На правах рукописи

Ермилов Алексей Валерьевич

Методы, алгоритмы и программы решения задач идентификации языка и диктора

Специальность 05.13.11 —
«Математическое обеспечение вычислительных машин, комплексов и компьютерных сетей»

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2014

Работа выполнена на кафедре Управления Разработкой Программного Обеспечения Федерального государственного автономного образовательного учреждения высшего профессионального образования Национальный Исследовательский Университет “Высшая Школа Экономики”.

Научный руководитель:
доктор технических наук, **Гостев Иван Михайлович**

Официальные оппоненты:

Харламов Александр Александрович, доктор технических наук, старший научный сотрудник (Федеральное государственное бюджетное учреждение науки “Институт Высшей Нервной Деятельности и Нейрофизиологии РАН”)

Гнеушев Александр Николаевич, кандидат физико-математических наук, научный сотрудник (Федеральное государственное бюджетное учреждение науки Вычислительный центр им. А.А. Дородницына Российской академии наук)

Ведущая организция: Лаборатория информационных технологий Объединенного института ядерных исследований

Защита диссертации состоится «30» октября 2014 г. в 15 ч. на заседании диссертационного совета Д 002.017.02 в ВЦ РАН по адресу 119333, Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в научной библиотеке и на официальном сайте (http://www.ccas.ru) ВЦ РАН.

Автореферат разослан «_____» _____ 2014 г.

Учѐный секретарь
диссертационного совета Д 002.017.02
доктор физико-математических наук Рязанов В.В.

Общая характеристика работы

Актуальность темы. В современном мире все большее значение уделяется интерфейсам, использующим речевой ввод и вывод для взаимодействия между пользователем и компьютером. Поэтому всё большее многообразие в голосовых сообщениях приходится принимать во внимание разработчику систем распознавания речи, реализующих акустический интерфейс.

Задача распознавания речи (во многих своих проявлениях: от транскрибирования слитной речи до верификации и идентификации диктора) в настоящее время является крайне актуальной. Свидетельством этому служит растущее число публикаций и конференций по данной тематике (таких как ICASSP, INTERSPEECH), а также то, что в крупнейших транснациональных корпорациях (таких как Microsoft, Google, IBM) открываются департаменты, ориентированные на исследования в данной тематике.

Исследовательские усилия в сфере речевых технологий привели к появлению большого числа коммерческих систем распознавания речи. Такие компании как Nuance, IBM, ScanSoft предлагают большой набор программных решений как для серверных, так и для десктопных приложений.

Улучшение существующих систем распознавания языка и диктора позволит существенно упростить взаимодействие человека с компьютером в том случае, когда использование классических интерфейсов невозможно (например, при управлении автомобилем или в сложных условиях, таких как ликвидация последствий чрезвычайных ситуаций) или затруднено (например, людям, обладающим слабым зрением, или с ограниченными физическими возможностями), а также сделать работу с компьютером или иной техникой более комфортной, например, для аутентификации пользователя. Также следует отметить, что применение систем распознавания диктора играет большую роль в работе правоохранительных органов.

Необходимость исследований по этой тематике объясняется малоудовлетворительными результатами существующих систем при уменьшении соотношения сигнал/шум, зависимостями результата от диктора и, в ряде задач, невысокой скоростью работы систем.

Существующие системы распознавания речи в основном построены на Скрытых Марковских Моделях (HMM), которые задают динамику перехода от одной фонемы в речи к другой, а моделирование вероятностного распределения признаков происходит посредством Гауссовой Смеси (GMM). Такой подход был предложен в 1989 Лоуренсом Рабинером и долгое время являлся основным для моделирования речевого сигнала. Для описания речевого сигнала в системах автоматического распознавания речи со времен работы Л. Рабинера используются так называемаы мел-частотные кепстральные коэффициенты

(MFCC Mel Frequency Cepstral Coefficients), начало развитию которых положил Пол Мермельстайн в 1976.

Также следует отметить, что в последнее время альтернативой используемым сейчас MFCC становятся признаки, устойчивые к вариабельности речевого тракта у диктора (например, bottleneck features), что позволяет строить робастные системы. В данной работе предлагается новая вероятностная модель, основанная на применении функции плотности распределения (расширении Грам-Шарлье) для дикторонезависимых признаков и использование Фишеровских ядер в алгоритме опорных векторов, а также используются новые вычислительные методы для оценки этих модели (алгоритм симуляции отжига), использующие преимущества параллельных вычислений. Применение этих моделей повышает точность распознавания языка и диктора, а также увеличивает быстродействие всей системы распознавания.

В течении длительного времени использование систем автоматического распознавания больших параллельных потоков речи было ограничено в виду недостаточного быстродействия оборудования, а именно - невозможности обработки online. Для функционирования в реальном времени системам, оперирующим с такими потоками речи, приходилось находить компромисс между объемом словаря (а значит, и потенциальной сферой применения), сложностью грамматики и точностью распознавания. Таким образом, повышение скорости работы распознавателя будет положительным образом сказываться на объеме тех задач, где необходима работа в реальном времени, а также на точности распознавания. Хорошим примером может служить работа сотовой станции или call – центра, где на обработку одновременно может приходить огромное количество заявок, требующих обработки в реальном времени.

Цель работы и задачи исследования. Целью диссертационной работы являлась разработка методов, алгоритмов и программ идентификации языка и диктора. Проведено исследование существующих методов распознавания, на основании которых была предложена система характерных признаков для распознавания языка с применением 4-х параметрического семейства распределений (Грам-Шарлье); модификация метода опорных векторов для повышения точности распознавания диктора, на основе введения в базовый алгоритм функционального преобразования (Фишеровских ядер), а также модификация алгоритма симуляции отжига для повышения быстродействия и точности получения признаков, применяемых для распознавания языка. Применение указанных методов позволило увеличить быстродействие и точность систем распознавания языка и диктора.

Вышеупомянутые методы, алгоритмы и программы были разработаны на основе физиологических особенностей человеческого языка и дикции, а также механизма восприятия звука человеком при распознавании речи.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Исследование моделей акустических сигналов, применяемых в системах распознавания языка и диктора.
2. Разработка математический модели дикторонезависимых акустических признаков на основе 4-х параметрического семейства распределений.
3. Модификация метода опорных векторов для решения задачи идентификации диктора по речевому сообщению фиксированной длины с целью повышения качества распознавания.
4. Модификация метода симуляции отжига для повышения быстродействия и качества признаков, применяемых для распознавания языка.
5. Анализ предложенных и существующий моделей и методов для сравнения их быстродействия и точности распознавания.

Методы исследования. При решении поставленных задач использовались методы и понятия теории вероятностей и математической статистики, теории случайных процессов, методы цифровой обработки сигналов, распознавания образов, алгоритмы и методы обработки данных, методы построения параллельных систем.

Научная новизна. Научная новизна заключается в том, что

1. Изучены информационные признаки идентификации языка и диктора на основе физиологических особенностей человеческого языка и дикции с учетом механизма восприятия звука человеком при распознавании речи.
2. Впервые предложена система характерных признаков для распознавания языка с применением 4-х параметрического семейства распределений (расширение Грам-Шарлье).
3. Разработана и обоснована теоретически модификация метода опорных векторов, основанная на применении фишеровских ядер, которая позволяет увеличить точность распознавания диктора.
4. Впервые проведён сравнительный анализ алгоритмов оптимизации для вычисления акустических дикторонезависимых признаков по скорости и точности.
5. Разработана модификация алгоритма симуляции отжига увеличивающая быстродействие системы при получении дикторонезависимых признаков за счет введения в алгоритм параллельно выполняющихся циклов.
6. Разработаны и теоретически обоснованы методы и алгоритмы получения параметров классификатора для решения задач идентификации языка

основанные на использовании метода опорных векторов повышающие точность распознавания.

7. Проведены экспериментальные исследования по оценке точности распознавания и быстродействию системы идентификации языка и диктора, которые показали преимущества разработанных методов по сравнению с применяемыми ранее.

Теоретическая значимость. Теоретическая значимость заключается в следующем.

1. Впервые разработаны методы идентификации диктора, основанные на методе опорных векторов с применением Фишеровских ядер.
2. Впервые была предложена и теоретически обоснована модель акустических дикторонезависимых признаков, использующая 4-х параметрическое распределение (расширение Грам-Шарлье) для моделирования речевых признаков, которая была использована для аутентификации и в системах безопасности и работе правоохранительных служб.
3. Впервые разработана модификация алгоритма симуляции отжига увеличивающая быстродействие системы при получении дикторонезависимых признаков за счет введения в алгоритм параллельно-выполняющихся циклов.

Практическая значимость. Полученные автором результаты имеют большое научное и народно-хозяйственное значение (имеется акт о внедрении) при создании человеко-машинных интерфейсов и идентификации личности и языка в работе различных государственных служб и органов.

Степень достоверности полученных результатов обеспечивается использованием строгих математических методов теории вероятностей и математической статистики, распознавания образов. Достоверность подтверждается моделированием и проведенными вычислительными экспериментами с использованием реальных и симулированных данных, а также путём сопоставления результатов, полученных в диссертации, с результатами, доступными в открытой печати.

Публикации и апробация работы. По материалам диссертации опубликовано 5 статей (3 из которых в журналах из списка ВАК, одна в международном реферируемом журнале), 6 тезисов на международных конференциях. Результаты настоящего исследования были представлены на следующих конференциях и семинарах: Конференции студентов, аспирантов и молодых специалистов МИЭМ в 2010 г; Конференции студентов, аспирантов и молодых специалистов МИЭМ в 2011 г; Международной конференции «Моделирование нелинейных процессов и систем» (СТАНКИН 2011 г.); 5-я Международной Конференции «Распределённые вычисления и Грид-технологии в науке и образовании» (GRID - 2012) (Дубна Московская обл. 2012 г.); X Международной научно-технической конференции «Оптико-

электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (Курск 2012); The First International Conference on Modern Manufacturing Technologies in Industrial Engineering “ModTech – 2013”, (Румыния, Синая 2013 г.); International Conference on Mathematic Modeling and Computing in Physics (MMCP’2013) (Дубна Московская обл., 2013 г.); XI Международной научно-технической конференции «Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации» (Курск 2013).

Объем и структура работы. Диссертация состоит из введения, четырёх глав и заключения. Полный объем диссертации составляет 135 страницы с 26 рисунками и 5 таблицами. Список литературы содержит 81 наименование.

Основное содержание работы.

Во введении обоснована актуальность работы, сформулированы цель и задачи диссертационного исследования, новизна и практическое значение полученных результатов, а также положения, выносимые на защиту, обоснованность, достоверность и апробация работы.

В первой главе приводится обзор физических аспектов акустического сигнала, а также характеристик и особенностей речевых сигналов. Приводятся модель речеобразования и схемы описания речи, например, фонемная. Рассмотрены общие принципы генерации и восприятия звукового сигнала, вводится понятие речевого тракта. Также даётся характеристика некоторых подходов к имплементации методов распознавания речи: акустико-фонетического, подхода с точки зрения распознавания образов, подхода с точки зрения искусственного интеллекта. Рассматриваются методы выделения акустических признаков, такие как спектральный анализ, модель банка фильтров. Приведены методы получения речевых признаков на основе коэффициентов линейного предсказания и кепстральных коэффициентов. В конце главы сформулированы выводы. На основе анализа физических аспектов звука были введены характеристики речевого сигнала, которые затем используются в диссертации для распознавания языка и диктора. Исследованы источники вариабельности в речевом сигнале, приводящие к следующим эффектам.

• Структура речевого сигнала может меняться под воздействием физиологических и эмоциональных факторов.

• Долговременные параметры речевого сигнала могут быть изменены диктором намеренно (эмоции).

• Акустическая реализация фонем может варьироваться (коартикуляция, акцент, спонтанная речь).

Во второй главе рассматриваются различные математические модели, использующиеся для построения систем распознавания языка и диктора с использованием Скрытых Марковских Моделей, особое внимание уделяется методам, применяемым для разработки системы распознавания языка, точность идентификации которой не зависит от диктора. Приводится способ построения дикторонезависимых признаков для описания речевого сигнала, опирающийся на психоакустическую модель восприятия человеком речевого сообщения.

В качестве базовой модели в работе используется Скрытая Марковская Модель (HMM – Hidden Markov Model, Lawrence Rabiner, Biing-Hwang Juang "Fundamentals of Speech Recognition"), которая определяется как двойной случайный процесс. Лежащий в основе случайный процесс представляет собой однородную Марковскую цепь с конечным числом состояний, каждое из которых производит свою последовательность наблюдений.

Определение 2.1. Пусть P_t - случайный процесс (Скрытая Марковская Модель), используемый в работе. Определим его с помощью следующих компонент:

1. Количество скрытых состояний N. Множество состояний модели обозначается $S = \{S_1, \ldots, S_N\}$. Состояния соединенны таким образом, что любое состояние S_i может быть достигнуто из любого другого состояния S_j за конечное число шагов (эргодическая модель).

2. Размер выходного алфавита M. Набор символов выходного алфавита обозначается через $V = \{v_1, \ldots, v_M\}$. Речевыми символами являются вектора из $\mathbb{R}^n$.

3. Матрица переходных вероятностей $A = (a_{ij})$, где

$$a_{ij} = P(q_t = S_j \,|\, q_{t-1} = S_i), \qquad i, j = 1, \ldots, M$$

4. Распределение вероятности выходных символов $B = \{b_j(k) : j = 1, \ldots, N,\ k = 1, \ldots, M\}$ для данного состояния j, где k -порядковый номер символа v_k, а $b_j(k) = P(v \in V \,|\, q_t = S_j), \quad j = 1, \ldots, N,\ k = 1, \ldots, M$, то есть, $b_j(k)$ - вероятность того, что в момент времени t система, находясь в состоянии S_j, выдаст символ v_k.

5. Вероятность нахождения в состоянии i в начальный момент времени π_i, формирующие начальное распределение Π.

Тогда набор компонент A, B, Π, задающих марковскую модель, обозначается $\lambda = \{A, B, \Pi\}$. Последовательность наблюдений, сгенерированных марковской моделью за время T, обозначают $O = O_1, O_2, \ldots, O_T$.

Теорема 2.1. Пусть Скрытая Марковсая Модель задаётся набором компонент $\lambda = \{A, B, \Pi\}$. Тогда для любого состояния S_k $P(q_{t+1} = S_k, \ldots, q_{t+T-1} = S_k, q_{t+T} \neq S_k \,|\, q_t = S_k) = a_{kk}^T(1 - a_{kk})$, то есть, время нахождения цепи в состоянии S_k распределено экспоненциально.

Рассмотрена общая постановка задач, решаемых с помощью HMM введенного типа. Для решения задачи идентификации языка были разработаны следующие алгоритмы.

1. Алгоритм вычисления вероятности наблюдения данной последовательности $P(O|\lambda)$ при заданной последовательности символов наблюдений $O = O_1, O_2, \ldots, O_T$ и модели $\lambda = \{A, B, \Pi\}$.

Вход: Последовательность наблюдений $O = O_1, O_2, \ldots, O_T$, параметры модели $\lambda = \{A, B, \Pi\}$.

Шаг 1. Инициализация: $\alpha_1(i) = \pi b_1(O_1), \quad 1 \le j \le N$.

Шаг 2. Индукция: $\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}$.

Шаг 3. Терминация: $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Выход: Вероятность $P(O|\lambda)$.

2. Алгоритм вычисления последовательности состояний $Q = q_1, q_2, \ldots, q_T$, оптимальной с точки зрения максимизации апостериорной вероятности $P(q_1, \ldots, q_T \mid O_1, \ldots, O_T, \lambda)$, при заданной последовательности символов наблюдений $O = O_1, O_2 \ldots, O_T$ и модели $\lambda = \{A, B, \Pi\}$.

Вход: Последовательность наблюдений $O = O_1, O_2, \ldots, O_T$, параметры модели $\lambda = \{A, B, \Pi\}$.]

Шаг 1. Инициализация:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N,$$
$$\psi_1(i) = 0.$$

Шаг 2. Рекурсия:

$$\delta_t(j) = b_j(O_{t+1}) \max_{1 \le i \le N} \delta_t(i) a_{ij}, \qquad 1 \le i \le N,$$
$$\psi_t(i) = \arg\max_{1 \le i \le N} \delta_t(i) a_{ij}, \qquad 2 \le t \le T$$

Шаг 3. Терминация:

$$\hat{P} = \max_{1 \le i \le N} \delta_T(j),$$
$$\hat{q}_T = \arg\max_{1 \le i \le N} \delta_T(j).$$

Шаг 4. Определение последовательности состояний:

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}), \qquad t = T-1, \ldots, 1.$$

Выход: Последовательность состояний $Q = q_1, q_2, \ldots, q_T$.]

3. Алгоритм поиска оптимальных параметров модели $\hat{\lambda} = \{\hat{A}, \hat{B}, \hat{\Pi}\} = \arg\max_\lambda P(O \mid \lambda)$ с точки зрения максимизации $P(O \mid \lambda)$.

Для описания алгоритма на множестве всех возможных моделей λ введена норма $|| \cdot ||$. Введены следующие обозначения. Совместная вероятность наблюдения последовательности, начиная с момента $t+1$ до момента времени T, при заданном в момент t состоянии S_i и модели λ: $\beta_t(i) =$

$P(O_{t+1}, O_{t+2}, \ldots, O_T \,|\, q_t = S_i, \lambda)$, вероятность нахождения в состоянии S_i в момент времени t и в состоянии S_j в момент $t+1$ при данной модели и последовательности наблюдений $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \,|\, O, \lambda)$, $\gamma_t(i)$ вероятность нахождения в состоянии S_i в момент времени t при заданной последовательности наблюдений и модели.

Вход: Последовательность наблюдений $O = O_1, O_2, \ldots, O_T$, начальные параметры модели $\lambda_0 = \{A_0, B_0, \Pi_0\}$, параметр точности ϵ.

Пока $||\lambda_{n+1} - \lambda_n|| > \epsilon$

Шаг 1. Вычисление вероятностей $\xi_t(i,j), \gamma_t(i)$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \,|\, \lambda)}, \gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

Шаг 2. Пересчёт параметров модели $\lambda_{n+1} = \{A_{n+1}, B_{n+1}, \Pi_{n+1}\}$:

$$\hat{\pi}_j^{n+1} = \gamma_1(j),$$
$$\hat{a}_{ij}^{n+1} = \frac{\sum_{j=1}^{N} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$
$$\hat{b}_k^{n+1} = \frac{\sum_{j=1, o_t = v_k}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}.$$

Выход: Параметры модели $\hat{\lambda} = \{\hat{A}, \hat{B}, \hat{\Pi}\}$

Далее в работе исследованы особенности способов решения задач идентификации языка и диктора. Задача идентификации решалась в следующей постановке. Пусть $\mathbb{X}$ - пространство объектов, $\mathbb{Y}$ - множество ответов, $f : \mathbb{X} \to \mathbb{Y}$ - целевая зависимость. Пусть $\mathbb{X}^t \in \mathbb{X} \times \mathbb{Y}$ - обучающее множество, то есть множество пар (X_i, y_i), где $y_i = f(X_i)$. По известному обучающему множеству требуется построить $\hat{f} : \mathbb{X} \to \mathbb{Y}$ аппроксимирующую f на всем $\mathbb{X}$.

Будем искать $\hat{f}$ в виде $\hat{f}(X) = sign(\mathbf{w}^T X + b)$, используя метод опорных векторов (В.Вапник, А.Червоненкис, Support Vector Machines, SVM)

Суть метода SVM заключается в построеннии параллельных разделяющих гиперплоскостей с максимальным расстоянием между ними.

Для формализации задачи построения SVM вводятся следующие обозначения. Данный для разделения набор точек-векторов в $\mathbb{R}^n$ обозначается как $\{X_i\}_{i=1}^N$, а линейная функция представляется в виде $\mathbf{w}^T X + b = 0$. Разделяемые классы обозначаются через A и B и вводятся значения из множества ответов для каждого вектора:

$$y_i = \begin{cases} 1, \ X_i \in A, \\ -1, \ X_i \in B. \end{cases}$$

Показывается, что построение оптимальной разделяющей полосы эквивалентно минимизации $||\mathbf{w}||$. Решение задачи построения оптимальной разделяющей полосы называется обучением. При этом параметры $\mathbf{w}, b$ оптимальной разделяющей полосы являются функциями только опорных векторов, то есть таких векторов, для которых $y_i(\mathbf{w}^T X_i + b) - 1 = 0$.

Поскольку в общем случае линейное разделение векторов может быть невозможно, то для разделения имеющееся пространство преобразуют таким образом, чтобы вектора классов после него стали линейно разделимыми. Пусть ϕ произвольное отображение пространства признаков в гильбертово пространство H. От отображения требуется, чтобы образы обучающих векторов были линейно разделимы в H (оно называется пространством вторичных признаков).

Свойства симметричности и положительной полуопределённости функции, называемой ядром, используются для получения преобразования ϕ. Показывается, что достаточно знать не само отображение ϕ, а только ядро $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, вычисляющее скалярное произведение в H образов пары векторов признаков $K(X_i, X_j) = (\phi(X_i), \phi(X_j))$.

В результате, обучение SVM представляет решение задачи квадратичного программирования с линейными ограничениями:

$$\begin{aligned} \frac{1}{2}(w, w) + C\sum_{i=1}^{N} p(e_i) &\to \min_{w,b} \\ y_i((w, \phi(X_i)) + b) &\geq 1 - e_i, \\ e_i &\geq 0, \\ i &= 1, \ldots, N \end{aligned}$$

где $p(e)$ - неотрицательная, монотонно неубывающая функция, такая, что $p(0) = 0$, а $C > 0$ и параметры ϕ определяются эмпирически. Для решения задачи идентификации языка в диссертации использовалось так называемое гауссово ядро $K(X, Y) = e^{-\gamma||X-Y||}$. На основании проведённого анализа разработан алгоритм оптимизации параметров функции распознавателя C и γ, имеющий следующий вид:

Вход: Набор векторов $\{X_i\}_{i=1}^N$

Шаг 1. Для фиксированного k представить обучающее множество $\mathcal{X} = \{X_i\}_{i=1}^N$ как $\mathcal{X} = \bigcup_{j=1}^k \mathcal{X}_j, \mathcal{X}_i \bigcap \mathcal{X}_j = \oslash, \forall i \neq j$. Зафиксировать точность решения задачи ϵ.

Шаг 2. Выбрать начальное значение $x_0 = (C_0; \gamma_0) \in \mathbb{R}^2$ и величину шага Δ_0.

Шаг 3. Выполнять пока $||x_k - x_{k+1}|| > \epsilon$

Подшаг 1. Решить задачу обучения SVM при $C = C_k, \gamma = \gamma_k$ и $X_i \in \mathcal{X}_1$.

Подшаг 2. Определить функцию $f(t) = \frac{1}{k-1}\sum_{j=2}^{k} E_j(t)$, где $E_j(t) = \frac{1}{|\mathcal{X}_j|}\sum_{X_i \in \mathcal{X}_j} \mathbb{I}\{\tilde{y}_{X_i}(t) \neq y_{X_i}\}$, где $\tilde{y}_{X_i}(t)$ - предсказанная метка вектора X_i, y_{X_i} - его настоящая метка.

Подшаг 3. Для $\forall t \in P_k = \{x_k \pm \Delta_k e_i : i = 1, 2\}$ вычислить $f(t)$

Подшаг 4. Если $\exists \hat{t} : f(\hat{t}) < f(x_k)$ установить $x_{k+1} = \hat{t}, \Delta_{k+1} = \Delta_k$; иначе $x_{k+1} = x_k, \Delta_{k+1} = \frac{\Delta_k}{2}$.

Выход: оптимальные значения параметров классификатора $\hat{C}, \hat{\gamma}$.

К преимуществам указанного алгоритма можно отнести следующее

• Не используются знания о градиенте функции, поскольку нет оснований считать, что эта функция будет дифференцируемой.

• Задача решается в параллельных процессах, так как сама процедура, описанная в подшаге 2, может быть выполнена параллельно, поскольку вычисление функции $E_j(t)$ может выполняться для каждого j независимо и нет никаких зависимостей по данным.

Наиболее часто используемые ядра, такие как полиномиальное $K(X, Y) = ((X, Y) + 1)^d + c$ или гауссово $K(X, Y) = e^{-\gamma||X-Y||}$, применённые к задаче идентификации диктора дают низкую точность распознавания, так как не позволяют использовать полное высказывание.

Для устранения этого недостатка был предложен метод, основанный на функциональном преобразовании (Фишеровских ядрах), которые отображеют всё озвученное диктором предложение целиком (полное высказывание) в единственную точку, что позволяет проводить их разделение.

В основе разработанного метода лежит применение в качестве ядра функции, вычисленной с помощью апостериорных вероятностей наблюдений, которые получены из порождающей модели появления векторов, в качестве которых могут выступать либо Скрытые Марковские модели, либо гауссовские смеси.

Теорема 2.2. Пусть $P(X|\lambda)$ апостериорная вероятность наблюдения X, полученная из модели λ. Зададим в пространстве всех возможных $P(X|\lambda)$ скалярное произведение как $U_X^T F^{-1} U_X$, где $F = \mathbb{E}_X U_X U_X^T$ - матрица информации Фишера и $U_X = \nabla \ln P(X|\lambda)$ фишеровская функция потерь. Тогда функция

$$K(X_i, X_j) = U_{X_i}^T F^{-1} U_{X_j}.$$

является ядром.

Вычисление преобразованных значений векторов происходит по следующему алгоритму.

Вход: Набор векторов $\{X_i\}_{i=1}^N$, оценки параметров порождающей модели θ и параметра γ.

Шаг 1. Для $\forall i \in 1, \ldots, N$ вычислить U_{X_i}.

Шаг 2. Получить оценку матрицы информации $\hat{F} = \frac{1}{N}\sum_{i=1}^{N} U_{X_i} U_{X_i}^T$ из порождающей модели и вычислить обратную к ней.

Шаг 3. Для $\forall i, j \in \{1, \ldots, N\}$ вычислить $K(X_i, X_j) = U_{X_i}^T F^{-1} U_{X_j}$.

Шаг 4. Вычислить $\hat{K}(X_i, X_j) = e^{\gamma K(X_i, X_j)}$

Выход: Значение ядра $\hat{K}(X_i, X_j)$ на всех парах векторов X_i, X_j

Далее в главе рассмотрен способ построения дикторонезависимых признаков речевого сигнала для применения в системах распознавания языка.

В работе была использована Auditory Image Model (AIM), которая разработана Р. Петерсоном для моделирования человеческой психоакустики. Эта модель даёт на выходе нормализванный спектр сигнала, пример которого представлен на рис. 1.

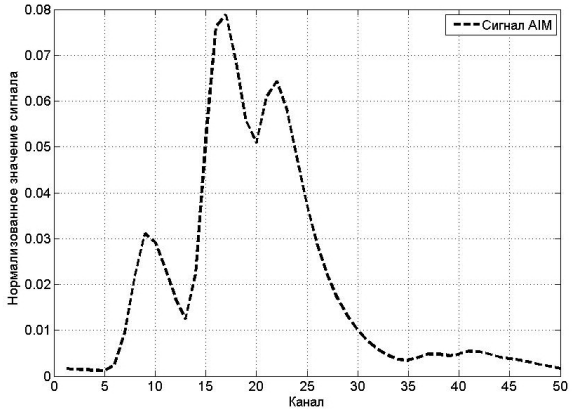

Рис. 1: Изображение огибающей спектра, полученного из модели AIM.

Для моделирования огибающей спектра (AIM значений) в главе предложено использование 4-х параметрического семейства распределений (расширения Грам-Шарлье), вместо обычно используемой гауссовской смеси.

Определение 2.2. Расширением Грам-Шарлье называется представление плотности распределения g случайной величины z в виде

$$g(z) = p_n(z)\psi(z), \tag{1}$$

где $\psi(z)$ – плотность стандартного нормального распределения, а $p_n(z)$ выбрана таким образом, чтобы $g(z)$ имела те же моменты, что и истинная плотность z.

Конструирование функции $p_n(z)$ основано на полиномах Эрмита H_i, которые образуют ортогональный базис относительно скалярного произведения, порожденного математическим ожиданием, взятым по плотности стандартного нормального распределения.

Представление (1) является необходимым для моделирования моментов высокого порядка, которые важны для распознавании языка.

Однако введённая функция не является в строгом смысле плотностью вероятности, так как может принимать отрицательные значения. Для устранения этого использовалась следующее преобразование:

$$g(z) = \psi(z)\frac{(1 + \sum_{i=1}^{n} c_i H_i(z))^2}{k},$$

где $k = 1 + \sum_{i=1}^{n} c_i^2 i!$, а c_i - коэффициенты.

Для моделирования огибающих полученного спектра с помощью предложенного семейства распределений необходимо получить оценки вектора неизвестных параметров $\theta = (c_1, \ldots, c_n)^T$, для чего автором найдено решение следующей оптимизационной задачи:

$$\begin{aligned} \ell(z, \theta) &= \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}(z_i) \to \max_{\theta}, \\ f(\theta) &\leq 0, \end{aligned} \tag{2}$$

где $\hat{\ell}(z_i) = \ln(\psi(z_i)) + \ln(1 + \sum_{j=1}^{n} c_j H_j(z_i))^2 - \ln(1 + \sum_{j=1}^{n} c_j^2 j!)$, θ - вектор неизвестных параметров, и $f(\theta)$ - функция ограничений, которая может быть добавлена для того, чтобы значения параметров удовлетворяли каким-либо априорно заданным ограничениям (например, некоторые оценки должны быть положительны). Пусть $\{\chi_i\}_{i=1}^{n}$ - выборка из распределения, зависящего от параметра $\theta \in \Theta$. Тогда оценка $\hat{\theta}$ называется состоятельной, если

$$\hat{\theta} \to \theta, \text{ по вероятности при } n \to \infty$$

асимптотически нормальной с дисперсией σ^2, если

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathbb{Z}, \text{ по распределению при } n \to \infty,$$

где $\mathbb{Z}$ - нормальная случайная величина с дисперсией σ^2 и средним 0.

Теорема 2.3. Решение задачи (2) дает состоятельные и асимптотически нормальные оценки параметра θ.

Для того, чтобы на практике получить значения параметров θ функции $\ell(z_i)$ необходимо численно решить оптимизационную задачу (2).

Существует множество методов численного решения задачи (2), которые можно разделить на градиентные и безградиентные. В диссертации использовались две модификации алгоритма симуляции отжига с использованием параллельных вычислительных процессов для увеличения скорости и качества работы алгоритма.

Первая модификация алгоритма, являющаяся более вычислительно затратной, приведена ниже. Суть модификации заключается в независимом старте k процессов отжига из разных начальных точек.

Вход: набор значений $\{z_i\}_{i=1}^{N}$

Шаг 1. Сгенерировать k начальных значений параметров $\{\theta_i\}_{i=1}^{k}$.

Шаг 2. К каждому значению θ_i применить алгоритм симуляции отжига, получив k финальных оценок $\{\hat{\theta}_i\}_{i=1}^{k}$.

Шаг 3. Вычислить $\ell(z, \hat{\theta}_i)$ для каждого $\hat{\theta}_i$, $i = 1, \ldots, k$.

Шаг 4. $\tilde{\theta} = \max_{i=1,\ldots,k} \hat{\theta}_k$.

Выход: Оптимальное значение $\tilde{\theta}$

При этом шаги 2-4 выполняются параллельно. Вторая модификация записывается следующим образом:

Вход: набор значений $\{z_i\}_{i=1}^{N}$

Шаг 1. Сгенерировать k начальных значений параметров $\{\theta_i\}_{i=1}^{k}$.

Шаг 2. Вычислить $\ell(z, \hat{\theta}_i)$ для каждого $\hat{\theta}_i$, $i = 1, \ldots, k$.

Шаг 3. Вычислить $\tilde{\theta} = \arg\max_{i=1,\ldots,k} \ell(z, \hat{\theta}_i)$.

Шаг 4. Применить алгоритм симуляции отжига к $\tilde{\theta}$, получив финальную оценку $\hat{\theta}$

Выход: Оптимальное значение $\hat{\theta}$

При этом шаги 2-3 выполняются параллельно.

В конце главы сформулированы выводы. Отмечена важность использования нормализации длины речевого тракта или применения психоакустических признаков при проектировании системы распознавания языка.

В третьей главе изложены различные аспекты реализации системы идентификации языка и диктора с применением предложенных методов.

Рассмотрена схема, содержащая этапы обработки речевого сигнала и идентификации языка и диктора, представленные на рис. 2. На первом этапе

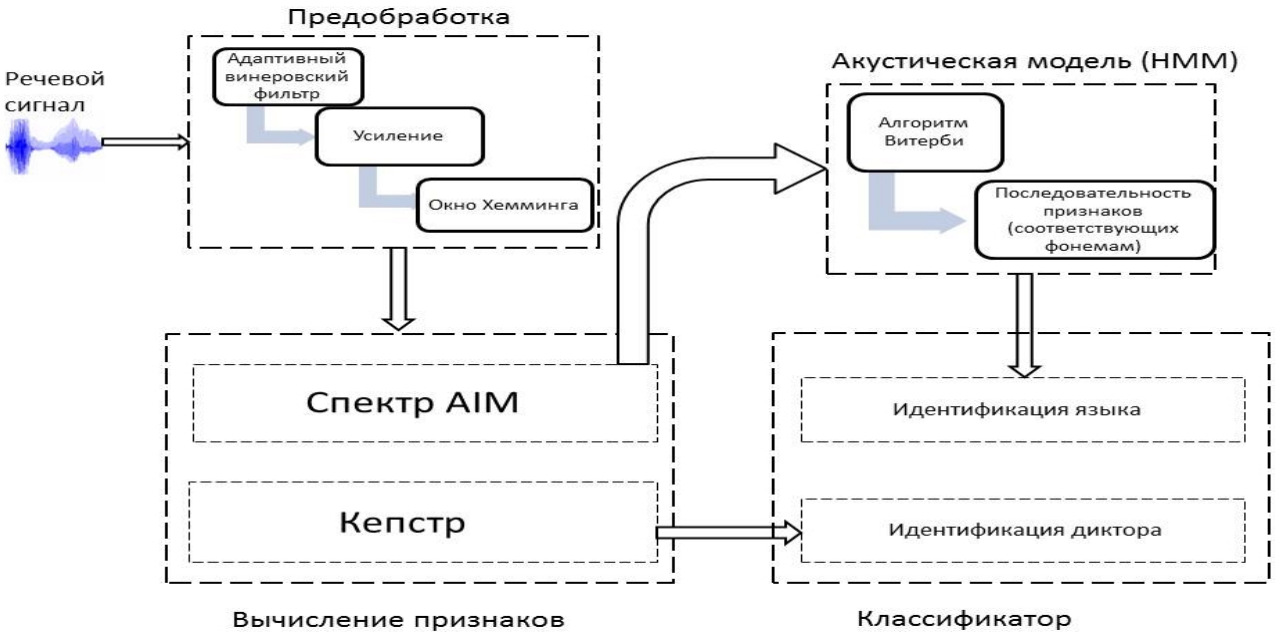

Рис. 2: Схема системы идентификации языка и диктора.

сигнал очищается от шумов с помощью адаптивного винеровского фильтра, усиливается и нарезается на участки (фреймы), посредством движущегося окна Хемминга.

На втором этапе происходит выделение акустических признаков. В предложенной системе используются мел - частотные кепстральные коэффициенты (описанные в главе 1) и модифицированные признаки из AIM (описанные в главе 2).

Для решения задачи идентификации диктора используются мел - частотные кепстральные коэффициенты, к которым применяется Фишеровское ядро по алгоритму, описанному в главе 2. Преобразованные признаки используются для идентификации диктора с помощью предварительно обученного классификотора на основе метода опорных векторов.

В случае решения задачи распознавания языка используются признаки из AIM, которые сначала подаются на вход акустической модели, основанной на HMM, изложенный в главе 2. Для получения списка фонемных признаков, которые применяется для распознавания языка, используется алгоритм Витерби, который также приведён в главе 2. Полученные признаки также используются для идентификации языка с помощью предварительно обученного классификатора на основе метода опорных векторов.

Далее приводится схема архитектуры реализованной системы идентификации языка и диктора на языке UML в виде диаграм классов.

На рис. 3 представлена диаграмма классов сущностей, которые являются объектными представлениями данных, которыми управляет система идентификции.

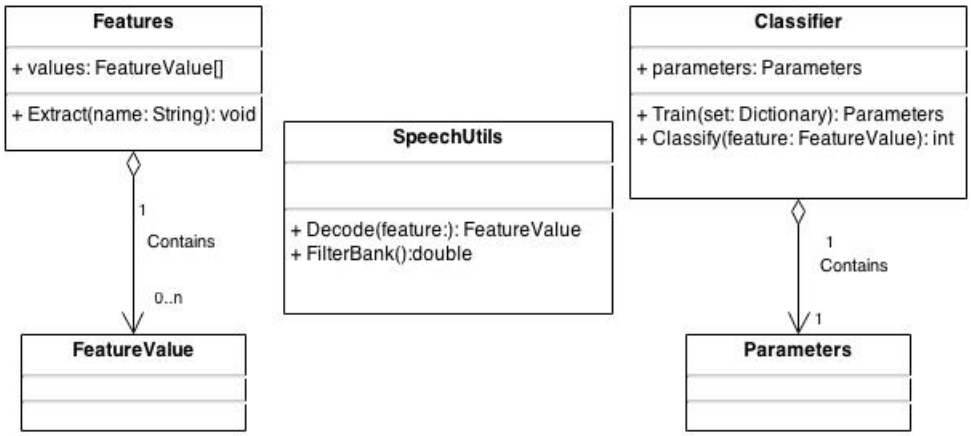

Рис. 3: Диаграмма классов - сущностей.

Абстрактный класс Features предназначен для хранения и вычисления признаков входного речевого сигнала. Класс состоит массива объектов FeatureValue и метода получения Extract, выполняющего извлечение признаков из полученного на вход речевого сигнала.

Абстрактный класс Classifier предназначен для реализации классифицирующего алгоритма опорных векторов. Класс состоит из методов Train и Classify, а также объекта Parameters, который содержит все необходимые для работы классификатора параметры. Метод Train принимает на вход словарь, в котором ключём является метка класса, а значением - объект типа Features, и возвращает объект Parameters. Метод Classify принимает объект FeatureValue и возвращает значение решающей функции, а также метку класса - решения.

Класс SpeechUtils содержит вспомагательные методы, необходимые для вычисления признаков и классификации, такие как, например, вычисление выхода банка фильтров и алгоритм Витерби.

Далее в диссертации описывается последовательность вызовов методов классов для идентификации языка и диктора. Сначала вызывается метод Extract у классов FeaturesMFCC и FeaturesAIM, которые являются наследниками класса Features. После этого вызываются метод Classify класса ClassifySpeaker, на вход которому подаётся объект FeaturesMFCC.FeatureValue, и метод Decode класса SpeechUtils, который реализует алгоритм Витерби, который принимает объект FeatureValue и возвращает объект Phonems, являющийся наследником FeatureValues. После этого происходит вызов метода Classify класса ClassifyLanguage, на вход которому подаётся объект Phonems. Результатом последовательности вызовов являются номер диктора и языка, к которым классификатор отнес входной речевой сигнал.

Далее рассматриваются особенности конвейерной обработки речевого сигнала, приведённой на рис. 4.

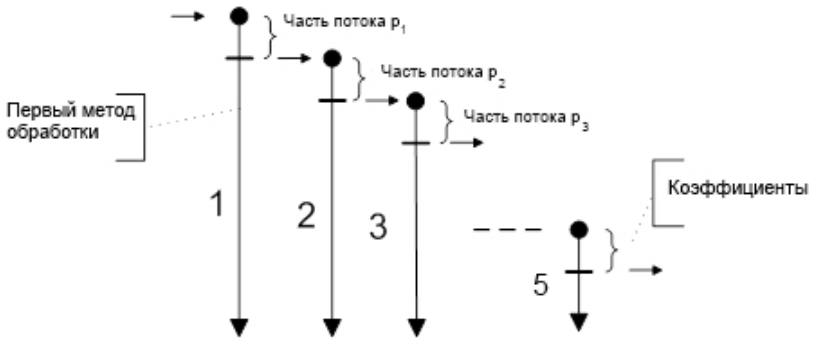

Рис. 4: Конвейерная схема процессов обработки речевого сигнала

В ряде случаев система должна работать в режиме реального времени. Например, сотовая станция, которая обрабатывает поток независимо приходящих в априори неизвестные моменты времени заявки. В этом случае увеличение времени ожидания обработки заявки недопустимо. Таким образом, предполагается, что узлы системы обрабатывают приходящие потоки речевых сообщений независимо. Это вызвано тем, что если вновь прибывшая заявка будет обрабатываться одним из уже задействованных узлов системы, то при большой плотности прихода новых заявок накладные расходы на переключение контекстов и синхронизацию различных узлов системы превысят выигрыш от использования дополнительных вычислительных мощностей на обработку этой заявки. Кроме того, синхронизация процессов на разных узлах системы и перенос данных между узлами увеличат нагрузку на сеть.

Каждый узел комплекса представляет собой реализацию системы, изображенной на рис. 2.

В четвёртой главе представлены результаты экспериментов с реальными и симулированными данными с применением моделей, описанных в главе 2. Также даётся объяснение полученных результатов с точки зрения теории оптимизации, минимизации структурного риска и практических ограничений.

В главе анализируются особенности практического применения Фишеровских ядер к задаче распознавания диктора и приводятся результаты экспериментов по распознаванию в различных каналах: микрофонном, телефонном и GSM.

При проведении экспериментов по распознаванию диктора в качестве входных данных использовалась база речевых отрезков различной длительности. В ней содержатся данные по 15 дикторам, записанные с помощью обычного телефона, телефона GSM и микрофона. Характеристики входного сигнала для каждого канала: битрейт 16, частота дискретизации 8 кГц, соотношение сигнал – шум в среднем 15дб. Продолжительность речи каждого диктора в обучающем множестве составляла примерно 120 минут. Следует отметить, что в случае микрофонного канала при записи использовались микрофоны с очень разными АЧХ, поэтому фонограммы сильно отличаются друг относительно друга, что сильно усложняет задачу распознавания.

Таблица 1: Сравнение точности распознавания диктора в различных каналах.

Длит., с	Точность, %					
	Микрофон		Телефон		GSM	
	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.
5	0.33	16.33	44.98	64.91	45.88	55.93
10	0	26.17	80.46	86.13	82.98	88.70
100	4.5	46.45	87.68	82.02	93.80	96.97

Результаты экспериментов, представленные в таблице 1, показывают, что применение Фишеровских признаков значительно увеличивает точность работы классификатора. Очевидно, что даже на небольшой выборке классификатор на новых признаках работает лучше, чем на MFCC признаках. Приведено объяснение этому с точки зрения размерности Вапника-Червоненкиса, так как применение Фишеровских признаков приводит к высокой VC – размерности пространства, и следовательно, большее количество точек может быть разделено гиперплоскостью.

Для выбора численного метода решения задачи (2) был произведен сравнительный анализ метода градиентного спуска, метода Нелдера-Мида и метода симуляции отжига. Для этого проведёно моделирование методом Монте-Карло по схеме марковской цепи случайной величины η, имеющей распределение (2) с параметрами $\theta_0 = (2, 3, 6, 10)$. Исследуются различные способы порождения марковской цепи, такие как алгоритмы Метрополиса - Хастингса, Гиббса и алгоритм срезов. Приведено объяснение выбора

метода для исследуемой задачи, а также описание проблем, возникающих при генерации марковской цепи, и путей их разрешения.

Результаты анализа, приведённые в таблице 2, подтверждают практическую применимость алгоритма симуляции отжига для решения задачи получения оценок параметров расширения Грам - Шарлье.

Таблица 2: Оценки параметров, полученные разными численными методами. В скобках приведены стандартные ошибки.

Параметр	Метод градиентного спуска	Метод Нелдера - Мида	Метод симуляции отжига
$c_1 = 2$	2.04 (0.07)	2.02 (0.07)	1.97 (0.07)
$c_2 = 3$	3.01 (0.05)	3.01 (0.05)	2.94 (0.05)
$c_3 = 6$	5.4 (0.84)	5.38 (0.85)	5.35 (0.84)
$c_4 = 10$	3.82 (5.1)	6.03 (5.12)	9.65 (5.84)

Далее в главе представлены результаты применения расширения Грам - Шарлье для моделирования фонемных признаков с применением двух модификаций алгоритма симуляции отжига, описанных в главе 2.

Проведён анализ предложенных модификаций алгоритма по скорости выполнения и точности подгонки. Как видно из таблиц 3 и 4 первый алгоритм оказался более затратным по времени. При этом, зависимость времени работы алгоритмов от количества процессоров почти линейная.

Для оценки качества подгонки была использована мера расхождения Кульбака-Лейблера $D_{KL}(P||Q)$ между двумя распределениями P и Q с плотностями $p(x)$ и $g(x)$, которая вычисляется по формуле

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} \ln \frac{p(x)}{g(x)} p(x) \mathrm{d}x$$

Первый алгоритм даёт лучшую степень подгонки: расхождение Кульбака-Лейблера меньше и значение целевой функции больше. Тем не менее, преимущество первого алгоритма нельзя назвать большим.

Таблица 3: Время работы алгоритмов.

Количество процессоров	Время 1го алгоритма, с	Время 2го алгоритма, с
1	9756	15
3	4465	13
6	2463	11

Таблица 4: Точность подгонки алгоритмов.

Мера подгонки	Результат 1го алгоритма	Результат 2го алгоритма
Расхождение Кульбака - Лейблера	0.117	0.140
Значение целевой функции	-32935	-34289

Из приведённых таблиц можно сделать вывод о том, что второй алгоритм, хотя и менее точный, является более предпочтительным при реальном

применении. При проведении экспериментов по распознаванию языка использовалась речевая база, содержащая данные по языкам (арабскому, английскому, мандарину, польскому, русскому и турецкому), записанным в телефонном канале. Характеристики входного сигнала: битрейт 16, частота дискретизации 8 кГц, соотношение сигнал – шум в среднем 15дб. Продолжительность речи на каждом языке в обучающем множестве составляла примерно 120 минут. Тестовое множество для каждого языка состояло из примерно 500 векторов.

Результаты экспериментов, приведённые в таблице 5, показывают преимущество в точности распознавания при применении предложенных методов.

Таблица 5: Точность распознавания языка (в процентах).

	Длительность, с							
	5		10		20		40	
	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.	Трад.	Разр.
Точность	34.11	48.35	22.72	46.91	28.42	33.74	25.71	36.91

В заключении подводятся итоги проделанной работы. Перечисляются основные результаты диссертации и следующие из них выводы.

Основные результаты работы, выносимые на защиту

1. Проведён анализ существующего состояния в сфере распознавания языка и диктора.
2. Выявлены дикторонезависимые признаки, основанные на 4-х параметрическом распределении, и доказана их оптимальность.
3. Разработана и теоретически обоснована модификация метода опорных векторов, основанная на применении фишеровских ядер, которая позволяет увеличить точность распознавания диктора.
4. Разработана модификация алгоритма симуляции отжига, увеличивающая быстродействие системы при получении дикторонезависимых признаков.
5. Проведён сравнительный анализ алгоритмов оптимизации для получения дикторонезависимых признаков по скорости и точности.
6. Разработаны и теоретически обоснованы методы и алгоритмы получения параметров классификатора для решения задач идентификации языка и диктора.
7. Создана программная реализация разработанной системы идентификации языка и диктора, фрагменты который внедрены на производстве.
8. Проведены экспериментальные исследования по оценке точности распознавания и быстродействию системы идентификации языка и диктора, которые показали преимущества разработанных методов по сравнению с применяемыми ранее.

Список работ, опубликованных по теме диссертации

Статьи в рецензируемых изданиях, рекомендованных ВАК РФ:

1. Ермилов А.В. Распознавание языка искаженного текста методом опорных векторов // Вестник РУДН. Серия Математика, Информатика, Физика. 2012. Т. 2. с. 126–130.
2. Ермилов А.В. Моделирование речевых признаков с помощью алгоритма симуляции отжига // Вестник РУДН. Серия Математика, Информатика, Физика. 2014. Т. 2. с. 354-358.
3. Гостев И.М., Ермилов А.В. О применении Фишеровских ядер в задаче распознавания диктора // Известия Юго-Западного Государственного Университета. Серия Вычислительная Техника, Информатика, Медицинское приборостроение. 2011. Т. 2. с. 15–20.
4. Ermilov A. V. Speech Technologies in human computer interactions // International Journal of Modern Manufacturing Technologies. 2013. Vol. 4. p. 52–57

Материалы международных, всероссийских, молодежных научных конференций

5. Ermilov A. V. Parallel Technologies in maximum likelihood estimation // Book of Abstracts of 5th International Conference “Distributed Computing and Grid-Technologies in Science and Education’’ (GRID-2012). 2012. p. 99.
6. Ermilov A. V. Speech technologies in human-computer interactions // Book of Abstracts of the First International Conference on Modern Manufacturing Technologies in Industrial Engineering “ModTech – 2013” 2013. p. 197.
7. Ermilov A. V. Modeling of speech features via simulated annealing algorithm //Book of Abstracts of the international Conference “Mathematical Modeling and Computational Physics - 2013” (MMCP’2013). 2013. p. 76.
8. Ermilov A. V. Fisher Kernels for speaker recognition // Book of Abstracts of Second International Symposium “Modeling of Nonlinear Processes and Systems (MNPS-2011)”. 2011. p. 291.
9. Ермилов А.В. Параллельные технологии в задаче максимизации правдоподобия // Труды 5-ой Международной конференции ”Распределенные вычисления и грид-технологии в науке и образовании”. 2012. с. 302-305.
10. Ермилов А.В. Применение расширения Грам-Шарлье для моделирования речевых признаков // Сборник материалов X Международной научно-технической Конференции “Оптико-электронные приборы и устройства в системах распознавания образов, обработки изображений и символьной информации. Распознавание - 2012”. 2012. с. 97.
11. Ермилов А.В. Математическая модель параллельных вычислений в системе автоматического распознавания речи // Сборник материалов XI Международной научно-технической Конференции “Оптико-электронные приборы и устройства в системах распознавания образов,

обработки изображений и символьной информации. Распознавание – 2013”. 2013. с. 252.

Личный вклад соискателя. В совместно опубликованных работах вклад автора диссертации является определяющим.

Ермилов А.В.

Методы, алгоритмы и программы решения задач идентификации языка и диктора

Аннотация

В диссертации разработаы методы идентификации диктора, основанные на методе опорных векторов с применением Фишеровских ядер. Кроме того, предложена и теоретически обоснована модель акустических дикторонезависимых признаков, использующая 4-х параметрическое распределение (расширение Грам-Шарлье) для моделирования речевых признаков. Разработана модификация алгоритма симуляции отжига, увеличивающая быстродействие системы распознавания при получении дикторонезависимых признаков за счет введения в алгоритм параллельно выполняющихся циклов.

Ermilov A.V.

Methods, algorithms and programms for language and speaker identification.

Abstract

In the dissertation elaborated methods of speaker identification, based on support vector machines with Fisher kernels. In addition developed and theoreticaly verified model of speaker-independent features, based on 4-parametric distribution (Gram-Charlier expansion). Modification of simulated annealing algorithm based on parallel cycles is developed. This modification allows to increase speed of the recognition system.