На правах рукописи

**Нгуен Чи Тхиен**

# Модели и алгоритмы распознавания коротких речевых команд на основе пробных спектральных преобразований входного сигнала

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Тула, 2014

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Тульский государственный университет».

Научный руководитель — **Двоенко Сергей Данилович**, доктор физико-математических наук, доцент

Официальные оппоненты: **Корсун Олег Николаевич**, доктор технических наук, профессор, Федеральное государственное унитарное предприятие «Государственный научно-исследовательский институт авиационных систем», начальник лаборатории
**Чучупал Владимир Яковлевич**, кандидат физико-математических наук, Федеральное государственное бюджетное учреждение науки «Вычислительный центр имени А.А. Дородницына Российской академии наук», ведущий научный сотрудник

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Институт проблем управления имени В.А. Трапезникова Российской академии наук»

Защита диссертации состоится « 11 » ноября 2014 г. в 14 часов на заседании диссертационного совета Д 212.271.05 при ФГБОУ ВПО «Тульский государственный университет» (300012, г. Тула, пр. Ленина, 92, 12 – 105).

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Тульский государственный университет» по адресу: 300012, г. Тула, пр. Ленина, 92.

Автореферат разослан «__» _____ 2014 г.

Ученый секретарь диссертационного совета — Соколова Марина Юрьевна

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность работы.** В настоящее время сохраняется большой интерес исследователей к задачам компьютерной обработки речи, таким как их кодирование (Gibson J., Chu W.), генерация (Лобанов Б.М., Taylor P.), а также распознавание (Woelfel M., Neustein A.).

Особый интерес к компьютерной обработке именно речи в значительной мере определяется тем фактом, что это естественный вид взаимодействия между людьми, а также между человеком и машиной.

Речь моделируется в компьютере как последовательность скалярных или векторных числовых значений, и эта совокупность упорядочена вдоль дискретной оси времени. Эту временную последовательность принято назвать речевым сигналом. К настоящему времени разработано большое количество численных методов обработки речевых сигналов.

Одной из известных задач обработки речевых сигналов является задача распознавания речевых команд. В данной задаче необходимо принять решение о том, к какому классу относится речевой сигнал, где классом назовём множество разных произношений одной и той же команды.

В классической теории распознавания образов объекты (Вапник В.Н., Ту Дж.), подлежащие распознаванию, описываются векторами фиксированной размерности и представляются точками в пространстве своих характеристик. Однако в задаче распознавания речевых команд фиксация размерности сигналов не вполне естественна. Например, одну и ту же речевую команду диктор произносит с разными скоростями. В результате, длины записанных речевых сигналов одной команды являются различными.

Речевые сигналы характеризуются большой вариабельностью. Они отличаются не только по длине, но и по высоте тона, тембру, которые зависят от характеристики голоса дикторов. В построении систем распознавания речевых команд для того, чтобы обеспечить репрезентативность обучающей выборки необходимо собрать речевые сигналы с многих разных дикторов. Собрание большого количества обучающих данных для необходимого набора речевых команд не всегда оказывается возможным, особенно в случае персонального пользователя системы распознавания.

Учитывая трудность в собрании обучающих речевых сигналов, в данной работе предлагается способ решения задачи распознавания речевых команд, который компенсирует малую обучающую выборку использованием имеющегося опыта из разных областей обработки речевых сигналов: кодирования, преобразования и распознавания. Когда обучающая выборка мала, построенная система распознавания дикторозависима (Fontaine V., Bourlard H.), т.е. она будет распознавать речевые команды «своих» пользователей (людей, которые обучали эту систему) с точностью распознавания, которая будет выше, чем точность, взятая по «чужим» пользователям. Поэтому для улучшения качества распознавания речевых команд в случае «чужого» пользователя предлагается преобразование речевых сигналов «чужого» пользователя к речевым сигналам «своего» пользователя перед тем, как подать сигнал на вход алгоритма распознавания. Такая

идея встречается в работе Загоруйко Н.Г. о подстройке под диктора при распознавании ограниченного набора устных команд, где преобразование и распознавание выполняются с помощью функций расстояния. В данной работе преобразование речевых сигналов и их распознавание реализованы с помощью функций правдоподобия (Pratt J.).

На практике результат распознавания сигналов как своего, так и чужого дикторов дополнительно ухудшается шумом. Обучающие речевые сигналы обычно являются незашумленными, а тестирующие речевые сигналы оказываются зашумленными. Присутствие шума приводит к сильному отклонению спектров тестирующих речевых сигналов от спектров их эталонов в обучающей выборке. Поэтому качество результата распознавания на фоне шумов резко падает (Wolfe J.).

Для уменьшения отклонений спектров тестирующих зашумленных речевых сигналов от спектров их незашумленных эталонов в обучающей выборке были предложены разные способы. Самый популярный подход – это удаление из спектров зашумленных сигналов шумовой составляющей (Haykin S., Vaseghi S.). Такой подход реализован в методе спектрального вычитания (spectral subtraction) и методе фильтрации Винера (Wiener filtering). Недостаток этих методов заключается в том, что перед удалением шума из спектров речевых сигналов должна быть известна априорная информация о шуме. Сам процесс выявления априорной информации о шуме вызывает трудности. Кроме того, если шум нестационарный, то его удаление сильно искажает спектр исходного сигнала, а в худшем случае нарушает формантную структуру его спектра.

Существует и другой подход, заключающийся в умножении значений отсчетов амплитудного спектра фрагментов каждого речевого сигнала на весовой параметр (Hung J.). Цель этого метода – подчеркнуть спектральное различие между речевыми и неречевыми (паузы) фрагментами сигнала. Этот метод был предложен для распознавания слитной речи.

Для задачи распознавания отдельных речевых команд, в которых нет пауз, этот метод не подходит. Необходимо найти метод, применимый к задаче распознавания отдельных речевых команд.

Поэтому необходимо обобщить уже имеющийся опыт из различных областей обработки речевых сигналов (кодирование, преобразование, распознавание) и применить его для решения актуальной задачи распознавания речевых команд, предложив новые и улучшенные подходы, обладающие элементами новизны на каждой из этапов ее решения.

**Цель работы.** Решение задачи распознавания речевых команд.

**Задачи исследования.** Для достижения указанной цели в диссертации поставлены следующие задачи:

1. сформулировать и исследовать задачу идентификации модели речевого сигнала с целью адекватного восприятия;
2. решить задачу распознавания речевых команд при недостаточном объеме обучающих данных;
3. решить задачу распознавания речевых команд на фоне шумов;
4. оценить предложенные решения процедурой скользящего контроля.

**Объект и предмет исследования.** Объектом исследования является задача распознавания речевых команд. Предметом исследования являются повышение качества работы алгоритма распознавания речевых команд и понижение требований к вычислительным ресурсам для алгоритма распознавания.

**Положения, составляющие научную новизну и выносимые на защиту:**

1. новый подход к решению задачи распознавания речевых команд как задачи идентификации модели речевого сигнала с целью адекватного восприятия, обобщающий имеющийся опыт обработки речевых сигналов;
2. новое решение задачи распознавания речевых команд при недостаточном объеме обучающих данных на основе упрощения процедуры распознавания слитной речи;
3. новое решение задачи распознавания речевых команд на фоне шумов, используя увеличение значений отсчетов амплитудных спектров речевых сигналов на константу;
4. процедуры оптимизации параметров полученной в итоге эвристической модели речевых сигналов с целью улучшения качества их распознавания;
5. упрощенные модели классов речевых сигналов как нормальных распределений или смесей нормальных распределений мел-частотных кепстральных коэффициентов.

**Методы исследования.** Теоретическое исследование основано на применении методов обработки цифровых сигналов, теории распознавания образов, методов оптимизации.

Экспериментальные исследования осуществлены на известных реальных данных (TIDigits).

**Достоверность полученных результатов** подтверждается процедурой скользящего контроля и экспериментами реальных данных.

**Практическая значимость работы.** Разработанные алгоритмы могут применяться для решения широкого класса прикладных задач анализа речевых сигналов, в частности, для распознавания речевых команд.

**Реализация результатов работы**. Результаты исследований реализованы в виде комплекса программ, использованного во вьетнамских компаниях «FTS», «Vietnam TeleCommunication Corporation».

**Апробация работы.** Основные результаты работы докладывались на VI магистерской конференции ТулГУ (г. Тула, 2011 г.), VIII Всероссийской конференции «Информационные системы и модели в научных исследованиях, промышленности, образовании и экологии» (г. Тула, 2011 г.), VII Конференции «Молодёжные инновации» (г. Тула, 2012 г.), XIV Международной конференции «Современные проблемы математики, механики, информатики» (г. Тула, 2013 г.), Конференции «Инновационные наукоемкие информационные технологии» (г. Тула, 2013 г.), XI Всероссийской конференции «Информационные технологии, системный анализ и управление» (г. Таганрог, 2013 г.), III Международной конференции «Информационные технологии и системы» (г. Челябинск, 2014 г.), XVI Международной конференции «Цифровая обработка сигналов и её применение» (г. Москва, 2014 г.), IV Всероссийской конференции «Актуальные

вопросы современной информатики» (г. Москва, 2014 г.), XII Всероссийском совещании по проблемам управления (г. Москва, 2014 г.).

**Публикации.** По материалам диссертации опубликованы 13 работ из них 3 в научных изданиях, рекомендованных ВАК при Минобрнауки РФ.

**Структура и объем работы.** Диссертация состоит из введения, четырех глав, заключения, списка литературы. Материал изложен на 162 страницах, содержит 95 рисунков, список литературы из 83 наименований.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

**Во введении** обоснована актуальность работы, описана общая структура диссертации.

**В первой главе** рассмотрены основные задачи обработки речевых сигналов.

Кодирование речевого сигнала представляет собой процесс сжатия речевого сигнала, устранение его избыточности, сохраняя приемлемое качество.

Пусть сигнал $Y = (y_1,..., y_T)$ означает произношение какой-то речевой команды, где $T$ – целое, положительное. Отсчеты $y_t, t = 1,..., T$ принимают вещественные значения.

Дискретные отсчеты речевого сигнала обрабатываются фрагментами с определенным периодом $L$. Фрагменты имеют длину $N$ – количество отсчетов речевого сигнала во фрагменте. Формально $i$-й фрагмент представлен следующим описанием:

$$Y_{t_i}^{t_i+N-1} = (y_t; t_i \le t \le t_i + N - 1),\ 1 \le t_i \le T - N + 1,\ L = t_{i+1} - t_i.$$

Для каждого фрагмента речевого сигнала строится его кратковременный спектр. Удобно считать последовательность кратковременных амплитудных спектров $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3,...)$ моделью речевого сигнала $Y$, где амплитудный спектр $\mathbf{a}_i$ представляет собой вектор отсчетов $\mathbf{a}_i = (a_i^k)$. Применив окно Хэмминга, отсчеты амплитудного спектра определяются по формуле:

$$a_i^k = \left| \sum_{n=1}^{N} y_{t_i+n-1} w_n e^{-j\frac{2\pi}{N}(n-1)(k-1)} \right|,\ k = 1,..., \frac{N}{2}.$$

где $w_n$ – это $n$-ый отсчет окна Хэмминга.

Таким образом, речевой сигнал представляется последовательностью $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3,...)$, где $i$-й кратковременный спектр представлен своими отсчетами $\mathbf{a}_i = (a_i^k, 1 \le k \le N/2)$.

Спектральный анализ речевых сигналов является очень развитой областью исследования (Котельников В. А., Пугачев В.С.). Разные методы и способы обработки речи в частотной области были предложены во многих работах. Но описание речевого сигнала последовательностью его кратковременных

спектров имеет недостаток в том, что количество отсчетов, необходимое для представления кратковременного спектра, оказывается большим. В задаче распознавания речевого сигнала такой недостаток считается нежелательным. Задача требует более компактного описания речевого сигнала, которое одновременно сохраняет его различительный характер. В данном разделе рассматривается кодирование кратковременного амплитудного спектра мел-частотными кепстральными коэффициентами (МЧКК), учитывая особенность слухового аппарата человека (Lieberman P.).

В работе описано построение $P$ перекрывающихся окон в доступном диапазоне частот.

Для построенных перекрывающихся окон МЧКК строятся как результат дискретного косинусного преобразования от логарифма сумм отсчетов амплитудного спектра в некотором окне:

$$x^m = \sum_{i=1}^{P} \ln\left( \sum_{k=1}^{N/2} w_i^k a^k \right) \cos\left( m\left( i - \frac{1}{2} \right) \frac{\pi}{P} \right),\ m = 1, \mathbf{K}, M/2,$$

где $w_i^k$ – это $k$-ый отсчет $i$-ого окна.

Вектор $M/2$ МЧКК используется для характеристики каждого кратковременного амплитудного спектра, т.е. статической характеристики речевого сигнала в рамке одного фрагмента. Для описания динамической характеристики речевого сигнала через фрагменты используются дифференциальные мел-частотные кепстральные коэффициенты (ДМЧКК), представляющие собой сумму различий МЧКК между фрагментами:

$$\Delta x_t^m = \frac{\sum_{d=1}^{D} d(x_{t+d}^m - x_{t-d}^m)}{2\sum_{d=1}^{D} d^2},\ m = 1, \mathbf{K}, M/2,$$

где $x_t^m$ – это $m$-й МЧКК $t$-ого фрагмента речевого сигнала, $D$ – уровень соседства МЧКК, $m$-й ДМЧКК $\Delta x^m$ описывает изменение $m$-го МЧКК $x^m$ через фрагменты, т.е. «время» $t$.

С целью упрощения предлагается, что $x_t^{m+M/2} = \Delta x_t^m,\ m = 1, \mathbf{K}, M/2$.

В целом, последовательность $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, ...)$ кратковременных амплитудных спектров $\mathbf{a}_i = (a_i^k, 1 \le k \le N/2)$ характеризуется последовательностью $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ...)$ векторов МЧКК $\mathbf{x}_t = (x_t^m, 1 \le m \le M)$, где обычно применяется $M << N/2$.

В работе описано преобразование речевых сигналов. Известно, что среднее расстояние между формантами (локальные максимумы огибающей спектра) зависит от характеристики голоса говорящего (Huber J.). Таким образом, можно выполнить преобразование речевого сигнала $Y$ преобразованием его ампли-

тудного спектра $\mathbf{a}$, используя имеющуюся в нашем распоряжении функцию преобразования $\varphi(\mathbf{a}, a)$, где $a$ – параметр преобразования.

Необходимо, чтобы амплитудный спектр расширялся, если $a<1$, и сжимался, если $a>1$. Напомним, что амплитудный спектр является функцией от угловой частоты $a^k = \phi(\omega_k)$, где $a^k$ – это $k$-ый отсчет спектра $\mathbf{a}$, $\omega_k$ – фиксированная нормированная частота $\omega_k \in [0, \pi]$.

Известно, что эффект расширения (сжатия) спектра может быть достигнут путём простого искажения оси частот (Huber J.). Расширенный (сжатый) спектр определяется выражением $\tilde{a}^k = \phi(\tilde{\omega}_k)$, где $\tilde{\omega}_k$ – искаженная частота, $\tilde{a}^k$ – $k$-ый отсчет искаженного спектра $\tilde{\mathbf{a}}$. В работе преобразование спектра сигнала путём искажения оси частот применяется следующим образом (Uebel L.):

$$\tilde{\omega} = \begin{cases} a\omega, & \omega \le b \\ ab + \dfrac{\pi - ab}{\pi - b}(\omega - b), & \omega > b \end{cases}, \quad \begin{cases} b = \dfrac{7\pi}{8} \text{ при } a < 1 \\ b = \dfrac{7\pi}{8a} \text{ при } a > 1 \end{cases}.$$

Рассмотрена задача распознавания речевых сигналов. Пусть применяются $V$ речевых команд, т.е. $V$ классов сигналов $v = 1,2,...,V$. Классом назовём множество разных произношений одной и той же команды. Обозначим каждый класс $\lambda^v$, $v = 1, \mathbf{K}, V$. Пусть поступил речевой сигнал $X = (\mathbf{x}_t, t = 1, \mathbf{K}, \tau)$. Необходимо принять решение о том, к какому классу сигналов он принадлежит. Предлагается использовать байесовский классификатор с решающим правилом вида (Ту Дж., Гонсалес Р.):

$$v^* = \arg\max_{\lambda^v} p(X \mid \lambda^v)\, p(\lambda^v), \quad v = 1,2,...,V.$$

Пусть априорные вероятности классов $p(\lambda^v)$ равны для всех $v = 1, \mathbf{K}, V$. Тогда решающее правило упрощается

$$v^* = \arg\max_{\lambda^v} p(X \mid \lambda^v), \quad v = 1,2,...,V.$$

Апостериорное распределение $p(X \mid \lambda)$ для класса $\lambda$ определяется на основе представления речевого сигнала в виде двухкомпонентного случайного процесса (Рабинер Л.). Двухкомпонентный случайный процесс содержит в себе наблюдаемую компоненту, представляющую собой векторы МЧКК, и скрытую компоненту, являющуюся скрытой марковской моделью. В работе рассмотрены эти компоненты двухкомпонентного случайного процесса. Построена процедура настройки параметров марковской модели классов речевых сигналов.

**Во второй главе** сформулирована и исследована задача идентификации модели речевого сигнала с целью адекватного восприятия. В данной задаче сле-

дует обобщить уже имеющийся опыт обработки речевых сигналов с целью увеличения качества их обработки.

Задачу идентификации модели речевого сигнала с целью адекватного восприятия по нашему мнению следует решать в три этапа.

Шаг 1 - идентификация. Предполагается, что речевой сигнал может кодироваться в соответствии с некоторой известной моделью. Предположив, что эта модель параметрическая, следует оценить значения его параметров для заданного речевого сигнала.

Шаг 2 - генерация. Если модель речевого сигнала идентифицирована, то предполагается, что можно варьировать параметры модели, добиваясь изменения речевого сигнала.

Шаг 3 - адекватное восприятие. Предполагается, что речевой сигнал воспринимается и интерпретируется. Будем считать, что в роли воспринимающего и интерпретирующего устройства выступает человек или группа лиц (испытуемых). Кроме того, в этой роли может быть использована и соответствующая техническая система. Будем считать, что восприятие речевого сигнала является адекватным, если он распознается испытуемыми (или технической системой).

Легко заметить, что задача идентификации модели речевого сигнала с целью адекватного восприятия обладает многими общими чертами с другими задачами речевой технологии: кодированием речи на первом этапе, преобразованием речи на втором этапе, распознаванием речи на третьем этапе.

Таким образом, этапы решения данной задачи означают выполнение вполне определенной обобщенной “процедуры” обработки, шаги которой определены рассмотренными выше этапами.

Если сгенерированный речевой сигнал адекватно воспринимается, то процедура заканчивается. В противном случае она возвращается ко второму шагу с другим набором параметров модели.

Формально, обобщенная “процедура” обработки представляет собой суперпозицию функций, отображающих одно описание речевого сигнала в его другое описание. Отображения формируются на основе экспертных знаний о природе речевого сигнала и цели его обработки. Цели таких отображений могут состоять в уменьшении объема сигнала в задаче кодирования, преобразовании сигнала в номер класса (число) заранее известных классов сигналов в задаче распознавания речевых сигналов и т.д.

В соответствии с поставленной задачей решение задачи распознавания речевых команд заключается в следующем.

Рассматривая задачу распознавания речевых команд как задачу идентификации модели речевого сигнала с целью адекватного восприятия, этапы решения данной задачи означают выполнение вполне определенной обобщенной “процедуры” обработки. Ниже приведен однократный алгоритм распознавания речевых команд (ОАРРК):

1. Построить последовательность $A=(\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,...)$ кратковременных спектров $\mathbf{a}_i=(a_i^k,1\le k\le N/2)$ из речевого сигнала $Y=(y_1,...,y_T)$ использованием параметрической функции отображения $A=G_Y(Y,N)$.

2. Получить последовательность $X=(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,...)$ векторов МЧКК $\mathbf{x}_t=(x_t^m,1\le m\le M)$ из последовательности $A=(\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,...)$ кратковременных амплитудных спектров использованием параметрической функции отображения $X=G_A(A,M,P,D)$.

3. Определить номер класса $v^*$ сигнала для последовательности $X=(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,...)$ векторов МЧКК использованием параметрической функции: $v^*=G_X\left(X,\lambda^v,v=1,2,...,V\right)$, где $G_X=\arg\max_{\lambda^v} p(X\mid\lambda^v)$, $v=1,2,...,V$, и параметр $\lambda^v$ описывает $v$-ый класс сигналов.

Структура функций $G_Y$, $G_A$, $G_X$ уже известна, но они полностью определены только после определения их параметров $N,M,P,D$ и $\lambda^v$, $v=1,2,...,V$. Возникает необходимость определения параметров функций отображения.

В работе предлагается новая процедура подбора параметров алгоритма распознавания речевых команд (ОАРРК) на основе скользящего контроля. При подборе параметров предполагается, что параметры $\lambda^v$, $v=1,2,...,V$ описывают классы сигналов не как двухкомпонентные случайные процессы, а как нормальные распределения МЧКК.

Предполагается, что МЧКК речевых сигналов каждого класса распределены нормально с параметрами, характерными для своего класса, что позволяет вычислить ковариационную матрицу $\Sigma_v$ и вектор средних значений $\boldsymbol{\mu}_v$ распределения каждого класса сигналов $v=1,2,...,V$. Такая модель проще двухкомпонентной модели речевого сигнала.

Вероятность того, что сигнал $X=(\mathbf{x}_t,t=1,...,\tau)$ принадлежит классу сигналов $\lambda^v$, вычисляется через вероятности их элементарных векторов МЧКК:

$$p(X\mid\lambda^v)=\prod_{t=1}^{\tau}p(\mathbf{x}_t\mid\lambda^v)$$

В свою очередь, условная плотность распределения $p(\mathbf{x}_t\mid\lambda^v)$ вычисляется по многомерному нормальному распределению:

$$p(\mathbf{x}_t\mid\lambda^v)=\frac{1}{\sqrt{(2\pi)^M\left|\Sigma_v\right|}}\exp\left(-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_v)^{\mathrm{T}}(\Sigma_v)^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_v)\right)$$

Численный метод подбора параметров $N,M,P,D$ является следующей процедурой:

1. Для каждого класса сигналов $v=1,2,...,V$ задать набор речевых сигналов. Каждый речевой сигнал со своей длительностью описывается соответствующей последовательностью отсчетов.

2. Задать диапазоны допустимых значений параметров $N$, $M,P,D$.

3. Оценка параметров методом скользящего контроля.

3.1 Набор речевых сигналов каждого класса сигналов $v=1,2,...,V$ разделяется на $k$, например 5, равных частей.

3.2 Взять очередные значения параметров $N$, $M,P,D$ из диапазонов допустимых значений и построить МЧКК сигналов всех частей.

3.3 Исключить одну часть и использовать МЧКК сигналов $k-1$ частей для построения модели классов сигналов как нормальных распределений МЧКК с параметрами $\Sigma_v, \boldsymbol{\mu}_v$, $v=1,...,V$.

3.4 Выполнить распознавание речевых сигналов с подсчетом числа ошибок распознавания на наборе речевых сигналов исключенной отдельной части.

3.5 Повторить шаги 3.3-3.4 для каждой отдельной части речевых сигналов и вычислить среднее число ошибок.

3.6 Повторить шаги 3.2-3.5 для всех возможных значений параметров $N$, $M,P,D$ и найти набор $\left(N^*,M^*,P^*,D^*\right)$, обеспечивший наименьшее число ошибок распознавания.

Процедура заканчивается, когда число ошибок распознавания оказывается приемлемым. Тогда для распознавания можно применить классификатор с классами сигналов $\lambda^v$, $v=1,2,...,V$, построенными как нормальные распределения МЧКК. В противном случае применяется более сложный классификатор с классами сигналов $\lambda^v$, $v=1,2,...,V$ как двухкомпонентными случайными процессами или смесями нормальных распределений МЧКК, генерирующимися при использовании подобранных значений параметров $\left(N^*,M^*,P^*,D^*\right)$.

Процедура подбора параметров алгоритмов распознавания применяется для диапазонов значений параметров: $N\in[64, 128, 256, 512, 1024]$, $M\in[16, 18, ..., 32]$, $P\in[20, 21, ..., 30]$, $D\in[1,2,3,4]$.

Оказалось, что для 11 команд, произнесенных по 100 раз (100 «своих» дикторов), оптимальные значения параметров $N^*=512$, $M^*=22$, $P^*=30$, $D^*=3$ доставляют среднее число ошибок $E(N^*,M^*,P^*,D^*)=0.73\%$.

В независимом тесте для тех же 11 команд, произнесенных по 80 раз (другими «чужими» 80 дикторами), однократный алгоритм распознавания (ОАРРК) с адекватным набором значений параметров обеспечивает малое число ошибок распознавания 1.25%, 0.8%, 0.57% соответственно, для классов сигналов как нормальных распределений, смесей нормальных распределений, двухкомпонентных случайных процессов.

**В третьей главе** описывается решение задачи распознавания речевых команд с недостаточным объемом обучающих данных.

Изучив влияние объёма и состава обучающей выборки на качество распознавания речевых команд, можно сказать, что для обеспечения низкого числа ошибок распознавания речевых команд «чужих» дикторов необходимо достаточно большое количество обучающих данных.

В случае малой обучающей выборки для улучшения качества распознавания речевых команд «чужих» дикторов в данной работе предлагается преобра-

зовать их речевые сигналы к сигналам «своих» дикторов перед тем, как подать сигнал на вход классификатора, построенного на речевых сигналах «своих» дикторов.

При выполнении алгоритма ОАРРК речевой сигнал не преобразуется, т.е. параметр преобразования $a=1$. Если учитывается преобразование речевого сигнала в процессе распознавания, т.е. параметр $a$ может меняться, то алгоритм распознавания речевых команд становится многократным (МАРРК):

1. Взять одно значение параметров преобразования $a$ из диапазона $[0.88, 0.9,..., 1.12]$.

2. Построить последовательность $A=(\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,...)$ кратковременных спектров $\mathbf{a}_i=(a_i^k, 1\le k\le N/2)$ из речевого сигнала $Y=(y_1,...,y_T)$ путём использования параметрической функции отображения $A=G_Y(Y,N,a)$ с очередным значением параметра $a$.

3. Получить последовательность $X=(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,...)$ векторов МЧКК $\mathbf{x}_t=(x_t^m, 1\le m\le M)$ из последовательности $A=(\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3,...)$ кратковременных амплитудных спектров использованием параметрической функции отображения $X=G_A(A,M,P,D)$.

4. Вычислить значения $p(X(a)|\lambda^v)$, $v=1,2,...,V$.

5. Повторить шаги 1-4 для всех значений параметра $a$.

6. Среди всех наборов $(a,v)$ найти набор $(a^*,v^*)$, обеспечивающий максимальную вероятность $p(X(a)|\lambda^v)$.

Таким образом, соответствующий номер классов сигналов $v^*$ является:

$$v^*=\arg\max_a\max_{\lambda^v} p(X(a)|\lambda^v),\ \ v=1,2,...,V$$

Обозначив функцию отображения $G_X=\arg\max_a\max_{\lambda^v} p(X(a)|\lambda^v)$, $v=1,2,...,V$, соответствующий номер классов сигналов $v^*$ получается использованием параметрической функции: $v^*=G_X\left(X(a),\lambda^v, v=1,2,...,V\right)$.

Таким образом, при выполнении алгоритма МАРРК подбирается некоторое значение параметра преобразования $a$ из диапазона $[0.88, 0.9,..., 1.12]$. В общем случае применение какого-то значения $a$ из этого диапазона вовсе не означает, что мел-частотное кепстральное представление входного сигнала становится ближе к мел-частотному кепстральному представлению сигналов «своих» дикторов. Но, в итоге, будет выбрано такое оптимальное значение $a$, которое все-таки улучшит качество распознавания входного сигнала, что и означает приближение к мел-частотному кепстральному представлению сигналов «своих» дикторов.

Если количество «своих» дикторов мало, но более одного, то предлагается для каждого из них подобрать свое значение параметра преобразования $a$ с целью уменьшения ошибок скользящего контроля на этапе обучения.

Процедура подбора значения параметра $a$ заключается в следующем:

1. Сгруппировать обучающие сигналы по дикторам, считая, что каждому из них соответствует единственное значение параметра преобразования $a$.

2. Подобрать значение $a$ для каждого диктора методом скользящего контроля.

2.1 Исключить сигналы текущего диктора из выборки. На сигналах остальных дикторов при $a=1$ подобрать параметры классов сигналов, например, как нормальных распределений с параметрами $\Sigma_v, \boldsymbol{\mu}_v$, $v=1,...,V$.

2.2 Подобрать значение параметра $a$ для текущего диктора из диапазона $[0.88, 0.9,..., 1.12]$, обеспечивающее минимальное число ошибок распознавания. Это делается путем одновременного применения модифицированного однократного алгоритма распознавания речевых команд (МОАРРК, где алгоритм МОАРРК такой же, как алгоритм ОАРРК за исключением того, что на первом шаге строятся искаженные кратковременные амплитудные спектры с параметром преобразования $a \neq 1$) к сигналам текущего диктора и вычисления общего числа ошибок из-за несоответствия оценок $v^*$ их истинным значениям $v$, где $v^* = \arg\max_{\lambda^v} p(X(a)|\lambda^v)$, $v=1,2,...,V$, $X(a) \in \mathbf{X}$, где $\mathbf{X}$ – множество мел-частотных кепстральных представлений сигналов текущего диктора.

3. Повторить шаг 2 со всеми «своими» дикторами в обучающей выборке

После такого обучения предполагается, что в алгоритме распознавания МАРРК снова используется, например, модель классов сигналов как нормальных распределений МЧКК с параметрами $\Sigma_v, \boldsymbol{\mu}_v$, $v=1,...,V$ с учетом значений параметра $a$ для каждого «своего» диктора.

В работе проведено сравнительное исследование алгоритмов распознавания речевых команд ОАРРК, МАРРК. Набор 220 речевых сигналов от 20 дикторов случайным образом делится на две выборки. Одна выборка играет роль обучающей, другая - роль тестовой. Обучается классификатор и выполняется распознавание тестовой выборки и расчет числа ошибок. Классы сигналов строятся как нормальные распределения векторов МЧКК. Процесс повторяется 20 раз, и находятся средние числа ошибок распознавания. В результате число ошибок распознавания для однократного, многократного и многократного с подбором параметра $a$ алгоритмов распознавания имеет значение 7%, 5.77%, 4.27%, соответственно.

Повторяется процедура скользящего контроля с построением классов сигналов как двухкомпонентных случайных процессов. Получаются числа ошибок для трех случаев 4.64%, 4.23%, 3.32%.

В работе были проведены эксперименты по разным схемам скользящего контроля. Все они подтверждают улучшение качества распознавания при использовании алгоритма распознавания речевых команд МАРРК.

**В четвертой главе** описывается решение задачи распознавания речевых команд на фоне шумов.

В данном разделе рассматривается другая проблема, не затрагивающая задачи устранения различия спектров «чужого» и «своего» дикторов. На прак-

тике результат распознавания сигналов как своего, так и чужого дикторов дополнительно ухудшается шумом. Обучающие речевые сигналы обычно являются незашумленными, а тестирующие речевые сигналы оказываются зашумленными. Шум сильно отклоняет спектры тестирующих речевых сигналов от спектров их эталонов в обучающей выборке. Поэтому качество результата распознавания на фоне шумов резко падает.

Если спектр зашумленного сигнала сильно отличается от спектра незашумленного сигнала, то очевидно, что степень связи таких спектров может оказаться достаточно малой. Для увеличения степени связи в данной работе предлагается увеличивать значения отсчетов амплитудных спектров обоих сигналов на константу.

После такого «усиления» кратковременного амплитудного спектра на величину $c \geq 0$ получается новая последовательность амплитудных спектров $\tilde{A} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \tilde{\mathbf{a}}_3, ...\}$, где $\tilde{\mathbf{a}}_i = \{\tilde{a}_i^k, 1 \leq k \leq N/2\}$, $\tilde{a}_i^k = a_i^k + c$.

Очевидно, что при неограниченном росте величины $c$ степень связи стремится к единице. Это означает, что в общем случае увеличение значений отсчетов амплитудных спектров на ограниченную константу приводит к увеличению степени связи спектров по сравнению с исходной.

Очевидно также, что величину $c$ не следует делать слишком большой, т.к. при этом устраняется различие в степени связи между похожими и непохожими спектрами.

Алгоритм распознавания речевых команд на фоне шумов такой же, как однократный алгоритм распознавания речевых команд (ОАРРК) за исключением того, что на первом шаге строятся усиленные кратковременные амплитудные спектры.

Чтобы проверить качество распознавания команд, были выполнены эксперименты по, как и ранее, разным схемам скользящего контроля. Одна из схем выглядит следующим образом:

1. Набор 440 речевых сигналов от 40 дикторов случайным образом делится на две выборки (каждая выборка содержит сигналы от 20 дикторов). Одна выборка играет роль обучающей, другая используется как тестовая выборка.

2. К тестовым речевым сигналам был искусственно добавлен аддитивный белый гауссовский шум с отношением сигнал/шум $R_{sn}$ дБ.

3. Увеличиваются значения отсчетов амплитудных спектров обучающих и тестовых сигналов на константу $c$. Значение $c$ взять из некоторого заданного диапазона.

4. По обучающей выборке строится модель классов сигналов, например как нормальных распределений МЧКК с параметрами $\Sigma_v, \boldsymbol{\mu}_v$, $v = 1, ..., V$.

5. Выполняется распознавание тестовой выборки алгоритмом распознавания речевых команд на фоне шумов и расчет числа ошибок распознавания.

6. Повторяются шаги 3-5 для всех значений параметра $c$ из заданного диапазона.

7. Опыт повторяется (шаги 1-6) несколько раз, например 10, с усреднением числа ошибок распознавания для каждого значения параметра $c$ из заданного диапазона.

Например, для уровня шума $R_{sn} = 6$ дБ, для другой модели классов сигналов (как двухкомпонентных случайных процессов) усиление амплитудных спектров на $c = 2$ приводит к уменьшению числа ошибок распознавания (10.11%) по сравнению с случаем без усиления $c = 0$ (76.02%) на $76.02 - 10.11 = 65.91\%$.

В работе были проведены эксперименты по разным схемам скользящего контроля. Все они подтверждают улучшение качества распознавания после усиления амплитудных спектров.

В работе подбирается константа $c$, обеспечивающая минимальное число ошибок распознавания. Процедура скользящего контроля выполняется для значения константы $c$ от 0 до 5 с шагом варьирования 0.1. В случае описания классов сигналов как нормальных распределений МЧКК оптимальное значение константы оказывается $c = 1.2$ с числом ошибок $E = 15.64\%$. В случае описания классов сигналов как двухкомпонентных случайных процессов оптимальное $c = 1.9$ с $E = 8.09\%$.

## ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Сформулирована и исследована задача идентификации модели речевого сигнала с целью адекватного восприятия речевого сигнала.
2. Решена задача распознавания речевых команд при недостаточном объеме обучающих данных, используя преобразование речевых сигналов.
3. Решена задача распознавания речевых команд на фоне шумов, используя увеличение значения отсчетов амплитудных спектров речевых сигналов на константу.
4. Оптимизированы параметры эвристической модели речевых сигналов с целью улучшения качества их распознавания.
5. Проведено экспериментальное исследование разработанных численных алгоритмов распознавания речевых команд.
6. Проведена оценка качества полученных решений процедурой скользящего контроля.

## СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

**1. Нгуен Ч.Т. Решение задачи распознавания речевых команд // Известия ТулГУ. Технические науки. 2013. Вып. 6., Ч. 2. С. 176–184.**

**2. Нгуен Ч.Т. Решение задачи распознавания речевых команд на фоне шумов // Известия ТулГУ. Технические науки. 2013. Вып. 11. С. 241–250.**

**3. Нгуен Ч.Т. Оптимизация параметров эвристической модели речевых сигналов с целью улучшения качества их распознавания // Известия ТулГУ. Технические науки. 2014. Вып. 1. С. 44–50.**

4. Нгуен Ч.Т. Комбинирование алгоритмов изменения высоты тона и алгоритмов изменения тембра речевых сигналов// Доклады VI магистерской конференции ТулГУ. Тула: Изд-во ТулГУ, 2011. С. 33–34.

5. Нгуен Ч.Т. Исследование алгоритмов искажения речевых сигналов изменением их высоты и тембра. // Доклады VIII Всероссийской конференции «Информационные системы и модели в научных исследованиях, промышленности, образовании и экологии». Тула: Изд-во «Инновационные технологии» , 2011. С. 120–121.

6. Нгуен Ч.Т. Группировка речевых сигналов, искаженных изменением высоты и тембра. // Доклады VII конференции «Молодёжные инновации», Ч. 2. Тула: Изд-во ТулГУ, 2013. С. 192–194.

7. Нгуен Ч.Т. Подстройка под диктора для улучшения распознавания речевых команд // Доклады XIV Международной конференции «Современные проблемы математики, механики, информатики». Тула : Изд-во ТулГУ, 2013. С. 558–565.

8. Нгуен Ч.Т. Влияние объёма и состава обучающей выборки на качество распознавания речевых команд // Доклады конференции «Инновационные наукоемкие информационные технологии». Тула: Изд-во ТулГУ, 2013. С. 6–8.

9. Нгуен Ч.Т. Изменение амплитудных спектров речевых сигналов с целью улучшения качества их распознавания // Доклады XI Всероссийской конференции «Информационные технологии, системный анализ и управление». Таганрог: Изд-во Южного федерального университета, 2013 – Т.2. С. 131–134.

10. Нгуен Ч.Т., Двоенко С.Д. Устранение зависимости от диктора и от влияния помех при распознавании речевых команд // Доклады III Международной конференции «Информационные технологии и системы ». Челябинск: Изд-во ЧелГУ, 2014. С. 79–80.

11. Нгуен Ч.Т. Применение эвристики для уменьшения влияния шума на качество распознавания речевых команд // Доклады XVI Международной конференции «Цифровая обработка сигналов и её применение». Москва, 2014. С.201–205.

12. Нгуен Ч.Т., Двоенко С.Д. Эвристические приемы улучшения качества распознавания речевых команд // Доклады IV Всероссийской конференции «Актуальные вопросы современной информатики». Коломна, 2014. С. 129–131.

13. Нгуен Ч.Т., Двоенко С.Д. Идентификация модели порождения речи с целью адекватного восприятия // Доклады XII Всероссийском совещании по проблемам управления. Москва, 2014 г. С. 8435–8443.

Формат бумаги 60 х 84 1/16. Бумага офсетная.
Усл. печ. л. 1,1. Уч.-изд. л. 1,0.
Тираж 100 экз. Заказ
Тульский государственный университет
300012, г. Тула, просп. Ленина, 92
Отпечатано в Издательстве ТулГУ
300012, г. Тула, пр. Ленина, 95