
$x(n)$ – выходной речевой сигнал, содержащий речевое сообщение с паузами; $x(l)$ – вектор признаков; $VAD_0(l)$ – выходной сигнал в виде принятого закодированного решения о наличии или отсутствии голосовой активности в сигнале $x(n)$; $VAD(l)$ – сглаженное окончательное решение о наличии или отсутствии голосовой активности.

Известны два принципиально разных подхода к определению голосовой активности в речевом сигнале:

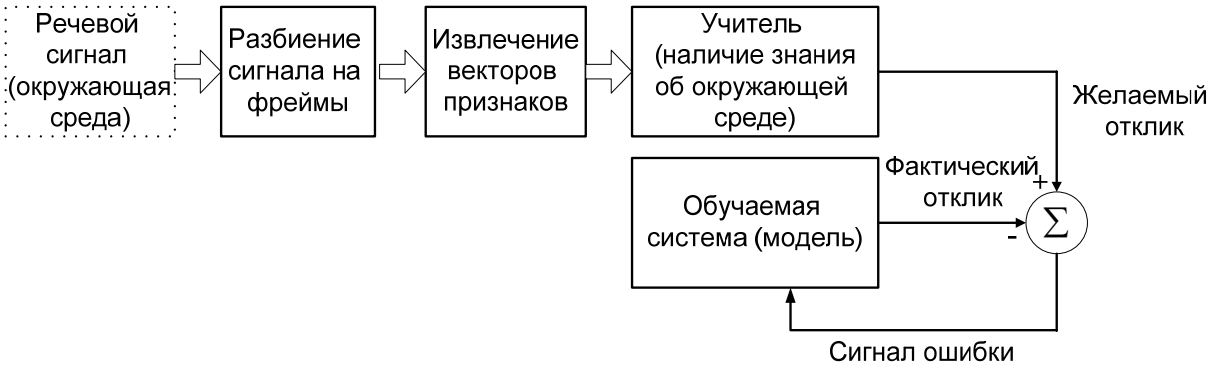
– «мягкий» VAD строится на основе схем обновления мощности шума [70Benyassine, 72AMR] (рекурсивное усреднение), поиска минимальной статистики мощности шума (в подполосах [73Martin]) либо же их объединения [74Cohen] (рис. 1.9,а). Такая схема встраивается в алгоритм, требующий оценки спектральной мощности шума и напрямую не определяет наличие пауз в речевом сигнале и считается предпочтительной для адаптации к изменяющемуся нестационарному шуму даже во время речевой активности [75Marzinzik];

– принятие решения на основе применения методов построения модели исследуемого процесса [69Atal] и принятие решения на основе построения соотношения правдоподобия [76Sohn], априори известного распределения

речевого и неречевого сигнала [77Ying], построения нейронной сети [78Pham] (рис. 1.9,б).



а - обновление параметров



б - схема обучения на основе коррекции ошибок

Рис. 1.9. Два подхода к выделению голосовой активности

Чрезвычайно важное и решающее достоинство VAD методов принятия решения на основе применение нейронных сетей заключается в том, что:

– сложность операций в ней можно определить наперед, путем изменения структуры связей между ее слоями в процессе ее проектирования.

– позволяет расширять выбор или заменять классификационные признаки, не меняя методы определения структуры связей между ее слоями.

– существует возможность адаптации параметров модели в процессе работы к изменениям окружающей среды, без изменения структуры выбранного метода принятия решения;

– принятие решения на основании анализа одного фрейма;

– исключение ограничений, налагаемых на выбранную модель, присущих традиционным методам, таких как:

– баланс ошибок и скорость отслеживания сигнала (tracking speed) для методов оценивающих спектральную мощности шума;

– оценивание параметров шума на речевых участках сигнала;

– упрощающего допущения статистическом законе распределения векторов признаков $x(l)$ [69Atal, 77Ying];

– оценивание статистических параметров шума на первых фреймах сигнала в предположении, что первые несколько фреймов не содержат речевого сигнала [70 Benyassine, 75Marzinzik, 72AMR];

– введения дополнительных модулей подавления шумовой помехи [79Ramírez], где используются методы оценки параметров шума как в случае «мягкого» VAD;

Формулирование «мягкого» принятия решения (рис. 1.9,а) существенно зависит от необходимости оценивать параметры инициализации алгоритма для шумовой помехи по нескольким первым фреймам сигнала для дальнейшего определения порогов и установки начального мульти-граничного принятия решения, проведения его сглаживания и обновления скользящего [70Benyassine,75Marzinzik] усреднения мощности шума.

Методы, основанные на оценке минимальной статистики (спектральной плотности шума) обновляют оценки параметров шума по фреймам. Основными недостатками этих методов являются: плохая реакция разделения на увеличение порога шума и уровня речевого сигнала [80Doblinger], внезапное увеличение порогового значения шума [81Hirch], зависимость оценки шума от оптимального сглаживания и минимальной статистики, что в свою очередь требует большего времени обновления спектра шума, когда порог шума резко возрастает [73Martin]

В статистических методах назначения класса (тон, шум пауза) вектору признаков [69Atal, 76Sohn 82Qi, 77Ying], вводят допущение о распределении значений вектора признаков по гауссовому закону, при этом инициализация

статистических моделей строится так же на предположении отсутствия речевого сигнала на начальных фреймах сигнала, что на практике не всегда верно. Главным недостатком этих методов является возможность неправильной оценки параметров шумовой помехи на фреймах анализа, содержащих речевую составляющую, что в свою очередь может приводить к неправильной корректировке речевого сигнала и его искажению.

Современные методы VAD для принятия решения используют статистические свойства мгновенного соотношения сигнал-шум (апостериорного SNR) в качестве критерия принятия решения [76Sohn]. При этом в качестве признаков выбирают Фурье область, а именно распределение (составная гауссова модель, распределенная по экспоненциальному закону) и зависимость Фурье коэффициентов. На участках речевых пауз среднее значение и дисперсия апостериорного SNR близки к единице и значительно больше апостериорного SNR для участков с речевой активностью [83 Vary].

Однако, главной проблемой построения таких VAD возникает при классификации нестационарного и не гауссового шума. В этом случае статистические свойства шума становятся похожими на статистику речевого сигнала и не совпадают с характеристиками гауссова шума [12Virtanen]. В этом случае необходимо строить более надежную статистическую модель, путем введения процедуры адаптации.

Статистические методы, основанные на обучении гауссовых смесей без учителя [77Ying], в которых уходят от предположения об отсутствии речевого сигнала в начальных фреймах сигнала, сводятся к упрощению вычислительной сложности оценки параметров речи и шума на основе модели смесей гауссиан. В частности, наиболее существенной проблемой алгоритмов построения классификатора на основе правила Байеса является требование наличия большого набора данных и метода для надежного обучения [82Qi].

Помимо этого существенным для построения детектора VAD является использование простой системы признаков. Существующий ряд алгоритмов

[84Ghasemi, 86Martin, 78Pham], используют сложную систему признаков, такие как спектральные или мел-частотные кепстральные коэффициенты MFCC. Главный недостаток использования процедуры вычисления системы MFCC признаков [78Pham, 86Kos] заключается в том, что она сравнима по вычислительной сложности с процедурой самого процесса распознавания и поэтому неэффективна для целей определения голосовой активности в методе получения новой параметризации речевого сигнала.

В других алгоритмах [69Atal, 71Архипов, 72AMR] используются простые системы признаков (энергия сигнала, число переходов через ноль, коэффициенты автокорелляционной функции), однако применяемые там правила приятия решения не позволяют обеспечить той точности, которая была бы возможна при использовании более сложных нелинейных классификаторов [78Pham, 86Kos].

В [82Qi, 69Atal] указано, что использование одного признака приводит к ограниченной точности метода VAD из-за того, что один отдельно взятый признак частично перекрывается между классами, в особенности, когда речь записана в условиях помех.

В связи с указанными ограничениями, присутствующими в традиционных VAD методах, для построения надежного детектора голосовой активности необходимо выполнить основные требования, предъявляемые к алгоритму PNCC:

– простая система признаков, которая может быть получена в реальном времени, что позволит не усложнить алгоритм PNCC;

– эффективный алгоритм принятия решения, управляемый ограниченным числом параметров;

– быстродействие для применения при работе с алгоритмами реального времени,

– работа на фреймах анализа системы APP.

Поэтому разработка устойчивой к нестационарному шуму модели детектора голосовой активности является задачей исключительной важности и составляет содержание одной из задач, решаемых в данной диссертации.

1.5. Выводы

Подводя итог проведенному обзору, можно сделать следующие выводы:

1. Значительное число современных методов повышения робастности систем АРР основаны на подавлении шумовой и реверберационной помехи методами предварительной обработки сигнала до поступления в систему АРР и в модуле предварительной обработки системы АРР. Предварительная обработка позволяет сохранять неизменной структуру и параметры системы АРР.

2. Наиболее простым в плане технической реализации, методом коррекции речевых сигналов является метод частотной коррекции. Преимущество метода частотной коррекции, по сравнению с методом обратной фильтрации, состоит в устойчивости результатов к перемещениям диктора относительно микрофона.

3. Оценивание амплитудной частотной характеристики корректирующего фильтра требует предварительной оценки спектра мощности поздней реверберации. Для расстояний между источником звука и микрофоном, больших критической дистанции, спектр мощности поздней реверберации может быть вычислен через спектр мощности наблюдаемого сигнала. К сожалению, изложенные в указанных работах рекомендации по выбору параметров степени усреднения спектра мощности наблюдаемого сигнала и времени реверберации недостаточно аргументированы Как следствие, это не гарантирует достижения максимального качества восстановленного сигнала и точности его распознавания.

3. В ряде случаев невозможно (отсутствует образец записи ИХ помещения, отсутствует доступ в помещение и т.п.) получить параметр времени реверберации. В этом случае необходимо произвести слепое измерение времени реверберации по речевому сигналу, с последующим восстановлением сигнала. Однако слепые измерения времени реверберации параметра неизбежно сопровождаются погрешностью измерений, снижающей эффективность подавления поздней реверберации. К сожалению, оценка степени снижения такой эффективности в литературе не указана.

4. Предварительное ослабление помех неизбежно сопровождается искажением речевых сигналов, что отрицательно сказывается на точности систем АРР. А используемые до сих пор методы параметрического представления речевых сигналов в системах АРР позволяют получать лишь усредненные значения спектральных характеристик речевого сигнала на выходе гребенки фильтров, что не дает возможности восстановить речевой сигнал во временной области или произвести коррекцию сигнала на частотах отличных от центральных частот критических полос слухового восприятия.

5. Разработка моделей слухового восприятия сопровождалась главным их недостатком, который заключался в большом объёме вычислений, необходимых для построения таких моделей. Метод представления речевого сигнала в системах АРР нормализованные по мощности кепстральные коэффициенты стал одной из упрощенных моделей слухового восприятия. Проблема вычисления коэффициентов PNCC заключается в несовершенстве процедуры выявления речевого сигнала на фоне шума. Таким образом формирование вектора коэффициентов PNCC полностью зависит от правильности определения речевого сигнала, которое на данный момент определяется на основе порогового энергетического детектора речевой активности. Определенный существенный недостаток вычисления PNCC коэффициентов может привести к некорректной обработки речевого сигнала в случае нестационарного шума окружения.

ГЛАВА 2.
ОПТИМИЗАЦИЯ ПРОЦЕДУРЫ ПОДАВЛЕНИЯ ПОЗДНЕЙ РЕВЕРБЕРАЦИИ ПРИ АВТОМАТИЧЕСКОМ РАСПОЗНАВАНИИ РЕЧИ

С целью разрешения в предыдущем разделе трудностей, связанных с оптимизацией параметров оценки спектра поздней реверберации, в данном разделе рассматриваются задачи оптимизации оценки спектра поздней реверберации.

Переходя к формулировке второй задачи, отметим, что поскольку в соотношении (1.5) фигурирует время реверберации $T_{60}(k)$, это означает, что время реверберации необходимо предварительно оценить по имеющейся импульсной характеристике помещения. Между тем, в ряде случаев такая оценка невозможна (отсутствует образец записи ИХ помещения, отсутствует доступ в помещение и т.п.). В этом случае необходимо произвести слепое измерение времени реверберации $T_{60}(k)$ по речевому сигналу $y(t)$, с последующим восстановлением сигнала $x(t)$. Однако слепые измерения времени реверберации параметра $T_{60}(k)$ неизбежно сопровождаются погрешностью измерений, снижающей эффективность подавления поздней реверберации. К сожалению, оценка степени снижения такой эффективности в литературе не указана. Кроме того, не известны рекомендации по минимизации такого снижения. Поиск ответов на данные вопросы и составляет содержание второй задачи, решаемой в данном разделе.

2.1. Оптимизация процедуры и параметров оценки спектра поздней реверберации

Если система подавления поздней реверберации (деревербератор) осуществляет предобработку речевого сигнала, подаваемого на вход системы АРР, как показано на рис. 2.1, тогда целесообразно использовать сквозной

(интегральный) показатель качества $Acc\%$, именуемый «точностью правильного распознавания слов» [87Young]:

$$Acc\% = \frac{N - D - S - I}{N} \times 100\%,$$

где N - общее количество распознаваемых слов; D - количество ошибочно удаленных слов; S - количество замененных слов; I - количество ошибочно вставленных слов.

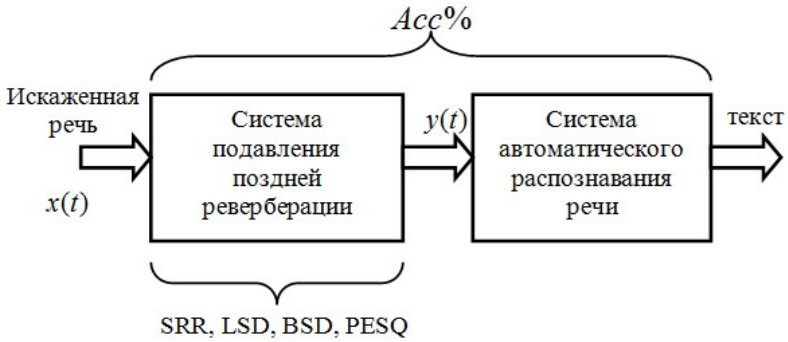


Рис. 2.1. Деревербератор как препроцессор системы АРР

Вместе с тем, понятно желание использовать более простые, с точки зрения объема вычислений при моделировании, «промежуточные» показатели качества. Например, для оценки деревербераторов часто используют такие показатели как отношение сигнал-реверберация (Signal-to-Reverberation Ratio - SRR) [51Habets]:

$$SRR = \frac{1}{L}\sum_{l=1}^{L} 10\lg\left[\frac{\sum_{n=Rl}^{Rl+N-1} x^2(l,n)}{\sum_{n=Rl}^{Rl+N-1}[x(l,n) - y(l,n)]^2}\right],$$

логарифмично-спектральное искажение (Logarithmic Spectral Distortion - LSD)

$$LSD = \frac{2}{KL}\sum_{l}\sum_{k=0}^{\frac{K}{2}-1}\left|G\{X(l,k)\} - G\{Y(l,k)\}\right|,$$

$$G\{X(l,k)\} = \max\{20\lg(|X(l,k)|), \delta\}, \quad \delta = \max_{l,k}\{20\lg(|X(l,k)|)\} - 50,$$

барк-спектральное искажение (Bark Spectral Distortion - BSD)

$$BSD = \frac{\sum_{l=1}^{L}\sum_{k=0}^{\frac{K}{2}-1}\left[B\{X(l,k)\} - B\{Y(l,k)\}\right]^2}{\sum_{l=1}^{L}\sum_{k=0}^{\frac{K}{2}-1}\left[B\{X(l,k)\}\right]^2},$$

где $x(l,n)$ и $y(l,n)$ - n-я выборка l-го фрейма входного и выходного сигналов $x(n)$ и $y(n)$, соответственно, системы подавления реверберации; $X(l,k)$ и $Y(l,k)$ - амплитудные спектры l-го фрейма сигналов $x(n)$ и $y(n)$, соответственно; $B\{X(l,k)\}$ и $B\{Y(l,k)\}$ - барк-спектры l-го фрейма сигналов $x(n)$ и $y(n)$, соответственно. Алгоритм вычисления показателя PESQ (Perceptual Evaluation of Speech Quality - перцептуальная оценка качества речи) более громоздок [88Loizou, 89Beerends], тем не менее, как и в случае с показателями SRR, LSD, и BSD, исходными данными для алгоритма вычисления PESQ служат выборки сопоставляемых сигналов $x(n)$ и $y(n)$.

2.1.1. Коэффициент передачи метода частотной коррекции.

Как отмечено в п. 1.2.1, метод частотной коррекции состоит в оптимальной фильтрации искаженного сигнала $y(t)$, при этом коэффициент передачи $G(l,k)$ можно вычислить различными способами. Одним из наиболее распространенным, в силу своей высокой эффективности, является алгоритм logMMSE (Logarithmic Minimum Mean-Square Estimation) [14Ephraim], согласно которому $G(l,k)$ вычисляют в соответствии с соотношением:

$$G(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\exp\left(\frac{1}{2}\int_{v(l,k)}^{\infty}\frac{e^{-t}}{t}dt\right), \qquad v(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\gamma(l,k),$$

где $\xi(l,k) = \lambda_x(l,k)/\lambda_n(l,k)$ и $\gamma(l,k) = \lambda_y(l,k)/\lambda_n(l,k)$ - априорное и апостериорное отношение сигнал-помеха, соответственно; $\lambda_n(l,k)$ - спектр

мощности l-го сегмента помехи; $\lambda_x(l,k)$ и $\lambda_y(l,k)$ - спектры мощности l-го сегмента сигналов $x(t)$ и $y(t)$, соответственно, на частоте $f_k = kF_s / N_{fft}$.

При этом принципиально важными и сложными являются две подзадачи [14Ephraim]:

– оценивание спектра помехи $\lambda_n(l,k)$;

– оценивание априорного отношения сигнал-помеха $\xi(l,k) = \lambda_x(l,k)/\lambda_n(l,k)$, где $\lambda_x(l,k)$ - спектр мощности сигнала $x(t)$.

Применительно к задаче подавления поздней реверберации, в роли спектра мощности помехи $\lambda_n(l,k)$ теперь выступает спектр поздней реверберации $\lambda_r(l,k)$.

2.1.3. Проверка целесообразности усреднения периодограмм фреймов. В пионерской работе К. Леберта [50Lebart] ключевыми являются два соотношения, используемыми для оценивания спектра поздней реверберации:

$$\lambda_r(l,k) = e^{-2\delta(k)T_l} \cdot \lambda_y(l - N_l, k), \tag{2.1}$$

$$\hat{\lambda}_y(l,k) = \eta_z \hat{\lambda}_y(l-1,k) + (1-\eta_z)\left|Y(l,k)\right|^2. \tag{2.2}$$

При этом параметр, регулирующий степень усреднения в (2.2), предложено принимать $\eta_z \approx 0{,}9$.

Отметим, что соотношение (2.2) в [50Lebart] было введено «волевым» путем. Аргументация при этом была следующей. Дисперсия периодограммы $\left|Y(l,k)\right|^2$ велика, поэтому для ее уменьшения следует усреднить несколько периодограмм. При этом следует учитывать, что чрезмерное усреднение может привести к росту смещения оценки (2.2).

При всей кажущейся логичности данных рассуждений, остается необоснованным выбор значения $\eta_z \approx 0{,}9$, хотя очевидно, что такой выбор

следует производить с учетом таких параметров оценивания как частота дискретизации F_s, время реверберации T_{60}, протяженность фреймов T_ϕ, степень перекрытия фреймов. Этот недостаток частично устранен в работе [51Habets], где параметр усреднения η_z предложено выбирать с учетом значений F_s, T_{60} и степени перекрытия фреймов (соотношения (1.8)-(1.10)). Однако и здесь почему-то совершенно не учтены такие базовые характеристики речевого сигнала как интервал стационарности и протяженность фонем.

Заметим также, что даже сама идея усреднения множества смежных периодограмм $|Y(l,k)|^2$ не свободна от критики. Чтобы продемонстрировать это, проделаем следующие выкладки.

Корреляционная функция реверберированного сигнала в момент времени $t-T_l$ описывается соотношением [50Lebart]:

$$r_y(t-T_l, t-T_l+\tau) = \sigma^2 e^{-2\delta(t-T_l)} \int_{-\infty}^{t-T_l} \langle x(\theta)x(\theta+\tau)\rangle e^{2\delta\theta} d\theta . \qquad (2.3)$$

Представим интеграл в соотношении (2.3) в виде суммы:

$$I = \int_{-\infty}^{t-T_l} \langle x(\theta)x(\theta+\tau)\rangle e^{2\delta\theta} d\theta = \int_{t-T_l-T_\phi}^{t-T_l} + \int_{t-T_l-2T_\phi}^{t-T_l-T_\phi} + \int_{t-T_l-3T_\phi}^{t-T_l-2T_\phi} + \ldots , \qquad (2.4)$$

где T_ϕ - протяженность фрейма.

Если размер фрейма T_ϕ настолько мал, что в его пределах можно считать речевой сигнал $x(t)$ стационарным случайным процессом, т.е.

$$\langle x(\theta)x(\theta+\tau)\rangle \approx r_x(\tau), \qquad (2.5)$$

тогда (2.4) можно приближенно представить в виде:

$$I \approx r_x(t-T_l,\tau)\int\limits_{t-T_l-T_\phi}^{t-T_l} e^{2\delta\theta}d\theta + r_x(t-T_l-T_\phi,\tau)\int\limits_{t-T_l-2T_\phi}^{t-T_l-T_\phi} e^{2\delta\theta}d\theta + r_x(t-T_l-2T_\phi,\tau)\int\limits_{t-T_l-3T_\phi}^{t-T_l-2T_\phi} e^{2\delta\theta}d\theta + \ldots =$$

$$= \frac{e^{2\delta(t-T_l)}}{2\delta}\cdot\left(1-e^{-2\delta T_\phi}\right)\cdot\left[r_x(t-T_l,\tau) + e^{-2\delta T_\phi}r_x(t-T_l-T_\phi,\tau) + e^{-4\delta T_\phi}r_x(t-T_l-2T_\phi,\tau) + \ldots\right] \tag{2.6}$$

Нетрудно видеть, что в соотношении (2.6) корреляционные функции отдельных фреймов подвергаются усреднению с помощью весовых коэффициентов

$$\left(1-e^{-2\delta T_\phi}\right),\ \left(1-e^{-2\delta T_\phi}\right)e^{-2\delta T_\phi},\ \left(1-e^{-2\delta T_\phi}\right)e^{-4\delta T_\phi}. \tag{2.7}$$

Эти коэффициенты являются членами убывающей геометрической прогрессии со знаменателем $q = e^{-2\delta T_\phi}$, сумма членов которой равна 1.

Таким образом, можно записать:

$$r_y(t-T_l,t-T_l+\tau) = \frac{\sigma^2}{2\delta}r_{x\,cp}(t-T_l,t-T_l+\tau), \tag{2.8}$$

где

$$r_{x\,cp}(t-T_l,t-T_l+\tau) = \left(1-e^{-2\delta T_\phi}\right)\cdot\left[r_s(t-T_l,\tau) + e^{-2\delta T_\phi}r_s(t-T_l-T_\phi,\tau) + \ldots\right]$$

Процедуру такого усреднения можно описать с помощью рекуррентного соотношения:

$$r_{x\,cp}(l,\tau) = e^{-2\delta T_\phi}\ r_{x\,cp}(l-1,\tau) + \left(1-e^{-2\delta T_\phi}\right)r_x(l,\tau), \tag{2.9}$$

где l - номер фрейма. Переходя от соотношений (2.8)-(2.9) к кратковременным спектрам, получим:

$$\lambda_y(l,f) = \frac{\sigma^2}{2\delta}\lambda_{x\,cp}(l,f), \qquad (2.10)$$

где

$$\lambda_{x\,cp}(l,k) = e^{-2\delta T_\phi}\ \lambda_{x\,cp}(l-1,k) + \left(1-e^{-2\delta T_\phi}\right)\left|Z_s(l,k)\right|^2, \qquad \delta = \frac{6{,}9}{T_{60}}. \qquad (2.11,\text{а})$$

Очевидно, если фреймы перекрываются и сдвинуты друг относительно друга на R выборок, тогда

$$\eta_z^d = e^{-2\delta R/f_s} \approx 1 - 2\delta R/f_s,$$

что при $2\delta R/f_s << 1$ совпадает с соотношением (4.6), поскольку

$$\frac{1}{1+2\delta R/f_s} \approx 1 - 2\delta R/f_s,$$

и тогда

$$\lambda_{x\,cp}(l,k) = \eta_z^d\ \lambda_{x\,cp}(l-1,k) + \left(1-\eta_z^d\right)\left|Z_s(l,k)\right|^2, \qquad \eta_z^d = e^{-2\delta R/f_s}, \quad \delta = \frac{6{,}9}{T_{60}}. \ (2.11,\text{б})$$

Соотношения (2.10) и (2.11) не только свидетельствуют, что оценку спектра мощности реверберированного сигнала следовало бы получать усреднением периодограмм предшествующих фреймов чистого сигнала, но и показывают, как выбрать коэффициент усреднения η_z. Между тем, в соотношениях (1.6)-(1.7) предлагается спектр реверберированного сигнала оценивать путем усреднения периодограмм этого же реверберированного сигнала. Таким образом, при оценивании спектра реверберированного сигнала происходит, по существу, двукратное, а не однократное усреднение по времени периодограмм чистой речи.

Оправданием соотношениям (1.6)-(1.7) может служить то обстоятельство, что эти соотношения можно трактовать как разновидность оценки Уэлча спектра мощности случайного процесса [90Дідковський]. Однако тогда, учитывая нестационарный характер речевого сигнала, интервал усреднения следовало бы выбирать, ориентируясь не только и не столько на время реверберации T_{60}, как это сделано в соотношениях (1.8)-

(1.20), сколько на интервал стационарности речевого сигнала (близкий 10-20 мс), а также учитывая статистику протяженностей звуков речи.

Чтобы проверить эффективность процедур усреднения (1.6)-(1.7), в данной работе испытаны три альтернативных варианта. Согласно первым двум вариантам, спектр мощности реверберированного сигнала оценивают в соответствии с соотношениями:

$$\hat{\lambda}_y(l,k) = |Y(l,k)|^2; \tag{2.12}$$

$$\hat{\lambda}_y(l,k) = \begin{cases} |Y(l,k)|^2, & \textit{для фрейма с гласным звуком речи}; \\ |Y(l,k)|^2 \otimes S_w(k), & \textit{в остальных случаях} \end{cases} \tag{2.13}$$

где $S_w(k)$ - спектральное сглаживающее окно.

Вариант (2.12) означает, что в качестве l-й оценки спектра мощности реверберированного сигнала используется l-я периодограмма одиночного фрейма этого сигнала. Вариант (2.13) более сложен: предварительно определяется принадлежность фрейма к гласным звукам, после чего периодограмма согласного звука, не содержащая дискретных компонентов, сглаживается по частоте спектральным окном $S_w(k)$, а периодограмма гласного звука оставляется без изменений. Заметим, что процедура сглаживания периодограммы спектральным окном $S_w(k)$, как и процедуры усреднения периодограмм по времени (1.6)-(1.7), направлена на уменьшение дисперсии оценки спектра случайного процесса и известна в литературе под названием «квадратично-модифицированной» периодограммы [90Дідковський].

Третий альтернативный вариант имеет вид двухэтапной процедуры:

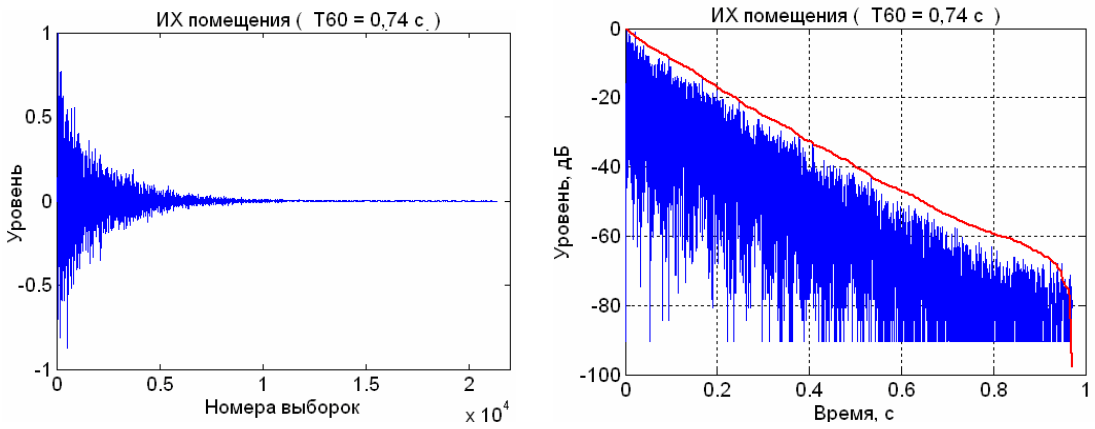
1) на первом этапе производится «первичная дереверберация» сигнала $y(t)$, при этом спектр реверберированного сигнала оценивают в соответствии с соотношением (2.12);

2) на втором этапе спектр реверберированного сигнала оценивают в соответствии с соотношениями (2.10)-(2.11), принимая в качестве «чистого» речевого сигнала результат выполнения первого этапа.

Экспериментальные исследования организованы следующим образом.

Моделирование реверберированного сигнала осуществлялось в соответствии с (1.4,а), т.е. путем вычисления свертки чистого речевого сигнала $x(t)$ с ИХ помещения $h(t)$ (рис. 2.2), полученной в результате записи звука лопнувшего резинового шарика, расположенного на расстоянии 3,5 м от микрофона. Красной линией на рис. 2.2,б показан огибающая функции $h(t)$, вычисленная в соответствии с соотношением Шредера [91Schroeder]:

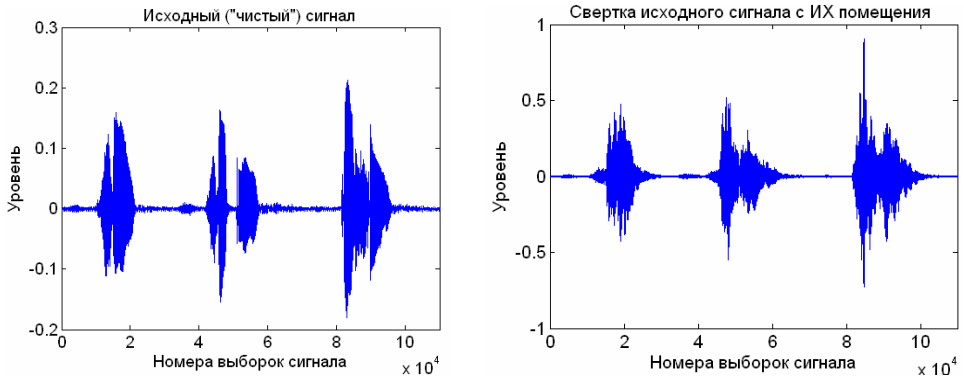
$$D(t) = c\int_{t}^{\infty} h^2(x)dx\,.$$



а б

Рис. 2.2. ИХ помещения в линейном (а) и логарифмическом (б) масштабах

Искажающее действие реверберации на форму исходного сигнала отчетливо наблюдается на рис. 2.3,б.

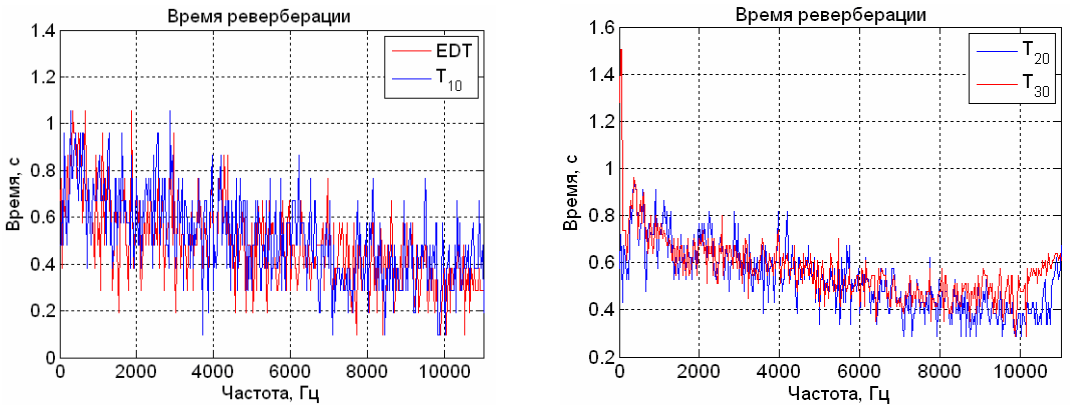


а б

Рис. 2.3. Исходный сигнал (а) и свертка его с ИХ помещения (б)

Значения параметров сигналов и алгоритмов обработки при деревербераци принимались следующими: частота дискретизации $F_s = 22050$ Гц; протяженность фрейма 32 мс; степень перекрытия фреймов 50%; фреймы взвешивались окном Хэмминга; момент начала поздней реверберации $T_l = 96$ *мс*; «интегральное» время реверберации $T_{60} = 0{,}74$ с.

Поскольку время реверберации зависит от частоты, требовалось оценить такую зависимость. Графики зависимости оценок времени реверберации от частоты для различных сочетаний порогов оценивания приведены на рис. 4.8. Анализ приведенных графиков позволяет заключить, что значения $EDT(f)$ и $T_{10}(f)$ существенно флюктуируют относительно средних значений, что поясняется малым расстоянием (10 дБ) между верхним и нижним порогами (0 дБ и -10 дБ для $EDT(f)$ и -5 дБ и -15 дБ для $T_{10}(f)$). Анализ графиков зависимостей $T_{20}(f)$ и $T_{30}(f)$ свидетельствует, что до частоты 8 кГц их средние значения совпадают, однако для более высоких частот для графика $T_{30}(f)$ наблюдается тенденция к завышению результатов оценивания, что можно пояснить ложными пересечениями нижнего порога (-35 дБ). Поэтому при экспериментальных исследованиях было принято $T_{60}(f) = T_{20}(f)$.



а б

Рис. 2.4. Время реверберации как функция частоты

Для моделирования системы АРР, а также для оценки показателя *Acc*%, применялся программный инструментарий НТК [9]. Обучение системы АРР производилось с использованием 269 образцов 27 слов украинской речи, произнесенных 2-мя дикторами женщинами. Тестовый сигнал представлял собой звуковой файл дискретной речи с записью поочередно зачитываемых всех 27 слов, использованных при обучении. Паузы между словами при этом составляли 0,2…0,5 с. Параметры оцифровки звуковых файлов: частота дискретизации 22050 Гц, равномерное квантование 16 бит. Фонемный словарь состоял из 27 фонем украинской речи. Использовались 39-мерные классификационные признаки вида MFCC_0_D_A.

Результаты экспериментальных исследований вариантов в виде значений точности автоматического распознавания *Acc*% приведены в табл. 2.1 и на графике рис. 2.5.

Из приведенных результатов следует, что качество автоматического распознавания речевых сигналов резко ухудшается из-за влияния реверберации: значение *Acc*% снижается с 93% для чистой речи до 22% для реверберированной речи. Процедура деревербераций, реализованная с учетом соотношения (2.12), позволила поднять *Acc*% до 44%.

Таблица 2.1

Вид сигнала	Чистая речь	Реверберр. речь	Дерев-ция по (4.17)	Дерев-ция по (4.18)	2-хэтапная дерев-ция	Дерев-ция по (4.2)-(4.3)
Эксперимент	1	2	3	4	5	6
$Acc\%$	93	22	44	41	67	74

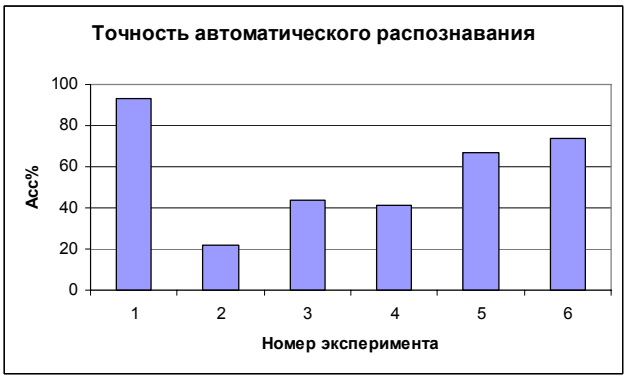


Рис. 2.5 4.9. Точность автоматического распознавания в различных экспериментах

Между тем, процедура деревербераций, реализованная с учетом соотношения (2.13), позволила поднять $Acc\%$ лишь до 41%, что можно пояснить сложностью принятия решения «гласный»-«согласный» по реверберированному сигналу $y(t)$.

Чтобы наглядно продемонстрировать эту сложность, рассмотрим графики рис. 2.6 и 2.7. На рис. 2.6, изображены спектрограмма и результат работы алгоритма идентификации «гласный-согласный» для фрагмента чистого звукового сигнала со словами: са, сапа, коса, си, син, сир. На рис. 2.7 изображены спектрограмма и результат работы алгоритма идентификации «гласный-согласный» для такого же фрагмента реверберированного звукового сигнала. Как следует из рис. 2.7, в случае реверберированной речи идентификация звуковых фреймов на «гласный-согласный» осуществляется со значительными ошибками: идентифицируются лишь шипящие согласные звуки. Эти ошибки в известной степени обусловлены простотой алгоритма идентификации, сущность которого состояла в оценке частоты нуль-

пересечений и в сравнении полученной оценки с заданным порогом (1 кГц). Вместе с тем, усложнять алгоритм идентификации не представлялось целесообразным ввиду нежелательного увеличения объема вычислений.

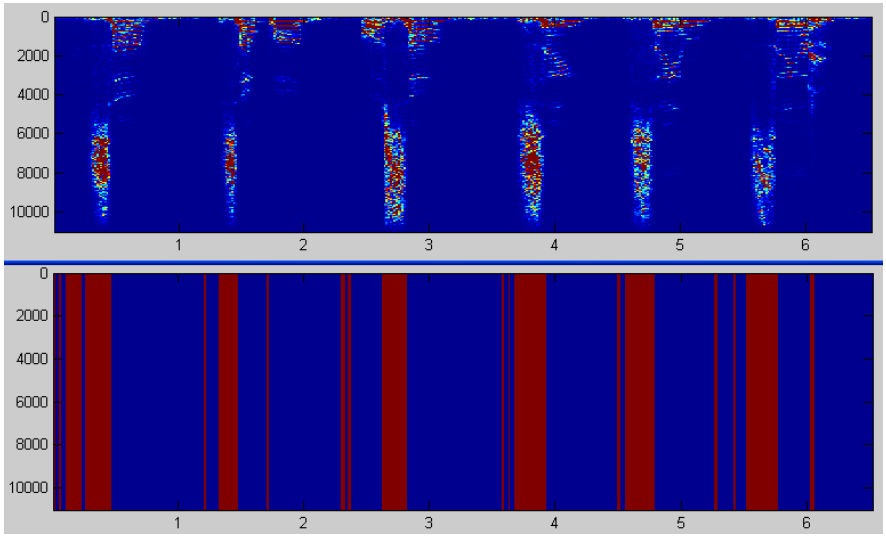


Рис. 2.6. Спектрограмма (а) и результат работы алгоритма идентификации согласных звуков (б) для чистого речевого сигнала

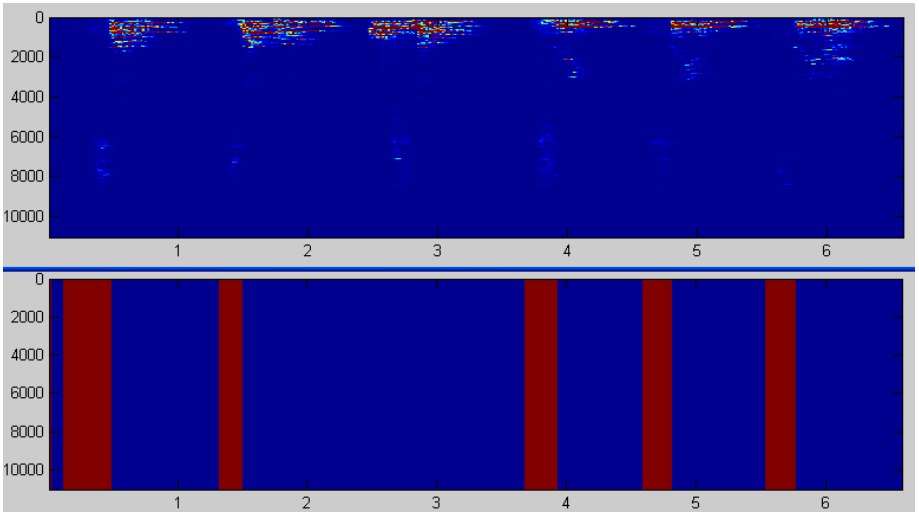


Рис. 2.7. Спектрограмма (а) и результат работы алгоритма идентификации согласных звуков (б) для реверберированного речевого сигнала

Как следует из табл. 2.1 и рис. 2.5, двухэтапная процедура оказалась существенно эффективнее одноэтапной, позволяя вплотную приблизиться к

точности автоматического распознавания, обеспечиваемой соотношениями (1.6)-(1.7). Однако двухэтапная процедура проигрывает одноэтапной почти в два раза по объему вычислений, что делает ее невыгодной в вычислительном плане.

Резюмируя приведенные результаты, заключаем, что процедура оценивания спектра реверберированного сигнала, реализуемая в соответствии с соотношениями (1.6)-(1.7), действительно наиболее эффективна как по объему вычислений, так и по качеству деревербераций.

При этом, однако, открытым остается вопрос о выборе параметров процедуры оценивания. Результаты таких исследований изложены ниже.

2.1.4. Оптимизация параметров процедуры оценивания спектра поздней реверберации. Как отмечалось в п. 1.2.1, к оценкам спектра поздней реверберации, описываемым соотношениями (1.5)-(1.10), можно предъявить ряд претензий, сущность которых вкратце сводится к трем вопросам:

1) каким должно быть оптимальное значение параметра T_l, используемого в соотношении (1.5)?

2) как выбрать оптимальное значение параметра сглаживания η_z при оценивании спектра $\lambda_y(l,k)$ реверберированного сигнала в соответствии с соотношением (1.6)?

3) действительно ли соотношения (1.7)-(1.10) предпочтительнее соотношения (1.6) при измерениях спектра $\lambda_y(l,k)$ реверберированного сигнала?

К сожалению, с помощью аналитических исследований невозможно ответить на поставленные вопросы, поскольку в данной работе деревербератор рассматривается как составная часть (препроцессор) системы АРР. Следствием такого подхода является то, что для адекватной оценки качества деревербератора необходимо использовать сквозной показатель

$Acc\%$, характеризующий точность распознавания речи. Что касается намного более «удобных» для аналитических исследований показателей SRR, LSD, BSD, PESQ, характеризующих искажение формы речевого сигнала, их согласованность с показателем $Acc\%$ можно проверить лишь экспериментальным путем.

Таким образом, исследования проводились путем компьютерного моделирования, а результаты экспериментов оценивались как качественно (на слух и путем визуального анализа спектрограмм), так и количественно, путем оценивания показателей $Acc\%$, SRR, LSD, BSD, PESQ.

Организация экспериментальных исследований в целом подобна таковой для п. 2.1.3, поэтому укажем лишь на отличительные особенности.

При качественных исследованиях анализировался реверберированный сигнал, записанный в помещении объемом 80 м^3, и временем реверберации 1,1 с. ИХ помещения $h(t)$ представляла собой запись звука лопнувшего резинового шарика (рис. 2.8). Расстояние между диктором и микрофоном составляло 2 м, что больше критической дистанции $D_c \approx 0{,}5$ м.

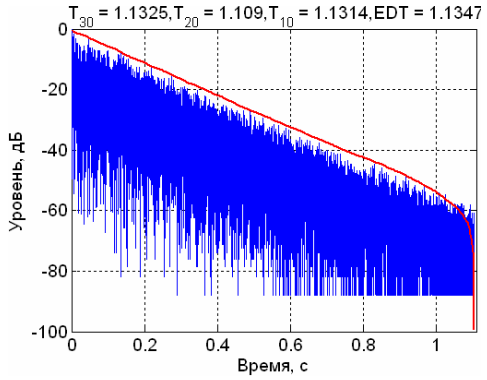


Рис. 2.8. ИХ помещения, T_{60}=1,1 с

Система подавления реверберации моделировалась в среде Matlab, с использованием модифицированной, в соответствии с рис. 4.4 и соотношениями (1.5)-(1.10), программы-функции ssubmmse.m, заимствованной из программного инструментария VoiceBox [93 Brooks]. При

этом фреймы речевого сигнала, протяженностью 32 мс, взвешивались окном Хэмминга и перекрывались на 50%.

Форма искаженного реверберацией и восстановленного сигналов показана на рис. 2.9, а соответствующие спектрограммы показаны на рис. 2.10. На слух искаженный реверберацией сигнал звучит весьма гулко, тогда как в восстановленном сигнале эта гулкость в значительной степени устранена, т.е. очевиден эффект подавления реверберации. Вместе с тем, при $T_l = 48\ мс$ на слух заметны небольшие искажения речевого сигнала, привнесенные процедурой дереверберации. Увеличение T_l до 100 мс привело к некоторому улучшению качества звучания восстановленного сигнала, что подтверждает закономерность постановки вопроса о поиске оптимального значения параметра T_l.

Рис. 2.9. Форма искаженных (а) и восстановленных (б) сигналов

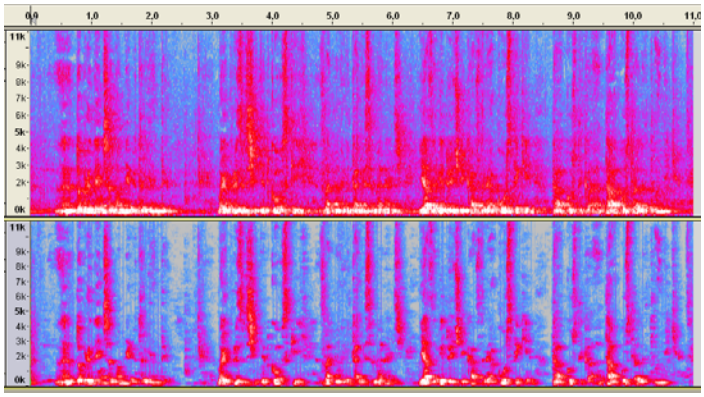


Рис. 2.10. Спектрограммы искаженных (а) и восстановленных (б) сигналов

Количественная оценка состояла в вычислении упомянутых выше показателей качества деревербераций. При этом реверберированные сигналы моделировались путем свертки чистого (записанного при отношении сигнал-шум примерно 40 дБ) речевого сигнала с ИХ трех помещений с временами реверберации 0,74 с, 0,89 с и 1,1 с. Оценивание показателей PESQ, SRR, LSD, BSD производилось в среде Matlab, при этом программа вычислений PESQ была заимствована из [88].

Для моделирования системы АРР, а также для оценки показателя $Acc\%$, применялся программный инструментарий НТК [89]. Описание процедуры обучение системы АРР приведено в п. 2.1.3.

Проведенные количественные эксперименты можно разделить на две группы. Эксперименты первой группы носили предварительный характер, а эксперименты второй группы были уточняющими.

Рассмотрим сначала результаты первой группы экспериментов, главной целью которых было выяснение существования оптимального, в смысле максимума критериев $Acc\%$ и PESQ, значения параметра T_l. Кроме того, безусловный интерес представляли интерес такие характеристики как степень разрушения сигнала из-за влияния реверберации и степень восстановления сигнала системой деревербераций.

В табл. 2.2 представлены результаты оценивания точности распознавания $Acc\%$ и качества речи PESQ в отсутствие реверберации и при ее наличии.

Таблица 2.2

Вид сигнала	T_{60}**, с**	$Acc\%$	**PESQ**
Чистый сигнал	0	92,59	4,5
Сигнал, искаженный реверберацией	0,74	22,22	2,281
	0,89	22,22	2,073
	1,10	29,63	2,030

Как следует из табл. 2.2, влияние реверберации существенно сказывается как на точности распознавания речи (*Acc*% снижается с 93% до 22…30 %), так и на качестве речевого сигнала (PESQ снижается с 4,5 до 2,03…2,28).

В табл. 2.3 и на рис. 2.11-2.12 представлены результаты оценивания *Acc*% и PESQ в отсутствие коррекции речевого сигнала, а также при его коррекции по способам 1 и 2.

Очевидно, коррекция по способу 1 (применение «традиционного» алгоритма logMMSE, предназначенного для подавления шумовой помехи) не привела к положительным результатам. Между тем, коррекция по способу 2 (применение модернизированного алгоритма logMMSE) позволила существенно повысить значения показателя *Acc*%. Интересно, что показатель PESQ при этом повысился не столь значительно.

Таблица 2.3

T_{60}, с	*Acc*%			PESQ		
	нет коррекции	есть коррекция (способ 1)	есть коррекция (способ 2)	нет коррекции	есть коррекция (способ 1)	есть коррекция (способ 2)
0,74 с	22,22	18,52	74,1	2,281	2,252	2,33
0,89 с	22,22	14,81	55,6	2,073	2,059	2,08
1,1 с	29,63	29,63	62,3	2,03	2,037	2,23

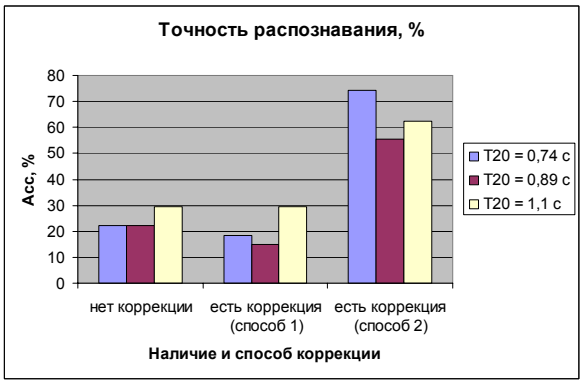


Рис. 2.11. Точность распознавания в отсутствие и при наличии коррекции

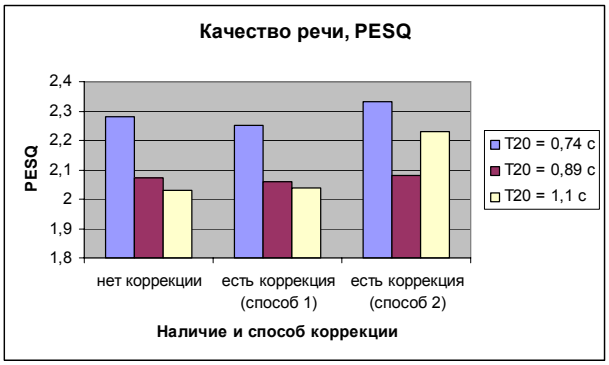


Рис. 2.12. Качество речи в отсутствие и при наличии коррекции

Результаты экспериментальных исследований зависимостей $Acc\%(T_l)$ и $\mathrm{PESQ}(T_l)$ представлены в табл. 2.4 и рис. 2.13-2.14.

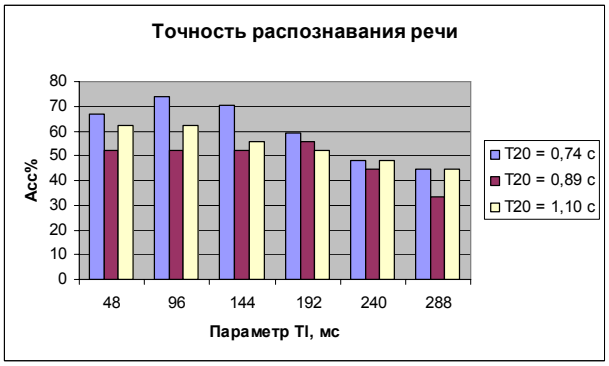


Рис. 2.13. Зависимость $Acc\%(T_l)$

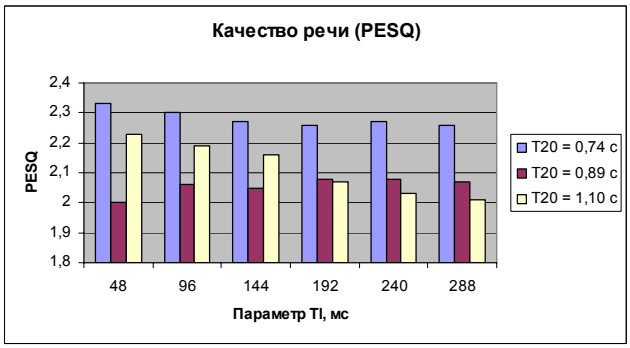


Рис. 2.14. Зависимость $\mathrm{PESQ}(T_l)$

Таблица 2.4. $Acc\%$ и PESQ для различных T_l

T_{60}, с	T_l, мс	$Acc\%$	PESQ
0.74	48	66.7	2.33
	96	74.1	2.30
	144	70.4	2.27
	192	59.3	2.26
	240	48.2	2.27
	288	44.4	2.26
0.89	48	51.9	2.00
	96	51.9	2.06
	144	51.9	2.05
	192	55.6	2.08
	240	44.4	2.08
	288	33.3	2.07
1.10	48	62.3	2.23
	96	62.3	2.19
	144	55.6	2.16
	192	51.9	2.07
	240	48.2	2.03
	288	44.4	2.01

Из представленных результатов следует, что оптимальное, в смысле максимума $Acc\%$, значение T_l находится в интервале 100…200 мс. Более неопределенной является ситуация с зависимостью $\mathrm{PESQ}(T_l)$. В двух из трех случаев качество речи понижается с ростом T_l, и лишь в одном случае наблюдается слабый максимум для $T_l \approx 200...240$ мс.

Рассмотрим теперь результаты серии уточняющих экспериментов, при постановке которых ставились следующие цели:

– уточнение оптимального, в смысле максимума $Acc\%$, значения параметра T_l;

– выяснение необходимости зависимости параметра усреднения η_z^d от переменной k, т.е. от частоты (см. соотношения (1.7)-(1.10));

– установление оптимальных, в смысле максимума $Acc\%$ и PESQ, значений параметров η_z^d и η_z^a;

– оценка максимально достижимых значений показателей $Acc\%$ и PESQ для дереверберированных сигналов;

– анализ поведения характеристик SRR, LSD, BSD в зависимости от значений параметров η_z^d и T_l.

Для более детального анализа зависимостей показателей $Acc\%$, PESQ, SRR, LSD и BSD от параметра T_l шаг изменения этого параметра был уменьшен с 48 до 16 мс. Приводимые ниже на графиках значения всех указанных показателей представляют собой результат усреднения по трем ситуациям, отличающихся временем реверберации: T_{60}=0,74 с, 0,89 с и 1,1 с.

На рис. 2.15 приведены графики зависимости $Acc\%(T_l)$ для различных значений параметра усреднения η_z^d, не зависящего от переменной k (при этом было установлено $\eta_z^a = 0{,}5\eta_z^d$). Как следует из рис. 2.15, максимальное значение $Acc\%(T_l) \approx 67\ldots69\%$ достигается при $\eta_z^d \approx 0{,}67\ldots0{,}75$ для $T_l \approx 100$ мс.

На рис. 2.16 сопоставлены зависимости $Acc\%(T_l)$, вычисленные при $\eta_z^d \approx 0{,}67\ldots0{,}75$, с зависимостью $Acc\%(T_l)$, вычисленной для η_z^d в соответствии с соотношениями (1.8)-(1.9). Как видим, задавая параметр η_z^d зависящим от времени реверберации и частоты, как того требуют соотношения (1.8)-(1.9), получаем максимальное значение $Acc\%(T_l) \approx 64\%$, что хуже на 5…7%, чем для $\eta_z^d \approx 0{,}67\ldots0{,}75$.

Зависимости $Acc\%(T_l)$, вычисленные при $\eta_z^d = 0{,}5$ для различных значений $\eta_z^a = k\,\eta_z^d$ $(k = 0{,}1;\ 0{,}5;\ 0{,}9)$ и представленные на рис. 2.17,

свидетельствуют о практической независимости результатов от параметра η_z^a.

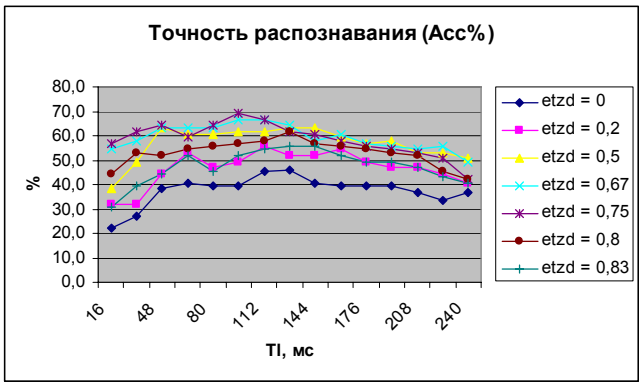


Рис. 2.15. $Acc\%(T_l)$ для различных значений η_z^d

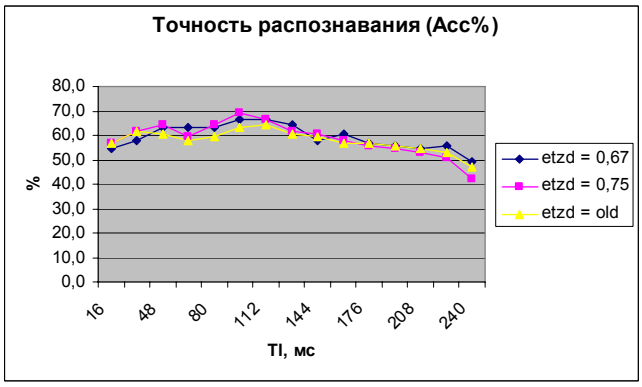


Рис. 2.16. $Acc\%(T_l)$ для η_z^d, определенных различными способами

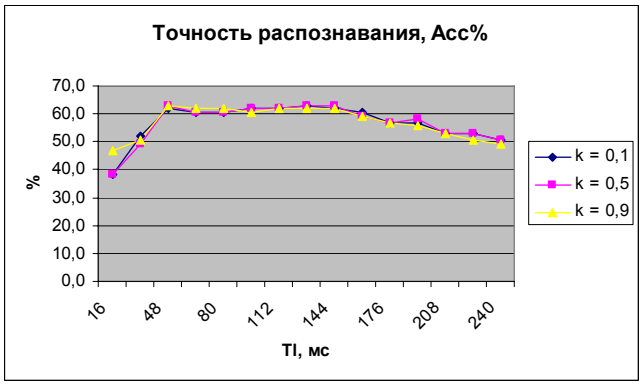


Рис. 2.17. $Acc\%(T_l)$ для различных значений $\eta_z^a = k\,\eta_z^d$

Более сложным, как следует из рис. 2.18, является поведение показателя $PESQ(T_l)$. Максимальное качество речи PESQ $\approx$ 2,3 получаем при условии $\eta_z^d \approx 0{,}5\ldots0{,}66$, однако зависимость $PESQ(T_l)$ при этом имеет монотонно ниспадающий характер, т.е. не содержит экстремума.

Экстремумы зависимости $PESQ(T_l)$ имеют место лишь при $\eta_z^d \approx 0\ldots0{,}2$, и соответствуют значениям $T_l \approx 50\ldots80$ мс, однако качество речи при этом ниже максимально достижимого: PESQ≈2,22-2,25.

Результаты вычисления зависимости $PESQ(T_l)$ для разных значений $\eta_z^a = k\,\eta_z^d$, показанных на рис. 2.19, свидетельствуют о практической независимости результатов от выбора параметра η_z^a или, иными словами, от согласования интервала скользящего усреднения периодограмм реверберированного сигнала с формой огибающей этого сигнала.

Результаты экспериментального оценивания усредненных по параметру T_{60} зависимостей $SRR(T_l)$, $LSD(T_l)$ и $BSD(T_l)$ для различных значений η_z^d, приведены на рис. 2.20-2.22, соответственно. Отсутствие экстремумов в зависимостях $SRR(T_l)$, $LSD(T_l)$ от параметра η_z^d не позволяет использовать их для выбора оптимального значения данного параметра. Вместе с тем, зависимости $BSD(T_l)$, как и $PESQ(T_l)$, достигают экстремальных значений при $\eta_z^d \approx 0{,}5$.

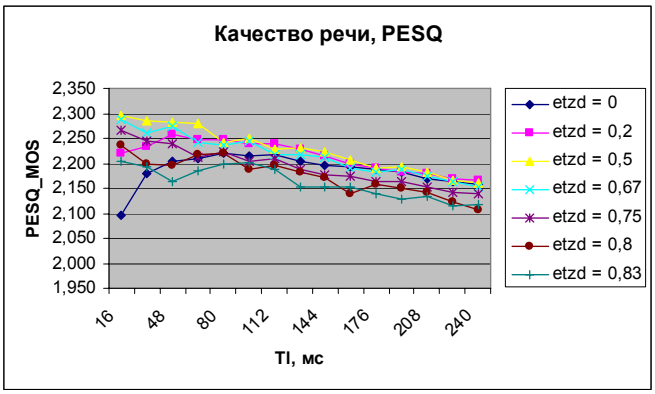


Рис. 2.18. Качество речи для различных значений η_z^d

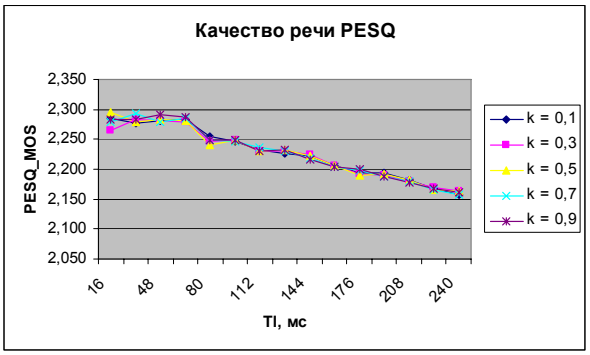


Рис. 2.19. Качество речи для различных значений $\eta_z^a = k\,\eta_z^d$

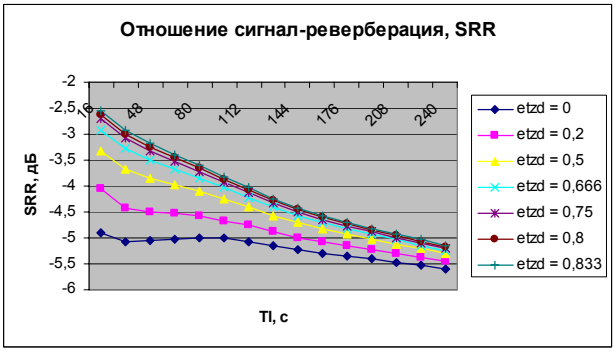


Рис. 2.20

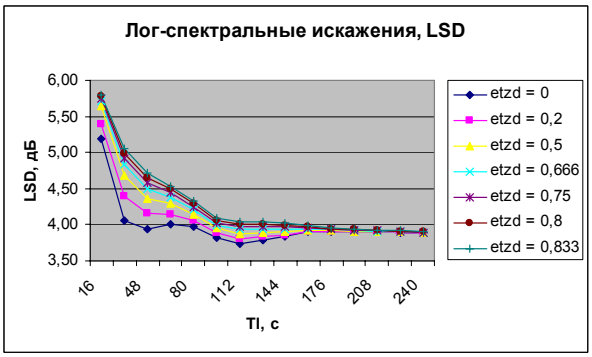


Рис. 2.21

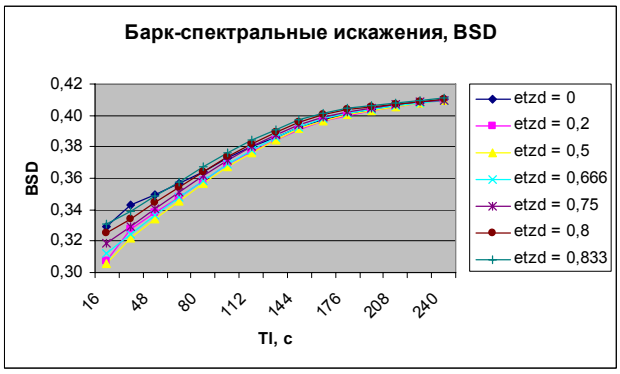


Рис. 2.22

Полученные результаты позволяют предложить эмпирическое соотношение для определения оптимальных значений параметра η_z, в зависимости от значений протяженности фреймов $T_{фр}$ и величины их сдвига $T_{сдв}$ (рис. 2.23):

$$\eta_z = \frac{T_{уср} - T_{фр}}{T_{уср} - T_{фр} + T_{сдв}}, \qquad (2.14)$$

Исходя из найденных в данной работе оптимальных, в смысле максимума $Acc\%(\eta_z)$, значений $\eta_z^d \approx 0{,}67\ldots0{,}75$, а также учитывая использовавшиеся при этом значения параметров $T_{фр}$ и $T_{сдв}$, заключаем, что в соотношении (2.14) следует принимать $T_{уср} = 60\ldots80$ мс. Объяснить данный результат можно, если сопоставить $T_{уср} = 60\ldots80$ мс с такой характеристикой речевого сигнала как протяженность фонемы, колеблющуюся в интервале 30-210 мс со средним значением 135 мс.

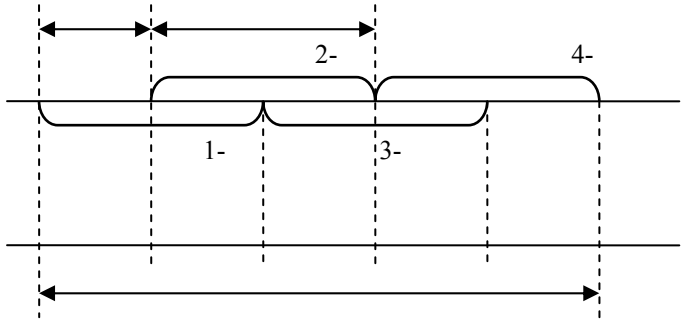


Рис. 2.23. Область усреднения и фреймы в ее пределах

Таким образом, ответы на поставленные в начале п. 2.1.4 вопросы таковы:

1) оптимальным, в смысле максимума $Acc\%$, является значение параметра $T_l \approx 100$ мс;

2) оптимальным, в смысле максимума $Acc\%$, является значение параметра сглаживания $\eta_z = \frac{T_{уср} - T_{фр}}{T_{уср} - T_{фр} + T_{сдв}}$, где $T_{уср} = 60\ldots80$ мс;

3) при измерениях спектра $\lambda_y(l,k)$ реверберированного сигнала, для времен реверберации $T_{60} = 0{,}7\ldots1{,}1$ с (типичных для офисных помещений, лабораторий и аудиторий), простое в вычислительном плане соотношение

(1.6) оказалось предпочтительнее, по точности вычислений, соотношений (1.7)-(1.10).

Кроме того, получены оценки потенциальных возможностей деревербератора, реализующего метод частотной коррекции в виде алгоритма logMMSE. В проведенных экспериментах точность распознавания Acc% повысилась с 22…30% до 56…75% при условии $T_l \approx 100$ мс. Показатель качества речи PESQ удалось повысить, в среднем, с 2,13 до 2,21 при условии $T_l \approx 16$ мс.

Показано также, что отсутствие экстремумов у зависимостей $SRR(\eta_z)$ и $LSD(\eta_z)$ внутри диапазона $0 \le \eta_z \le 0{,}83$ свидетельствует о непригодности применения соответствующих критериев для выбора оптимальных значений параметра усреднения η_z.

2.2. Влияние погрешностей слепого измерения времени реверберации на качество реверберации

Обычно время реверберации T_{60} оценивают «напрямую», используя импульсную характеристику $h(t)$ канала передачи [53]. Если импульсная характеристика неизвестна, приходится прибегать к так называемым «слепым» измерениям, в соответствии с которыми информацию о времени реверберации T_{60} извлекают из реверберированного речевого сигнала [54, 55].

Результаты исследований свидетельствуют, что слепые измерения T_{60} менее точны, нежели прямые измерения. Другим недостатком слепых измерений является приближенный характер оценивания частотной зависимости параметра T_{60} [56, 57]. Оба упомянутых фактора негативно сказываются на качестве функционирования системы АРР, снижая точность распознавания речи. К сожалению, степень такого снижения, как и меры противодействия такому снижению, до последнего времени оставалась

неизвестной, что в значительной степени поясняется как сложностью теоретического аспекта данной проблемы, так и сложностью построения соответствующих экспериментальных программных комплексов. В данном подразделе указанные трудности преодолены путем проведения экспериментальных исследований с использованием программного комплекса, построенного на базе современных алгоритмов оценивания спектра поздней реверберации и слепого измерения времени реверберации.

Как было проанализировано в главе 1, существует многообразие методов слепого измерения времени реверберации. В данной работе использован иной метод, впервые предложенный в работе [54] и названный методом максимального правдоподобия (МП). Основанием для такого выбора является хорошая помехоустойчивость метода МП, не уступающего (и даже несколько превосходящего) конкурентным методам в точности измерений [62]. Кроме того, организацию и проведение соответствующих экспериментальных исследований существенно облегчает наличие в открытом доступе программы Matlab для слепого оценивания времени реверберации методом МП [94].

Сущность измерений времени реверберации T_{60} методом МП вкратце состоит в следующем. Информацию о параметре T_{60} извлекают преимущественно в паузах речевого сигнала, где действие реверберации проявляется в виде звуковых «шлейфов», тянущихся за последними звуками слов. Структура этих шлейфов подобна структуре ИХ канала передачи:

$$y(n) = \xi(n)a(n), \qquad a(n) = e^{-\delta n/F_s},$$

где $\xi(n),\ n \geq 0; n = tF_s$ - дискретный гауссовский белый шум с параметрами $[0, \sigma]$.

Если наблюдается N-мерный вектор $\mathbf{y}$, тогда в силу статистической независимости выборок процесса $\xi(n)$, многомерная плотность вероятностей

этого вектора, именуемая также функцией правдоподобия, имеет вид произведения N одномерных гауссовских распределений:

$$L(\mathbf{y};\mathbf{a},\sigma)=\frac{1}{a(0)\cdots a(N-1)}\left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\exp\left(-\frac{\sum_{n=0}^{N-1}(y(n)/a(n))^2}{2\sigma^2}\right) \qquad (2.15)$$

Используя соотношение (2.15), необходимо оценить параметры σ и $\mathbf{a}$, где $\mathbf{a}$ - N-мерный вектор, по имеющимся N значениям вектора $\mathbf{y}$. Учитывая экспоненциальный характер функции $a(n)$:

$$a(n)=a^n, \quad a=\exp(-\delta/F_s), \qquad (2.16)$$

из (2.15) с учетом (2.16) следует

$$L(\mathbf{y};a,\sigma)=\left(\frac{1}{2\pi a^{(N-1)}\sigma^2}\right)^{N/2}\exp\left(-\frac{\sum_{n=0}^{N-1}a^{-2n}y(n)^2}{2\sigma^2}\right) \qquad (2.17)$$

Логарифмируя (2.17), получим

$$\ln L(\mathbf{y};a,\sigma)=-\frac{N(N-1)}{2}\ln(a)-\frac{N}{2}\ln(2\pi\sigma^2)-\frac{\sum_{n=0}^{N-1}a^{-2n}y(n)^2}{2\sigma^2} \qquad (2.18)$$

Приравнивая нулю частные производные от выражения (2.18) по параметрам σ и a, получаем систему уравнений для неизвестных σ и a:

$$\frac{\sum_{n=0}^{N-1}na^{-2n}y(n)^2}{a\sigma^2}-\frac{N(N-1)}{2a}=0, \qquad (2.19)$$

$$\sigma^2=\frac{1}{N}\sum_{n=0}^{N-1}a^{-2n}y(n)^2. \qquad (2.20)$$

Подставляя (2.20) в (2.19), получаем соотношение, определяющее параметр a:

$$\frac{\sum_{n=0}^{N-1} n a^{-2n} y(n)^2}{\sum_{n=0}^{N-1} a^{-2n} y(n)^2} - \frac{(N-1)}{2} = 0. \qquad (2.21)$$

В силу нелинейности уравнения (2.21) алгоритм нахождения неизвестного параметра a должен быть итерационным.

Формируя таким образом оценку a_k^* в λ-м фрейме речевого сигнала, для множества фреймов получаем множество оценок, по которым можно построить гистограмму. Часть оценок хороши, поскольку соответствуют фреймам, захватывающим паузы, а часть оценок неудачны, т.к. захватывают речевой сигнал. Кроме того, различные участки пауз обладают неодинаковой информативностью. Очевидно, предпочтительными являются те участки пауз, которые примыкают к окончаниям слов. Поэтому необходимо решающее правило для выбора информативных участков пауз, а также для выбора хорошей оценки параметра a с использованием гистограммы.

Такое решающее правило может базироваться на следующих рассуждениях. Если внутри текущего фрейма, в пределах двух смежных участков времени (субфреймов) энергия сигнала снижается, тогда принимают решение о наличии участка паузы, на котором замечено действие реверберации. Субфреймы объединяют в так называемый «сегмент данных», по которому вычисляют «текущее» время реверберации T_{60}, решая уравнение (2.21) и используя соотношения

$$T_{60} = 6{,}91/\delta, \qquad \delta = -F_s \ln a. \qquad (2.22)$$

После этого обновляют гистограмму оценок T_{60}, сформированную из последних оценок T_{60}. Оценку $\hat{T}_{60}^{(1)}$, соответствующую максимуму обновленной гистограммы, считают текущей оценкой времени реверберации. Для снижения дисперсии оценки времени реверберации используют процедуру сглаживания:

$$\hat{T}_{60}(\lambda) = \beta(\lambda) \cdot \hat{T}_{60}(\lambda - 1) + (1 - \beta(\lambda)) \cdot \hat{T}_{60}^{(1)}(\lambda), \qquad (2.23)$$

где $0 < \beta(\lambda) < 1$; λ - номер фрейма.

Организация экспериментальных исследований в целом подобна таковой для пп. 2.1.3 и 2.1.4, поэтому укажем лишь на отличительные особенности.

Слепое измерение времени реверберации производилось в среде Matlab с использованием пакета программ ML_RT_estimation, реализующего измерение параметра T_{60} методом МП [94]. Подпрограмма ML_RT_estimation_init.m этого пакета подверглась незначительной модернизации: значение параметра $\beta(\lambda)$, используемого в соотношении (2.23), было снижено с 0,996 до 0,9, а значение начального времени $\hat{T}_{60}^{(1)}$ было повышено 0,3 с до 0,5 с. Такая модернизация позволила ускорить сходимость процедуры оценивания.

Было проведено две группы экспериментов, в каждой из которых сравнивались результаты дереверберации, осуществленные с использованием оценок T_{60}, полученных прямым и слепым способами.

В первой группе экспериментов слепое измерение времени реверберации производилось в предположении, что параметр T_{60} не зависит от частоты. Во второй группе экспериментов при слепом измерении времени реверберации приближенно оценивалась зависимость $T_{60}(f)$.

Результаты первой группы экспериментов таковы. На слух результаты дереверберации, полученные путем измерения T_{60} по ИХ помещения, практически не отличаются от таковых, где T_{60} оценивалось вслепую.

На рис. 2.24 приведены спектрограммы исходного (а) и реверберированного (б) сигналов, а также сигналов, подвергнутых процедуре дереверберации: для T_{60}, измерявшегося по ИХ помещения (в) и для T_{60}, измерявшегося вслепую (г). Хорошо видно, что реверберация действует на сигнал как НЧ фильтр, подавляя высокочастотные составляющие согласных звуков. Вместе с тем видно, что после дереверберации затянутость сигнала по времени действительно уменьшилась, хотя мощность высокочастотных компонентов так и не восстановилась полностью.

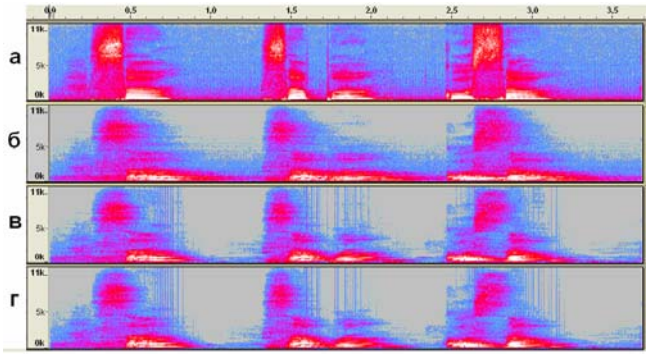


Рис. 2.24 4.28. Спектрограммы речи до и после дереверберации

Результаты слепого оценивания параметра T_{60} приведены в табл. 2.5. Как видим, абсолютная и относительная погрешности слепой оценки составили 0,1-0,3 с и 14%-27%, соответственно.

Результаты оценивания точности распознавания *Acc*% сведены в табл. 2.5 и представлены графиком рис. 2.25. Здесь «способ 1» - попытка коррекции реверберированной речи с помощью традиционного алгоритма подавления шумовой помехи; «способ 2» - коррекция путем подавления поздней реверберации при оценивании T_{60} по ИХ помещения; «способ 3» - коррекция путем подавления поздней реверберации при слепом оценивании T_{60}.

Таблица 2.5

Оценка по ИХ T_{60}, с	Слепая оценка T_{60}, с	Абсолют. погрешн., с	Относит. погрешн., %
0,74	0,65	0,09	13,8
0,89	0,7	0,19	21,3
1,1	0,8	0,3	27,3

Как видим, использование усредненной оценки T_{60}, измеренной вслепую (способ 3), заметно ухудшило качество распознавания для помещений с временем реверберации 0,74 с и 0,89 с. Исключением явилось помещение с временем реверберации 1,1 с. В этом случае точность распознавания не ухудшилась, а даже несколько повысилась. Тем не менее, в среднем слепое измерение T_{60} приводит к снижению *Acc*% примерно на 20%.

Таблица 2.6

T_{60}, с	Точность распознавания, *Acc*%			
	нет корр-ции	корр-ция спос. 1	корр-ция спос. 2	корр-ция спос. 3
0,74	22,22	18,52	74,1	37,04
0,89	22,22	14,81	55,6	37,04
1,1	29,63	29,63	62,3	62,96
Сред.	**24,69**	**21,00**	**64,00**	**45,68**

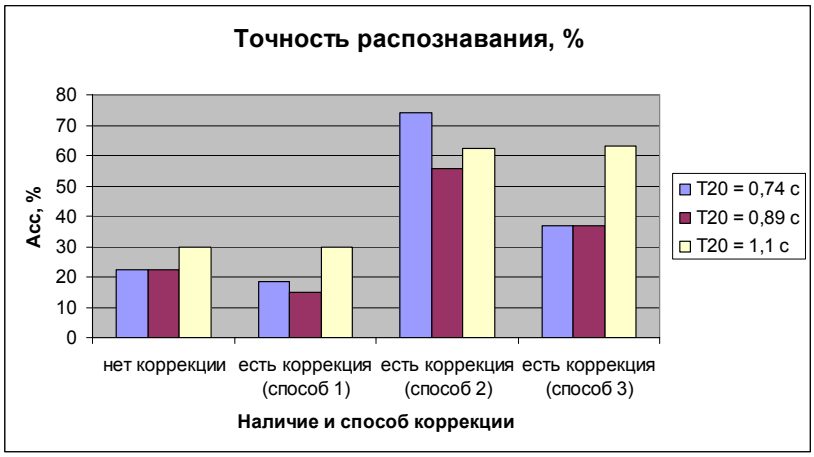


Рис. 2.25. Точность распознавания в отсутствие и при наличии коррекции

Как следует из табл. 2.7 и рис. 2.26, качество речи, характеризуемое показателем PESQ, при слепом оценивании T_{60} также несколько ухудшилось по сравнению с ситуацией, когда T_{60} оценивалось по ИХ помещения. Здесь также не обошлось без исключения – для помещения с временем реверберации 0,89 с качество речи повысилось. Однако в среднем качество речи осталось практически неизменным, что свидетельствует о низкой эффективности показателя PESQ при оценке качества деревербераторов в системах APP.

Таблица 2.7

T_{60}, с	PESQ			
	нет коррекции	корр. способ 1	корр. способ 2	корр. способ 3
0,74	2,281	2,252	2,33	2,308
0,89	2,073	2,059	2,08	2,111
1,1	2,03	2,037	2,23	2,212
Сред.	2,128	2,116	2,213	2,210

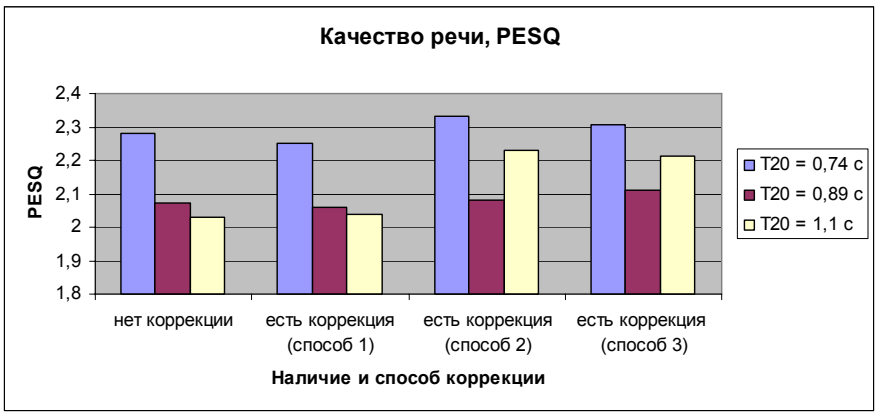


Рис. 2.26. Качество речи (PESQ) в отсутствие и при наличии коррекции

Во второй группе экспериментов при слепом измерении времени реверберации приближенно оценивалась зависимость $T_{60}(f)$. С этой целью реверберированный речевой сигнал подвергался фильтрации гребенкой фильтров. Время реверберации оценивалось вслепую по речевому сигналу на выходе каждого из фильтров, затем по полученным результатам, путем интерполяции и экстраполяции, формировалась зависимость $T_{60}(f)$. На рис. 2.27 результаты прямого оценивания $T_{60}(f)$ по спектрограмме имеющейся ИХ показаны синей линией, а красной линией представлены результаты слепого оценивания $T_{60}(f)$ для трех разновидностей гребенок фильтров.

Результаты рис. 2.27,а соответствуют гребенке из семи октавных фильтров с центральными частотами $f_0 = 125, \ldots, 8000$ Гц, рис. 2.27,б – гребенке из трех октавных фильтров с центральными частотами 2 кГц, 4 кГц и 8 кГц, рис. 4.31,в - гребенке из двух октавных фильтров с центральными частотами 2 и 8 кГц. Практически те же результаты получаются при использовании гребенок третьоктавных фильтров.

Результаты оценивания точности распознавания *Acc*% во второй группе экспериментов сведены в табл. 4.8 и представлены графиком рис. 2.28. Нетрудно видеть, что гребенка из двух фильтров обеспечивает вполне приемлемое качество слепого оценивания зависимости $T_{60}(f)$. При этом учет зависимости времени реверберации от частоты позволил повысить

точность распознавания речи в среднем на 15-17% по сравнению со случаем, когда такой учет не производился.

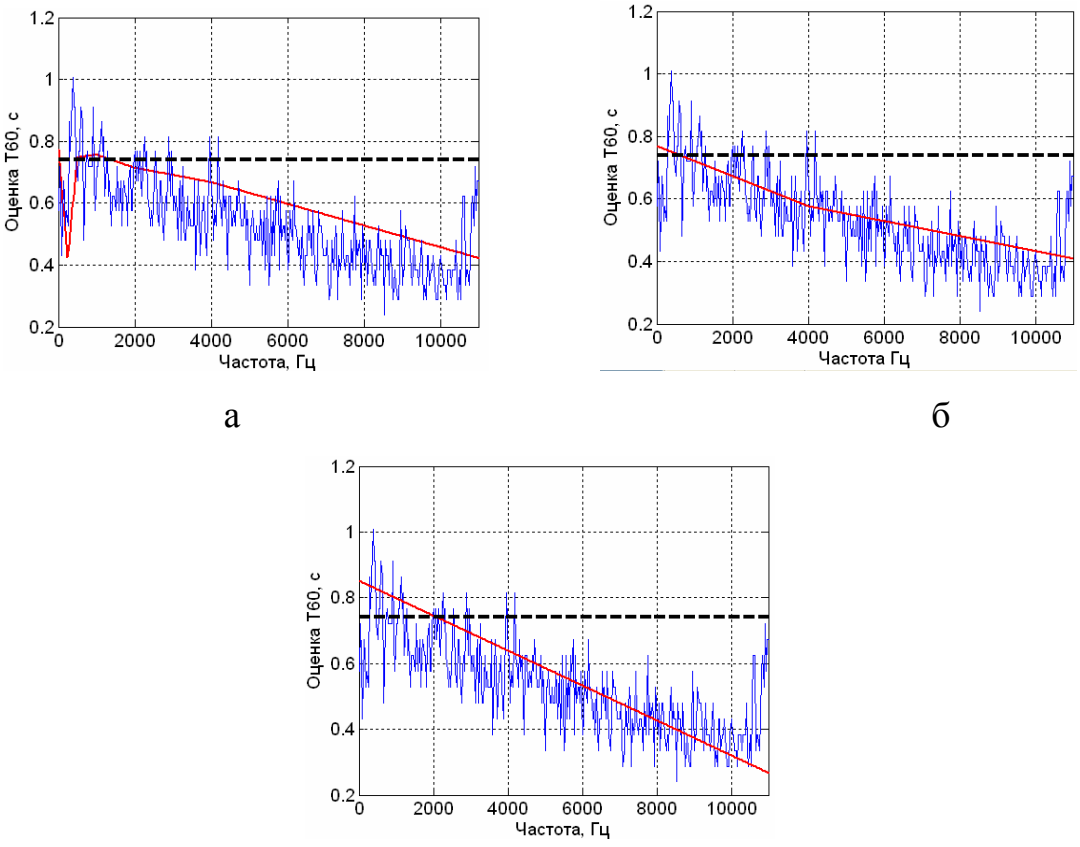


Рис. 2.27. Зависимости $T_{60}(f)$ для гребенки из семи (а), трех (б) и двух (в) октавных фильтров

Таблица 4.8

T_{60}, с	Точность распознавания, Acc%				
	Нет коррекции	T60(f) =const	T60(f), 7 фильтров	T60(f), 3 фильтра	T60(f), 2 фильтра
0,74	22,22	37,04	70,37	70,37	59,26
0,89	22,22	37,04	51,85	51,85	66,67
1,1	29,63	62,96	55,56	59,26	59,26
Сред.	**24,69**	**45,68**	**59,26**	**60,49**	**61,73**

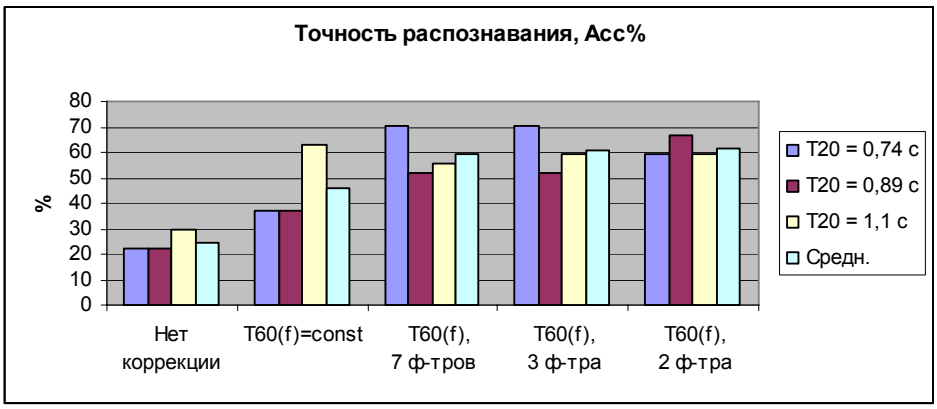


Рис. 2.28. Точность распознавания в отсутствие и при наличии коррекции

Сравнивая между собой полученные результаты, заключаем, что при слепом измерении времени реверберации можно достичь точности распознавания речи, всего лишь на 3-5% уступающей таковой для прямого измерения времени реверберации, если при слепом измерении учитывать зависимость времени реверберации T_{60} от частоты. Достаточная для приложений точность измерений зависимости $T_{60}(f)$ может быть достигнута путем линейной аппроксимации оценок времен реверберации T_{60} для процессов на выходе двух октавных или третьоктавных фильтров с центральными частотами 2 кГц и 8 кГц.

2.4. Выводы

1. Впервые показано, что интервал времени скользящего усреднения периодограмм реверберированного сигнала, производимого при оценивании спектра реверберированного сигнала, должен быть больше интервала стационарности и меньше средней длительности фонемы. Для среднего темпа речи это интервал времени 50-80 мс. Данный вывод позволяет достичь максимальных результатов как в плане качества дереверберированного речевого сигнала, так и в плане точности его распознавания в системе АРР.

2. Впервые показано, что, при оценивании спектра реверберированного сигнала путем скользящего усреднения периодограмм реверберированного сигнала, интервал усреднения нецелесообразно делать зависимым от частоты и от формы огибающей речевого сигнала. Благодаря отказу от такой зависимости, точность функционирования системы АРР удается повысить на 5-7%, а объем вычислений - существенно сократить.

3. Выработаны рекомендации по оптимизации параметра T_l, характеризующего момент начала поздней реверберации, при оценивании спектра поздней реверберации. Показано, что в смысле максимальной точности АРР оптимальным является значение $T_l \approx 100$ мс, тогда как в смысле максимального качества речевого сигнала оптимальным является значение $T_l \approx 16$ мс.

4. Получены оценки потенциальных возможностей деревербератора, реализующего метод частотной коррекции на основе алгоритма logMMSE. Показано, что при условии $T_l \approx 100$ мс точность распознавания Acc% может быть повышена на 30-45%, а качество речевого сигнала, характеризуемой показателем PESQ, может быть повышено примерно на 0,1 при условии $T_l \approx 16$ мс.

5. Показано, что при слепом измерении времени реверберации T_{60} можно достичь точности распознавания речи, всего лишь на 3-5% уступающей таковой для прямого измерения времени реверберации, если при слепом измерении учитывать зависимость времени реверберации от частоты. Достаточная для приложений точность измерений зависимости $T_{60}(f)$ может быть достигнута путем линейной аппроксимации оценок времен реверберации T_{60} для процессов на выходе двух октавных или третьоктавных фильтров с центральными частотами 2 кГц и 8 кГц.

6. Показано, что при использовании деревербератора как препроцессора системы АРР, сквозной показатель качества в виде точности распознавания речи Acc% нецелесообразно заменять более простыми

показателями SRR, LSD, BSD и PESQ, характеризующими форму речевого сигнала. Причиной тому является неудовлетворительная согласованность этих показателей, рассматриваемых как функции параметров системы деревербераций.

-

ГЛАВА 3
ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ЦЕЛЕСООБРАЗНОСТИ ИСПОЛЬЗОВАНИЯ ПАРАМЕТРИЗАЦИИ PNCC

3.1. Оптимизация параметров алгоритмов параметризации речевого сигнала в системах АРР

Поскольку в данной работе блок параметризации речевого сигнала рассматривается как составная часть (модуль извлечения признаков) системы АРР (рис. 2.10), с помощью аналитических исследований, к сожалению, невозможно ответить на поставленный вопрос.

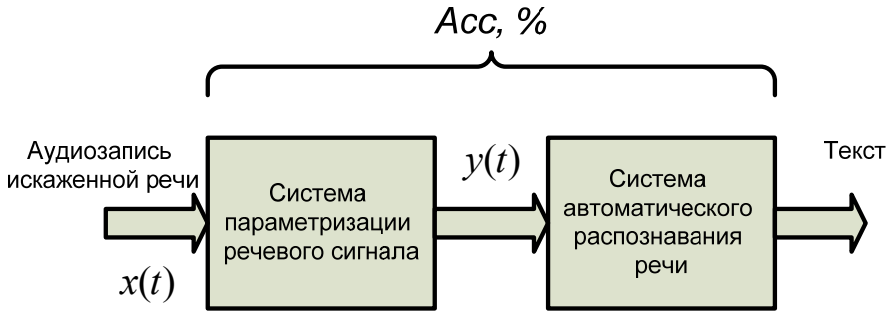


Рис. 3.1. Модуль извлечения параметризации речевого сигнала как модуль предобработки системы АРР

Следствием такого подхода является то, что для адекватной оценки качества различных видов параметризации необходимо использовать сквозной показатель $Acc\%$, характеризующий точность распознавания речи [8]:

$$Acc\% = \frac{N - D - S - I}{N} \times 100\%$$

где N - общее количество распознаваемых единиц речевого сигнала; D - количество ошибочно удаленных единиц речевого сигнала; S - количество замененных единиц речевого сигнала; I - количество ошибочно вставленных

единиц речевого сигнала; $WER\% = (1 - Acc) \times 100\%$ – ошибка правильного распознавания единиц речевого сигнала.

Оптимизация параметров алгоритмов параметризации в задаче АРР описывается для моделей сигналов на выходе телефонного канала связи, которые представлены в виде корпусов TIMIT, NTIMIT и STC-TIMIT.

TIMIT – это классический речевой корпус, содержащий свыше 5 часов цифровых звукозаписей различных английских фраз, произнесённых 630 дикторами (мужчинами и женщинами) на 8 диалектах американского английского. Все звукозаписи имеют временную пофонемную разметку, выполненную профессиональными фонетистами. Речевой корпус разбит на два непересекающихся множества: обучающее и тестовое [25Zue]. Соотношение среднего SNR для TIMIT, вычисленного как соотношение максимальной энергии сигнала к минимальному пороговому значению энергии для предложения, составляет около 40 дБ [26Zhimin] -53 дБ [27Reynolds].

Речевые корпуса NTIMIT и STC-TIMIT построены на основе корпуса TIMIT. Звукозаписи речевого корпуса TIMIT были пропущены через телефонные каналы американской телефонной компании NYNEX и заново оцифрованы. Это позволило представить в речевом корпусе NTIMIT звукозаписи с нестационарным случайным шумом, характерным для естественного телефонного канала связи искажения включая влияние характеристик, внесенных телефонными трубками, которые представленные в NTIMIT [28Jankowski]. Корпус STC-TIMIT[29Morales] получен путем пропускания TIMIT через один телефонный канал связи, при этом сигнал подавался напрямую в коммутатор без использования телефонной трубки.

Базы STC-TIMIT и NTIMIT являются аналогами речевого корпуса TIMIT и имеют похожие частоты среза и средний SNR [29Morales] в диапазоне от 36 дБ [27Reynolds] до 25 дБ [26Zhimin], с пониженной пропускной способностью 300-3400Гц.

Базы TIMIT и NTIMIT имеет частоту дискретизации 16кГц и эффективную полосу пропускания 6,4кГц [2]. Корпус STC-TIMIT имеет два варианта записей с частотой дискретизации 16кГц и 8Гц и эффективную полосу пропускания 6,4кГц и 3,4кГц соответственно. Таким образом, корпус STC-TIMIT при 8кГц представляет реальный сигнал в телефонном канале связи с пониженной частотой дискретизации.

Таким образом, исследования проводились путем компьютерного моделирования, а результаты экспериментов оценивались как качественно (на слух и путем визуального анализа спектрограмм), так и количественно, путем оценивания показателей $Acc\%$ и $WER\%$.

3.1.1. Выявление целесообразности использования параметризации речевого сигнала нормализованными по мощности кепстральными коэффициентами.

В работе [20Kim] получены результаты распознавания фонем для метода (рис.2.16) PNCC, ETSI AFE, MFCC-VTS, MFCC, RASTA-PLP в условиях искажения речевого сигнала белым шумом при различных SNR. Применение метода PNCC дало существенное понижение ошибки распознавания (фонем) $Acc\%$ по отношению к традиционным методам (MFCC, RASTA-PLP) (рис.2.16).

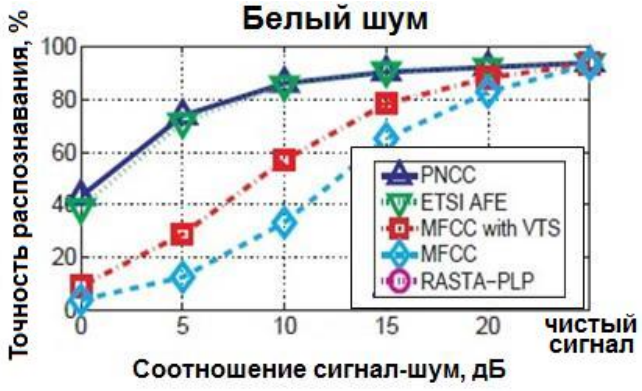


Рис. 2.16. Результаты распознавания Kim

Однако в работе [20, 21Kim] в качестве результатов распознавания с использованием параметризации речевого сигнала методом PNCC

представлены только графики без точных значений, кроме того не указаны параметры системы АРР, используемые в этих экспериментах. В связи с этим невозможно проверить свои выводы без непосредственного моделирования собственной системы АРР на основе параметризации PNCC.

К сожалению, чтобы проверить эффективность алгоритма PNCC по сравнению с традиционными методами параметризации MFCC и PLP для речевых сигналов, искаженных телефонным каналом связи, невозможно аналитическим путем, поскольку алгоритм параметризации (PNCC, MFCC либо PLP) рассматривается как составная часть системы АРР (рис.2.3, 2.10). Следствием такого подхода является то, что для адекватного сравнения таких алгоритмов необходимо использоваться сквозной и общий для них показатель оценки качества $Acc\%$, характеризующий точность распознавания речи (п. 2.2.1).

Таким образом, исследования проводились путем компьютерного моделирования систем АРР и проведения серии экспериментов. Согласно первым двум вариантам точность распознавания фонем при использовании методов MFCC и PLP оценивалась экспериментально при моделировании системы АРР при помощи средств разработки НТК систем АРР [15HTK,30Script].

Третий вариант распознавания при использовании метода PNCC потребовал моделирования системы АРР, в которой данный алгоритм реализован и внедрен сторонними разработчиками [31Shmyrev], при помощи средств разработки Sphinx [14Lee, 32CMU]. Сравнение результатов работы по методу PNCC проводилось с параметризацией MFCC, поскольку это единственный метод, реализованный в Sphinx [32CMU]. Для проведения экспериментов по методу PNCC необходимо определить его оптимальные определить его базовые параметры, а именно:

1. коэффициенты забывания λ_a и λ_b блока ассиметричной фильтрации;

2. коэффициент забывания λ_t и коэффициент подавления μ_t для блока временного маскирования.

Рассмотрим сначала результаты серии первых двух экспериментов над традиционными методами параметризации MFCC и PLP по схемам рис. 2.4-2.5. Главной целью, которых было выяснения влияния помех телефонного канала связи на точность распознавания $Acc\%$ в системе АРР при двух традиционных методах параметризации MFCC и PLP. Моделирование системы АРР и оценка показателя $Acc\%$ осуществлялась по схеме рис. 2.17.

Для моделирования системы АРР использовались программные средства разработки современных систем АРР HTK [13HTK, 30 Cantab] и Sphinx [14Lee, 15CMU, 31Shmyrev]. Модель параметрического представления сигнала представляла собой модуль вычисления MFCC и PLP коэффициентов (рис. 2.17 выделен серым цветом).

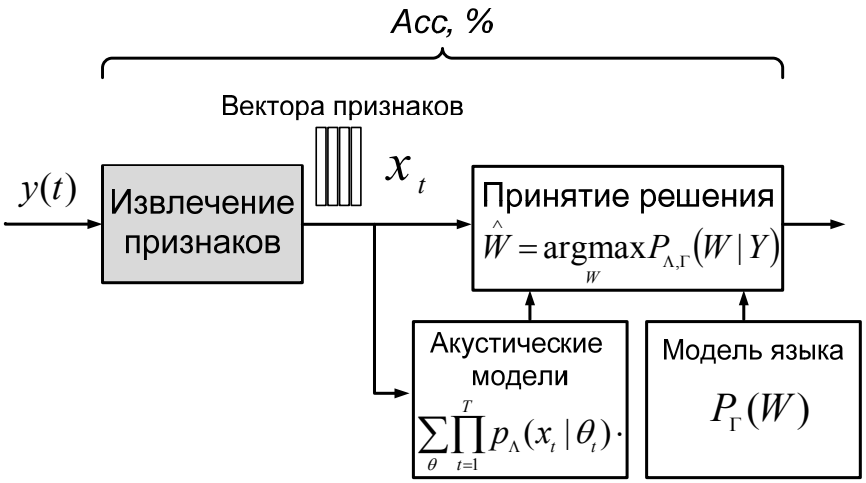


Рис. 2.17. Схема моделируемой системы АРР

Для построения акустических моделей использовались лево-правые HMM модели, состоящие из 3-х состояний без пропуска с непрерывными гауссовыми смесями. В качестве языковых моделей использовались биграммы. Элементарной единицей распознавания были выбраны монофоны (элементарная единица распознавания – одна фонема языка) и трифоны (составная модель распознавания, образованная монофонами слева и справа от текущего монофона).

Анализ речевого сигнала проводиться с помощью окна Хэмминга длительностью 25 мс, с шагом анализа 10 мс. Вектор признаков извлекался из окна анализа на основе MFCC и PLP параметризации. Количество треугольных окон для проведения анализа на нелинейной мел-шкале частот равно 26. Вычислялись 12 кепстральных коэффициентов, дополненные логарифмом энергии.

С целью учета изменения параметров во времени коэффициенты MFCC кепстра, и логарифм энергии были дополнены первой (префикс _D) и второй производными (префикс _A) [13HTK]. К PLP коэффициентам, вместо дополнения логарифмом энергии, к вектору параметров добавлялся нулевой кепстральный коэффициент (префикс _0). Путем добавления префикса _Z, проводилась нормализация кепстрального среднего (CMN – cepstral mean normalization) , позволяющая устранить различные эффекты, связанные с искажениями частотных характеристик записывающих устройств или каналов передачи, путём вычитания среднего значения кепстральных коэффициентов, вычисленных за длительный интервал, из последовательности кепстральных коэффициентов каждого кадра анализа [13HTK].

Тестирование системы APP проводилось на базовом множестве для тестирования таких систем Core Test Set [30Cantab] для исследуемых баз (п. 2.2.2). Данное подмножество Core Test Set содержало 192 предложения (24 диктора 2 мужчин и 1 женщина). Специфические настройки параметров системы APP на основе HTK при использовании методов параметризации речевого сигнала MFCC и PLP приведены в приложении А.

Результаты экспериментальных исследований метода MFCC в виде значений автоматического распознавания $Acc\%$ при параметризации сигнала для различных частот дискретизации F_d для различного числа гауссовых смесей приведены в табл. 2.1-2.2. На графике рис.2.18 изображены усредненные результаты для данных из табл. 2.1-2.2. Как следует из табл. 2.1-2.2 и сводной диаграммы результатов рис.2.18 в случае влияния

телефонного канала связи точность распознавания речи падает для всех корпусов искаженной речи (NTIMIT16, STC8, STC16) при MFCC_E_D_A_Z параметризации сигнала:

– для NTIMIT-NTIMIT $Acc\%$ снижается в среднем с 59,9-63,1% (TIMIT) до 43,9-47,3%;

– для STC-STC с $F_d = 8$ кГц $Acc\%$ снижается в среднем с 59,9-63,1% (TIMIT) до 30,4-29,2;

– для STC-STC с $F_d = 16$ кГц $Acc\%$ снижается в среднем с 59,9-63,1% (TIMIT) до 41,5-47%;

– Снижение частоты дискретизации с $F_d = 16$ до 8кГц снижает точность распознавания $Acc\%$ с 41,5-47% до 30,4-29,2% при одинаковых условиях обучения и тестирования моделей STC-STC (рис.2.18).

Таблица 2.1. Точность распознавания монофонов $Acc\%$. MFCC_E_D_A_Z

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
TIMIT-TIMIT	16	55,6	58,0	59,3	59,9	60,9	61,0	61,6	61,4	61,8
NTIMIT-NTIMIT	16	38,6	41,2	42,3	44,1	45,2	45,5	45,8	46,1	46,4
TIMIT-NTIMIT	16	21,5	22,2	22,7	22,8	22,9	23,6	24,2	24,3	24,7
STC-STC	8	29,1	29,3	29,8	30,2	30,4	30,8	31,2	31,2	31,3
STC-STC	16	36,9	38,6	40,3	41,9	42,2	43,0	43,2	43,7	44,1

Таблица 2.2. Точность распознавания трифонов. MFCC_E_D_A_Z

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
TIMIT-TIMIT	16	62,9	63,8	64,0	63,8	63,7	63,1	62,7	62,0	62,1
NTIMIT-NTIMIT	16	45,8	47,4	47,7	47,7	47,7	47,6	47,8	47,3	46,6
TIMIT-NTIMIT	16	25,8	25,5	24,5	24,4	23,9	24,5	23,5	23,4	23,1
STC-STC	8	28,8	29,8	30,5	30,5	29,1	29,0	28,7	28,5	27,6
STC-STC	16	44,6	45,9	46,8	47,3	48,1	47,7	47,7	47,8	47,1

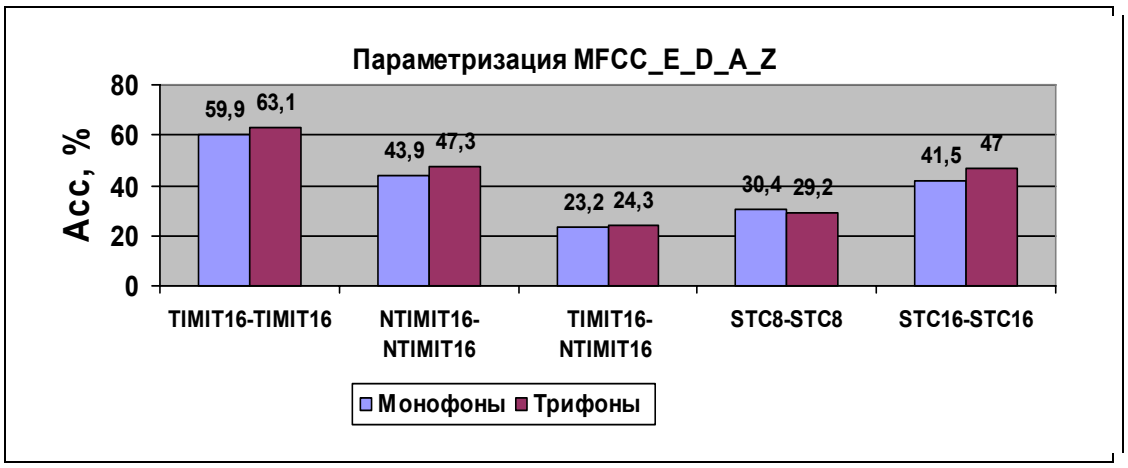


Рис. 2.18. Точность распознавания $Acc\%$. MFCC_E_D_A_Z

В условиях, типичных для эксплуатации системы АРР, когда модели обучаются на неискаженной речи TIMIT, а тестируются в условиях влияния телефонного канала связи NTIMIT. При этом точность распознавания $Acc\%$ в среднем из табл. 2.1-2.2 и рис.2.18 составила 23,2-24,3%, что в 2,5 раза хуже случая отсутствия помех 59,9-63,1% (TIMIT) в обучающей и тестовой выборке.

Как следует из табл. 2.3-2.4 и сводной диаграммы результатов рис.2.19 в случае влияния телефонного канала связи точность распознавания речи падает для всех корпусов искаженной речи (NTIMIT16, STC8, STC16) при PLP_0_D_A_Z параметризации сигнала (рис. 2.5):

– для NTIMIT-NTIMIT $Acc\%$ снижается в среднем с 59,8-62,9% (TIMIT) до 44,2-47,0%;

– для STC-STC с $F_d = 8$ кГц $Acc\%$ снижается в среднем с 59,9-63,1% (TIMIT) до 42,7-48,3%;

– для STC-STC с $F_d = 16$ кГц $Acc\%$ снижается в среднем с 59,9-63,1% (TIMIT) до 51,5-55,0%;

– Снижение частоты дискретизации с $F_d = 16$ до 8кГц снижает точность распознавания $Acc\%$ с 51,5-55,0%; до 42,7-48,3%;при одинаковых условиях обучения и тестирования моделей STC-STC (рис. 2.19).

Таблица 2.3. Точность распознавания монофонов. PLP_0_D_A_Z

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
TIMIT-TIMIT	16	55,7	57,4	59,1	60,2	60,5	60,7	61,2	61,7	61,9
NTIMIT-NTIMIT	16	40,1	41,8	42,7	44,0	44,5	45,2	45,7	46,2	47,2
TIMIT-NTIMIT	16	22,9	23,2	23,0	22,9	23,1	23,6	23,6	23,4	23,5
STC-STC	8	38,2	40,1	41,5	42,7	43,1	44,1	44,8	44,5	44,9
STC-STC	16	47,6	49,1	50,9	51,8	52,4	52,7	52,8	53,2	53,1

Таблица 2.4. Точность распознавания трифонов. PLP_0_D_A_Z

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
TIMIT-TIMIT	16	62,6	63,6	63,9	63,5	62,8	62,8	62,7	62,6	61,9
NTIMIT-NTIMIT	16	46,3	47,1	47,5	47,7	47,1	46,8	47,0	47,2	46,0
TIMIT-NTIMIT	16	26,9	26,5	27,0	26,6	26,9	26,2	25,5	25,4	25,1
STC-STC	8	45,7	47,4	48,5	49,1	49,2	49,1	48,8	48,6	48,3
STC-STC	16	54,1	54,9	55,1	55,1	55,4	55,0	55,3	55,1	54,8

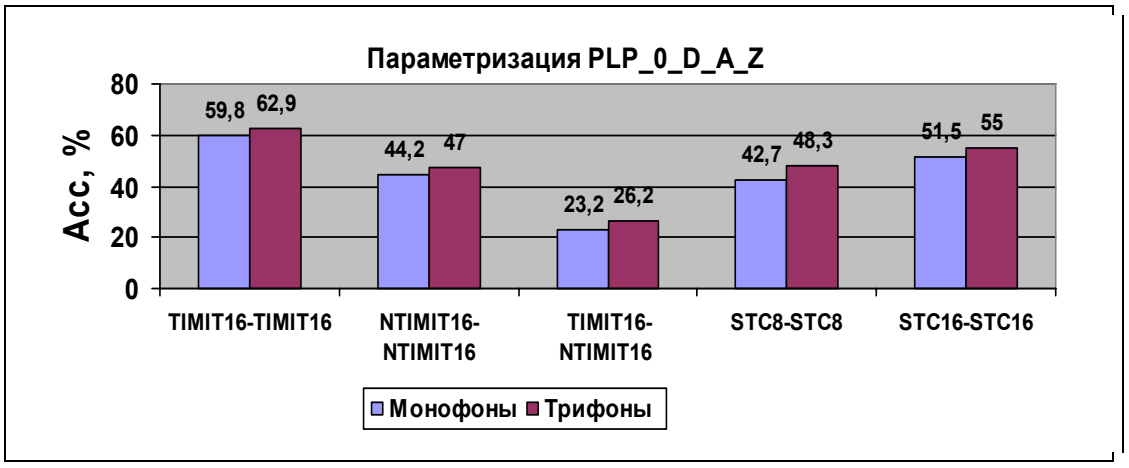


Рис. 2.19. Точность распознавания $Acc\%$. PLP_0_D_A_Z

Как следует из результатов табл. 2.1-2.4 и рис.2.18-2.19 существенную роль при распознавании играет частота дискретизации F_d для корпусов STC-TIMIT. При $F_d =$ 8кГц показатель $Acc\% =$ 30,4-29,2% (MFCC) и $Acc\% =$ 42,7-48,3 (PLP) для 8 кГц ниже, чем для 16 кГц $Acc\% =$ 41,5-47% (MFCC) и $Acc\% =$ 51,5-55,0 (PLP). Это понижение $Acc\%$ можно пояснить резким ухудшением качества сигнала, его информативности при извлечении параметризации, при понижении частоты дискретизации F_d сигнала.

Результаты сравнения точности распознавания для моделей трифонов (составные модели фонем) для MFCC_E_D_A_Z и PLP_0_D_A_Z параметризации речевого сигнала изображены на рис. 2.20

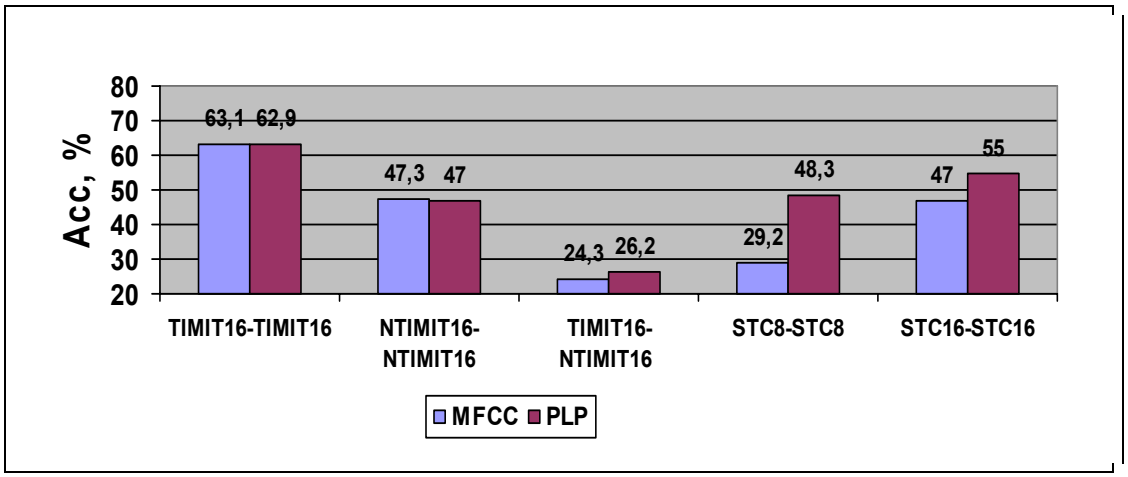


Рис. 2.20. Точность распознавания $Acc\%$. Сравнение MFCC_E_D_A_Z и PLP_0_D_A_Z параметризации

Анализ графиков рис.2.20 свидетельствует о том, что для корпуса чистой речи TIMIT (63,1% и 62,9%), а так же для корпуса NTIMIT (47,3% и 47%) точность распознавания существенно не изменяется для обоих видов параметризации. Существенное увеличение показателя $Acc\%$ наблюдается для различных условий создания моделей и их тестирования TIMIT-NTIMIT ($Acc\% =$ 24,3% увеличился до 26,3%), а так же для корпуса STC-TIMIT для обоих $F_d = 8$ и 16кГц.

Из приведенных результатов табл. 2.1-2.4 можно заключить о необходимости проведения обучения акустических моделей при

распознавании телефонной речи на выборке, записанной в аналогичных условиях. Тем не менее, обучение моделей в условиях аналогичных условиям тестирования не позволяет обеспечить приемлемой точности распознавания $Acc\%$ системы APP 59,8-63,1%, соответствующей условиям чистой речи TIMIT-TIMIT.

Как следует из рис. 2.20. PLP-параметризация позволяет улучшить точность $Acc\%$ (29,2% до 48,3% и 47% до 55%) в случае аналогичных условий создания и тестирования моделей STC-TIMIT, однако при использовании корпуса данных NTIMIT точность распознавания изменяется незначительно при изменении метода параметризации. Такой результат может быть объяснен тем, что корпус STC-TIMIT является упрощенным вариантом корпуса NTIMIT, состоящим лишь из записей одной телефонной линии непосредственно с коммутатора без использования (п. 2.2.2). Тем не менее, в условиях реальной работы системы APP TIMIT-NTIMIT это улучшение составляет порядка 2%.

Поскольку телефонный канал ограничивает эффективную полосу частот сигнала, в данной работе целесообразно исследовать вариант ограничения сигнала в диапазоне полосы пропускания телефонного канала связи от 300 до 3400Гц. В табл. 2.5 представлены результаты точности распознавания при MFCC_E_D_A параметризации и ограничении полосы пропускания от 300 до 3400Гц.

Базы TIMIT и NTIMIT имеет частоту дискретизации 16кГц и эффективную полосу пропускания 6,4кГц [24Zue,22Jankowski]. Проходя через телефонный тракт, полоса от 3,4-6,4 кГц не несет речевой информации. На стадии предобработки треугольные фильтры покрывают весь частотный диапазон от нуля вплоть до частоты Найквиста [15Young]. Чтобы убрать нежелательные компоненты в полосе от 3,4-6,4 кГц из дальнейшего анализа заданное число каналов гребёнки фильтров было распределено равномерно по мел-шкале от края до края результирующей полосы пропускания от 300 до 3400 Гц.

Таблица 2.5. Точность распознавания трифонов при MFCC_E_D_A параметризации и ограничении полосы пропускания от 300 до 3400Гц

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
NTIMIT- NTIMIT	16	46,8	47,5	47,8	47,7	48,0	47,1	46,4	45,9	45,8
STC- STC	8	33,2	35,1	34,6	34,9	35,1	35,9	36,0	35,8	36,1
STC- STC	16	33,3	34,3	35,7	35,9	36,5	36,8	36,4	36,2	36,5

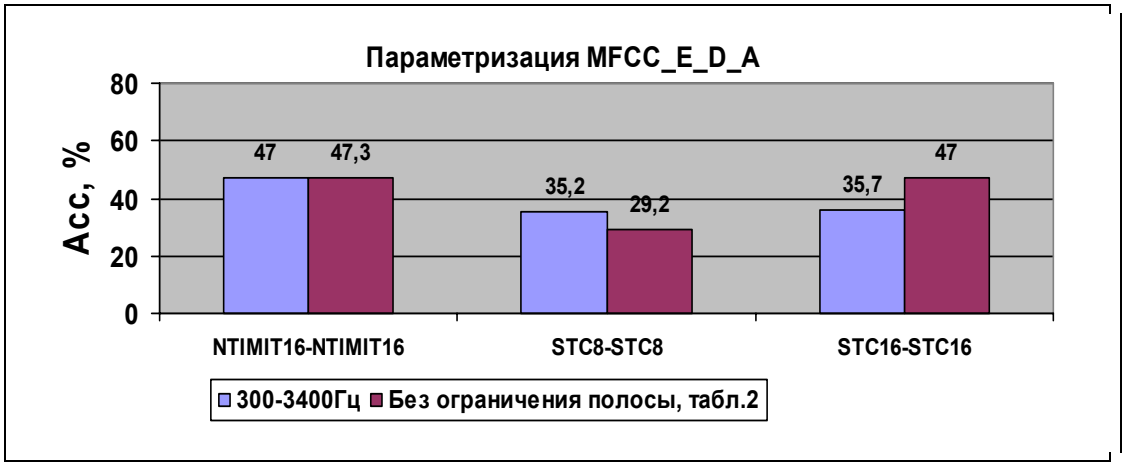


Рис. 2.21. Точность распознавания $Acc\%$ при ограничении полосы частот сигнала

Результаты экспериментальных исследований в виде значений автоматического распознавания $Acc\%$ при параметризации сигнала MFCC_E_D_A для различных частот дискретизации F_d для различного числа гауссовых смесей приведены в табл. 2.5.

Из табл. 2.5 видно, что полосовая фильтрация позволяет повысить надежность распознавания $Acc\%$ даже при использовании стандартного подхода на основе MFCC коэффициентов для корпуса STC-TIMIT $F_d = 8$ кГц (сравнивая значения табл. 2.5 и табл. 2.2 на графике рис. 2.21) в среднем с 29,2% до 35,5%.

Сравнивая результаты рис. 2.20 и рис. 2.21, следует, что при отсутствии нормализации кепстрального среднего коэффициентов MFCC точность для метода MFCC не улучшилась даже в случае ограничения полосы сигнала и составляет порядка 47-47,3%.

Незначительное улучшение результатов распознавания путём ограничения полосы частот может быть объяснено тем, что в случае NTIMIT вся неинформативная часть сигнала выше 4 кГц эффективно отфильтрована самим телефонным каналом. Кроме того в эксперименте без ограничения полосы существенное улучшение результатов во всех случаях объясняется применением нормализации кепстрального среднего (CMN – cepstral mean normalization) (см. префикс _Z). Данная операция позволяет устранить эффекты, связанные с искажениями частотных характеристик записывающих устройств или каналов передачи, путём вычитания среднего значения, вычисленного за длительный интервал, из последовательности кепстральных коэффициентов. К сожалению, вычисление среднего за длительный интервал не позволяет эффективно использовать этот метод в системах реального времени обработки сигнала, так как требует вычисления кепстрального среднего, вычисленного за длительный интервал [15 Young].

Результаты экспериментальных исследований в виде значений автоматического распознавания $Acc\%$ при ограничении полосы частот сигнала от 300 до 3400Гц для параметризации MFCC_E_D_A_Z при различном числе гауссовых смесей приведены в табл. 2.6, а на графике рис.2.22 изображены усредненные результаты для данных из табл. 2.6

Результаты сравнения точности распознавания $Acc\%$ в зависимости от применения процедуры нормализации кепстрального среднего (использовался флаг _Z) для моделей трифонов при MFCC параметризации речевого сигнала и ограничении сигнала в полосе частот от 300 до 3400Гц изображены на рис. 2.22.

Таблица 2.6. Точность распознавания трифонов при MFCC_E_D_A_Z параметризации и ограничении полосы пропускания от 300 до 3400Гц

Обучение-Тестирование	F_d, Гц	Количество моделирующих смесей								
		4	6	8	10	12	14	16	18	20
NTIMIT_Z	16	46,5	47,6	47,9	48,8	48,6	48,2	47,9	47,6	48,0
STC_Z	8	33,3	34,3	35,7	35,9	36,5	36,8	36,4	36,2	36,5
STC_Z	16	34,3	35,7	35,9	36,5	36,8	36,4	36,2	36,5	33,3

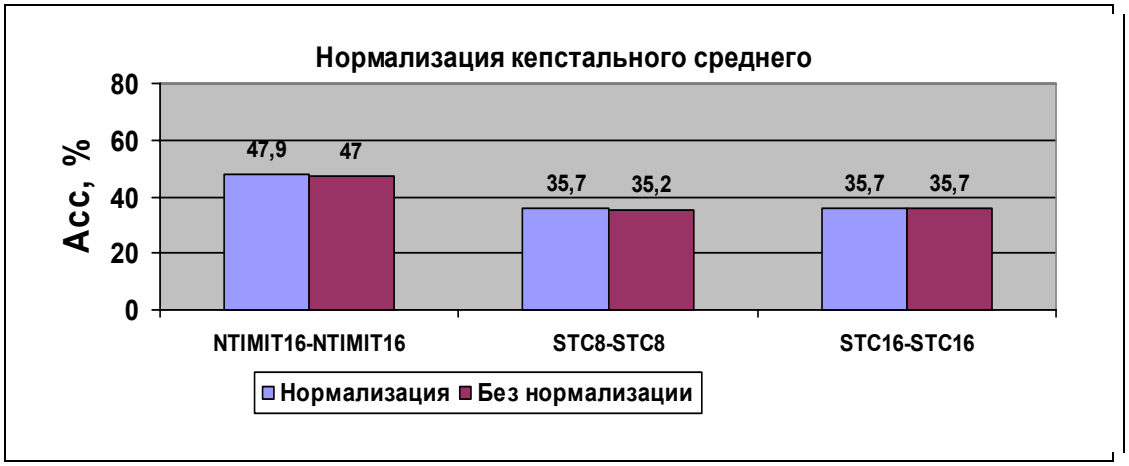


Рис. 2.22. Точность распознавания $Acc\%$ при нормализации кепстрального среднего MFCC_E_D_A_Z

Из результатов рис. 2.22 следует, что при ограничении полосы частот применение нормализации кепстрального среднего дает улучшения точности распознавания для NTIMIT абсолютное изменение составляет 0,9%, а для STC-TIMIT 8кГц – 0,5%, для корпуса STC 16 кГц изменение точности $Acc\%$ не произошло.

Резюмируя приведенные результаты, заключаем, что процедура параметризации сигналов PLP действительно эффективна по сравнению с MFCC вместе с применением нормализации кепстрального среднего (рис.2.22). Однако такое улучшение достигается не для всех рассмотренных речевых данных. Процедура простой фильтрации в полосе частот не дала существенного улучшения точности $Acc\%$ (рис. 2.21).

При этом, открытым остается вопрос применения MFCC и PLP параметризации в условиях реальной работы системы АРР, т.е. случае различных условий создания и тестирования акустической модели (TIMIT-NTIMIT). В этом случае улучшение при смене метода параметризации MFCC на PLP составило 2%, при MFCC точность составила при $Acc\% = 24{,}3$, а при PLP параметризации показатель точности распознавания $Acc\% = 26{,}2$ не является приемлемым для систем АРР. Отсюда следует вывод о целесообразности исследования метода PNCC для параметризации речевого сигнала в телефонном канале связи.

3.1.2. Оптимизация параметров параметрического представления речевого сигнала нормализованными по мощности кепстральными коэффициентами.

Как отмечалось в п. 2.1.2. для выяснения эффективности параметризации речевого сигнала методом PNCC необходимо определять медленно изменяющуюся нижнюю огибающую кратковременной оценки спектра мощности сигнала рис. 2.7. По сути эта огибающая является моделью для оценки среднего уровня шума, для которого любая активность выше этого уровня рассматривается как речевая активность. Для получения нижней огибающей применяют ассиметричный нелинейный фильтр (2.2), параметры которого необходимо установить. Для решения этой задачи обходимо ответить на такие вопросы:

– какими должны быть значения коэффициентов забывания λ_a и λ_b блока ассиметричной фильтрации (2.2)?

– какими должны быть значения коэффициента забывания λ_t и коэффициента подавления μ_t для блока временного маскирования (2.3) ?

– какая эффективность алгоритма PNCC для искажений и шумов, связанных с наличием телефонного канала связи?

– какая точность распознавания метода PNCC по сравнению с традиционным методом MFCC на искаженном и не искаженном телефонным каналом сигнале речи?

Для решения поставленной задачи автоматического распознавания речи в телефонном канале следует определить оптимальные параметры алгоритма PNCC. Экспериментальные исследования были организованы следующим образом. Критерием качества в этом случае выступает точность автоматического распознавания $Acc,\%$.

Организация экспериментальных исследований подобна таковой для п. 2.2.3, поэтому укажем лишь на отличительные особенности. Для моделирования системы APP использовались программные средства разработки современной системы APP и Sphinx [14Lee] с учетом рекомендаций работ [31Shmyrev]. Модель параметрического представления сигнала представляла собой модуль вычисления MFCC (рис. 2.4) и PNCC (рис. 2.6) коэффициентов. Реализация алгоритма параметризации PNCC речевого сигнала для экспериментальных исследований была взята из [31Shmyrev,30 Климков].Специфические параметры настройки системы APP на основе Sphinx приведены в приложении Б. Для создания акустической модели использовалось 3658 файлов. Тестирование проводилось на 1344 файлах. Данные для создания акустической модели и тестирования не пересекались.

Поскольку увеличение количества моделирующих гауссовых смесей не приводит лишь к постепенному увеличению точности распознавания $Acc\%$ (см. табл. 2.1-2.4) поэтому в дальнейших экспериментах для моделирования фонем использовалось только 8 гауссовых смесей.

Статистическая языковая модель timit.lm.DMP была создана с помощью средств CMU-Cambridge Statistical Language Modeling Toolkit v2 [29Rosenfeld] с использованием словаря в 6070 слов. Словарь фонем состоял из 40 фонем английского языка корпуса TIMIT.

В работе [25, 26Kim] широкий диапазон значений λ_b был ограничен между 0,25 и 0,75, а выбор $\lambda_a \approx 0{,}9$ был экспериментально установлен как лучший в смысле точности распознавания $Acc,\%$ для условий белого шума 5дБ, музыкального шума 5дБ и реверберационной помехи с временем реверберации $T_{60} = 0.5\text{c}$.

Чтобы определить оптимальные параметры алгоритма PNCC, для случая телефонной речи, в данной работе испытаны два альтернативных варианта согласно соотношению $1 > \lambda_a > \lambda_b > 0$, при котором в методе PNCC получают нижнюю огибающую мощности в качестве пороговой модели оценки уровня шума:

1. при фиксированных параметрах $\lambda_a = \{0{,}995 \quad 0{,}9\}$ и $\mu_t = 0{,}2$ изменяем λ_b и определяем оптимальный параметр λ_b;

2. при фиксированных $\lambda_a = 0{,}9$ и $\lambda_b = 0{,}5$ и $\mu_t = \{0{,}1 \quad 0{,}2\}$ определяем оптимальный λ_t.

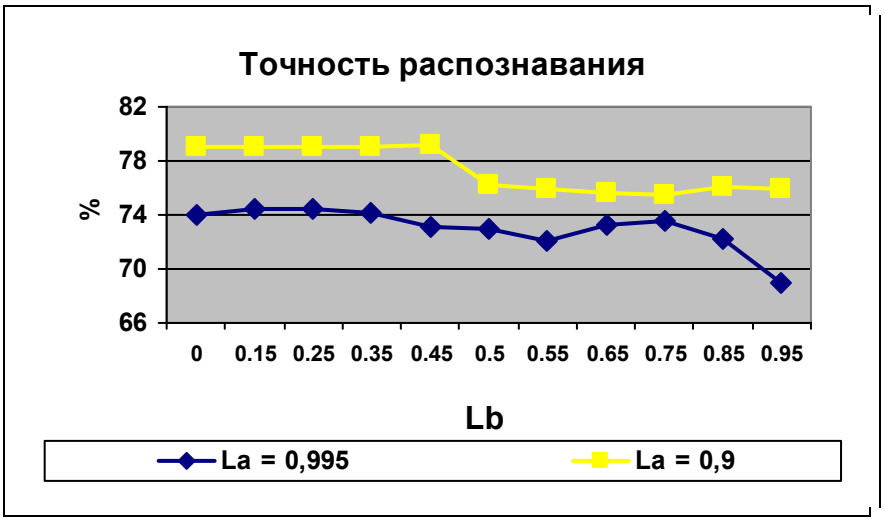


Рис.2.25. Оптимизация параметра λ_b

На рис.2.25 показаны графики зависимости точности распознавания слов в зависимости от параметра λ_b (выражение (2.1)). Параметр λ_a принят равным $0{,}995$ и $0{,}9$ для выполнения условия $1 > \lambda_a > \lambda_b > 0$ для выделения

нижней огибающей входной мощности сигнала. $\lambda_t = 0{,}85$ было взято из работы [25, 26Kim]

Как следует из экспериментально полученных графиков рис. 2.25 обе зависимости λ_a от точности распознавания $Acc,\%$ графика имеют плоскую форму для $0 \le \lambda_b \le 0{,}45$. Максимальное значение точности распознавания достигается для $\lambda_b = 0{,}25$ и 0,45 при $\lambda_a = 0{,}9$, поэтому в качестве оптимальных значений будут приняты эти значения.

Перейдем к рассмотрению вопроса оптимизации коэффициент забывания λ_t, входящего в выражение (2.3) для блока временного маскирования, схемы подавления шума (рис. 2.8) для случая присутствия речевого сигнала. При фиксированных $\lambda_b = 0{,}25$, $\lambda_a = 0{,}9$ значениях изменение параметра μ_t более 0,2 не показало существенного изменения точности распознавания, поэтому в качестве оптимальных значений рассматривались два значения $\mu_t = \{0{,}1 \quad 0{,}2\}$, при этом изменялся параметр λ_t.

На рис. 2.26 показаны графики зависимости точности распознавания в зависимости от параметра λ_t (выражение (2.3)).

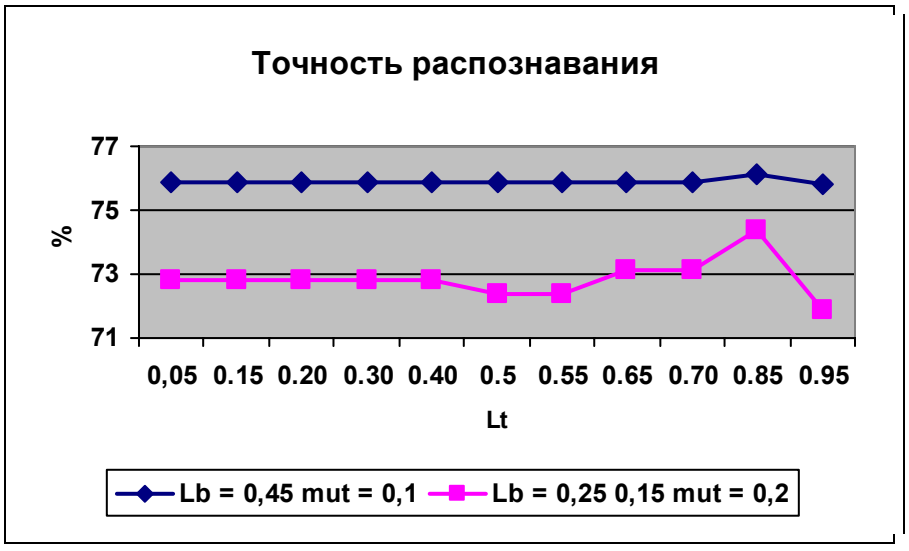


Рис.2.26. Оптимизация параметра λ_t

Как следует из рис. 2.26 графики имеют плоскую форму для $\lambda_b = 0{,}45$ $\mu_t = 0{,}1$, $\lambda_a = 0{,}9$ при $0 \le \lambda_b \le 0{,}7$ и для $\lambda_b = 0{,}25$ и $\mu_t = 0{,}2$, $\lambda_a = 0{,}9$ при $0 \le \lambda_b \le 0{,}85$.

Максимальное значение точности распознавания достигается для двух зависимостей при $\lambda_t = 0{,}85$ и, поэтому оптимальное значение λ_t принято $\lambda_t = 0{,}85$.

Перейдем теперь к вопросу выяснения эффективности алгоритма PNCC для искажений и шумов, связанных с наличием телефонного канала связи. Чтобы определить точность распознавания при использовании метода параметризации PNCC был проведен эксперимент, при котором система APP моделировалась средствами программной разработки Sphinx по схеме рис. 2.17.

Спектральная мощность получена в 40 полосах анализа, путем взвешивания амплитудно-квадратированного результата кратковременного преобразования Фурье для положительных частот 40-канальной гребенкой гамматон-фильтров [35Patterson, 36Slaney's], у которых центральных частоты линейно расположены на шкале ERB (Equivalent Rectangular Bandwidth) между 200 Гц и 8000 Гц.

В качестве оптимальных параметров асинхронного нелинейного фильтра взяты выше установленные значения параметров $\lambda_b = 0{,}25$ и $0{,}45$, $\lambda_a = 0{,}9$, а коэффициент забывания (2.3) блока временного маскирования (рис. 2.8) $\lambda_t = 0{,}85$ при $\mu_t = 0{,}1$.

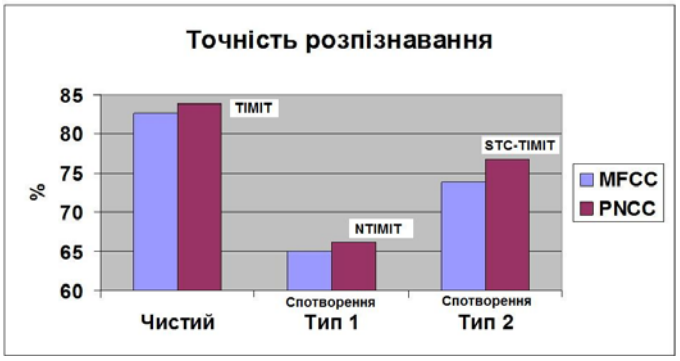


Рис. 2.27. Сравнение точности распознавания методов MFCC и PNCC

Результаты экспериментальных исследований в виде значений автоматического распознавания $Acc\%$ при MFCC и PNCC параметризации сигнала представлены на рис. 2.27. На рис. 2.27 показаны графики зависимости точности распознавания $Acc\%$ в зависимости от корпуса искаженных речевых сигналов (обучение-тестирование).

Таблица 2.7. Точность распознавания методов MFCC и PNCC

Обучение-Тестирование	F_d, Гц	$Acc\%$, параметризация		Абсолютная разность, %
		MFCC	PNCC	$\lvert Acc_{MFCC} - Acc_{PNCC} \rvert$
TIMIT - TIMIT	16	82,7	83,9	1,2
NTIMIT - NTIMIT	16	64,9	66,2	1,3
STC-STC	16	73,8	75,7	1,9
TIMIT - NTIMIT	16	82,7	83,9	1,2
TIMIT - STC	16	82,7	83,9	1,2

Как следует из рис. 2.27 для одинаковых условий создания модели и её тестирования в отсутствии искажающего влияния телефонного канала связи (TIMIT-TIMIT), точность распознавания методов MFCC и PNCC отличается на 1,2% (см. табл. 2.7) при одинаковых условиях и общих параметрах проведения экспериментов.

Очевидно, что параметризация сигнала по методу PNCC привела к положительным результатам (точность $Acc\%$ для метода PNCC во всех случаях рис. 2.27 выше на 1,2-1,9%). Совпадение точности $Acc\%$ распознавания для условий TIMIT – TIMIT, TIMIT – NTIMIT и TIMIT – STC можно объяснить использованием статистической модели языка, которая в свою очередь «выравнивает» незначительные изменения точности $Acc\%$, т.е. влияет на оцениваемый показатель, $Acc\%$ изменяя его согласно языковой модели общей для всех корпусов.

К сожалению, в системе APP Sphinx не предусмотрена возможность оценки точности распознавания $Acc\%$ без учета влияния языковой модели. Поэтому точность $Acc\%$ оценивалась как точность распознавания слов, а не фонем.

Таким образом, ответы на поставленные вначале п. 2.2.5 вопросы для решения первой задачи таковы: оптимальными, в смысле максимума $Acc,\%$, являются значения коэффициентов забывания блока ассиметричной фильтрации $\lambda_b = 0{,}25$ и 0,45 и $\lambda_a = 0{,}9$; оптимальными, в смысле максимума $Acc,\%$, являются значения коэффициента забывания $\lambda_t = 0{,}85$ и коэффициента подавления для блока временного маскирования $\mu_t = 0{,}1$; эффективность алгоритма PNCC для искажений и шумов, связанных с наличием телефонного канала связи по сравнению с MFCC параметризацией составляет 1,2-1,9% для различных корпусов, данных из табл. 2.7;

3.3. Выводы

1. Впервые получены значения точности распознавания для алгоритмов параметризации речевого сигнала PNCC для искажений и шумов, связанных с наличием телефонного канала связи NTIMIT и STC-TIMIT и показано, что метод PNCC превосходит по эффективности традиционный метод MFCC. Очевидно, что параметризация сигнала по методу PNCC привела к положительным результатам и позволила поднять точность распознавания $Acc\%$ для метода PNCC для всех речевых корпусов на 1,2-1,9%.

2. Выработаны рекомендации по оптимизации параметров алгоритма PNCC для речевого сигнала, искаженного телефонным каналом связи. Оптимальными, в смысле максимума $Acc,\%$, являются значения коэффициентов забывания блока ассиметричной фильтрации $\lambda_b = 0{,}25$ и 0,45 и $\lambda_a = 0{,}9$; коэффициента забывания $\lambda_t = 0{,}85$ и коэффициента подавления для блока временного маскирования.

ГЛАВА 4
РАЗРАБОТКА РАЗДЕЛЬНОЙ ПРОЦЕДУРЫ ОБРАБОТКИ ГОЛОСОВОЙ АКТИВНОСТИ В ПРИСУТСТВИИ НЕСТАЦИОНАРНОГО ШУМА

Четвертый раздел посвящен усовершенствованию процедуры раздельной обработки речевого сигнала при определении голосовой активности и при определении параметров шума методом PNCC. Результаты моделирования проверены экспериментальным путем на искусственных и реальных сигналах. Для этого был разработан программную модель (язык программирования С ++) нейросетевого детектора голосовой активности (рис. 10) системы АРР.

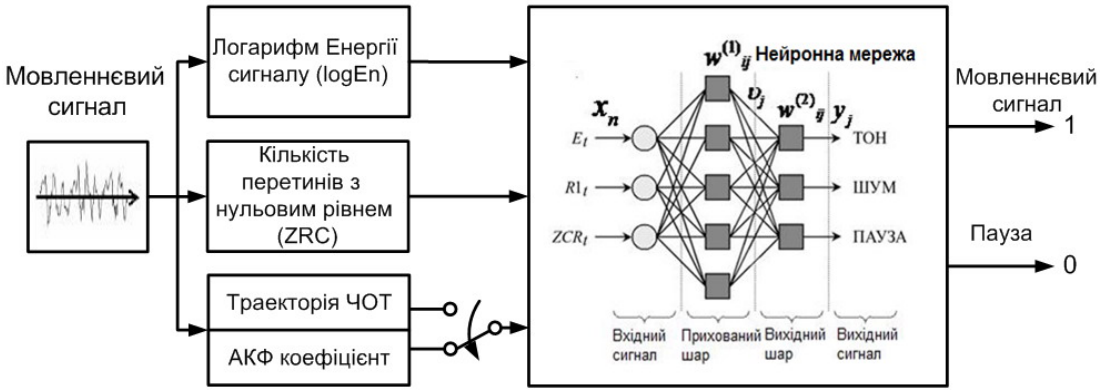


Рис. 10. Нейросетевой детектор голосовой активности (VAD)

В таком детекторе, который в дальнейшем обозначать аббревиатурой VAD (Voice Activity Detector), использовались 3 классификационные признаки: логарифм кратковременной энергии, количество пересечений амплитуды сигнала с нулевым уровнем и автокорреляционная коэффициент с единичной задержкой между соседними отсчетами сигнала. Обоснование структуры детектора голосовой активности, обеспечивает робастность системы АРР при использовании PNCC признаков, обусловленное основными требованиями: простая система признаков, которая может быть вычислена в реальном времени; эффективный алгоритм принятия решения; работа на интервалах стационарности (фреймах) системы АРР.

В качестве классификатора в пространстве признаков избран классическую стационарную нейронную сеть в виде многослойного персептрона, который в дальнейшем обозначать аббревиатурой MLP (Multilayer Perceptron). Одной из важных свойств MLP сети является то, что сложность операций в ней можно определить заранее, путем изменения структуры связей между ее слоями в процессе ее проектирования. Кроме того, предложенная MLP сеть позволяет расширять выбор или заменять классификационные признаки, не меняя методы определения структуры связей между ее слоями.

4.1. Применение классической стационарной нейронной сети к задаче детектирования голосовой активности

На рис. 3.3. представлена двухслойная сеть с одним скрытым слоем. Входной сигнал $x = [x_0, x_1, ..., x_N]^T$, где $x_0 = 1$, распространяется по сети в прямом направлении, от слоя к слою.

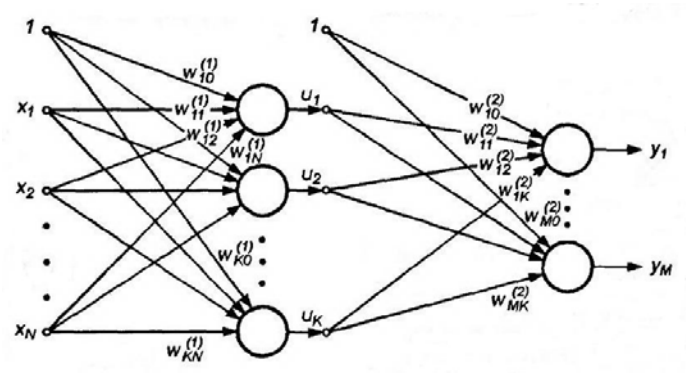


Рис. 3.3.Обобщенная структура двухслойной нейронной сети

Аналитическая модель выходного сигнала i – го нейрона скрытого слоя представлена на рис. 3.3.1 и описывается функцией (3.2) для выходного сигнала первого скрытого слоя:

$$v_j = \varphi\left(\sum_{j=0}^{N} w_{ij}^{(1)} x_j\right), \qquad (3.2)$$

В выходном слое k – й нейрон вырабатывает выходной сигнал:

$$y_k = \varphi\left(\sum_{j=0}^{N} w_{kj}^{(2)} v_j\right) = \varphi\left(\sum_{l=0}^{K} w_{kl}^{(2)} \varphi\left(\sum_{j=0}^{N} w_{ij}^{(1)} x_j\right)\right) \qquad (3.3)$$

где

v_j – выходные сигналы нейронов скрытого слоя, $(j = 1,2,...,K)$;

$w_{ij}^{(1)}$, $w_{ij}^{(2)}$ – вектора весов связей между слоями;

y_k – выходные сигналы нейросети, $(k = 1,2,...,M)$;

$\varphi(\cdot)$ – функция активации нейрона, описывает нелинейную взаимосвязь входного и выходного сигналов этого нейрона.

С вектором x связаны два входных вектора – фактических выходных сигналов $y = [y_0, y_1,..., y_M]^T$ и ожидаемых сигналов $d = [d_0, d_1,..., d_M]^T$.

Таким образом, для построения модели VAD необходимо настроить весовые коэффициенты $w_{ij}^{(1)}$ и $w_{ij}^{(2)}$ для всех слоёв сети нейронной сети (рис. 3.4) так, чтобы минимизировать целевую функцию ошибок вида:

$$E(w) = \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{M}\left(y_k^j - d_k^j\right)^2, \qquad (3.1)$$

при использовании p обучающих векторов $\{(x(n), d(n))\}_{j=1}^{p}$ для обучения сети, включающей M выходных нейронов (рис.3.3).

Процесс обучения нейронной сети основывается на стратегии обратного распространения ошибки (Back Propagation) с использованием одного из обучающих алгоритмов на множестве обучающих данных [11Осовски, 12Хайкин].

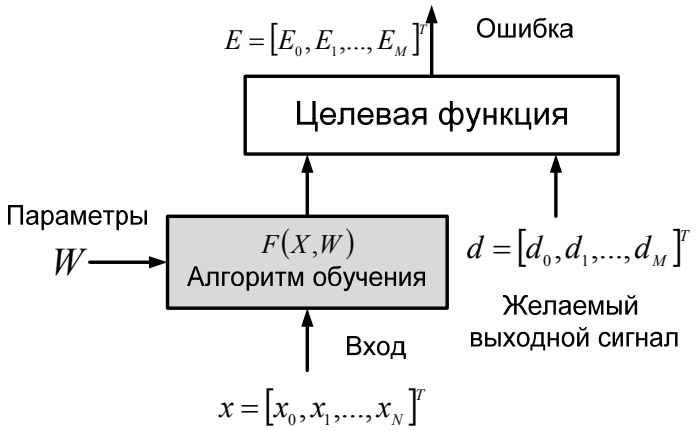


Рис. 3.4. Обобщенная схема обучения нейронной сети

Известно, что на данный момент не существует универсального алгоритма обучения, подходящего для всех архитектур нейронных сетей. Существует лишь набор средств, представленный множеством алгоритмов обучения, каждый из которых имеет свои достоинства[11Осовский,13LeCun].

3.1.4. Алгоритм обратного распространения ошибки обучения MLP сети. Обучение MLP сети реализуется с применением наиболее эффективных методов оптимизации [12Хайкин]. Как следует из (3.2) каждый нейрон j MLP сети вычисляет нелинейную функцию $\varphi_j(\cdot)$, поэтому значения оптимальных весов невозможно найти аналитическим путем, вместо этого применяют градиентный метод для некоторой глобальной (целевой) функции ошибок:

$$E(w)=\frac{1}{2}\sum_{k=1}^{M}\left(y_k-d_k\right)^2 \tag{3.4}$$

В свою очередь градиентные методы требуют определения вектора градиента:

$$\nabla E(w)=\frac{\partial E(n)}{\partial w}$$

относительно весов w всех слоёв сети. Уточнение весов (обучение) производиться по формуле

$$w(n+1) = w(n) + \Delta w,$$

где

$$\Delta w(n) = -\eta \cdot \nabla E(w), \tag{3.5}$$

где η – коэффициент обучения или скорость обучения.

Очевидное решение эта задача имеет лишь для весов выходного слоя, поскольку для выходного слоя известны вектора ожидаемых значений. Для других слоёв используется специальная стратегия – алгоритм обратного распространения ошибки [14Haykin]. Уточнение весов проводиться после предъявления каждой обучающей выборки (x,d) из всего обучающего множества (эпохи) $\{(x(n),d(n))\}_{n=1}^{N}$.

При анализе сети в прямом направлении получают реакцию сети на выходной образ $\{x(n)\}_{n=1}^{N}$ из подаваемых на выход и выход сети примеров

$$\{(x(n),d(n))\}_{n=1}^{N} = \{(x(1),d(1)),\ldots,(x(N),d(N))\}, \tag{3.0}$$

при этом рассчитываются значения выходных сигналов $\{y(n)\}_{n=1}^{N}$, далее фактические сигналы выходов сети $y^{(L)}(n) = o(n)$. вычитаются из желаемого отклика $\{d(n)\}_{n=1}^{N}$, в результате чего формируется сигнал ошибки $e(n)$.

Это процесс происходит следующим образом. Выходной сигнал нейрона j слоя l получают путем расчета функциональных сигналов на выходе сети при прямом проходе:

$$y_j^{(l)}(n) = \varphi_j(v_j^{(l)}(n)), \qquad v_j^{(l)}(n) = \sum_{i=0}^{m_0} w_{ji}^{(l)}(n) y_i^{(l-1)}(n) \tag{3.6}$$

где

$y_i^{(l-1)}(n)$ – выходной сигнал нейрона i, предыдущего слоя $(l-1)$ на итерации n (для $i=0$ $y_o^{(l-1)}(n)=+1$);

$v_j^{(l)}(n)$ – локальное поле нейрона j;

$w_{ij}^{(l)}(n)$ – вес связи нейрона j слоя l с нейроном i слоя $l-1$;

$w_{j0}^{(l)}(n)=b_j^l(n)$ – порог, применяемый к нейрону j слоя l;

$\varphi_j(\cdot)$ – сигмоидальная функции активации;

n – соответствует n-му обучающему примеру, поданному на выход сети.

Если нейрон j находиться в первом скрытом слое $l=1$, то

$$y_j^{(0)}(n)=x_j(n) \tag{3.7}$$

где $x_j(n)$ – j-й элемент входного вектора $x(n)$. Если нейрон j находиться в выходном слое, то есть $l=L$, то

$$y_j^{(L)}(n)=o_j(n). \tag{3.8}$$

Сигнал ошибки определяется из соотношения

$$e_j(n)=d_j(n)-o_j(n), \tag{3.9}$$

где $d_j(n)$ – j-й элемент вектора желаемого отклика $d(n)$.

В работах [12Хайкин,13LeCun] показано, что при обратном проходе вычисляют локальные градиенты узлов сети по следующей формуле:

$$\delta_j^{(l)} = \begin{cases} e_j^{(L)}(n)\varphi_j'(v_j^{(L)}(n)) & \text{для нейрона } j \text{ вых.слоя } L \\ \varphi_j'(v_j^{(l)}(n))\sum_k \delta_k^{l+1}(n) w_{kj}(n) & \text{для нейрона } j \text{ скрытого слоя } l \end{cases} \quad (3.10)$$

где $\varphi_j'(\cdot)$ – дифференцирование по аргументу. Изменение весов w_{ji} осуществляется согласно выражению

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \Delta w_{ji}(n), \;\; \Delta w_{ji}(n) = \eta \delta_j^{(l)}(n) \cdot y_j^{(l-1)}(n) \quad (3.11)$$

где

$w_{ji}(n)$ – вес, связывающий выход нейрона i со входом нейрона j на итерации n;

$\Delta w_{ji}(n)$ – коррекция, применяемая к этому весу;

η – параметр скорости обучения;

$\delta_j^{(l)}$ – локальный градиент;

$y_j^{(l-1)}(n)$ – выходной сигнал нейрона j;

4.2. Разработка метода повышения помехоустойчивости (робастности) детектора голосовой активности за счет оценивания признака «траектория основного тона»

Предложено расширить пространство признаков распознавания классов «громкий-согласный-пауза» за счет введения признака «траектория частоты основного тона». Целью такого расширения является повышение помехоустойчивости предложенного детектора.

Траектория оценок частоты основного тона (ЧОТ) определяется для гласных звуков как непрерывная линия, объединяющая оценки, полученные для группы фреймов. ЧОТ $F0 = 1/T_0$ связана с периодом основного тона (ОТ)

$T_0 = p_{F0}/F_d$, F_d - частота дискретизации сигнала, p_{F0} - период ОТ в отсчетах сигнала. Оценки ОТ получают путём поиска таких периодов ОТ p_m, де m - количество отобранных оценок, которые соответствуют $p_m = \arg\max R(p)$, де $R(p)$ - функция нормированной автокорелляции (ФНАК) для сдвигов в ограниченном диапазоне $p \in [p_{\min}, p_{\max}]$.

Традиционно при автокорреляционном анализе ищется максимальное значение автокорреляционной функции (АКФ) при ненулевом смещении, считая, что АКФ имеет пики на сдвигах, кратных периода ОТ. Однако для речевых сигналов такой подход может привести к ошибочному выбору максимумов, соответствующих частотам, кратным $F0$. Для учета возможности выбора ошибочных максимумов АКФ предлагается проводить совокупный анализ полученных оценок. На рис. 11 схематично изображена сетку поиска наиболее вероятной траектории ОТ.

Вероятность выбора оценки p_m в траекторию ЧТО пропорциональна величине $q_N(m_k) = R_k(p(m_k))$. Для уменьшения ложного выбора максимумов ФНАК, обусловленных действием шумовой помехи, предлагается ограничивать границы поиска траектории ОТ в смежных фреймах k та k_{k+1} таким образом: $q_T(m_k, m_{k+1}) = 0$, если $|p(m_k) - p(m_{k+1})| \le \alpha \cdot p(m_k)$, где α - ограничивает возможные отклонения траектории ОТ для смежных кадров.

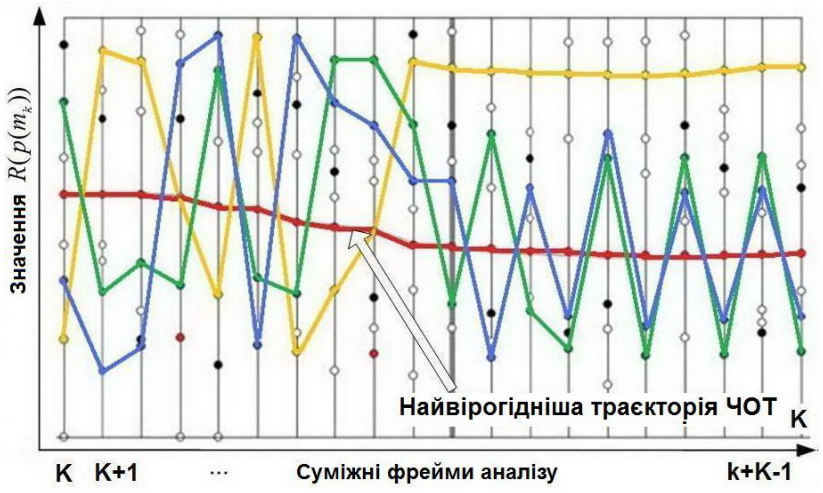


Рис. 11. Сітка пошуку найбільш ймовірної траєкторії ОТ

Объединение полученных оценок в смежных фреймах в траекторию ОТ выполняется для группы смежных фреймов только при условии $q_T(m_k, m_{k+1}) = 0$, суммированием $q_N(m_k)$ для группы кадров K и выбирая ту траекторию, сумма вероятностей выбора оценок которой максимальна.

Разработанный в диссертационной работе алгоритм для получения «траектории частоты основного тона» был экспериментально проверен на эталонных сигналах ларингограм и на тестовых сигналах. Процент грубых ошибок (GPE) полученной траектории ОТ, то есть отклонения эталонной и тестовой траектории, для женских голосов составил GPE = 0,8%, а для мужских GPE = 1%. Для оценки помехоустойчивости полученной траектории ОТ проведены эксперименты для чистых и искаженных белым шумом (с отношением сигнал-шум = 0дБ) сигналов, при этом погрешность составила всего GPE = 1,4%.

4.3. Разработка алгоритма инкрементного изменения параметров стационарной нейронной сети

Классическим методом подбора оптимальных весов сети (корректировки) является метод наискорейшего спуска, представленного в данном случае алгоритмом обратного распространения ошибки.

Метод наискорейшего спуска работает на основе линейного приближения целевой функции:

$$E(w + p) = E(w) + g^T(w)p, \qquad (3.2)$$

при её разложении в ряд Тейлора в ближайшей окрестности точки имеющегося вектора решения w:

$$E(w+p) = E(w) + g^T(w)p + \tfrac{1}{2}p^T H(w)p + O(\|p\|^3) \quad (3.2.3.1)$$

где $g(w)$ – вектор локального градиента;

$$g(w) = \nabla E(w) = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, ..., \frac{\partial E}{\partial w_n}\right]$$

$H(w)$ – матрица производных второго порядка, Гессиан:

$$H(w) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1 \partial w_1} & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 E}{\partial w_n \partial w_1} & \cdots & \frac{\partial^2 E}{\partial w_n \partial w_n} \end{bmatrix},$$

$O(\|p\|^3)$ – локальная погрешность отсеченной части.

Точкой решения считают точку w_k, в которой достигается минимум $\min = E(w_k)$ и $g(w_k) = 0$, где k – номер цикла обучения.

В случае метода наискорейшего спуска вектор выбора направления, отвечающий условию $g(w_k)^T p < 0$ уменьшения значения целевой функции $E(w_{k+1}) < E(w_k)$, получают в виде [Осовский] $p = -g(w_k)$.

В процессе минимизации $E(w)$ направление p и шаг η должны подбираться таким образом, чтобы для каждой очередной точки выполнялось условие $E(w_k + \eta_k p_k) < E(w_k)$. В этом случае корректировка, применяемая к весу w_k, определяется согласно выражению (5) или в общем виде:

$$\Delta w_k = -\eta_k \cdot p_k, \qquad w_{k+1} = w_k + \Delta w_k \qquad (100)$$

где

η – коэффициент обучения;

p_k – направляющий вектор, зависящий от значений вектора $w = [w_1, w_2, ..., w_n]^T$.

Поиск минимума продолжается, пока норма градиента не упадет ниже априори заданного значения допустимой погрешности либо пока не будет превышено максимальное время вычислений (количество итераций).

Ограничение слагаемым первого порядка (3.2.3.0) не позволяет использовать информацию о кривизне $E(w)$, поскольку единственным источником информации о поверхности ошибок является градиент $g(w)$. Это обуславливает медленную (линейную) сходимость метода. Указанный недостаток, а так же резкое замедление минимизации в ближайшей окрестности точки оптимального решения, когда градиент принимает очень малые значения, делают алгоритм наискорейшего спуска низкоэффективным. Тем не менее, с учетом его простоты, невысоких требований к объёму памяти и относительно небольшой вычислительной сложности этот метод остается базовым методом обучения нейронных сетей.

Как было сказано в п. 3.2.4. включение слагаемого момента $\alpha \cdot w_{ji}^{(l)}(n-1)$ в уравнение корректировки вектора весов

$$\Delta w_{ji}(n) = \eta \delta_j^{(l)}(n) \cdot y_j^{(l-1)}(n) + \alpha \cdot w_{ji}^{(l)}(n-1) \qquad (3.31)$$

было грубой попыткой использования информации второго порядка о поверхности ошибок, что сделало процесс корректировки более управляемым. Тем не менее, его использование сделало процесс обучения более чувствительным к подбору такого значения α, которое наилучшим образом отображало специфику решаемой проблемы. Независимость коэффициента α от фактического значения градиента приводит к тому, что вблизи минимума α может вызвать слишком большое изменение весов, а

при малых значениях градиента показатель α начинает доминировать в (3.31), что приводит к изменению $\Delta w_{ji}(n)$, которое выводит значение целевой функции ошибок, позволяющему пропустить локальный минимум.

Ключевым моментом этого алгоритма является предотвращение влияния показателя момента α на протяжении процесса обучения, чтобы избежать выше описанной нестабильности алгоритма.

В случае использования методов квадратичного приближения целевой функции $E(w)$ (3.2.3.1), вектор направления, который гарантирующий достижение минимального для конкретного шага значения целевой функции определяют из выражения:

$$p_k = -\left[H(w_k)\right]^{-1} g(w_k) \qquad (3.34)$$

На практике выражение (3.34) требует положительной определенности гессиана $H(w_k)$ на каждом шаге, что в общем случае не осуществимо. По этой причине вместо точно определенного $H(w_k)$ используется его приближение $G(w_k)$ в методе переменной метрики [Осовский].

Достоинство метода переменной метрики заключается в более быстрой сходимости, по сравнению с методом наискорейшего спуска. Тем не менее, этот метод требует относительно большой вычислительной сложности, связанный с необходимостью расчета в каждом цикле n^2 элементов $H(w_k)$, а так же использовании значительных объёмов памяти для хранения элементов $H(w_k)$. Поэтому применимость метода переменной метрики с использование персонального компьютера допустима для сети, содержащей не более 1000 в связей.

Другим вариантом ньютоновской стратегии оптимизации является алгоритм Левенберга-Марквардта [Осовский, Pham]. При его использовании точное значение гесианна $H(w_k)$ заменяется аппроксимированным значением

$G(w_k)$, которое требует грамотного подбора регуляризационнного фактора v_k. В алгоритме сопряженных градиентов при выборе направления минимизации не используется информация о гесианне. Направление поиска p_k выбирается таким образом, чтобы оно было ортогональным и сопряженным ко всем предыдущим направлениям $p_0, p_1, ..., p_{k-1}$. Этот метод имеет сходимость, близкую к линейной и он менее эффективен, чем метод переменной метрики, однако заметно быстрее метода наискорейшего спуска.

Алгоритмы, описанные выше, позволяют определить только направление, в котором уменьшается целевая функция, но ничего не говорят о величине шага, при котором эта функция может получить минимальное значение. Необходимо подобрать такое значение η_k, чтобы новое решение $w_{k+1} = w_k + \eta_k p_k$ лежало как можно ближе к минимуму функции $E(w)$ в направлении p_k. Грамотный подбор коэффициента η_k оказывает существенное влияние на сходимость алгоритма. Чем сильнее величина η_k отличается от значения, при котором $E(w)$ достигает минимума в выбранном направлении p_k, тем большее количество итераций потребуется для поиска оптимального решения. Слишком малое значение η_k не позволяет минимизировать целевую функцию за один шаг и вызывает необходимость повторно двигаться в том же направлении. Слишком большой шаг приводит к «перепрыгиванию» через минимум функции и фактически заставляет возвращаться к нему.

Существуют различные способы подбора значения коэффициента обучения η_k. Простейший из них основан на фиксации постоянного значения η_k на весь период оптимизации. Этот способ практически используется только совместно с методом наискорейшего спуска. Он имеет низкую эффективность, поскольку значение η_k никак не зависит от вектора фактического градиента и, следовательно, от направления на данной итерации. Величина η_k подбирается, как правило, раздельно для каждого

слоя сети с использованием различных эмпирических зависимостей. Один из подходов состоят в определении минимального значения коэффициента η_k для каждого слоя по формуле [72]

$$\eta \le \left(\frac{1}{n_i} \right),$$

где n_i обозначает количество входов i-го нейрона в слое.

Другой более эффективный метод основан на адаптивном подборе коэффициента η_k с учетом фактической динамики величины целевой функции в результате обучения. В соответствии с этим методом стратегия изменения η_k определяется путем сравнения суммарной погрешности ε на $i-$й итерации с её предыдущим значением, причем ε рассчитывается по формуле:

$$\varepsilon = \sqrt{\sum_{j=1}^{M} (y_i - d_i)^2}$$

В случае когда $\varepsilon_i > k_w \varepsilon_{i-1}$, где i – текущая итерация; k_w – коэффициент допустимого прироста погрешности, то η должно уменьшиться в соответствии с формулой $\eta_{i+1} = \eta_i \rho_d$, где ρ_d – коэффициент уменьшения η. В противном случае, когда $\varepsilon_i \le k_w \varepsilon_{i-1}$ принимается $\eta_{i+1} = \eta_i \rho_i$, где ρ_i – коэффициент увеличения η.

Однако необходимо подчеркнуть, что адаптивный метод подбора η сильно зависит от вида целевой функции и значений коэффициентов k_w, ρ_d, ρ_i. Значения, оптимальные для функции одного вида, могут замедлять процесс обучения при использовании другой функции. Поэтому при практической реализации этого метода следует обращать внимание на механизмы контроля и управления значениями коэффициентов, подбирая их в соответствии со спецификой задачи.

Наиболее эффективный, хотя и наиболее сложный, метод подбора коэффициента обучения связан с направленной минимизацией целевой

функции в выбранном заранее направлении p_k. Необходимо так подобрать скалярное значение η_k, чтобы новое решение $w_{k+1} = w_k + \eta_k p_k$ соответствовало минимуму целевой функции в данном направлении p_k. В действительности получаемое решение w_{k+1} только с определенным приближением может считаться настоящим минимумом. Это результат компромисса между объемом вычислений и влиянием величины η_k на сходимость алгоритма.

Серди наиболее популярных способов направленной минимизации можно выделить безградиентные и градиентные методы. В безградиентных методах используется только информация о значениях целевой функции, а её минимум достигается в процессе последовательного уменьшения диапазона значений вектора w.

Однако лучшим решением считается применение градиентных методов, в которых кроме значения функции, учитывается также и её производная вдоль направляющего вектора p_k. Они позволяют значительно ускорить достижение минимума, поскольку используют информацию о направлении уменьшения величины целевой функции.

Помимо алгоритмов обучения, реализующих апробированные методы оптимизации целевой функции, рассмотренные выше, существуют методы эвристического типа. Зачастую они представляют модификацию методов наискорейшего спуска или сопряженных градиентов. Подобные модификации связаны с внесением в них некоторых изменений, ускоряющих процесс обучения [33Quickprop, 133PROP], однако в таких алгоритмах реализуется личный опыт работы авторов с нейронными сетями.

Из выше перечисленных возможных алгоритмов корректировки весов сети следует, что ключевым моментом правильного изменения весов сети является необходимость применения различных значений параметров η и α, то есть локальных для каждой конкретной связи в сети.

В данной диссертации предложен алгоритм, позволяющий применить метод инкрементной изменения дельт (ошибки на выходе отдельного нейрона) Incremental Delta-Bar-Delta (IDBD), разработанный Р. С. Саттоном, для выбранного типа стационарной нелинейной MLP сети.

Предлагаемый алгоритм Incremental Delta-Bar-Delta формулируется для взвешенной суммы входов в этот узел $x_i(t)$, где $i = \overline{1,n}$, n – количество входов в линейный нейрон:

$$y(t) = \sum_{i=1}^{n} w_i(t) \cdot x_i(t) \tag{3.13}$$

Корректировка весов выполняется согласно обновляемому выражению:

$$w_i(t+1) = w_i(t) + \alpha_i(t+1) \cdot \delta(t) \cdot x_i(t), \tag{3.14}$$

где

t – индекс, который показывает порядок вычисления всех параметров;

$w_i(t)$ – изменяемое в момент времени t значение веса;

$\delta(t) = d(t) - y(t)$ – разность между желаемым и полученным выходом;

$x_i(t)$ – входные сигналы в линейный нейрон, где $i = \overline{1,n}$ – индексы n входных сигналов в один линейный нейрон;

$\alpha_i(t) = e^{\beta_i(t)}$ – коэффициент скорости обучения;

$$\beta_i(t+1) = \beta_i(t) + \theta \cdot \delta(t) \cdot x_i(t) \cdot h_i(t), \tag{3.15}$$

$\beta_i(t)$ – параметры адаптации;

θ – положительная константа – мета-скорость обучения;

$h_i(t)$ – дополнительный параметр памяти для каждого входного сигнала:

$$h_i(t+1) = h_i(t) \cdot \left[1 - \alpha_i(t+1) \cdot x_i^2(t)\right]^+ + \alpha_i(t+1) \cdot \delta(t) \cdot x_i(t), \qquad (3.16)$$

где $\left[x\right]^+$ – это x, если $x > 0$, иначе 0.

Важным достоинством IDBD метода является возможность учета параметра скорости обучения $\alpha_i(t)$, адаптированного к каждому входу $x_i(t)$, при котором потребуется настраивать всего один свободный параметр θ для корректировки весов [15Sutton, 16Hampson]. В то время как для подобного алгоритма Delta-Bar-Delta в работе [17Jacobs] требуется настраивать три параметра.

С вычислительной точки зрения Incremental Delta-Bar-Delta возможно применять в «он-лайн» режиме, в отличие от Delta-Bar-Delta [17Jacobs], который работает только после обработки всего обучающего множества целиком.

4.5. Предлагаемая модель выделения речевых и не речевых участков в речевом сигнале на основе применения искусственной нейронной сети

Предлагаемая модель учитывает недостатки традиционных [3,4,10,8] и стандартизированных [2, 6] VAD методов, перечисленных выше. Принципиальная схема предлагаемой модели VAD отображена на рис. 3.6.

Процедура принятия решения VAD строится на основе проектирования структуры многослойной персептронной сети (MLP – Multilayer Perceptron) [11Осовский] с полными последовательными связями. При создании MLP сети строится нелинейная модель физического процесса, обеспечивающая обобщение примеров типа «вход-выход», использованных при обучении сети.

Модуль VAD (рис. 3.6) принимает цифровой аудио сигнал, обрабатывает этот сигнал скользящим оконным анализом с длиной окна 20 мсек и шагом 10 мсек. В результате такого анализа речевой сигнал

разбивается на T речевых фреймов. Для каждого фрейма с помощью нейросетевого классификатора определяется принадлежность к одному из классов – тон, шум или пауза.

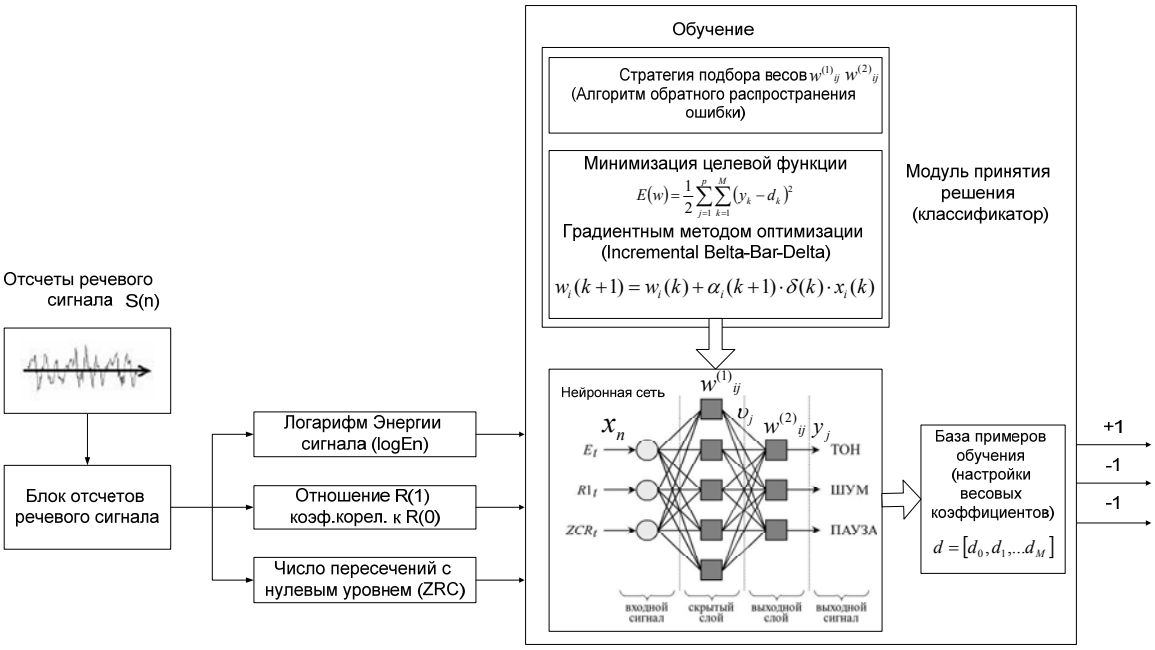


Рис. 3.6. Предлагаемая модель выделения речевых и не речевых участков в речевом сигнале

Классификатор тренируется методом обратного распространения ошибки, последовательно представляя все классы образцов, подлежащих распознаванию. В режиме воспроизведения классифицируемый образ (вектор признаков) подается на выход сети, возбуждая тот выходной нейрон, который соответствует требуемому классу.

Таким образом, на выходе нейросетевого классификатора появляется кодовая комбинация из трёх сигналов [32] (последовательность из трёх чисел +1 и -1), определяющая принадлежность фрейма к одному из классов – тон (+1 -1 -1), шум (-1 +1 -1) или пауза (-1 -1 +1) (рис. 3.6).

Существенным для предлагаемой модели VAD является процесс выбора архитектуры MLP сети, то есть определение минимального количества слоёв сети и нейронов в каждом слое с позиции решения поставленной задачи.

Помимо выбора архитектуры ключевым моментом построения MLP сети является выбор алгоритма коррекции (установления связей между слоями) её весовых коэффициентов.

В данной диссертационной работе предлагается в качестве алгоритма коррекции весовых коэффициентов MLP стратегия обратного распространения ошибки, функционирующая в режиме «он-лайн» [13LeCun,11Осов, 12Хайкин] с использованием адаптивного алгоритма Incremental Delta-Bar-Delta корректировки весов по методу градиентного спуска [15Sutton], определяемые из выражений (3.14)-(3.16) п. 3.2.5.

Отметим, что в отличие от традиционных подходов (метод наискорейшего спуска [12Хайкин, 17Jacobs]), предлагаемый к использованию метод Incremental Delta-Bar-Delta находит распределение параметров скорости обучения по всем входам в каждый нейрон каждого слоя сети, что позволяет ускорить процесс обучения.

Сформированный вектор из трёх признаков – E_t, $R1_t$, ZRC_t компонент – параметров речевого сигнала, вычисляемых на каждом t-м фрейме, $t = 1...T$ является входным сигналом MLP сети.

Пример исходного звукового сигнала, полученного при произнесении диктором-мужчиной английской фразы «At twilight on the twelfth day we'll have Chablis» [30Zue], значения трёх признаков звукового сигнала, по которым проводится классификация, а так же желаемый результат классификации, размеченный вручную на тональные (V – voiced), шумовые (U –unvoiced) и паузные (S – silence) участки, приведёны на рис. 3.7.

Существенным для предлагаемой модели VAD является процесс выбора архитектуры MLP сети, то есть определение минимального количества слоёв сети и нейронов в каждом слое с позиции решения поставленной задачи. При этом количество входных узлов сети равно размеру вектора признаков. Каждый выходной нейрон представляет единственный класс, поэтому количество нейронов выходного слоя выбрано равным числу классов (тон, шум, пауза).

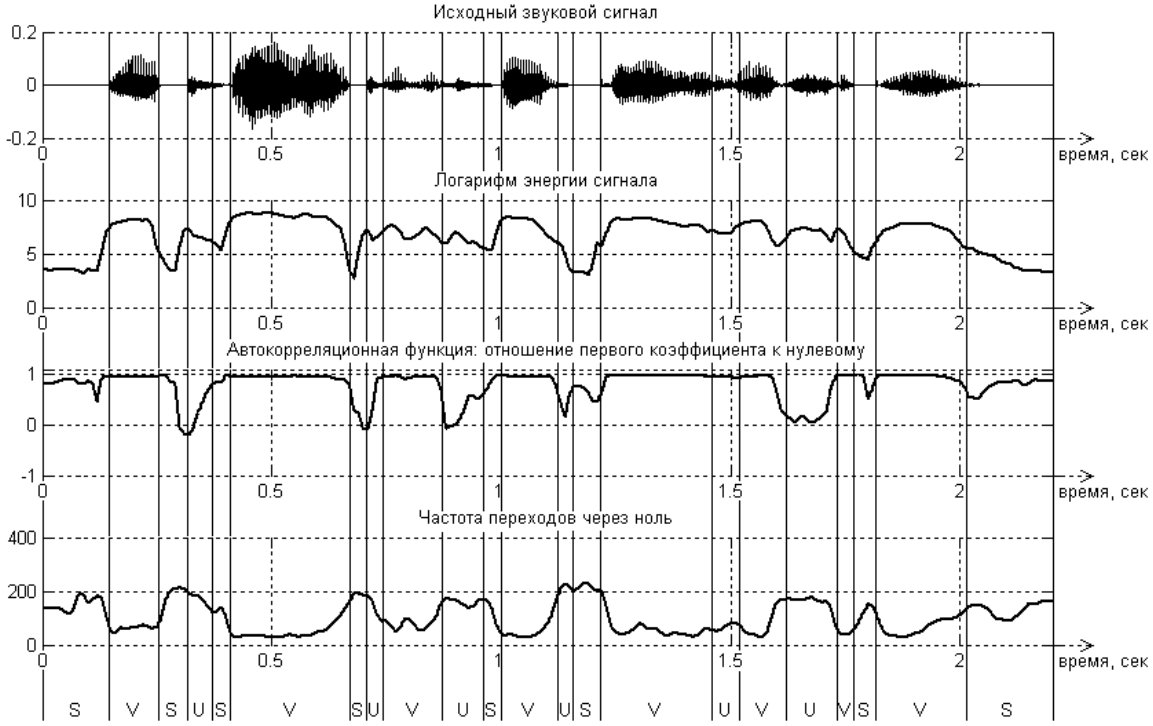


Рис.3.7. Звуковой сигнал и признаки этого сигнала, используемые для классификации

Помимо выбора архитектуры ключевым моментом построения MLP сети является выбор алгоритма коррекции её весовых коэффициентов, то есть установления связей между слоями.

В данной диссертационной работе в качестве алгоритма коррекции весовых коэффициентов MLP сети выбрана, реализуется стратегия обратного распространения ошибки, функционирующая в режиме «он-лайн» [13LeCun,11Осов, 12Хайкин] с использованием адаптивного алгоритма Incremental Delta-Bar-Delta корректировки весов по методу градиентного спуска [15Sutton], определяемые из выражений (3.14)-(3.16) п. 3.2.5.

Отметим, что в отличие от традиционных подходов (метод наискорейшего спуска [12Хайкин, 17Jacobs]), предлагаемый к использованию в нелинейной MLP сети метод Incremental Delta-Bar-Delta находит распределение параметров скорости обучения по всем входам в каждый нейрон каждого слоя сети, что позволяет ускорить процесс обучения.

3.3. Обоснование выбора структуры детектора голосовой активности.

Истинная цель обучения состоит в подборе архитектуры и параметров сети, которые обеспечат минимальную погрешность распознавания тестового подмножества данных, не участвовавших в обучении [11Осов].

Содержание задачи оптимизации заключается в поиске архитектуры оптимальной MLP сети, определяемой такими параметрами:

1) минимальное число скрытых слоёв
2) оптимальное число нейронов в каждом их скрытых слоев;

оптимальный размер обучающей выборки.

Поскольку оптимизация аналитическими методами невозможно ввиду сложности аналитики, в данной работе она проводится экспериментальным путем.

3.3.1. Критерии оптимальности. Обучение модели MLP сети производилось методом обратного распространения ошибки. В общем случае не существует доказательства сходимости этого алгоритма (рис. 3.5), как не существует и какого-либо четкого определенного критерия его останова. Известно лишь несколько обоснованных критериев, которые можно использовать для прекращения корректировки весов [12Хайкин].

Существенным моментом процесса обучения MLP сети является определение наступления стадии излишнего переобучения, при котором теряется свойство обобщения MLP сети – распознавать данные, не участвовавшие в процессе её обучения.

В данной работе предложено останавливать обучение на основе проверки, в ходе которой данные разбиваются на два подмножества – оценивания (обучения – вычисляется ошибка обучения $E_T(w)$ – Training Error) и проверки (контроля – вычисляется ошибка обобщения $E_G(w)$ – Generalization Error). Множество оценивания используется для обучения сети, но сеанс обучения периодически останавливается через каждые K эпох

(в нашем случае 11 эпох). После чего сеть тестируется на проверочном подмножестве. Ошибки вычисляются как среднее на всём множестве входных и выходных сигналов (множество обучения)

$$\{(x(n),d(n))\}_{n=1}^{N} = \{(x(1),d(1)),\ldots,(x(N),d(N))\} \tag{3.17}$$

$$E(w) = \frac{1}{N}\sum_{n=1} E^{n}(w) \tag{3.18}$$

На практике оказывается, что для хорошего обобщения достаточно, чтобы размер обучающего множества N (3.17) удовлетворял соотношению [12Хайкин]:

$$N = O(W/\varepsilon), \tag{3.19}$$

где

W – общее количество весов сети;

ε – допустимая точность ошибки классификации;

$O(\cdot)$ – порядок величины в скобках.

Таким образом, для хорошего качества аппроксимации размер обучающего множества N должен превышать отношение общего количества весов сети W к среднеквадратическому значению ошибки оценивания ε. В экспериментах используется обучающее множество размером 1410379 (1680 файлов) примера, которое превышает общее количество весов рассматриваемых сетей (см. табл. 3.1).

Уже сформированное обучающее множество случайным способом делится на два: собственно обучающее и контрольное для проверки критерия раннего останова. Определение доли примеров, отведённых для контрольного множества, осуществляется на основе информации о числе свободных параметров w_{ji} (весов) обучаемой нейронной сети:

$$r_{opt} = 1 - \frac{\sqrt{2 \cdot W - 1} - 1}{2 \cdot (W - 1)},$$

поэтому архитектура нейросети должна быть уже определена. Доля примеров, отведённых для контрольного множества, задаётся коэффициентом разбиения $r = 0{,}05$. Это значит, что 95% данных обучения составляет подмножество оценивания и только 5% приходится на проверочное подмножество. Размер контрольного множества составляет $0{,}05 \times 1410379 = 70518$. Если в предлагаемом алгоритме останова коэффициент r не задан, доля примеров контрольного множества должна составлять не менее 1% от общего числа примеров. Для определения момента завершения обучении используется критерий раннего останова. Если критерий раннего останова не используется, то обучение завершится по истечению максимально заданного количества эпох обучения

Останавливать обучение в точке минимума кривой тестирования (рис. 3.9) предложено путем сравнения медианного значения буфера ошибок обобщения на текущем шаге $E_G(w)$ с предыдущим значением медианного значения буфера ошибок обобщения $E_{G\_предыдущее=}(w)$:

$$E_G(w) > E_{G\_предыдущее}(w), \quad E_{G\_предыдущее=}(w) = E_G(w) \tag{3.20}$$

В случае, когда ошибка обобщения $E_G(w)$ (Generalization Error) на следующем этапе больше чем на предыдущем выполняется останов обучения. При этом веса w_{ji} сети и Δw_{ji} смещения выбираются для этой минимальной $E_{G\_\min}(w)$ ошибки обобщения.

На каждом шаге значения буфера ошибок сдвигаются на одно значение вперед, а в его конец добавляется новое значение. Экспериментальным путем было установлено, что размер буфера ошибок равный 11 элементов хорошо

подходит для отслеживания изменений ошибок обобщения на текущем и предыдущем шаге.

В режиме тестирования используется 1680 звуковых файлов (воспроизведения) ранее обученная MLP сеть классифицирует образ на тоновые (V), шумовые (U) и паузные (S) участки устной речи (рис. 3.7) и ошибка классификации «тон/шум/пауза» рассчитывалась по следующей формуле:

$$Err = \frac{T_{err}}{T_{all}} \cdot 100\%, \qquad (3.21)$$

где

T_{err} – количество неправильно классифицированных звуковых фреймов;

T_{all} – общее количество звуковых фреймов.

Err –усредненная суммой ошибок классификации по отдельным классам тон, шум и пауза.

3.3.2. Оптимизация архитектуры MLP сети.

В связи с тем, что однозначного метода подбора оптимальной архитектуры сети не существует [12Хайкин,13LeCun], исследования проводились путем компьютерного моделирования и экспериментальной оценки параметров нейронной сети [25LadBond]. При таком подходе наилучшей выбиралась наиболее простая архитектура с минимальной ошибкой классификации (3.21).

Чтобы подобрать оптимальную архитектуру для реализации модели VAD в телефонном канале связи эксперименты были проведены на материалах речевых корпусов TIMIT, NTIMIT и STC-TIMIT следующим образом.

Моделирование предлагаемой модели VAD на основе MLP сети (рис.3.6) осуществлялось при помощи исходных программных кодов

[33БондФедяев] для реализации алгоритма обратного распространения, а так же программной реализации предлагаемого модуля VAD.

Учитывая результаты работ [20Qi, 9Pham] изначально была выбрана MLP сеть с двумя скрытыми слоями (рис. 3.3, табл. 3.1), а после количество слоёв уменьшили до одного (рис. 3.10 2, табл. 3.1).

Размер входного слоя MLP сети соответствует размеру входного вектора признаков $x = [x_0, x_1, ..., x_N]^T$, $N = 3$ а выходного – количеству классов, ожидаемых на выходе $d = [d_0, d_1, ..., d_M]^T$, $M = 3$.

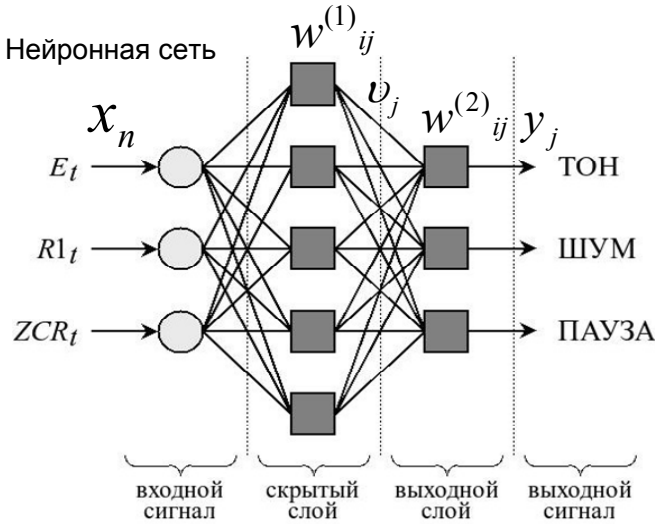


Рис. 3.10. Структура нейросетевого классификатора с пятью нейронами в скрытом слое

Архитектуры всех исследуемых сетей (число слоёв и количество скрытых нейронов в каждом слое) и соответствующее количество свободных параметров, а так же оцененная согласно выражению (3.19) допустимая точность ошибки классификации при фиксированном размере обучающего множества (410379 примера) описаны в табл. 3.1.

Обозначение архитектуры сети 3-100-50-3 говорит о том, что сеть состоит из 4-ёх слоёв, входной слой содержит $N = 3$ нейрона, первый скрытый слой содержит 100 нейронов, второй скрытый слой содержит 50 нейронов, а выходной соответствует количеству классов, ожидаемых на выходе, $M = 3$.

Таблица 3.1

Архитектура сети	W, число весов	ε, % погрешность,
3-100-50-3	5603	$39 \cdot 10^{-4}$
3-20-3	143	10^{-4}
3-40-3	283	$2 \cdot 10^{-4}$
3-60-3	423	$2.9 \cdot 10^{-4}$
3-80-3	563	$4 \cdot 10^{-4}$
3-100-3	703	$5 \cdot 10^{-4}$

В качестве функции активации нейронов используется рациональная сигмоида следующего вида:

$$\varphi(x) = \frac{2 \cdot x}{1 + |x|}. \qquad (3.22)$$

Во-первых, такая функция активации, во-первых, обеспечивает биполярность всех сигналов внутри сети и тем самым повышает эффективность обучения [13LeCun]. Во-вторых, функция (3.14) быстро вычисляется (например, в отличие от биполярной сигмоиды – гиперболического тангенса).

Нейронная сеть обучается с учителем [25BondLad, 33Бондаренко]. Обучающее множество формируется на основе списка речевых сигналов и их временных разметок на тональные, шумовые и паузные участки, выполненных вручную в соответствующем корпусе речевых данных [25BondLad]. Входные сигналы всех обучающих примеров нормализуются так, чтобы мат. ожидание по всем компонентам входного сигнала было нулевым, а среднеквадратичное отклонение – единичным [13LeCun]. Фрагмент подготовленного таким образом обучающего множества,

включающий в себя по два обучающих примера для каждого из классов, приведён в табл. 3.2.

Таблица 3.2 1. Фрагмент обучающего множества

Входной сигнал			Желаемый выходной сигнал		
E_t	$R1_t$	ZCR_t	Тон	Шум	Пауза
1,434	0,609	–1,071	+1	–1	–1
1,274	0,618	–1,158	+1	–1	–1
–0,232	–1,185	1,076	–1	+1	–1
–0,144	–0,981	1,041	–1	+1	–1
–1,939	0,360	1,024	–1	–1	+1
–2,144	–0,240	0,937	–1	–1	+1

В качестве алгоритма обучения используется алгоритм обратного распространения ошибки (backprop), функционирующий в режиме «онлайн» [13LeCun] схема реализуемого в модели VAD алгоритма изображена на рис. 3.11-3.12 7-8.

Существенным преимуществом алгоритма обратного распространения ошибки по сравнению с другими методами более высокого порядка, такими как метод Ньютона, квазиньютоновские методы, метод сопряженных градиентов является то, что последние, при незначительном ускорении процесса сходимости, требуют существенно больше вычислительных ресурсов [12ХайкинС.317-319]. Помимо того алгоритм обратного распространения ошибки является более точным, чем варианты пакетного обратного распространения [34Wilson], и более быстрым, чем алгоритмы глобальной оптимизации, такие как алгоритм имитации отжига или генетические алгоритмы [11Осовский,25БондаЛад].

Одним из основных отличий предлагаемого в качестве алгоритма корректировки весов IDBD метода от классического обратного распространения ошибки методом наискорейшего спуска [12Хайкин, 13LeCun] является то, что коэффициент скорости обучения $\alpha_i(t)$ метода IDBD (12) распределен по всем входам в нейроны в зависимости от знака

предыдущего изменения веса $h_i(t)$ из (14). В свою очередь первое слагаемое выражения (14) $h_i(t)\cdot\left[1-\alpha_i(t+1)\cdot x_i^2(t)\right]^+$ уменьшает значение $h_i(t)$ в сторону нуля. Таким образом, параметр памяти $h_i(t)$ это убывающий «след» (память) кумулятивной суммы предыдущих изменений весов сети w_i из выражения (11).

Итак, предлагаемый алгоритм адаптивной корректировки весов MLP сети при обратном проходе MLP сети схематически изображен на рис. 3.12 состоит из выполнения следующих действий:

1) Инициализация параметров $\beta(t)$ и $h(t)$, определяющих процесс адаптации коэффициентов скорости обучения $\alpha(t)$, соотношения (3.15) и (3.16), для каждого весового коэффициента $w(t)$ нейронной сети. Правила инициализации таковы:

1.2) Начальное значение параметр памяти для каждого входного сигнала $h(t)$ для каждого весового коэффициента нейронной сети всегда равно нулю.

1.3) Поскольку в работе [15Sutton] не указано, каким образом следует выполнять инициализацию обновляемых параметров $\beta(t)$ в данной работе начальное значение $\beta(t)$, для каждого весового коэффициента $w(t)$ нейронной сети, подбирается индивидуально.

Сначала определяется оптимальный начальный коэффициент скорости обучения *rate* для этого весового коэффициента, исходя из того, в каком слое находится нейрон, к которому относится весовой коэффициент, и сколько входов этот нейрон имеет.

Если обозначить $M_i^{сред}$ – среднее число входов в нейрон i для каждого нейрона всех слоёв сети, а $m_i=\sqrt{n}$ – текущее значение количества входов для каждого нейрона каждого слоя, где n – число входных сигналов в текущий нейрон, то оптимальный начальный коэффициент скорости обучения $rate_k$ для текущего слоя k определяется:

$$rate_k = \frac{m\_rate \cdot \sqrt{M_i^{сред}}}{m_i}$$

где m_rate – начальный коэффициент скорости. Затем вычисляется $\beta(t) = \ln(rate)$.

2. При прямом ходе на вход нейросети подаётся входной сигнал из i-го примера обучающего множества, и вычисляются выходные сигналы (3.3) (3.6) и производные выходных сигналов всех нейронов этой нейросети согласно полученным выражениям для локального градиента выходного и произвольного скрытого слоя (3.24)-(3.25). Производные $\varphi_j'(v_j^{(L)}(n))$ и $\varphi_j'(v_j^{(l)}(n))$ выходных сигналов нейронов выходного L и произвольного скрытого l слоёв входят в соотношения для вычисления локальных градиентов узлов сети (3.10) при обратном проходе для уточнения связей (3.11)-(3.12) в MLP сети.

Теперь выходы нейронов последнего скрытого слоя $(l-1)$ становятся входами нейронов выходного слоя. Вычислив выходные сигналы нейронов выходного слоя:

1) цикл по всем входам в нейрон, включая вход смещения

2) пропускаем сигнал через функцию активации

3) считаем разность между реальным и желаемым выходами

4) возводим разность в квадрат и накапливаем,

образующие выходной сигнал нейросети, сразу же вычисляем сумму квадратов разностей (3.4) между полученными y_k на выходе сети сигналами и соответствующими им желаемыми выходными сигналами d_k.

Ошибку обобщения считаем аналогичным образом с отличием в том, что тестовое множество берется согласно установленному в п. 3.2.1. критерию раннего останова. Очередная эпоха обучения выполняется для всех примеров обучающего множества.

3. При обратном проходе алгоритма обратного распространения ошибки, по схеме рис. 3.12, выполняется корректировка весов сети после предъявления очередного примера обучающего множества по алгоритму Incremental Delta Bar Delta (блоки рис. 3.12 выделены серым цветом) при условии, что на сеть распространено входное воздействие $x_i(n)$ из соотношения (3.8) текущего примера обучающего множества.

Для каждого нейрона i выходного слоя K (рис. 3.12):

1) вычисляем сигнал ошибки i-го нейрона K выходного слоя как разность между реальным $o_i(n)$ и желаемым $d_i(n)$ выходами, соотношение (3.9)

$$e_i^K(n) = d_i(n) - o_i(n),$$

2) вычисляем локальный градиент (по формуле для выходного слоя K) из соотношений (3.8) и (3.10):

$$\delta_i^{(K)} = e_i^{(K)}(n) \cdot \varphi_i'(v_i^{(K)}(n)) \qquad (3.26)$$

2) записываем его в соответствующее место массива LocalGradients2 локальных градиентов $\delta_i^{(K)}$;

3) корректируем параметр $\beta_i(t)$ для смещения нейрона i

$$\Delta\beta_i(t) = \theta \cdot \delta_i^{(K)} \cdot y_j^{K-1}(t) \cdot h_i(t), \qquad \beta_i(t+1) = \beta_i(t) + \Delta\beta_i(t),$$

где $Y^K = \{y_1^k, y_2^k, ..., y_{N_k}^k\}$ – выходной сигнал $k = K-1$-го слоя (рис. 3.11), который является входным для $k = K$ слоя ;

Ограничиваем $\beta_i(t)$ так, чтобы изменения $\beta_i(t)$ на одном шаге составляли $\beta_i(t) = \pm 2$. Если $\beta_i(t) \geq 2$, то устанавливаем $\beta_i(t) = 2$. Если $\beta_i(t) \leq -2$, то устанавливаем $\beta_i(t) = -2$.

4) на основе $\beta_i(t)$ вычисляем для смещения нейрона $\alpha_i(t) = e^{\beta_i(t)}$;

5) корректируем смещение нейрона

$$w_i(t+1) = w_i(t) + \alpha_i(t) \cdot \delta_i^{(K)} \cdot y_j^{K-1}(t),$$

где $y_j^{(K-1)}(n)$ – входной сигнал нейрона i, предыдущего слоя $(K-1)$, с индексами нейронов j.

$$y_j^{(K-1)}(t) = \varphi_j(v_j^{(K-1)}(t)), \qquad v_j^{(K-1)}(t) = \sum_{i=0}^{m_0} w_{mj}^{(K-1)}(t) y_m^{(K-2)}(t),$$

где $y_m^{(K-2)}(t)$ – выходной сигнал нейрона m $(K-2)$ слоя.

Вычисляем новое значение параметра $h_{ji}(t)$ для связи между i-м нейроном K-го выходного слоя и j-м нейроном $(K-1)$-го слоя:

$$h_{ji}(t+1) = 1 - \alpha_{ji}(t+1) \cdot y_j^{(K-1)}(t),$$

Если $h_{jk}(t+1) \leq 0$, то

$$h_{jk}(t+1) = \Delta w_{ji}(t) = \alpha_i(t+1) \cdot \delta_i^{(K)}(t) \cdot y_j^{K-1}(t),$$

Иначе

$$h_i(t+1) = h_i(t) \cdot \left[1 - \alpha_i(t+1) \cdot x_i^2(t)\right]^+ + \alpha_i(t+1) \cdot \delta(t) \cdot x_i(t),$$

где

$x_i(t) = y_i^{(K-1)}(t)$ – входные сигналы в этот нейрон;

$\alpha_i(t+1)$ – вычисленный ранее, индекс $t+1$ коэффициент скорости обучения для смещения нейрона

Далее, чтобы распространить корректировку весов по методу IDBD для всех нейронов всех скрытых слоёв нелинейной MLP сети по методу обратного распространения ошибки реализуется цикл по всем скрытым слоям от последнего до первого, таким образом:

Для каждого нейрона i-го скрытого слоя (цикл по j от 1 до N_{k+1}):

1) с помощью обратного распространения вычисляем ошибку j нейрона скрытого слоя (рис. 3.12)

$$e = \delta^{(K)} \cdot w_{jk}$$

где

$\delta^{(K)}$ – значения локальных градиентов LocalGradients2, вычисленных из выражения (3.26) для нейронов выходного слоя;

$\sum_i e_i^K(n)$ – суммарный сигнал ошибки i-го нейрона K выходного слоя

w_{jk} – вес связывающий нейрон k выходного слоя K и нейрон j $K-1$ слоя.

Накапливаем ошибку для всех нейронов по всем слоям $e = e + \delta^{(K)} \cdot w_{jk}$ согласно:

$$e_j(n) = -\sum_k e_k(n) \cdot \varphi_k'(v_k(n)) \cdot w_{jk}(n) = -\sum_k \delta^{(K)}(n) \cdot w_{jk}(n), \qquad (3.27)$$

где

$\delta^{(K)}(n)$ – требует значения сигналов ошибки $e^{(K)}(n)$ для всех нейронов слоя, находящегося правее скрытого нейрона j, связанных с ним (в данном случае это выходной слой K).

$w_{jk}(n)$ – веса этих связей, т.е. вес между k-м нейроном $(i+1)$-го слоя и j-м нейроном i-го слоя;

2) вычисляем локальный градиент для нейрона j по формуле для скрытого слоя (3.10)

$$\delta_j^{(i)} = \varphi_j'(v_j)\sum_k \delta_k(n)\cdot w_{jk}(n) = \varphi_j'(v_j^{(i)})\cdot e_j(n),$$

где $e_j(n)$ определяется согласно (3.27).

Записываем его в соответствующее место массива локальных градиентов LocalGradients1;

4) корректируем веса связей по методу IDBD между данным нейроном и всеми нейронами следующего, $(i+1)$-го слоя (теперь эти веса уже можно менять):

4.1) вычисляем новое значение параметра $\beta_{jk}(t+1)$ для связи между j-м нейроном i-го слоя и k-м нейроном $(i+1)$-го слоя:

$$\Delta\beta_{jk}(t) = \theta\cdot\delta^{(K)}\cdot x_i(t)\cdot h(t), \qquad (3.28)$$

где

$\delta^{(K)}$ – значения локальных градиентов LocalGradients2, вычисленных из выражения (3.26) для нейронов выходного слоя;

$x_i(t)$ – i-й входной сигнал в k-й нейрон;

Корректируем $\beta_k(t+1) = \beta(t) + \Delta\beta_{jk}(t)$ и ограничиваем $\beta_k(t+1)$ так, чтобы изменения на одном шаге составляли $\beta(t) = \pm 2$.

4.2) вычисляем коэффициент скорости обучения для веса этой межнейронной связи; $\alpha_{jk}(t) = e^{\beta_{jk}(t)}$

4.3) корректируем вес межнейронной связи, вычисляем новое значение веса связи между j-м нейроном i-го слоя и k-м нейроном (i+1)-го слоя:

$$\Delta w = \alpha_{jk}(t) \cdot \delta^{(K)}(t) \cdot x_i(t). \qquad (3.29)$$

где

$\delta^{(K)}$ – значения локальных градиентов LocalGradients2, вычисленных из выражения (3.26) для нейронов выходного слоя;

$x_i(t)$ – i-й входной сигнал в k-й нейрон или выходной сигнал j-го нейрона $y_j(t)$;

$$w(t+1)_{jk} = w\ (t) + \Delta w \qquad (3.30)$$

4.4) корректируем параметр $h(t)$ для межнейронной связи, вычисляем новое значение параметра $h(t)$ для связи между j-м нейроном i-го слоя и k-м нейроном $(i+1)$-го слоя;

$$h_{jk}(t+1) = 1 - \alpha_{jk}(t+1) \cdot y_j^2(t), \qquad (3.31)$$

Если $h_{jk}(t+1) \le 0$, то

$$h_{jk}(t+1) = \Delta w_{ji}(t) = \alpha_i(t+1) \cdot \delta_i^{(K)}(t) \cdot y_j^{K-1}(t), \qquad (3.32)$$

Иначе

$$h_i(t+1) = h_i(t) \cdot \left[1 - \alpha_i(t+1) \cdot x_i^2(t)\right]^+ + \alpha_i(t+1) \cdot \delta(t) \cdot x_i(t), \quad (70)$$

5) снова корректируем параметр $\beta(t)$ для смещения k-го нейрона аналогично выражению (3.28)

$$\Delta\beta_{jk}(t) = \theta \cdot \delta^{(K)} \cdot x_i(t) \cdot h(t)$$

$$\beta_k(t+1) = \beta\ (t) + \Delta\beta_{jk}(t)\,,$$

и ограничиваем $\Delta\beta_{jk}(t)$ так, чтобы изменения на одном шаге составляли $\Delta\beta_{jk}(t) = \pm 2$.

6) вычисляем коэффициент скорости обучения для смещения нейрона $\alpha_{jk}(t) = e^{\beta_{jk}(t)}$;

7) корректируем смещение данного нейрона аналогично (3.29)-(3.30);

8) корректируем параметр $h_i(t)$ для смещения нейрона аналогично (3.31)-(3.32).

На следующей итерации осуществится переход на слой назад (рис.3.12), и локальные градиенты нейронов текущего слоя LocalGradients1 станут локальными градиентами следующего слоя LocalGradients2. Меняем местами массивы LocalGradients1 и LocalGradients2 локальных градиентов текущего и следующего слоев. Когда в массиве LocalGradients2 содержатся локальные градиенты для нейронов первого слоя $\delta^{(1)}$, необходимо скорректировать веса всех нейронов первого слоя, и на этом завершить обратный проход. Для этого реализуются аналогичные вычисления:

1) цикл по нейронам 1го слоя и по входам в каждый нейрон. Ввычисляем новое значение параметра $\beta(t)$ для связи между k-м входом и j-м нейроном 1-го слоя, ограничиваем $\Delta\beta_{kj}(t)$ так, чтобы изменения на одном шаге составляли $\Delta\beta_{kj}(t) = \pm 2$, и корректируем

$$\beta_{kj}(t+1) = \beta\ (t) + \theta \cdot \delta^{(1)} \cdot x_i(t) \cdot h_{kj}^{(1)}(t)$$

2) на основе нового значения $\beta_{kj}(t+1)$ вычисляем коэффициент скорости обучения $\alpha_{kj}(t) = e^{\beta_{kj}(t)}$,

3) вычисляем новое значение веса связи между k-м входом и j-м нейроном 1-го слоя $w(t+1)_{jk} = w(t) + \alpha_{kj}(t) \cdot \delta^{(1)}(t) \cdot x_i(t)$, где $\delta^{(1)}(t)$ – соответствующий локальный градиент, участвующий в корректировке веса первого слоя.

4) вычисляем новое значение параметра h_j для связи между k-м входом и j-м нейроном 1-го слоя:

$$h_j(t+1) = h_j(t) \cdot \left[1 - \alpha_{jk}(t+1) \cdot x_i^2(t)\right]^+ + \alpha_{jk}(t+1) \cdot \delta^{(1)}(t) \cdot x_i(t)$$

На этом обратный проход завершен. Предложенный алгоритм корректировки весов сети выполняется до выполнения условий критерия раннего останова, описанного в п. 3.2.1. После выполнения останова алгоритма обучения, весовые коэффициенты сети фиксируются, и данная архитектура MLP сети с оптимизированными весовыми коэффициентами выбирается в качестве модели принятия решения.

Отметим, что существенной особенностью предложенного адаптивного алгоритма Incremental Delta-Bar-Delta корректировки весов для многослойной нелинейной MLP сети [15Sutton] заключается в (новизна) его применении для разработанной нелинейной MLP сети (рис. 3.1), в отличие от работы [15Sutton], где IDBD описан и экспериментально исследован для случая единственного линейного нейрона;

В рамках решаемой задачи VAD для предложенного алгоритма адаптивной корректировки весов MLP сети необходимо экспериментально на реальных речевых сигналах оптимизировать параметры согласно выбранным в п. 3.2.1 критериям.

Поскольку метод IDBD содержит один управляемый параметр скорости адаптации θ изменяемых для каждого веса MLP сети параметров $\beta_i(t)$, выражение (3.15):

$$\beta_i(t+1) = \beta_i(t) + \theta \cdot \delta(t) \cdot x_i(t) \cdot h_i(t),$$

поэтому оптимизация параметров алгоритма IDBD заключается в поиске такого параметра скорости адаптации θ, при котором среднеквадратическая ошибка обучения модели $E_T(w)$ будет наименьшей.

Экспериментальные исследования в целом подобны таковым, как описано в п. 3.2.2 укажем лишь отличительные особенности. реализованный в п. 3.2.2. алгоритм тестируется с различными параметрами адаптации θ. Согласно установленным критериям п. 3.2.1 определяется среднеквадратические ошибки обучения для модели MLP сети с архитектурой i3_40sig_3sig. Останов алгоритма происходил по истечению максимально установленного числа эпох обучения, в данном случае 100 эпох обучения было выбрано для исследования ошибки обучения в зависимости от параметра адаптации θ.

При качественных исследованиях анализируются зависимости среднеквадратической ошибки обучения $E_T(w)$ для архитектуры MLP сети i3_40sig_3sig (3 входных, 40 скрытых и 3 выходных нейрона) в зависимости от регулируемого параметра скорости адаптации θ.

На рис. 3.13- 3.14 приведены графики ошибок обучения $E_T(w)$ в течение 100 эпох обучения MLP сети для различных значений параметра скорости адаптации $\theta = \{0{,}0001 \quad 0{,}001 \quad 0{,}001 \quad 0{,}01 \quad 0{,}5 \quad 0{,}9\}$, первое из которых $\theta = 0{,}0001$ было взято за основу для тестирования алгоритма IDBD в работе [15Sutton].

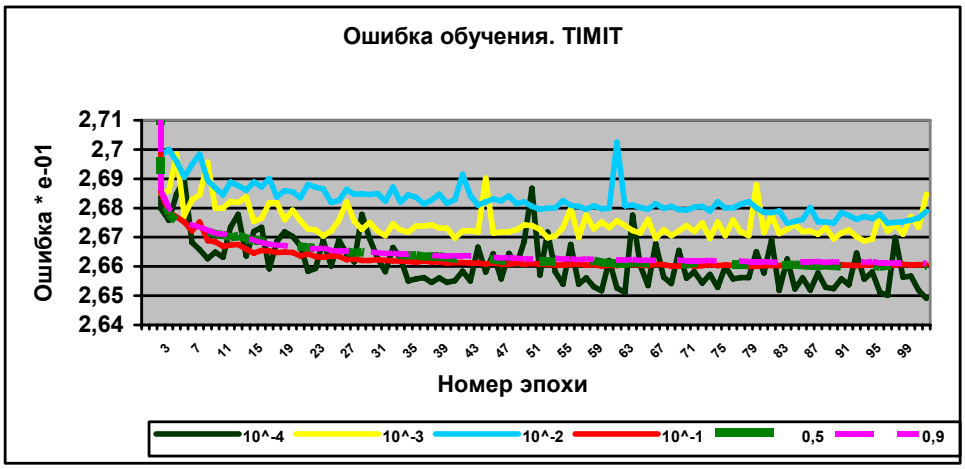


Рис. 3.13 $E_T(w)$ для различных параметров скорости адаптации θ. TIMIT

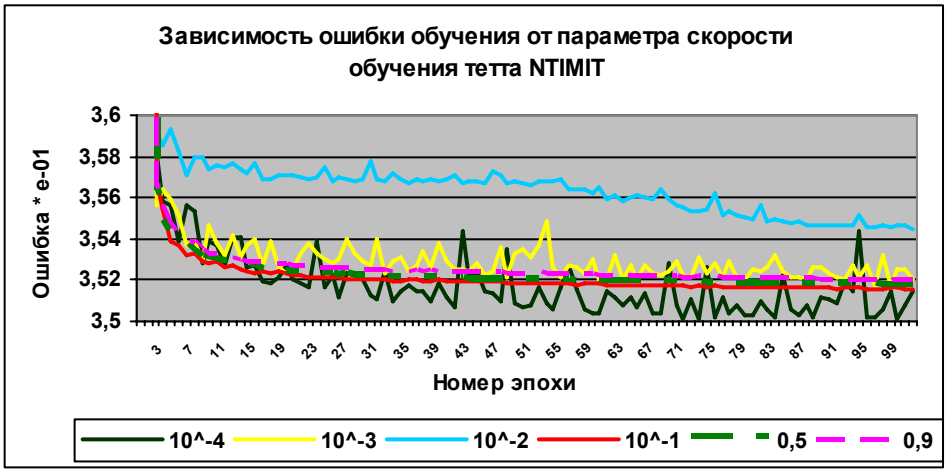


Рис. 3.14 $E_T(w)$ для различных параметров скорости адаптации θ. NTIMIT

Как видим из полученных экспериментальным путем графиков рис. рис. 3.13 - 3.14, что оптимальным значением θ при построении модели TIMIT (чистый речевой сигнал) является $\theta = 0{,}0001$, при котором ошибка обучения на 100 эпохе обучения составляет $E_T(w) = 26{,}5\%$ для речевых данных корпуса TIMIT и $E_T(w) = 35{,}1\%$.

Сопоставляя зависимости для различных параметров адаптации θ (рис. 3.13 - 3.14) ошибок $E_T(w) = 26{,}5\%$ видно, что значения $E_T(w)$ график для $\theta = 0{,}0001$ существенно флюктуирует относительно средних значений.

Аналогичные флюктуации наблюдаются для графиков $\theta = 0{,}001$ и $\theta = 0{,}01$ (рис. 3.13 - 3.14).

Зависимости ошибки обучения $E_T(w)$ для графиков $\theta = \{0{,}1 \quad 0{,}5 \quad 0{,}9\}$ превышают эту ошибку при $\theta = 0{,}0001$. Тем не менее, $E_T(w)$ для $\theta = \{0{,}1 \quad 0{,}5 \quad 0{,}9\}$ практически неизменны для всех эпох обучения. Поэтому при экспериментальных исследованиях было принято $\theta = 0{,}0001$.

Для того чтобы выбрать оптимальную архитектуру MLP сети во-первых необходимо определить алгоритм корректировки её весов (параметров, при которых среднеквадратическая ошибка будет минимальной). Поэтому проведенные далее количественные эксперименты можно разделить на две группы. Эксперименты первой группы носили предварительный характер для определения алгоритма корректировки весов, а эксперименты второй группы проводились с целью определения оптимальной архитектуры MLP сети.

Рассмотрим сначала результаты первой группы экспериментов, главной целью которых было выяснение существования эффективного алгоритма корректировки весов, в смысле средней ошибки классификации Err,% «тон/шум/пауза». В ходе экспериментов исследуются два метода:

1) классический backprop или метод наискорейшего спуска (steepest descent) [12Хайкин,13LeCun]. В данном случае η коэффициент скорости обучения одинаков для всех весов сети и не меняется в процессе обучения. Определяется согласно выражениям (3.5) и (3.12);

2) метод Incremental Delta-Bar-Delta [15Sutton] (IDBD). Коэффициент скорости обучения $\alpha_i(t)$ автоматически подбирается для каждого настраиваемого веса и изменяется в процессе обучения, соотношения (3.15)-(3.16).

Рассмотрим сначала вопрос, связанный с выбором метода корректировки весов MLP сети в процессе её обучения.

Известно, что в классическом алгоритме обратного распространения ошибки методом наискорейшего спуска изменение параметра скорости обучения η в сторону его уменьшения приводит к меньшей корректировке весов на каждой итерации, выражение (3.5) [12Хайкин]:

$$w(n+1) = w(n) + \eta \cdot \nabla E(w),$$

Такое улучшение замедляет процесс обучения. В случае увеличения параметра η система может перейти в неустойчивое состояние («осциллированию» вектора весов w), что не позволит определить их оптимальных значения.

Принципиально важным моментом корректировки весов является медленное их изменение (малое значение коэффициента обучения η) возле точки оптимума (минимума на поверхности ошибок $E(w)$ в пространстве весов сети w) и быстрое их изменение вдали от точки оптимума (большие значения η) для увеличения скорости сходимости алгоритма.

Таким образом, ключевыми моментами правильной корректировки весов в алгоритмах градиентного спуска являются выбор:

1) направления корректировки (увеличение или уменьшение параметра скорости обучения)
2) величины коррекции текущего веса Δw (амплитуда Δw изменения веса).

В связи с этим целесообразно проверить насколько предлагаемый метод Incremental Delta-Bar-Delta реализованный для MLP сети по методу градиентного спуска выигрывает по сравнению с классическим методом наискорейшего спуска, согласно выбранным в п. 3.2.1 критериям.

К сожалению, с помощью аналитических исследований невозможно однозначно ответить на поставленный вопрос, поскольку в данной работе метод корректировки весов Incremental Delta-Bar-Delta используется как

часть нелинейной системы, MLP сети, работающей в качестве алгоритма принятия решения в модуле VAD (рис. 3.6). Следствием такого подхода является то, что для адекватной оценки качества работы алгоритма Incremental Delta-Bar-Delta необходимо использовать сквозной показатель $Err,\%$, характеризующий среднюю ошибку классификации «тон/шум/пауза» предлагаемой модели VAD.

Таким образом, исследования проводились путём компьютерного моделирования, а результаты экспериментов оценивались качественно (путем анализа зависимостей среднеквадратической ошибки обучения для предложенных моделей MLP сети из табл. 3.1) и количественно путём оценивания показателя $Err,\%$.

Организация экспериментов исследований в целом подобна таковой, как описано в п. 3.2.2. Укажем лишь отличительные особенности.

В классическом методе наискорейшего спуска использовалось постоянное значение параметра скорости обучения $\eta = 0{,}001$ аналогичное для управляющего параметра скорости-скорости обучения метода IDBD $\theta = 0{,}001$. Корректировка, применяемая к весу w_k, определяется согласно выражению (3.11) или в общем виде без указания слоя:

$$\Delta w_k = -\eta_k \cdot p_k, \qquad w_{k+1} = w_k + \Delta w_k \tag{3.32}$$

Для поиска оптимальной архитектуры в методе IDBD использовались значения параметров алгоритма с учетом работы [15Sutton], при этом $\theta = 0{,}001$.

При качественных исследованиях анализируются зависимости среднеквадратической ошибки обучения для предложенных моделей MLP сети из табл. 3.1.

На рис. 3.14 - 3.14 приведены графики ошибок обобщения $E_G(w)$ для различных архитектур MLP сетей (см. табл. 3.1), обученных на корпусах

TIMIT, NTIMIT в зависимости от оптимального числа эпох обучения MLP сети с использованием классического алгоритма [12Хайкин, 13LeCun] корректировки весов, выражение (3.32), метода наискорейшего спуска (steepest descent).

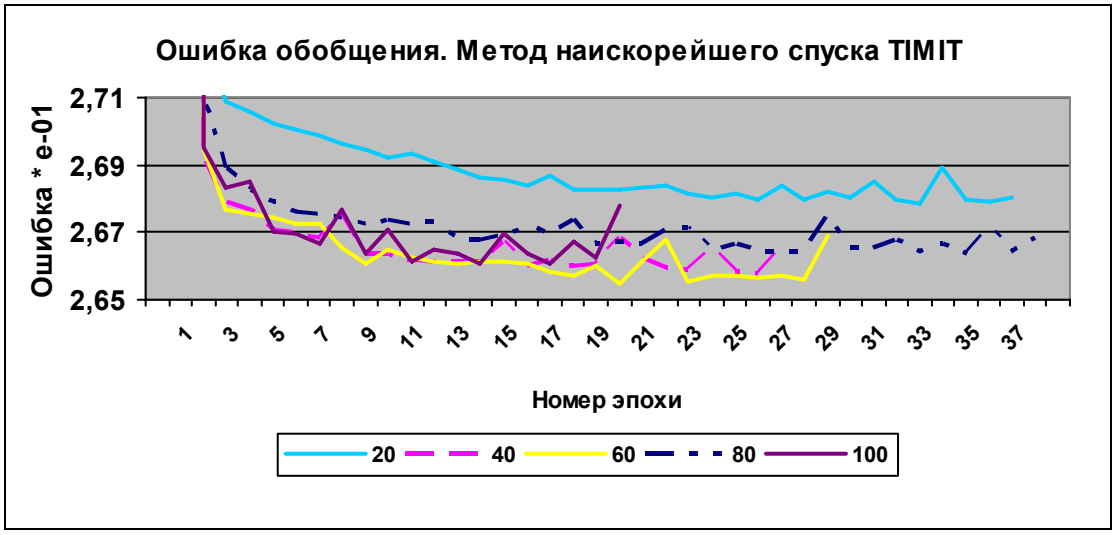


Рис. 3.14 Кривые $E_G(w)$ для различных архитектур сети TIMIT

Как видим из рис. 3.14 (модель TIMIT), что сеть с максимальным количеством скрытых нейронов (100 скрытых нейронов) 3i-100sig-3sig не является оптимальной, в смысле ошибки обобщения $E_G(w) = 26{,}8\%$ при 19 эпохе обучения. Модель 3i-40sig-3sig при меньшем числе скрытых нейронов (40 нейронов) и при увеличении числа эпох обучения до 28 имеет ошибку обобщения $E_G(w) = 26{,}6\%$.

Сопоставляя зависимости (рис. 3.14) ошибок $E_G(w)$ между моделями с архитектурами из табл. 3.1, видно, что при числе скрытых нейронов 20 (модель 3i-20sig-3sig) обучаемая MLP обладает наихудшей обобщающей способностью с ошибкой обобщения $E_G(w) = 26{,}8\%$ (36 эпоха обучения). При этом ошибки обобщения для моделей 3i-60sig-3sig и 3i-80sig-3sig соответственно составляют $E_G(w) = 26{,}7\%$ (для 28 эпохи) и $E_G(w) = 26{,}7\%$ (для 37 эпохи).

Анализируя зависимости ошибок $E_G(w)$ рис. 3.15, модели NTIMIT следует, что сеть со 100 скрытыми нейронами (3i-100sig-3sig) является оптимальной, в смысле ошибки обобщения $E_G(w) = 35{,}1\%$ при 27 эпохе обучения. Однако сеть с архитектурой 3i-80sig-3sig имеет наибольшую ошибку обобщения $E_G(w) = 35{,}4\%$ при максимальном значении эпох обучения – 37 эпох. Быстрее всего (17 эпох) обучается модель 3i-60sig-3sig при ошибке $E_G(w) = 35{,}3\%$.

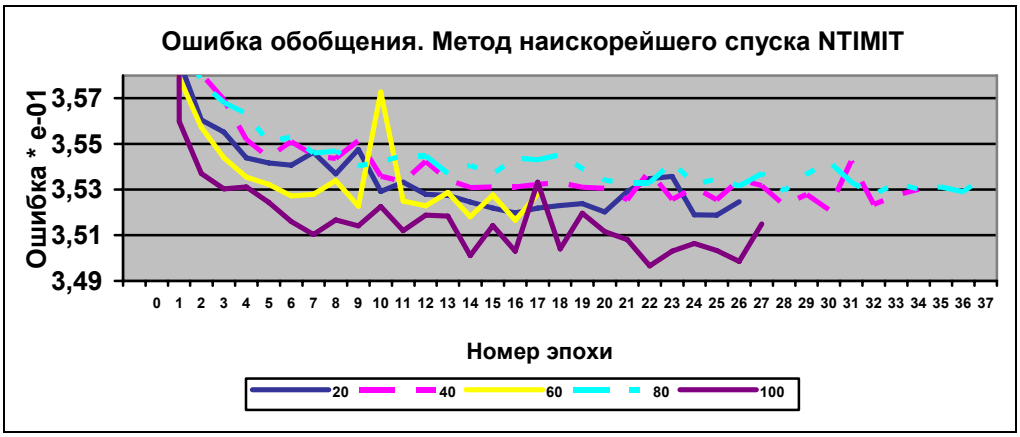


Рис. 3.15. Кривые $E_G(w)$ для различных архитектур сети NTIMIT

Сопоставляя зависимости ошибок $E_G(w)$ рис. 3.15 между моделями с архитектурами 3i-40sig-3sig и 3i-100sig-3sig видно, что разница между значениями $E_G(w)$ на последних эпохах обучения составляет около 0,002, поэтому в качестве оптимальной модели TIMIT используем модель с наименьшим количеством нейронов скрытого слоя (40sig). Поскольку для случая модели TIMIT оптимальной выбрана архитектура 3i-40sig-3sig, выберем её в качестве оптимальной и для модели NTIMIT.

Чтобы наглядно сравнить два метода корректировки весов (классический метод наискорейшего спуска и адаптивный алгоритм Incremental Delta-Bar-Delta) по экспериментальным графикам рис. 3.14-3.15 приведем сводную табл. 3.2 общих результатов для каждого из корпусов данных TIMIT и NTIMIT.

Использовались такие обозначения для различных методов корректировки весов: η – соответствует классическому методу наискорейшего спуска, а $\alpha(t)$ – методу Incremental Delta-Bar-Delta, $E_G(w)$ – соответствующая ошибка обобщения на последней эпохе.

Таблица 3.2. Метод наискорейшего спуска – $\eta = 0{,}001$, IDBD – $\theta = 0{,}001$

Архитектура	TIMIT				NTIMIT			
	η, эпох	$E_G(w)$,%	$\alpha(t)$, эпох	$E_G(w)$,%	η, эпох	$E_G(w)$,%	$\alpha(t)$, эпох	$E_G(w)$,%
3i-20sig-3sig	36	26,8	31	26,9	26	35,2	18	35,9
3i-40sig-3sig	26	26,7	24	27,1	34	35,3	21	35,3
3i-60sig-3sig	28	26,7	25	26,9	17	35,3	29	35,5
3i-80sig-3sig	37	26,7	24	26,7	37	35,4	26	35,5
3i-100sig-3sig	19	26,8	26	26,5	27	35,1	17	35,3

Как следует из приведенных результатов табл. 3.2 и рис. 3.16-3.17 следует, что корректировка весов MLP сети по методу Incremental Delta-Bar-Delta существенно сокращает количество эпох обучения модели MLP сети. Лишь в одном случае для TIMIT (модель 3i-100sig-3sig) и NTIMIT (модель 3i-60sig-3sig) экспериментов количество эпох обучения для метода Incremental Delta-Bar-Delta превышает число эпох обучения для классического метода наискорейшего спуска.

Результаты экспериментальных исследований исследуемых моделей MLP сети в виде распределения ошибок классификации «тон/шум/пауза» по отдельным классам при обучении и тестировании классификаторов на различных корпусах в виде средней ошибки классификации по всему тестовому множеству звуковых данных приведены в приложении В, табл. В.3.1-В.3.9.

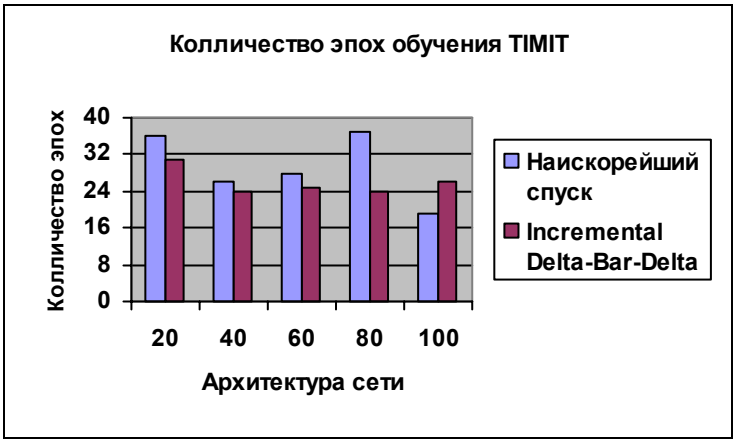


Рис. 3.16. Количество эпох обучения в различных экспериментах TIMIT

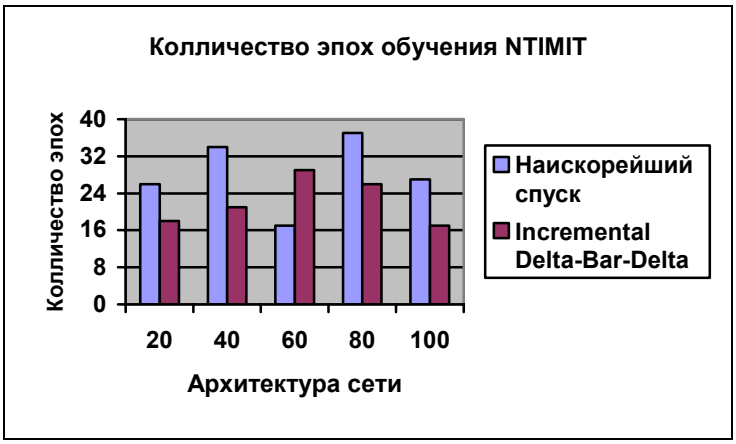


Рис. 3.17. Количество эпох обучения в различных экспериментах NTIMIT

Чтобы наглядно продемонстрировать распределение ошибок классификации «тон/шум/пауза» по экспериментальным данным корпусов TIMIT и NTIMIT и сравнить результаты тестирования моделей, построенных на основе двух методов корректировки весов сети в табл. 3.3 введены следующие обозначения: η – соответствует классическому методу наискорейшего спуска, а $\alpha(t)$ – методу Incremental Delta-Bar-Delta.

В табл. 3.3 представлены результаты распределение ошибок классификации «тон/шум/пауза» по экспериментальным данным корпуса NTIMIT при корректировке по классическому методу наискорейшего спуска.

Таблица 3.3. Обучение модели TIMIT и тестирование TIMIT.

Архитектура	Ошибка выделения тона, %		Ошибка выделения шума, %		Ошибка выделения паузы, %		Средняя ошибка, %	
	η	$\alpha(t)$	η	$\alpha(t)$	η	$\alpha(t)$	η	$\alpha(t)$
3i-20sig-3sig	6,1	6,3	1,1	4,2	21,2	5,9	28,4	16,4
3i-40sig-3sig	6,1	6,2	0,9	4,1	22,0	5,8	29,0	16,1
3i-60sig-3sig	6,2	6,1	1,1	4,3	21,1	5,7	28,3	16,1
3i-80sig-3sig	6,1	6,1	1,1	4,2	21,5	6,0	28,7	16,3
3i-100sig-3sig	6,6%	6,2	0,9%	4,3	20,8	5,6	28,3	16,1
3i-100sig-50sig-3sig	-	4,6	-	4,2	-	5,7	-	16,1

Таблица 3.4. Обучение модели NTIMIT и тестирование NTIMIT

Архитектура	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
3i-20sig-3sig	7,12	5,08	9,87	22,07
3i-40sig-3sig	7,09	4,64	10,37	22,09
3i-60sig-3sig	7.07	5,21	9,57	21,86
3i-80sig-3sig	7,32	4,86	9,65	21,84
3i-100sig-3sig	7,1	5,16	9,94	22,21

Из результатов табл. 3.3-3.4 видно, что средняя ошибка классификации для случая NTIMIT- NTIMIT составляет порядка 22%, в то время как для TIMIT- TIMIT – 28%, что обусловлено меньшей ошибкой выделения паузы для случая NTIMIT- NTIMIT (порядка 10%), чем для TIMIT- TIMIT – 21%.

В табл. 3.5. представлены результаты сравнения алгоритмов наискорейшего спуска и метода адаптивной корректировки весов Incremental Delta-Bar-Delta для архитектуры модели 3i-40sig-3sig.

Таблица 3.5. Обучение модели 3i-40sig-3sig NTIMIT и тестирование NTIMIT

Алгоритм корректировки весов	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
Наискорейшего спуска	7,09	4,64	10,37	22,09
IDBD	6,98	5,13	9,85	21,96

Очевидно, что коррекция весов MLP сети по адаптивному методу Incremental Delta-Bar-Delta приводит к меньшим ошибкам «тон/шум/пауза», а так же средняя ошибка классификации снижается с 22,09% до 21,96%.

Как следует из табл. 3.3-3.4 для метода наискорейшего спуска:

1. для модели TIMIT (табл. 3.3) ошибка выделения пауз составляет наибольшее значение 21%. При этом наименьшей является ошибка выделения шума 1%.
2. для модели NTIMIT (табл. 3.4) ошибка выделения шума повысилась до 5% по сравнению с моделью для TIMIT (1%), однако ошибка выделения паузы уменьшилась с 21% TIMIT до 10% NTIMIT.
3. для архитектуры (табл. 3.3) с двумя скрытыми слоями ошибка выделения тона наименьшая 4,56%, однако, средняя ошибка классификации не изменилась по сравнению с более простыми архитектурами MLP сети.

В табл. 3.3 и 3.4 сопоставлены результаты работы алгоритмов корректировки весов MLP сети. Очевидно, что коррекция по методу IDBD (применение адаптивного алгоритма корректировки весов, обозначение $\alpha(t)$) привела к положительным результатам, средняя ошибка классификации Err,% уменьшилась с 28% до 16%. Такое понижение ошибки Err,% обусловлено уменьшением ошибки классификации пауз с 21% до 6% и небольшим повышением ошибки выделения шума с 1% до 4%.

Исходя из полученных результатов распределения ошибок классификации «тон/шум/пауза» (табл. 3.3, 3.4), а так же средней ошибки классификации для различных архитектур моделей (табл. 3.1), обученных и тестируемых в условиях телефонного канала связи NTIMIT, заключаем, что корректировку весов MLP сети следует проводить по методу IDBD. Объяснить данный результат можно тем, что фиксированный параметр скорости обучения (метод наискорейшего спуска) подходит не для всех областей поверхности ошибок [12Хайкин]. Если производная функции стоимости по отдельному весу на нескольких последовательных итерациях имеет один и тот же знак, то значение параметра скорости обучения для данного веса должно увеличиваться. В противном случае, если производная имеет разные знаки, то значение скорости обучения для данного веса должно уменьшаться. Такие особенности поведения локальных градиентов позволяет учесть адаптивный метод IDBD.

Введение зависимости параметра обучения от времени конкретного веса приводят к фундаментальному изменению алгоритма обратного распространения. Таким образом, алгоритм IDBD уже не осуществляет поиск методом наискорейшего спуска. Корректировка конкретных весов основывается на частных производных поверхности ошибок по конкретным весам и на оценке кривизны поверхности ошибок в текущей точке относительно конкретных изменений пространства весов.

Рассмотрим теперь результаты второй группы экспериментов по выбору оптимальной архитектуры MLP сети. Целью этих экспериментов было определить оптимальную архитектуру MLP сети, в смысле:

1) критерия корректной классификации п. 3.2.1;
2) минимального количества эпох обучения;
3) минимального числа скрытых слоёв и нейронов в них;
4) исследование выбранной архитектуры модели, построенной на адаптационных данных из телефонного канала связи NTIMIT на корпусах TIMIT, NTIMIT, STC-TIMIT.

Для более детального анализа насколько предложенная модель VAD надёжна, т. е. насколько снижается точность её работы, выбранная MLP сеть, тестировалась в условиях шума двух телефонных каналов связи NTIMIT и STC-TIMIT [35Morales, 36Jankowski].

Перейдем к вопросу поиска оптимальной архитектуры MLP сети. На графиках рис. 3.18 - 3.20 представлены кривые обучения и обобщения в зависимости от оптимального числа эпох по критерию раннего останова для обучения моделей с архитектурой сети 3i-100sig-50sig-3sig (табл. 3.1) на корпусах данных TIMIT, NTIMIT и STC-TIMIT.

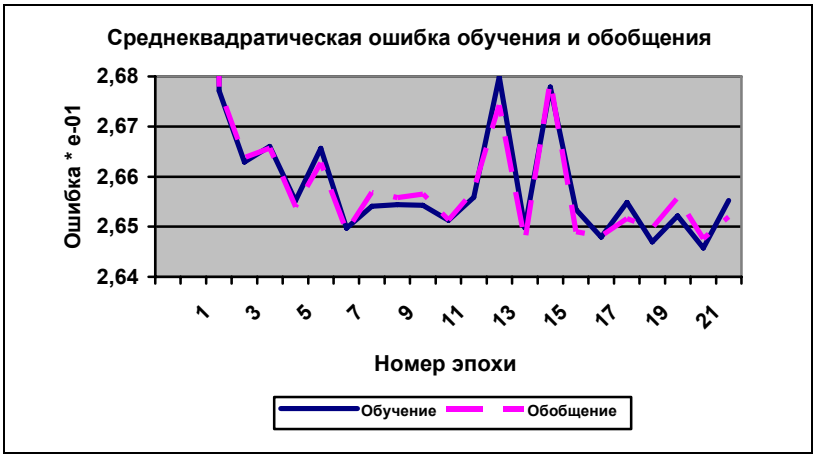


Рис. 3.18. $E_T(w)$ и $E_G(w)$, архитектура 3i-100sig-50sig-3sig, TIMIT,

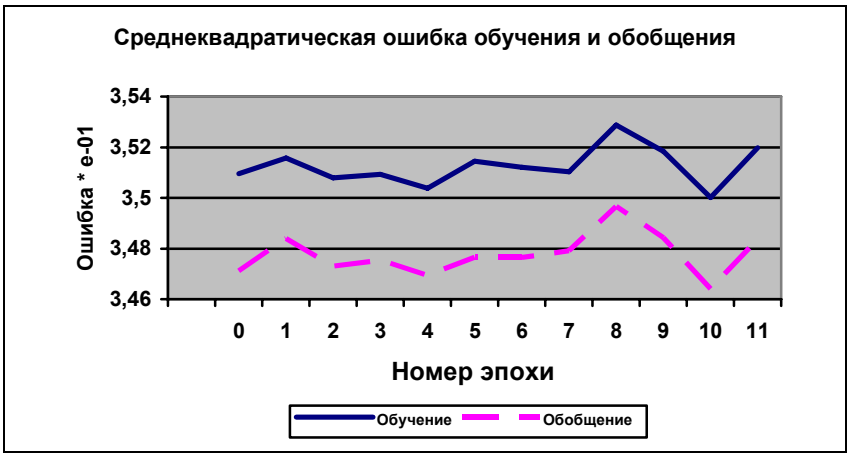


Рис. 3.19. $E_T(w)$ и $E_G(w)$, архитектура 3i-100sig-50sig-3sig, NTIMIT

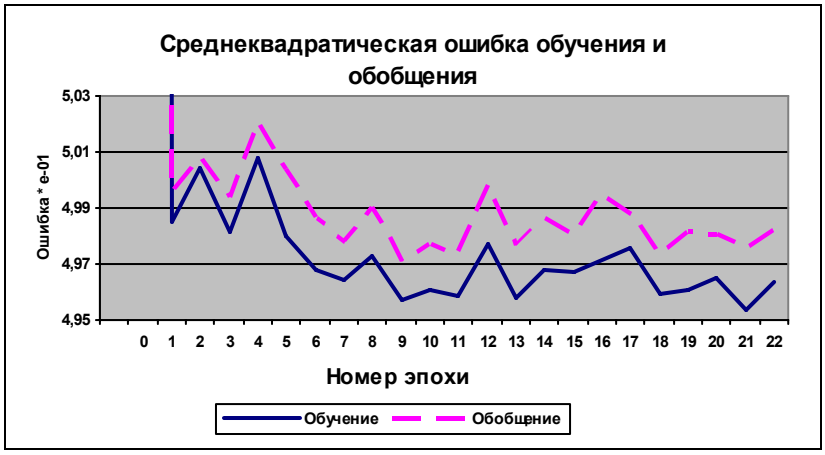


Рис. 3.20 2.3 $E_T(w)$ и $E_G(w)$, архитектура 3i-100sig-50sig-3sig, STC-TIMIT

Анализ приведенных графиков позволяет заключиться, что значения ошибки обучения и ошибки обобщения рис.3.18 - 3.20. повторяют форму друг друга, однако, ошибка обобщения ведет себя несколько хуже, чем ошибка на подмножестве обучения. Наименьшая среднеквадратическая ошибка обучения и обобщения (27%) получена для модели TIMIT, 35% для NTIMIT. и 50% для STC-TIMIT.

Из графиков рис.3.18 - 3.20 видно, что обе ошибки флуктуируют и убывают с увеличением количества епох обучения. При этом наблюдается быстрая сходимость обучения для корпуса NTIMIT по критерию раннего останова за 11 эпох обучения.

Чтобы проверить эффективность обученных моделей архитектура 3i-100sig-50sig-3sig (рис.3.18 - 3.20) в данной работе испытаны все перекрестные альтернативы тестирования на материалах корпусов TIMIT, NTIMIT и STC-TIMIT. Подбор количества скрытых нейронов (и связанный с ним подбор количества весов) предлагается выполнять путём тренировки нескольких сетей с последующим выбором той из них, которая содержит наименьшее количество скрытых нейронов при допустимой точности обучения (табл. 3.1)

Результаты экспериментальных исследований предложенной модели VAD в виде распределения ошибок классификации «тон/шум/пауза»,

выражение (3.21), по отдельным классам при обучении и тестировании классификаторов на различных корпусах в виде средней ошибки классификации по всему тестовому множеству звуковых данных приведены в табл. 3.6-3.14. Целью этих экспериментальных исследований было определить, как точно разработанный нейросетевой алгоритм выделяет тоновые, шумовые и паузные участки устной речи и насколько этот алгоритм надежен, т.е. насколько снижается точность работы в условиях помех двух телефонных каналов связи NTIMIT STC-TIMIT.

Чтобы наглядно продемонстрировать распределение ошибок классификации «тон/шум/пауза» по всем экспериментальным данным, приведем сводную табл. 3.15.

Из приведенных результатов табл. 3.15 и табл. 3.6-3.14 следует, что наиболее типичным видом ошибки для предложенной модели VAD стало то, что тон (вокализованная речь) чаще всего классифицируется как шум, кроме случая обучение NTIMIT и тестирование STC-TIMIT, где тон был распознан как пауза табл. 3.11 в 22.69% случаев.

Таблица 3.15

Данные	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %
TIMIT-TIMIT	4,56	4,18	5,69
TIMIT-NTIMIT	6,53	0,77	21,61
TIMIT-STC	19,3	3,88	9,42
NTIMIT-NTIMIT	18,76	0.01	7,85
NTIMIT-TIMIT	8,5	1,1	16,63
NTIMIT-STC	39,48	0.01	0,25
STC-TIMIT	25,24	0.00	5,14
STC-NTIMIT	5.19	0.64	24.15
STC-STC	8,15	0,48	23,03

Между тем, ошибка выделения шума достигла наибольшего значения для случая TIMIT-TIMIT – 4,18% и TIMIT-STC – 3,88%, что поясняется отсутствием в обучающем TIMIT множестве примеров шума, в результате наблюдается тенденция определения шума как тона в 2,85% случаев (табл. 3.6 – TIMIT-TIMIT) и шума как паузы в 2,24% случаев (табл. 3.8).

Ошибка выделения паузы достигла наибольшего значения для экспериментов STC-TIMIT (24,2%), STC-NTIMIT (23.0%), TIMIT-NTIMIT (21,6%) NTIMIT-TIMIT (16,6%). Эти ошибки в известной степени обусловлены простотой выбранных признаков, сущность которых состояла в их оценке непосредственно в режиме реального времени по речевому сигналу, не прибегая к вычислению дополнительных оценок параметров и их обновления. Вместе с тем, усложнять алгоритм вычисления вектора признаков не представлялось целесообразным ввиду нежелательного увеличения объёма вычислений.

Поскольку вектор признаков для участков пауз из обучающего множества не в полной мере соответствуют вектору признаков из тестового множества, возникают ошибки связанные с неполным соответствием условий обучения и тестирования. В случае необходимости эти ошибки могут быть устранены путем адаптации (механизм адаптивного обучения [20Qi]) моделей на небольшом множестве адаптации из тестовых данных посредством ранее использованных алгоритмов обучения модели MLP сети.

В связи с тем, что адаптация модели обычно проводится на меньшем количестве обучающих данных, возникает интерес проверить, насколько измениться ошибка обучения, если предложенная модель будет обучаться на множестве обучающих данных существенно меньшего размера.

Организация экспериментов исследований в целом подобна таковой как описано в п. 3.2.2. Укажем лишь отличительные особенности. Для обучения моделей использовалась не полная обучающая выборка (4614 файлов), а 380 файлов с условным названием DR1, 760 файлов для данных

DR2 и DR3. Обучение останавливалось после завершения 100 числа эпох обучения.

В связи с тем, что качество адаптации модели в зависимости от объема обучающего множества не зависит от архитектуры MLP сети и использования различных корпусов эксперименты проводились только для корпуса TIMIT для архитектуры сети i3_40sig_3sig.

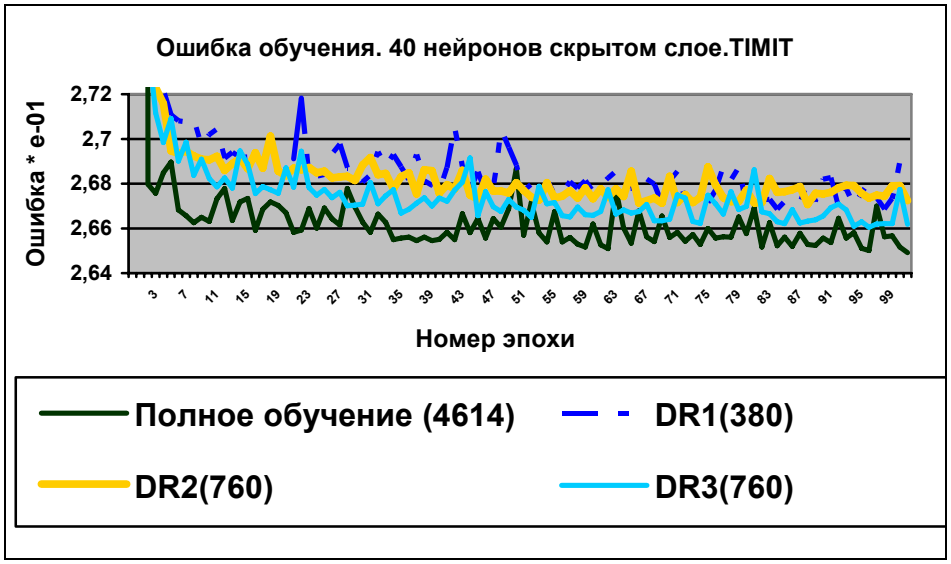


Рис. 3.21 7.7. $E_T(w)$ в зависимости от размера обучающего множества

Графики зависимости ошибки обучения $E_T(w)$ в зависимости от размера обучающего множества, соответственно полного множества данных из 4614 файлов и наборов из 380 файлов (DR1), 760 файлов (DR2 и DR3), представлены на графиках рис. 3.21.

Как следует из рис. 3.21 ошибка обучения модели меняется не значительно:

1. в пределах 0,05% при уменьшении обучающего множества в 12 раз для набора обучающих данных DR1 (380 файлов)
2. в пределах 0,04% в случае уменьшения обучающего множества в 6 раз для набора обучающих данных DR2 и DR3 (760 файлов).

Из приведенных результатов следует, что предложенная модель MLP сети может быть обучена на небольшом множестве обучающих примеров и

при этом сохранять эффективность классификации, ошибку обучения модели $E_T(w)$ в пределах 0,04-0,05%.

В связи с тем, что в тестовую выборку попали дикторы женского и мужского полов в соотношении 560/1120, в сводной таблице 3.16 представлены средние ошибки классификации раздельно для каждого пола диктора, а так же общая средняя ошибка классификации для всего тестового множества.

Таблица 3.16

Данные обучение-тестирование	Средняя ошибка классификации "тон/шум (женщины -560)	Средняя ошибка классификации "тон/шум (мужчины - 1120)	Средняя ошибка, %
TIMIT-TIMIT	15,3	15,9	16,1
TIMIT-NTIMIT	29.2	27.9	28,9
TIMIT-STC	31.9	31.9	32.3
NTIMIT-NTIMIT	26.5	25.5	26,5
NTIMIT-TIMIT	27.0	25.1	26,2
NTIMIT-STC	39.3	39.4	39,7
STC-TIMIT	31.4	28.4	30,4
STC-NTIMIT	33.3	29.9	29,9
STC-STC	30.3	29.6	31,7

Анализ приведенных результатов позволяет заключить, что для обучения модели MLP сети с архитектурой 3i-100sig-50sig-3sig для распознавания речевого сигнала телефонного канала NTIMIT, лучше всего использовать речевой корпус TIMIT (табл. 3.16 – средняя ошибка 28,9%.) или же NTIMIT (табл. 3.16 – ошибка 26,5%.).

Вариант обучения модели на NTIMIT, а тестировании на TIMIT даёт среднюю ошибку классификации (26,2%) соответствующей варианту обучения модели на NTIMIT (26,5%), а тестировании на NTIMIT. В остальных случаях ошибка превышает уровень 30%, что говорит о существенном несоответствии данных обучения и тестирования. Поэтому

дальнейшие эксперименты по выбору оптимальной архитектуры проводились для корпусов TIMIT и NTIMIT.

На рис. 3.22.-2.23 приведены графики ошибок обобщения $E_G(w)$ для различных архитектур MLP сетей (см. табл. 3.1), обученных на корпусах TIMIT, NTIMIT в зависимости от оптимального числа эпох с использованием алгоритма адаптивной корректировки весов сети IDBD [15Sohn].

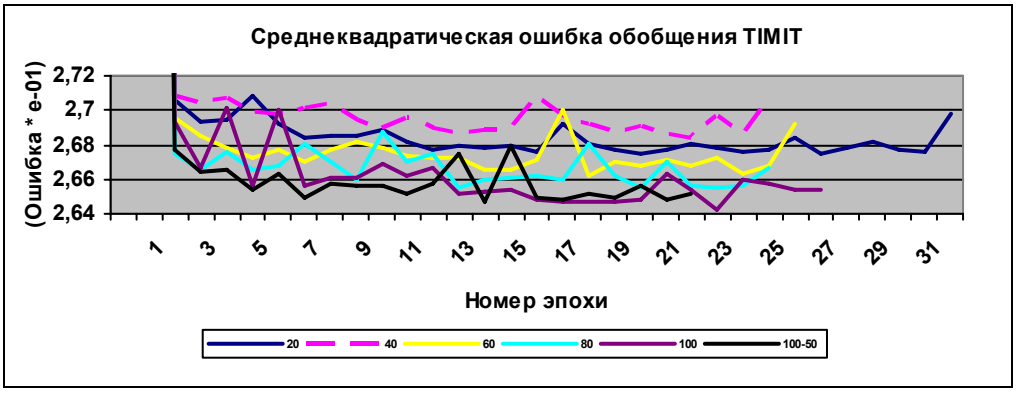


Рис. 3.22. Кривые $E_G(w)$ для различных архитектур сети TIMIT

Проведенные количественные эксперименты можно разделить на две группы. Эксперименты первой группы носили предварительный характер, а эксперименты второй группы были уточняющими.

Целью экспериментов первой группы было выяснение, какая из рассматриваемых архитектур табл. 3.1 является оптимальной, в смысле ошибки обобщения для решения задачи VAD.

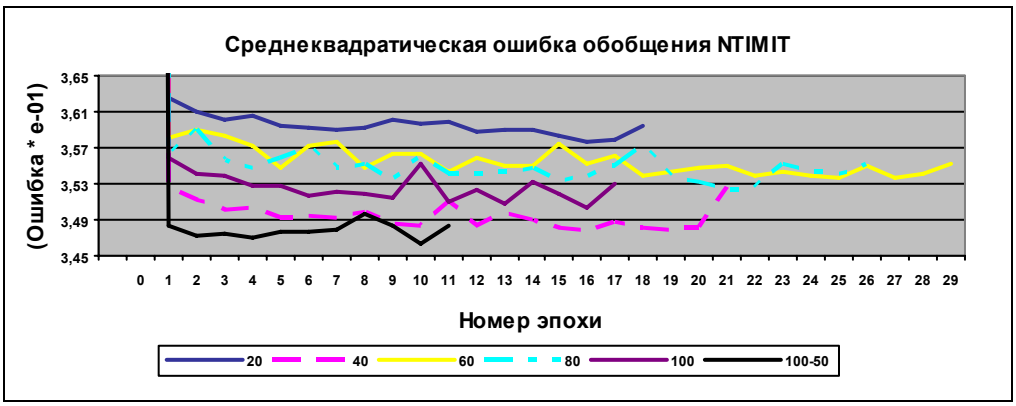


Рис. 2.23 2.5 Кривые $E_G(w)$ для различных архитектур сети NTIMIT

Как видим, что сеть с максимальным количеством скрытых нейронов (100 скрытых нейронов) 3i-100sig-3sig не является оптимальной, в смысле ошибки обобщения $E_G(w) = 35{,}1\%$, для модели сети на основе NTIMIT, в то время как для сети 3i-100sig-3sig TIMIT наблюдается наименьшая $E_G(w) = 26{,}4\%$ ошибка.

Из представленных результатов следует, что оптимальной в смысле среднеквадратической ошибки $E_G(w)$ является архитектура 3i-100sig-3sig ($E_G(w) = 26{,}4\%$) для модели TIMIT и 3i-40sig-3sig (40 скрытых нейронов $E_G(w) = 34{,}8\%$) для модели NTIMIT.

На рис. 3.22. сопоставлены зависимости ошибок $E_G(w)$ между моделями с архитектурами 3i-40sig-3sig и 3i-100sig-3sig. Разница между значениями $E_G(w)$ составляет около 0,006, поэтому в качестве оптимальной модели TIMIT необходимо использовать модель 3i-40sig-3sig с наименьшим количеством нейронов скрытого слоя (40sig).

Для более детальной проверки эффективности обученных моделей различных архитектур (рис. 3.22-3.23) в виде распределения ошибок классификации «тон/шум/пауза» по отдельным классам при обучении и тестировании классификаторов на различных корпусах TIMIT, NTIMIT и STC-TIMIT и представлены в приложении А.

Проведенные количественные эксперименты можно разделить на две группы. Эксперименты первой группы носили предварительный характер, а эксперименты второй группы были уточняющими.

Рассмотрим сначала результаты первой группы экспериментов, главной целью которых было выяснение оптимальной архитектуры сети, в смысле минимального количества скрытых нейронов в одном скрытом слое при минимальной ошибке классификации $Err,\%$. В качестве используемой модели сети для предложенной VAD модели была выбрана модель, построенная на корпусе TIMIT (без влияния телефонного канала связи).

Кроме того, безусловный интерес представляла надежность предложенного алгоритма, т.е. насколько снижается ошибка классификации «тон/шум/пауза» в условиях помех двух телефонных каналов связи NTIMIT STC-TIMIT, а так же при их отсутствии (корпус TIMIT).

Чтобы наглядно продемонстрировать распределение ошибок классификации «тон/шум/пауза» по всем экспериментальным данным из приложения А (табл. А. 3.1 - А.3.15), приведем сводные табл. 3.1, 3.2 и 3.3 общих результатов по каждому классу для исследуемых архитектур сети, а так же среднюю ошибку классификации «тон/шум/пауза».

Таблица 3.1. Обучение модели TIMIT и тестирование TIMIT

Архитектура	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
3i-20sig-3sig	6,32	4,17	5,93	16,4
3i-40sig-3sig	6,18	4,08	5,84	16,1
3i-60sig-3sig	6,1	4,31	5,65	16,1
3i-80sig-3sig	6,12	4,2	6,03	16,3
3i-100sig-3sig	6,2	4,31	5,55	16,1
3i-100sig-50sig-3sig	4,56 %	4,18 %	5,69 %	

Таблица 3.2. Обучение модели TIMIT и тестирование NTIMIT

Архитектура	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
3i-20sig-3sig	6,34	0,97	21,2	28,5
3i-40sig-3sig	6,27	0,9	21,25	28,4
3i-60sig-3sig	6,1	1,11	21,27	28,5
3i-80sig-3sig	6,13	0,96	21,98	29,1
3i-100sig-3sig	6,21	1,2	21,41	28,8
3i-100sig-50sig-3sig	6,53	0,77	21,61	

Таблиц 3.3. Обучение модели TIMIT и тестирование STC-TIMIT

Архитектура	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
3i-20sig-3sig	18,55	4,42	10,66	33,6
3i-40sig-3sig	18,15	4,32	10,32	33,8
3i-60sig-3sig	17,68	5,16	11,02	33,9
3i-80sig-3sig	18,34	4,6	10,37	33,3
3i-100sig-3sig	18,18	5,15	11,35	34,7
3i-100sig-50sig-3sig	19,3	3,88	9,42	

Как следует из табл. 3.1, 3.2 и 3.3 помехи канала связи существенно влияют на среднюю ошибку классификации «тон/шум/пауза» *Err*,% (ошибка повышается с 16% до 29% для корпуса NTIMIT и с 16% до 34% для корпуса STC-TIMIT)

Как следует из табл. 3.1, 3.2 ошибка выделения тона для условий обучения на TIMIT и тестирования на TIMIT (табл. 3.1) и NTIMIT (табл. 3.2) меньше (составляет 6%.), чем для тестирования на STC-TIMIT (составляет 18%). Ошибка выделения шума минимальна для модели TIMIT при тестировании на NTIMIT (табл. 3.2) и составляет 1%, а на TIMIT и STC-TIMIT составляет 4-5%. Ошибка выделения паузы достигает наибольшего значения для условий обучения на TIMIT и тестирования на NTIMIT (табл. 2.2) и составляет 21%.

Очевидно, что для табл. 2.13 архитектура 3i-40sig-3sig содержит минимальные ошибки классификации «тон/шум/пауза», соответственно 18,15% - ошибка выделения тона, 4,32 – ошибка выделения шума и 10,32% – ошибка выделения паузы. Как следует из табл. 3.1 и 3.2 минимальные значения ошибки классификации «тон/шум/пауза» достигаются при различных архитектурах MLP сети и изменяются не значительно для. Поэтому оптимальной была выбрана архитектура 3i-40sig-3sig из табл. 2.13, в

смысле минимальных ошибок классификации «тон/шум/пауза» по каждому из классов.

Из представленных результатов следует, с учетом оптимальной архитектуры сети NTIMIT в качестве оптимальной выбрана сеть с 40 скрытыми нейронами 3i-40sig-3sig в одном скрытом слое (40 нейронов) и минимальной ошибке классификации $Err,\%$ (TIMIT – $Err = 16{,}1\%$, NTIMIT – $Err = 28{,}4\%$, STC-TIMIT – $Err = 33{,}8\%$).

Такой выбор обусловлен тем, что количество вычислений при обучении модели пропорционально числу весов сети, которые связывают все скрытые нейроны при незначительном изменении ошибки $E_G(w)$ и средней ошибки классификации $Err,\%$ для каждого из поставленных экспериментов из табл. 3.1, 3.2, 3.3.

Рассмотрим теперь результаты уточняющих экспериментов, при постановке которых ставились следующие цели:

- уточнение оптимального числа эпох обучения сети без применения критерия останова;

- определить распределение ошибок классификации «тон/шум/пауза» для выбранного числа эпох обучения;

- определить среднюю ошибку классификации предлагаемой MLP модели на множестве тестовых данных;

- определить насколько снижается точность работы в условиях помех двух телефонных каналов связи NTIMIT STC-TIMIT

Для более детального анализа оптимального числа эпох обучения сети без применения критерия останова было принято осуществлять останов алгоритма обучения по окончании 229 эпох. Такую продолжительность мы посчитали достаточной для исследования достижения алгоритмом обратного распространения ошибки некоторого локального минимума на поверхности ошибок.

Цель этих исследований заключалась в подтверждении достижения минимальной ошибки обучения $E_T(w)$ MLP модели сети при заданном числе эпох обучения, выбранных в соответствии с критерием раннего останова (рис. 3.12).

Следует отметить, что в рамках обучения MLP сети слишком большое число эпох обучения приводит к потере свойства MLP сети обобщения при тестировании на данных, не участвующих в процессе обучения MLP сети. Одновременно слишком малое число эпох приводит к недостаточному обучению, что в свою очередь влияет на качество классификации такой MLP сетью.

На рис. 3.24. сопоставлены зависимости количества эпох обучения, определенных по критерию раннего останова, MLP сети в зависимости от различных её архитектур (см. табл. 3.1).

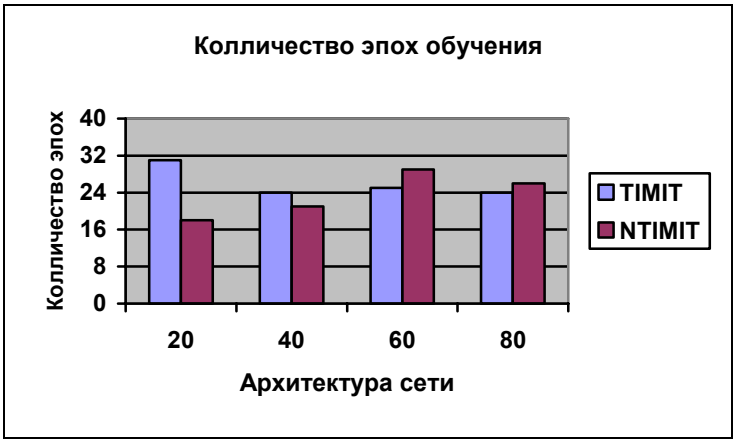


Рис. 3.24 2.6. Зависимость количества эпох обучения от архитектуры сети

Очевидно, что оптимальной относительно количества эпох обучения для обеих MLP сетей TIMIT и NTIMIT является MLP сеть с 40 скрытыми нейронами.

Поскольку поиск оптимального числа эпох обучения является существенным моментом для построения эффективной MLP модели сети,

необходимо исследовать оптимальность полученной модели при останове алгоритма обучения по максимальному числу эпох обучения.

На рис. 3.25, 3.26 показаны графики процесса обучения моделей MLP сети при максимально установленном числе эпох обучения (229 эпох) и при использовании критерия раннего останова. Эксперименты проводились на обучающих множествах TIMIT и NTIMIT соответственно.

Как следует из рис. 32.5, 3.26 при использовании критерия раннего останова для архитектуры модели потребовалось 24 (TIMIT с $E_T(w) = 26{,}8\%$) и 21 (NTIMIT с $E_T(w) = 35{,}5\%$) эпохи обучения соответственно.

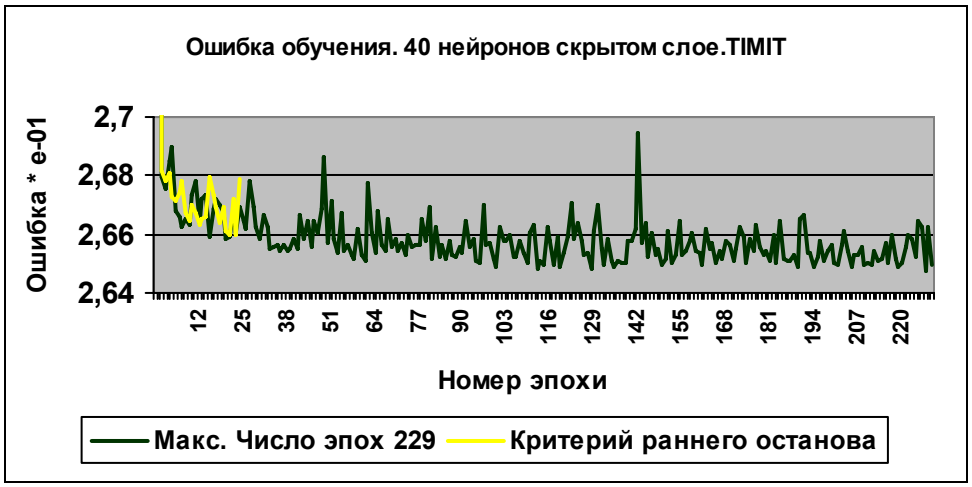


Рис. 3.25 2.7. Кривые обучения модели TIMIT

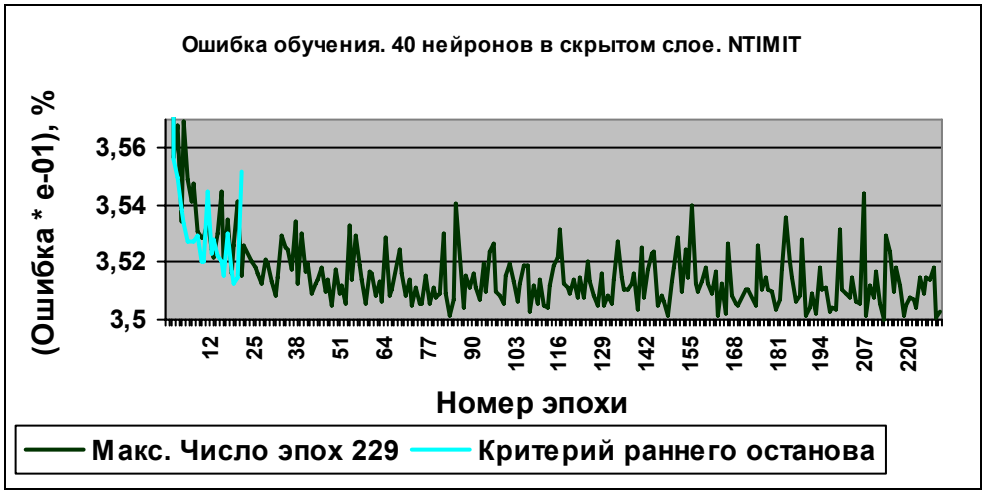


Рис. 3.26 2.8. Кривые обучения модели NTIMIT

Анализ приведенных графиков рис. 3.25, 3.26 позволяет заключить, что достижение локального минимума на поверхности ошибок по окончанию 229 эпох обучения было получено при ошибке обучения $E_T(w) = 26{,}5\%$ для модели TIMIT и $E_T(w) = 35{,}0\%$ для модели NTIMIT.

Заметим, что время, затраченное на одну эпоху пропорционально одному полному проходу по обучающей выборке размером в 1410379 примеров обучающего множества и пропорционально количеству оцениваемых весов MLP сети (для архитектуры 3i-40sig-3sig оценивается 283 весовых параметра). Очевидно, что время, затраченное на одну эпоху обучения, составляет $1410379 \times 283 = 399 \cdot 10^6$ операций. Поэтому ключевым моментом обучения MLP модели является выбор оптимального количества эпох обучения.

Из рис. 3.25 очевидно, что остановка обучения по критерию раннего останова позволяет получить оптимальную модель быстрее (соответственно за 24 и 21 эпоху). В этом случае ошибка обучения $E_T(w)$ незначительно отличается от обучения при числе эпох обучения равным 229 (разница для TIMIT составляет 0,3%, а для NTIMIT – 0,5%) . Таким образом, используя критерий раннего останова проводиться оптимизация процедуры обучения модели по времени (соответственно за 24 и 21 эпоху).

Результаты экспериментальных исследований в виде ошибок классификации «тон/шум/пауза» по каждому из классов архитектуры 3i-40sig-3sig модели MLP сети, при её обучении до момента наступления максимального числа эпох (229 эпох), приведены приложении Б (табл. Б.3.1 - Б.3.4).

Как следует из табл. Б.3.2, Б.3.4, ошибка выделения паузы наибольшая и составляет соответственно 20,60% (обучение NTIMIT и тестирование TIMIT) и 21,7% (обучение на TIMIT, тестировании на NTIMIT).

В табл. 3.4, представлены результаты экспериментов табл. Б.3.1-Б.3.4 в виде распределения ошибок классификации по классам «тон/шум/пауза» при останове по максимальному количеству эпох обучения (229 эпох) и по критерию раннего останова (3.20) (п. 3.2.1).

Из приведенных результатов следует, что ошибка выделения тона уменьшается с 18,76% до 7,45% при использовании максимального числа эпох обучения вместо критерия раннего останова для случая обучения модели на корпусе NTIMIT и тестировании на NTIMIT. Однако ошибка выделения шума возрастает для всех случаев кроме обучения модели на корпусе TIMIT и тестировании на TIMIT, где она остается постоянной (4,2%). Ошибка выделения паузы возрастает для случаев обучения модели на корпусе NTIMIT и тестировании на TIMIT (с 16,63% до 20,6%) и NTIMIT (с 7,85% до 9,89%).

Таблица 3.4

Данные	Ошибка выделения тона, %		Ошибка выделения шума, %		Ошибка выделения паузы, %	
	$Err_{Останов}$	Err_{229}	$Err_{Останов}$	Err_{229}	$Err_{Останов}$	Err_{229}
TIMIT-TIMIT	4,56	6,14	4,18	4,15	5,69	5,69
NTIMIT-TIMIT	8,5	6,84	1,1	3,23	16,63	20,6
NTIMIT-NTIMIT	18,76	7,45	0.01	4,71	7,85	9,89
TIMIT-NTIMIT	6,53	6,34%	0,77	0,95%	21,61	21,7%

Возрастание ошибок в известной степени обусловлено различными условиями создания и тестирования модели. На практике для создания модели наиболее часто используют корпус неискаженных внешними помехами данных (TIMIT), а тестирование выполняют в различных условиях помех (TIMIT, NTIMIT, STC- TIMIT).

Из табл. 3.4 видно, что ошибка для тестирования модели TIMIT увеличивается на 1,58% при тестировании на TIMIT в случае использования максимального числа эпох обучения. Такое увеличение ошибки классификации тона говорит о том, что модель MLP сети на 229 эпохе обучения «переучена», т.е. теряет свою способность к обобщению на выходных данных, отличных от данных использованных для обучения модели.

В табл. 3.5 представлены результаты экспериментов табл. Б.3.1 -Б.3.4 в виде средней ошибки классификации Err,% на всем множестве тестовых данных для остановки алгоритма по критерию раннего останова $Err_{Останов}$,% и по завершению максимального числа эпох обучения Err_{229},%.

Таблица 3.5

Данные	$Err_{Останов}$,%, Ранний останов	Err_{229},% 229 эпох
TIMIT-TIMIT	15,98	16,1
NTIMIT-TIMIT	30,67	26,2
NTIMIT-NTIMIT	20,06	26,5
TIMIT-NTIMIT	28,99	28,9

Как следует из табл. 3.5 в случае обучения модели MLP сети на корпусе неискаженных данных TIMIT средняя ошибка классификации существенно не ухудшается при использовании обоих критериев останова алгоритма обучения. В то время как для модели MLP сети NTIMIT в случае NTIMIT- TIMIT ошибка уменьшается при останове на 229 эпохе обучения с 30,67% до 26,2%, однако при одинаковых условиях обучения и тестирования модели (NTIMIT- NTIMIT) ошибка классификации повышается для 229 эпох обучения с 20,06% до 26,5%.

Чтобы наглядно продемонстрировать работу выбранной оптимальной архитектуры модели 3i-40sig-3sig, построенной на адаптационных данных из телефонного канала связи NTIMIT проведены три тестовых эксперимента на корпусах TIMIT, NTIMIT, STC-TIMIT.

Вариант построения модели на данных NTIMIT возможен в случае необходимости проведения адаптации построенной модели на данных искаженных телефонным каналом связи. В связи с этим безусловный интерес представляет влияние степени разрушения сигнала под воздействием телефонного канала связи, в смысле средней ошибки классификации Err,% при тестировании на корпусах TIMIT, NTIMIT, STC-TIMIT, на качество построенной модели в зависимости от её архитектуры.

Результаты экспериментальных исследований выбранной ранее оптимальной архитектуры (3i-40sig-3sig) модели MLP сети для обучающих данных корпуса телефонной речи NTIMIT в виде ошибок классификации «тон/шум/пауза» по каждому из классов приведены в Приложении табл.3.6-3.8.

Чтобы наглядно продемонстрировать распределение ошибок классификации «тон/шум/пауза» по всем экспериментальным данным из табл.3.6-3.8, приведем сводную табл. 3.9 общих результатов по каждому классу для исследуемых архитектур сети, а так же среднюю ошибку классификации «тон/шум/пауза».

Таблица 3.9. Обучение модели NTIMIT, архитектура 3i-40sig-3sig

Тестовые данные	Ошибка выделения тона, %	Ошибка выделения шума, %	Ошибка выделения паузы, %	Средняя ошибка, %
NTIMIT	6,98	5,13	9,85	21,96
TIMIT	53,93	5,54	40,52	30,17
STC	14,17	20,04	5,12	36,69

Из табл. 3.9 очевидно, что модель NTIMIT наилучшим образом подходит для тестирования в соответствующих условиях, т.е. на корпусе NTIMIT, при этом средняя ошибка классификации *Err*,% составила 21,96%, ошибка выделения тона 6,98%, шума 5,13% и паузы 9,85%.

Процедура тестирования на корпусе речевых данных чистых от влияния телефонного канала не привела к положительным результатам (*Err*,% составила 36,69%). В случае тестирования модели NTIMIT на тестовых записях одноканального телефонного коммутатора корпуса STC-TIMIT средняя ошибка классификации *Err*,% приняла наихудшее значения 36,69%. В последнем случае увеличение ошибки *Err*,% тестирования модели NTIMIT на STC-TIMIT можно пояснить не соответствием условий записи корпусов NTIMIT. В случае NTIMIT запись производилась в многоканальном телефонном режиме с различных абонентских точек [36Jankowski], а в случае STC-TIMIT использовался искусственно смоделированный одноканальный режим записи непосредственно с коммутатора, без учета влияния записывающих устройств [35Morales].

Анализ ошибок классификации по каждому классу позволяет заключить, что модель NTIMIT может быть использована лишь для выделений участков шума при тестировании в условиях корпуса TIMIT с ошибкой классификации 5,54%, а так же для выделения пауз при тестировании в условиях корпуса STC-TIMIT с ошибкой классификации 5,12%.

Такой результат обусловлен, тем, что в корпусе TIMIT участки пауз представляют так же образцы участков шума, что в результате приводит к неправильной классификации при использовании искаженной телефонным каналом модели MLP сети.

Для выделения тона в условиях, несоответствующих модели NTIMIT наилучшим образом подходят тестовые данные STC-TIMIT, ошибка классификации тона составила 14,17%. Такой результат может быть объяснен тем, что NTIMIT и STC-TIMIT влияют на речевой сигнал

(тональную составляющую) физически похожим способом, однако различие характеристик этих телефонных каналов связи обуславливают полученную ошибку классификации (14,17%.).

Из представленных результатов следует, что модель, построенную на адаптационных данных из канала связи следует использовать лишь в тестовых условиях, соответствующих условиям построения модели для получения минимальной ошибки классификации *Err*,% «тон/шум/пауза» 21,96%.

4.6. Выводы

1. В частности для выбранной модели с 40 нейронов в скрытом слое для сигнала, искаженном телефонным каналом связи (рис. 12, б), количество циклов удается уменьшить с 34 до 21 циклов, то есть в 1,6 раза.

2. Из представленных результатов следует, что модель, построенную на адаптационных данных из канала связи следует использовать лишь в тестовых условиях, соответствующих условиям построения модели для получения минимальной ошибки классификации *Err*,% «тон/шум/пауза» 21,96%.

3. Анализ ошибок классификации по каждому классу позволяет заключить, что модель NTIMIT может быть использована лишь для выделений участков шума при тестировании в условиях корпуса TIMIT с ошибкой классификации 5,54%, а так же для выделения пауз при тестировании в условиях корпуса STC-TIMIT с ошибкой классификации 5,12%.

4. Как следует из табл. 3.5 в случае обучения модели MLP сети на корпусе неискаженных данных TIMIT средняя ошибка классификации существенно не ухудшается при использовании обоих критериев останова алгоритма обучения.

ЗАКЛЮЧЕНИЕ

С целью решения актуального научно-технического задания разработки новых методов обеспечения помехоустойчивости систем АРР методами обработки сигналов получены следующие результаты.

1. Впервые решена задача построения нейросетевого детектора голосовой активности для системы автоматического распознавания речи, что является основой для использования таких признаков как нормализованные по мощности кепстральные коэффициенты PNCC при работе с нестационарными шумами телефонного канала связи в диапазоне соотношений сигнал-шум от -12 до +18 дБ.

2. Обоснован выбор структуры детектора голосовой активности, которая обеспечивает робастность системы АРР при использовании PNCC признаков на основе выбора алгоритма коррекции параметров нейронной сети, поиска оптимальных, по критерию погрешности классификации распознанных речевых фреймов, значений его параметров и выбора рациональной архитектуры нейронной сети, позволяет ограничить сложность предложенной модели и использовать разработанный детектор в режиме реального времени работы системы АРР.

3. Предложен алгоритм адаптивной коррекции параметров стационарной нелинейной MLP сети, что позволило усовершенствовать нейросетевой детектор голосовой активности и ускорить процедуру обучения, сократив число циклов обучения с 34 до 21 циклов для сигнала, искаженного телефонным каналом связи.

4. Впервые предложен метод повышения помехоустойчивости детектора голосовой активности за счет оценивания признака «траектория основного тона», что позволило повысить робастность его работы, снизив процент грубых ошибок до 1,4% при соотношении сигнал-шум 0 дБ.

5. Усовершенствован метод подавления поздней реверберации путем оптимизации значений его параметров по критерию точности

автоматического распознавания речи, при этом установлено, что оптимальным значением момента начала поздней реверберации является 100 мс, а параметр сглаживания периодограммы при оценке спектра поздней реверберации не должен зависеть от частоты.

6. Установлено, что при слепом оценке времени реверберации методом максимального правдоподобия можно достичь точности автоматического распознавания речи, которая на 3-5% уступает таковой для измерения времени реверберации по имеющейся импульсной характеристикой помещения.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев, Наук. думка, 1987. – 264 с.

2. Сажок М. М. Автоматизовані засоби формування баз даних і знань для озвучення українських текстів: дис. ... канд. техн. наук: 05.13.06 / Сажок Микола Миколайович; НАН України, МОН України, Міжнар. наук.-навч. центр інформ. технологій та систем. - К., 2004. – 168 с.

3. Людовик Т. В. Інформаційна технологія синтезу індивідуалізованого мовлення за текстом: дис. ... канд. техн. наук: 05.13.23 / Людовик Тетяна Владленівна; НАН України, МОН України, Міжнар. наук.-навч. центр інформ. технологій та систем. - К., 2006. – 176 с.

4. Пилипенко В. В. Автоматизированный стенограф украинской речи / В. В. Пилипенко, В. В. Робейко // Искусственный интеллект. – 2010. – № 3. – С. 238-248.

5. Карпов О. М. Методи аналізу і розпізнавання складних сигналів в автоматизованих системах мовного діалогу: автореф. дис... д-ра техн. наук: 05.13.06 / Карпов Олег Миколайович ; Національний авіаційний ун-т. - К., 2003. – 36 с

6. Ермоленко Т. В. Применение вейвлет-анализа для предварительной обработки речевых голосовых сигналов в задачах сегментации, классификации и пофонемного распознавания: дис. ... канд. техн. наук: 05.13.23 / Ермоленко Татьяна Владимировна; Ин-т проблем искусств. интеллекта. – Донецк, 2008. – 173 с.

7. Л. Рабинер, Шафер Р.В. Цифровая обработка речевых сигналов. Пер. с англ. Под ред. М. В. Назарова и Ю. Н. Прохорова. – М.: Радио и связь, 1981. – 496 с.

8. The HTK Book / S. Young, G. Evermann, M. Gales, et al. – Cambridge: University Engineering Department, 2009. – 375 p.

9. H. Hermansky. Perceptual Properties of Current Speech Recognition Technology // Hermansky H., Cohen J., Stern R. / Proceedings of IEEE – 2013. – P.1968-1985.

10. D. Ellis Tandem acoustic modelling in large-vocabulary recognition // Ellis D., Singh R., Sivadas S. / Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). – 2001. – P.309-312.

11. Jinyu L. An overview of Noise-Robust Automatic Speech Recognition / L. Jinyu, L. Deng, Y. Gong, R. Haeb-Umbach // IEEE Trans. Acoust., Speech, Signal Processing. – 2014, Apr. – Vol. 22. – №4. – P. 745-776.

12. Techniques for Noise Robustness in Automatic Speech Recognition / T. Virtanen, R. Singh, B. Raj. – Wiley, 2013, 496 p.

13. Springer Handbook of Speech Processing / J. Benesty, M. Mohan Sondhi, Y. Huang. – Springer-Verlag Berlin Heidelberg 2008, 1176 p.

14. Ephraim Y. Speech enhancement using a minimum mean square error Log-spectral amplitude estimator / Y. Ephraim, D. Malah // IEEE Trans. Acoust., Speech, Signal Processing. – 1985, Apr. – Vol. ASSP-33. – P. 443-445.

15. Davis S. Comparison of parametric representation of monosyllabic word recognition in continuously spoken sentences / S. Davis, P. Mermelstein // IEEE Trans. Acoust., Speech, Signal Processing. – 1980. – Vol. 28. – №4. – P. 357-366.

16. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech / JASA, 1990. – Vol. 87. – №4. – P1635-1638s.

17. Kim C. Signal Processing for Robust Speech Recognition Motivated by Auditory Processing: dissert. Doctor of Philosophy In Language and Information Technologies / Chanwoo Kim. – Language Technologies Institute School of Computer Science Carnegie Mellon University, 2010. – 210 p.

18. Kim C., Stern R. Power-normalized cepstralcoefficients (PNCC) for robust speech recognitions / C. Kim // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). –2012. – P. 4101 - 4104.

19. Ладошко О. М. Дослідження акустичних особливостей вокалізованих пауз спонтанної мови / О. М. Ладошко // Міжнародна науково-технічна конференція «Штучний інтелект. Інтелектуальні системи ШІ-2011», 19-23 вересня 2011 р.: тези доп., Т. 3. – Донецьк, 2011. – С. 82-86.

20. Ладошко О. М. Дослідження характеристик вокалізованих пауз спонтанної мови / О. М. Ладошко // Акустичний симпозіум «Консонанс-2011», 27-29 вересня 2011 р.: тези доп. – К. – 2011. – С.188-193.

21. Ладошко О. М. Моделювання виділювача частоти основного тону для досліджень спонтанного мовлення / О. М. Ладошко // Міжнародна науково-технічна конференція "Моделювання і комп'ютерна графіка", 5-8 жовтня 2011 р.: тези доп. – Донецьк, 2011. – С. 304-308.

22. Ладошко О. М. Дослідження впливу характеристик телефонного каналу зв'язку на надійність розпізнавання фонем / О. М. Ладошко // ІІІ Міжнародна науково-технічна конференція студентів, аспірантів та молодих вчених. Інформаційні управляючі системи та комп'ютерний моніторинг (ІУС і КМ-2012), 16-18 квітня, 2012р.: тези доп. – Донецьк. – 2012. – С. 143-148.

23. Ладошко О. М. Дослідження впливу параметризації мовленнєвого сигналу і характеристик каналів зв'язку на надійність розпізнавання фонем / О. М. Ладошко // Акустичний симпозіум «Консонанс-2013», 1-2 жовтня 2013.: тези доп. – К. – 2013. – С. 169-174.

24. Ладошко О. М. Оптимізація алгоритмів системи розпізнавання мови з використанням інструментарію НТК / О. М. Ладошко, А. М. Продеус // Електроніка та зв'язок. – 2007. – № 4 (39). – С. 53-60.

25. Ладошко О. М. Анотація та врахування мовленнєвих збоїв в задачі автоматичного розпізнавання спонтанного українського мовлення / О. М. Ладошко, В. В. Пилипенко // Штучний інтелект. – 2010. – № 3. – С. 238-248.

26. Ладошко О. М. Розмітка спонтанної української мовлення / О. М. Ладошко, А. М. Продеус // Електроніка та зв'язок. Тематичний випуск «Електроніка і нанотехнології». – 2011. – С. 97-103.

27. Ладошко О. М. Оцінка надійності виділення частоти основного тону для акустичного аналізу мови / О. М. Ладошко, А. М. Продеус // Інформатика, кібернетика та обчислювальна техніка. – 2012. – № 15. – С.162-169.

28. Ладошко О. М. Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок мовлення / О. М. Ладошко, І. Ю. Бондаренко // Електроніка та зв'язок. – 2012. – №6 (71). – С. 19-25.

29. Ладошко О. М. Залежність показників систем ослаблення реверберaційної завади від ступеня спотворення сигналу / О. М. Ладошко, А. М. Продеус // Стандартизація, сертифікація, якість. – 2014. – №3(88). – С. 45-49.

30. Оцінка ефективності захисних конструкцій за критерієм розбірливості мови / В. С. Дідковський, А. М. Продеус, О. М. Ладошко, Н. О. Самійленко // Вісник вищих навчальних закладів. Радіоелектроніка. – 2014, Том 57. – № 2 (620). – С.55-60. (Включено до наукометричної бази SCOPUS).

31. Пат. на кор. мод. 150553 Україна МПК (2015.01) G10L 17/20. Спосіб визначення тонових, шумових та паузних ділянок мовного сигналу. / Ладошко О. М., Дідковський В. С, Продеус А. М. – № u201511068; заявл. 12. 11. 2015.

32. Ладошко О. М. Анотація та облік мовленнєвих збоїв в задачі автоматичного розпізнавання спонтанної української мови / О. М. Ладошко, В. В. Пилипенко // Міжнародна науково-технічна конференція «Штучний інтелект. Інтелектуальні системи ШІ-2010», 20-24 вересня 2010 р.: тези доп., Т. 1. – Донецьк, 2010. – С. 223-227.

33. Ладошко О. М. Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок усного мовлення /О. М. Ладошко, І. Ю. Бондаренко // Одинадцята всеукраїнська міжнародна конференція «Оброблення сигналів і зображень та розпізнавання образів»

(УкрОбраз’2012), 15-19 жовтня 2012 р.: тези доп. – Київ: УАсОІРО. – 2012. – С. 55-58.

34. Ladoshko O. On existence of optimal boundary value between early reflections and late reverberation /O. Ladoshko, A. Prodeus // Proc. of IEEE 34th International Scientific Conference «Electronic and Nanotechnology», 15-18 April 2014: Proceedings. – Kyiv, 2014. – P. 442-446.

35. S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust.Speech Signal Process. ASSP-27, 113–120 (1979)

36. J.S. Lim, A.V. Oppenheim: Enhancement and bandwidth compression of noisy speech, Proc. IEEE 67, 1586–1604 (1979)

37. R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Trans. Speech Audio Process. 9, 504–512 (2001)

38. J. Chen, J. Benesty, Y. Huang, S. Doclo: New insights into the noise reduction Wiener filter, IEEE Trans. Speech Audio Process. 14, 1218–1234 (2006)

39. Y. Ephraim, H.L. Van Trees: A signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process. 3, 251–266 (1995)

40. H. Lev-Ari, Y. Ephraim: Extension of the signal subspace speech enhancement approach to colored noise, IEEE Trans. Speech Audio Process. 10, 104–106 (2003)

41. Y. Ephraim, D. Malah: Speech enhancement using aminimum-mean square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121 (1984)

42. Ephraim Y. Speech enhancement using a minimum mean square error Log-spectral amplitude estimator / Y. Ephraim, D. Malah // IEEE Trans. Acoust., Speech, Signal Processing. – 1985, Apr. – Vol. ASSP-33. – P. 443-445.

43. K.K. Paliwal, A. Basu: A speech enhancement method based on Kalman filtering, Proc. IEEE ICASSP 1987, 177–180 (1987)

44. K.K. Paliwal, A. Basu: A speech enhancement method based on Kalman filtering, Proc. IEEE ICASSP 1987, 177–180 (1987)

45. Y. Ephraim, D. Malah, B.-H. Juang: On the application of hidden Markov models for enhancing noisy speech, IEEE Trans. Acoust. Speech Signal Process. 37, 1846–1856 (1989)

46. Y. Ephraim: Statstical-model-based speech enhancement systems, Proc. IEEE 80, 1526–1555 (1992)

47. Noise Reduction in Speech Processing / Benesty J., Chen J., Huang Y., Cohen I. – Springer-Verlag: Berlin, Heidelberg, 2009.

48. Naylor P. Speech Dereverberation / Naylor P., Gaubitch N. – Springer-Verlag: London, 2010.

49. Yoshioka T. Making Mashine Understand Us in Reverberant Rooms / T. Yoshioka, A. Sehr, M. Delcroix et al. // IEEE Signal Processing Magazine. – 2012, Nov. – Vol. 29, No. 6. – P. 114-126.

50. Lebart K. A new method based on spectral subtraction for speech dereverberation / K. Lebart, J. Boucher, P. Denbigh //Acta Acoustica. – 2001. – Vol. 87, No. 3. – P. 359–366.

51. Habets E. Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement: dissert. … Doctor of Philosophy Electrical Engineering / Emanuel Habets. –Technische Universiteit Eindhoven, 2007. – 241 p.

52. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука. Главная редакция физико-математической литературы, 1979.

53. Schroeder R.M. New Method of Measuring Reverberation Time / JASA, 1964. – P.409-412.

54. Ratnam R., Jones D.L., Wheeler B.C. at al. Blind estimation of reverberation time / J. Acoust. Soc. Am. – 2003. – Vol. 114, No. 5. – P.2877-2892.

55. Lollmann H., Yilmaz E., Jeub M., and Vary P. An improved algorithm for blind reverberation time estimation / Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC). – 2010, Aug.

56. Lollmann H. Estimation of the Frequency Dependent Reverberation Time by Means of Warped Filter-Banks // Lollmann H., Vary P. / Proceedings of

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). – 2011. – P.309-312.

57. Jeub M. Joint Dereverberation and Noise Reduction for Binaural Hearing Aids and Mobile Phones: Master of Sc. Degree Dissert. / Jeub M. – Aahen, Germany, 2012. – 174 p.

58. Nannariello J. The prediction of reverberation time using neural network analysis // Nannariello J., Fricke F. / Appl. Acoust. – 1999. – No.58. – P.305–325.

59. Cox T.J. Extracting room reverberation time from speech using artificial neural networks // Cox T.J., Li F., Darlington P. / J. Audio Eng. Soc. – 2001. – No.49. – P.219–230.

60. Neely S.T. Invertibility of room impulse response // Neely S.T., Allen J.B. / J. Acoust. Soc. Am. – 1979. – Vol.66. – P.165-169.

61. Miyoshi M. Inverse filtering of room impulse response // Miyoshi M., Kaneda Y. / IEEE Trans. Acoust., Speech, Signal Process. – 1988. – Vol.36. – P.145-152.

62. Gaubitch N. Performance comparison of algorithms for blind reverberation time estimation from speech // Gaubitch N., Lollmann H.W., Jeub M. et al. / Acoustic Signal Enhancement; Proc. of IWAENC 2012; Internat. Workshop on 4-6 Sept. 2012. – P.1-4.

63. Gunawan T. Perceptual speech enhancement exploiting temporal masking properties of human auditory system / T. Gunawan, E. Ambikairajah, J. Epps // Speech Commun. – 2010 . – No.52. – P.381-393.

64. Stern R. Hearing is Believing. Biologically inspired methods for robust automatic speech recognition // Stern R., Morgan N. / IEEE Signal Processing Magazine. – 2012. – №29. – P.34-43.

65. Hermansky H. RASTA processing of speech // H. Hermansky, N. Morgan / IEEE Trans. Acoust., Speech, Signal Process. – 1994. – Vol. 2. – P.145-152.

66. Hilger A. Quantile based histogram equalization for noise robust large vocabulary speech recognition // F. Hilger, H. Ney / IEEE Trans. Audio, Speech, Lang. Process. – 2006. – Vol. 14. – №3. – P.845-854.

67. X. Zhang, M. Heinz, I. Bruce, L. Carney. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression / JASA, 2001. – Vol. 109. – №2. – P.648-670.

68. Kim C. Physiologically motivated synchrony-based processing for robust automatic speech recognition / C. Kim, Y. Chiu, Stern R // Proceedings of Interspeech. –2006. – P. 1975 - 1978.

69. Atal B., Rabiner L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition / B. Atal // Acoustics, Speech and Signal Processing. – 1976. – Vol. 24, № 3. – P. 201-212.

70. Benyassine A., Shlomot E., Su H.-Y. et. al. ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications / A. Benyassine // IEEE Communications Magazine. – 1997. – Vol. 35. – P.64-73.

71. Архипов И. А., Гитлин В. Г., Лузин Д. А. адаптивный алгоритм принятия решения «ТОН-НЕ ТОН», синхронный с основным тоном / И. А. Архипов // Речевые технологии. – 2009. –№1. – С. 80-93.

72. Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, ETSI EN 301 708 Rec., ETSI, 1999.

73. Martin R. Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics /R. Martin // IEEE Transactions On Speech and Audio Processing. – 2001. – vol. 9. – No.5. – P.504-512.

74. Cohen I. Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging /I. Cohen // IEEE Transactions On Speech and Audio Processing. – 2003. – vol. 11. – No.5. – P.466-475.

75. Marzinzik M., Kollmeier B. Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics/ M. Marzinzik // IEEE

Transactions On Speech and Audio Processing. – 2002. – vol. 10. – No.2. – P.109-118.

76. Sohn J., Kim N., Sung W. A Statistical Model-Based Voice Activity Detection / J. Sohn // IEEE Signal Processing Letters. – 1999. – vol. 6. – P.1-3.

77. Ying D., Yan Y. Dang J., Soong F. K. Voice Activity Detection Based on an Unsupervised Learning framework / D. Ying // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). –2011. – Vol.19, № 8. – P. 2624 -2633.

78. Pham T., Tang C. Using Artificial Neural Network for Robust Voice Activity Detection Under Adverse Conditions / T. Pham // Proceedings of IEEE International Conference on Computing and Communication Technologies. RIVF '09. – 2009. – P. 1 - 8.

79. Ramírez J., Segura C., Benitez C., Torre A, Rubio. An Effective Subband OSF-Based VAD With Noise Reduction for Robust Speech Recognition /J. Ramírez // IEEE Transactions On Speech and Audio Processing. – 2005. – vol. 13. – No.6. – P.1119-1128.

80. Doblinger G. Computationally efficient speech enhancement by spectral minima tracking in subbands / G. Doblinger // Proceedings of Eurospeech. – 1995. – Vol.2. – P.1513-1516.

81. Hirch H., Ehrlicher C. Noise estimation techniques for robust speech recognition / H. Hirch // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). –1995. – P. 152 -156.

82. Qi Y., Hunt B. R. Voiced-Unvoiced-Silence Classifications of speech Using Hybrid features and a network / Y. Qi // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). –1993. – Vol.1, № 2. – P. 250 -255.

83. Vary P., Martin R. Digital Speech Transmission: Enhancement, Coding and Error Concealment. Chichester, England: John Wiley & sons, 2006.

84. Ghasemi J., Afzalian A., Karami Mollaei. A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform / J. Ghasemi // Signal Processing. – 2010. – Vol.4, Issue 1. – P.54-61.

85. Martin A., Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC / A. Martin // Proceedings of ICASSP'01. –2001. – Vol.1. – P. 237-240.

86. Kos M. Noise Reduction Algorithm for Robust Speech Recognition Using Minimum Statistics Method and Neural Network VAD / M. Kos // Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on. – 2007. – P. 284-287.

87. Young S. et al, The HTK Book // Cambridge University Engineering Department, 2005, 354 p.

88. Loizou P. Speech enhancement: Theory and Practice / P. Loizou // Boca Raton: CRC Press, 2007. – 648 p.

89. Beerends J. Measurement of speech intelligibility based on the PESQ approach / J. Beerends, E. Larsen, N. Iyer, J. van Vugt // Measurement of Speech and Audio Quality in Networks (MESAQIN): int. conf., 2 June 2004, Prague, Czech Republic. – Prague, 2004.

90. Дідковський В.С., Дідковська М.В., Продеус А.М. Комп'ютерна обробка акустичних сигналів. Навчальний посібник – Київ, "Імекс-ЛТД", 2010. – 420 с.

91. Schroeder R.M. New Method of Measuring Reverberation Time / JASA, 1964. – P.409-412.

92. The HTK Book / S. Young, G. Evermann, M. Gales, et al. – Cambridge: University Engineering Department, 2009. – 375 p.

93. Brooks M. VOICEBOX: Speech Processing Toolbox for MATLAB / Brooks M. // Imperial College London, Electrical Engineering Department. – Режим доступа: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. – Дата доступа: 12.05.2014. – Imperial College London.

94. Jeub M. Blind Reverberation Time Estimation / Jeub M. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation.

105. Осовский С. Нейронные сети для обработки информации.: Пер. с польского Рудницкого И. Д. – М.: Финансы и статистика, 2002. – 344 с.

106. Хайкин С. Нейронные сети: полный курс, 2-е издание: Пер.с англ.. – М.: Издательский дом «Вильямс», 2006. – 1104с.

107. LeCun Y., Bottou L., Orr G., Muller K. Efficient BackProp / Y. LeCun // Neural Networks: Tricks of the trade. – Springer Verlag. – 1998. – P. 5-50.

108. Haykin S. Neural networks, a comprehensive foundation /S. Haykin // Macmillan College Publishing Company. – N.Y. – 1994.

109. Sutton R. Adapting Bias by Gradient desent: an Incremental Version of Delta-Bar-Delta/ R. Sutton // Proceedings of the 10th National Conf. on Artificial Intelligence. – MIT Press. – 1992. – P.171-176.

110. Hampson S., Volper D. Linear function neurons: Structure and training/ S. Hampson // Biological cybernetics. – No.53. – 1986. – P. 203-217.

111. Jacobs R. Increased Rates of Convergence Through Learning Rate Adaptation / R. Jacobs // Neural Networks. – 1988. – vol. 1. – P.295-307.

112. Rangachari S., Loizou P. A noise-estimation algorithm for highly non-stationary environments / S. Rangachari // Speech Commun. – 2006 . – No.48. – P.220-231.

113. Martin R. Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics /R. Martin // IEEE Transactions On Speech and Audio Processing. – 2001. – vol. 9. – No.5. – P.504-512.

114. Zue V., Sneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, №4. – P. 391-395.

115. Методы обработки речевых сигналов во временной области / Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. Пер. с англ. – М.: Радио и связь, 1981. – С.110-160.

116. Wilson D., Martinez T. The general inefficiency of batch training for gradient descent learning /D. Wilson // Neural Networks. – 2003. – vol.16. Issue 10. – P.1429-1451.

117. Morales N., Javier T., Javier G., Colas J., Toledano D.T. STC-TIMIT: Generation of Single-channel Telephone Corpus // Proc. of LREC-2008 – 2008. – P. 391-395.

118. Jankowski C., Kalyanswamy A., Basson S., Spitz J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database // Proc. of ICASSP-90. – 1990. – P. 109-112.

119. Kola J., Espy-Wilson C., Pruthi T. Voice Activity Detection. MERIT BIEN. – 2011.

120 Davis A., Nordholm S.,Togneri R. Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold / A. Davis // IEEE Transactions On Speech and Audio Processing. – 2006. – vol. 14. – No.2. – P.412-423.

121. Ferroni G., Bonfigli R., Principi E., et al. Neural Networks Based Methods for Voice Activity Detection in a Multi-room Domestic Environment // XIII AI*IA Symposium on Artificial Intelligence. – Pisa, Italy. – 2014.

122. Mauch M., Ewert S. The Audio Degradation Toolbox and its Application to Roubustness evaluation / M. Mauch // Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013), 2013.

ПРИЛОЖЕНИЯ

Приложение А

Акты внедрения результатов диссертационной работы

Державний концерн
УКРОБОРОНПРОМ
Державне підприємство
«Київський науково-дослідний
Інститут
ГІДРОПРИЛАДІВ»

Україна, 03035, Київ, вул. Сурікова, 3
Тел. (380-44) 239-90-18,
факс(380-44) 239-90-17
e-mail: office@hydrodevices.kiev.ua
ryba@ukrpack.net
WWW.HYDRODEVICES.KIEV.UA

The State Concern
UKROBORONPROM
State Enterprise
«Kyiv Scientific Research
Institute
Of HYDRODEVICES»

3, Surikova str., Kyiv, 03035, Ukraine
Tel.: (380-44) 239-90-18,
fax:(380-44) 239-90-17
e-mail: office@hydrodevices.kiev.ua
ryba@ukrpack.net
WWW.HYDRODEVICES.KIEV.UA

01.12.2015 № 714/дон.
на № _____ від _____

ЗАТВЕРДЖУЮ
Заступник директора з наукової роботи Державного підприємства
"КНДІ ГІДРОПРИЛАДІВ"
К.В. Ковальчук
01.12.2015 р.

АКТ ВПРОВАДЖЕННЯ

результатів дисертаційної роботи **Ладошко Ольги Миколаївни**
«Підвищення робастності систем автоматичного розпізнавання мовлення методами обробки сигналів»
на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.09.08 – «Прикладна акустика та звукотехніка»

При виконанні дослідної конструкторської роботи «Зірниця-58250» використані наступні результати дисертаційної роботи:

1) Удосконалено метод оцінювання спектру реверберації:

1. Ладошко О. М. Залежність показників систем ослаблення реверберaційної завади від ступеня спотворення сигналу/ Продеус А. М., Ладошко О. М. // Стандартизація, сертифікація, якість – 2014 – №3(88). – С. 45-49.
2. Ladoshko O. On existence of optimal boundary value between early reflections and late reverberation / Prodeus A., Ladoshko O. // Proc. of IEEE 34th International Scientific Conference «Electronic and Nanotechnology», 15 – 18 April 2014: Proceedings. – Kyiv, 2014. – P. 442-446.

2) Запропоновано використання нейромережевого алгоритму прийняття рішення в системі виявлення сигналу:

1. Ладошко О. М. Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок мовлення / Ладошко О. М., Бондаренко І.Ю. // Електроніка та зв'язок. - 2012. - №6 (71). - С. 19-25.
2. Ладошко О.М. Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок усного мовлення / Ладошко О.М., Бондаренко І.Ю. // Одинадцята всеукраїнська міжнародна конференція «Оброблення сигналів і зображень та розпізнавання образів» (УкрОбраз'2012), 15-19 жовтня 2012 р.: тези доп. – Київ: УАсОІРО. – 2012. – С. 55 – 58

Використання результатів дисертаційної роботи дозволило підвищити надійність розв'язання задачі виявлення сигналу на тлі завад, а також дозволило підвищити точність вимірювання інформативних параметрів сигналу шляхом пригнічення дії шумової та

реверберaційної завад. Таким чином, результати дисертаційної роботи є науково та практично цінними та можуть бути використані в подальшому при розробці перспективних систем виявлення та класифікації акустичних сигналів.

Головний конструктор ДКР «Зірниця-58250» **І.М. Фалєєв**

ЗАТВЕРДЖУЮ
Перший проректор Національного технічного
університету України "КПІ"
академік НАН України, професор
_____ **Ю. І. Якименко**
"04" 03 2016 р.

Д О В І Д К А

про впровадження в учбовий процес кафедри
Акустики та акустоелектроніки
Національного технічного університету України "КПІ"
результатів дисертаційної роботи Ладошко Ольги Миколаївни
на тему "Підвищення робастності систем автоматичного
розпізнавання мовлення методами обробки сигналів",
поданої на здобуття ученого ступеню кандидата технічних наук
за спеціальністю 05.09.08 – прикладна акустика та звукотехніка

Наукові положення та результати дисертаційної роботи Ладошко Ольги Миколаївни на тему "Підвищення робастності систем автоматичного розпізнавання мовлення методами обробки сигналів" використовуються для викладання навчальних курсів спеціалістам і магістрам напряму підготовки "Акустотехніка" за спеціальністю "Акустичні засоби та системи", а саме: у курсі лекцій з дисципліни "Пристрої реєстрації та відображення інформації", у розділі «Кодування акустичних сигналів», а також у курсі лекцій з дисципліни «Комп'ютерні акустичні системи», у розділі «Корекція мовленнєвих сигналів».

Декан факультету електроніки
д.т.н., професор Жуйков В. Я.

Завідувач кафедри АтаАЕ
д.т.н.,професор Дідковський В. С.