МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

имени Н. Э. Баумана (национальный исследовательский университет)

На правах рукописи

ДУ ЦЗЯНЬМИН

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДОВ ЗВУКОВОГО ПОИСКА В БАЗАХ ДАННЫХ НА ОСНОВЕ ФОНЕТИЧЕСКОГО КОДИРОВАНИЯ И ИХ ИСПОЛЬЗОВАНИЕ ДЛЯ УСКОРЕНИЯ РАСПОЗНАВАНИЯ РЕЧИ

Специальность: 05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

(технические науки)

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель: Выхованец Валерий Святославович

д.т.н., доцент

Москва — 2019

2

СОДЕРЖАНИЕ

Стр.

ВВЕДЕНИЕ 5

ГЛАВА 1. АЛГОРИТМ ФОНЕТИЧЕСКОГО КОДИРОВАНИЯ 11

1.1. Обзор алгоритмов фонетического кодирования 11

1.2. Алгоритм SoundEx 12

1.3. Другие производные алгоритмы 15

1.3.1. Алгоритм NYSIIS 15

1.3.2. Алгоритм Daith-Mokotoff SoundEx 17

1.3.3. Алгоритм Metaphone 19

1.4. Фонетическое расстояние 21

1.4.1. Расстояние Левенштейна 22

1.4.2. Расстояние на основе N-грамм 23

1.4.3. Расстояние Джаро 24

1.5. Исследование фонетических алгоритма русского языка 25

1.5.1. Алгоритм хеширования по сигнатуре 26

1.5.2. Алгоритм MetaphoneRu 28

1.5.3. Алгоритм Polyphone 30

1.5.4. Анализ важных характеристик русских слов 31

1.6. Выводы по первой главе 34

ГЛАВА 2. ТЕОРИЯ МЕТОДА ПОИСКА СЛОВ ДЛЯ

УСКОРЕНИЯ РАСПОЗНАВАНИЯ РЕЧИ 35

2.1. Обзор системы распознавания речи 35

2.2. Выбор набора инструментов для исследования 36

3

Стр.

2.3. Скрытая Марковская модель в области распознавания речи 38

2.3.1. Обзор HMM 38

2.3.2. Архитектура системы распознавания речи на основе HMM 39

2.3.3. Развитие модели HMM 49

2.4. Теория поиска слов на основе фонетического кодирования 50

2.4.1. Проблемы поиска слов в словаре 50

2.4.2. Исследование фонем русского языка с помощью CMUSphinx 51

2.4.3. Основная схема метода поиска слова 54

2.4.4. Функция «Аллофон» и его использование 57

2.4.5. Фонетический алгоритм и анализ результата последовательности фонем 60

2.4.6. Анализ правила и строения кодирования 72

2.5. Разработка метода звукового поиска слов в базах данных 85

2.5.1. Построение базы данных слов 85

2.5.2. Таблица перекодирования 87

2.5.3. Поиск из словаря и строение нового словаря 89

2.6. Применение 91

2.7. Выводы по второй главе 94

ГЛАВА 3. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ 95

3.1. Модуль «поиск слов на основе фонетического

кодирования» 95

3.2. Эксперименты системы и анализ результатов 96

3.2.1. Эксперимент эффективности 96

4

Стр.

3.2.2. Эксперимент правильности 98

3.3. Выводы по третьей главе 99

ОБЩИЕ ВЫВОДЫ И ЗАКЛЮЧЕНИЕ 100

СПИСОК ЛИТЕРАТУРЫ 101м

ОБЩИЕВЫВОДЫИЗАКЛЮЧЕНИЕ

ВпервойглаверассмотренопонятиеалгоритмовфонетическогокодированияОписанатеоретическаячастьработыссамымизвестнымалгоритмомфонетическогокодированияОсновнойпринципкодированиявалгоритмесостоитвтомчтоблизкиепозвучаниюбуквыкодируютсяоднойитойжецифройОписаныдругиепроизводныеалгоритмыиалгоритмыопределениярасстояниямеждусловамиприпоискепозвучаниюПредложеноиспользованиефонетическихалгоритмовдляпоискасловвбазахданных

ВовторойглаверассмотренатеорияалгоритмапоискасловвбазахданныхдляускоренияраспознаванияречинаосновефонетическогокодированияВпервойчастиглавыпредставленобзорсистемыраспознаванияречиВовторойчастиглавырассмотренатипичнаяархитектурасистемраспознаванияречиВтретейчастиглавыразработанэффективныйалгоритмфонетическогокодированиядляпоследовательностифонемрусскогоязыкапозволявшийнаходитьблизкиепопроизношениюсловаРазработанметодпоискасловвбазахданныхнаосновеалгоритмафонетическогокодированияВчетвертойчастиглавыописанпроцесспоискасловвбазахданныхнаосновефонетическогокодированияВпоследнейчастиглавыописаныспособыпримененияразработанногометодазвуковогопоисканаосновефонетическогокодирования

ВтретейглавеосвещеноэкспериментальноеисследованиеОпределенаэффективностьполученныхрезультатовМодифицированнаясистемаобладаетболеевысокойэффективностьюиправильностьюраспознаванияслов

ТакимобразомвсепоставленныевработезадачирешеныацельдостигнутаРезультатынаучныхисследованийипроведенныхиспытанийрекомендованыдлядальнейшегоиспользованиявсистемахраспознаванияречинабольшимсловаресцельюускоренияпроцессараспознаванияиповышениякачестваполучаемыхрезультатов