На правах рукописи

НИЦЕНКО АРТЁМ ВЛАДИМИРОВИЧ

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ ДЛЯ АНАЛИЗА УСТНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ ДИФОНОВ НА ОСНОВЕ АПРИОРНОЙ СЕГМЕНТАЦИИ

Специальность 05.13.17 – «Теоретические основы информатики»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Таганрог – 2018

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Южный федеральный университет» (ЮФУ) на кафедре систем автоматического управления.

Научный руководитель: **Финаев Валерий Иванович,** доктор технических наук, профессор, ФГАОУ ВО ЮФУ, кафедра «Системы автоматического управления», заведующий кафедрой (г. Таганрог).

Официальные оппоненты: **Харламов Александр Александрович,** доктор технических наук, Федеральное государственное бюджетное учреждение науки Институт высшей нервной деятельности и нейрофизиологии РАН, старший научный сотрудник (г. Москва).

Хусаинов Айдар Фаилович, кандидат технических наук, Институт прикладной семиотики Академии наук Республики Татарстан, старший научный сотрудник (г. Казань).

Ведущая организация: Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации РАН.

Защита состоится 4 июля 2018 г. на заседании диссертационного совета Д 999.065.02 Южного федерального университета по адресу: 347928, г. Таганрог, пер. Некрасовский, 44, ауд. Д-406.

С диссертацией можно ознакомиться в Зональной научной библиотеке ЮФУ по адресу: г. Ростов-на-Дону, ул. Зорге 21, и на сайте: https://hub.sfedu.ru/diss/announcement/

Автореферат разослан «___»_____ 2018 г.

Ученый секретарь диссертационного совета
Д 999.065.02, доктор технических наук,
профессор А.Н. Целых

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. С момента формирования и в процессе развития информатики как науки, а также совершенствования компьютерной техники как технической базы информатики, одной из важнейших проблем является проблема теоретического обоснования и практической реализации средств человеко-машинного интерфейса. В рамках этой проблемы важное место занимают задача автоматического распознавания речи. В числе последних достижений появились достаточно успешно работающие практически применяемые системы распознавания речи с большими словарями – прежде всего, голосовой ввод в поисковых интернет-системах Google и Yandex. Однако их применение связано с работой в сети Internet и использованием облачных технологий. Проблема же распознавания речи на локальных компьютерах остается актуальной.

Появление ЭВМ привело к необходимости развития методов цифровой обработки устной речи. Важную роль в этой области сыграли работы Б. Голда, Д. Маркела, А. Оппенгейма, Л. Рабинера, Д. Рейди, Р. Шафера и др. Значительный вклад в развитие технологий распознавания речи внесли известные ученые Х. Сакоэ и С. Чиба в Японии, Ф. Итакура в США, В.М. Величко, Н.Г. Загоруйко, В.М. Сорокин, Т.К Винцюк в Советском Союзе, О.Н. Карпов, Н.Н. Сажок, Ю.В. Крак в Украине, А.А. Карпов, Р.К. Потапова, А.Л. Ронжин, А.А.Харламов в России. Для решения задачи распознавания устной речи было разработано большое количество методов, однако в общем виде задача до сих пор не решена. Поэтому возникает необходимость в разработке и реализации усовершенствованных методов и алгоритмов анализа речевых данных. Анализ существующих в настоящее время систем распознавания речи, работающих на локальных устройствах, показывает, что они не удовлетворяют современным требованиям. Это обстоятельство определяет актуальность исследований в этом направлении.

Одним из первых методов распознавания, которому уделяется внимание и сейчас, является метод сравнения исследуемого образца речи с эталоном на основе алгоритмов динамического программирования. Однако считается, что данный метод пригоден только для распознавания малого словаря команд в силу больших затрат времени на создание эталонов и значительных вычислительных и временных затрат при распознавании. Тем не менее, он остается простым в реализации, открытым для улучшений и подходящим для приложений, которым требуется простое распознавание слов: телефоны, автомобильные компьютеры, системы безопасности и т.д. Поэтому актуальной задачей является повышение его эффективности путем разработки методов анализа речи, использующих автоматическую сегментацию. Это осуществимо за счёт синтеза эталонных образов из дифонов – сравнительно небольшого числа речевых данных, содержащих межфонемные переходы. Такое усовершенствование позволит выполнять распознавание со словарями большого объема без предварительного создания голосовых эталонов для цифровых данных всех слов из словаря и существенно повысить скорость распознавания по сравнению с классическим методом.

Тема диссертационной работы является актуальной, т.к. она посвящена решению задачи разработки и исследования методов анализа устной речи на основе априорной сегментации и алгоритма динамического программирования, использующего эталоны слов, автоматически синтезируемые из эталонов дифонов. Это позволяет во много раз сократить количество базовых эталонов и, как следствие, время обучения системы распознания, а также обеспечить возможность работы в режиме реального времени с большим объемом словаря.

Цель диссертационной работы состоит в разработке методов и алгоритмов для анализа цифровых данных устной речи на основе априорной сегментации и модифицированного алгоритма динамической трансформации временной шкалы (DTW-алгоритма), обеспечивающего повышение эффективности процесса распознавания данных речи за счет использования эталонов, синтезируемых из эталонов дифонов по транскрипциям слов.

В соответствии с поставленной целью в диссертационной работе решаются следующие задачи:

– анализ известных методов и алгоритмов распознавания устной речи, выявление и обоснование подходов, наиболее пригодных для достижения поставленной цели;

– разработка метода автоматической априорной сегментации речевых данных;

– разработка метода автоматического извлечения дифонов из речевых данных на основе априорной сегментации;

– разработка метода анализа данных устной речи на основе априорной сегментации и модифицированного алгоритма DTW, использующего эталоны слов, автоматически синтезируемые из эталонов дифонов по транскрипциям слов словаря;

– реализация и тестирование разработанных методов путем разработки специализированного программного обеспечения для сравнения разработанных методов с известными методами распознавания устной речи.

Методы исследований. При решении сформулированных в работе задач использовались методы динамического программирования, методы сегментации речевых данных, методы объектно-ориентированного проектирования и программирования.

Достоверность и обоснованность научных положений, выводов и результатов, сформулированных в диссертационной работе, подтверждается результатами теоретических исследований и логическими выводами, публикациями, апробацией работы на международных научно-технических конференциях, актами о внедрении.

Объектами исследования в диссертационной работе являются методы, алгоритмы и системы распознавания речи.

Научная новизна. В диссертации получены следующие новые научные **результаты, которые выносятся на защиту**:

– разработан метод анализа устной речи, отличающийся тем, что эталонные образы целых слов синтезируются из эталонов дифонов (сравнительно небольшого числа речевых данных, содержащих межфонемные переходы), что

позволяет выполнять распознавание речи со словарями большого объема без предварительного создания голосовых эталонов всех слов; за счет использования дифонного дерева обеспечивается повышение скорости распознавания в 3-4 раза по сравнению с классическим методом на основе DTW;

– разработан метод автоматической априорной сегментации речевых данных, отличающийся тем, что для определения границ между фонемами в речи с заранее неизвестным фонемным составом применяется новый способ анализа структуры коротких участков речи, позволяющий с использованием численного аналога полной вариации и полной вариации с переменным верхним пределом автоматически определять межфонемные переходы в данных речи;

– разработан метод автоматического извлечения эталонов дифонов из речи, отличающийся использованием информация о границах между фонемами, полученной с помощью априорной сегментации, и позволяющий автоматизировать процесс создания базы дифонов при обучении системы распознавания.

Практическая ценность результатов исследований состоит в использовании полученных результатов для создания программного обеспечения, с применением которого решается задача автоматизации создания речевых баз данных, проектирования систем распознавания речи со словарем большого объема и интеллектуальных систем взаимодействия пользователя и компьютера.

Соответствие специальности. Тематика работы соответствует следующим пунктам паспорта специальности 05.13.17 – Теоретические основы информатики:

– п.5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечения; разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»

– п.6 «Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке».

Результаты работы внедрены:

– при выполнении госбюджетных научно-исследовательских работ в институте проблем искусственного интеллекта (ИПИИ) МОН и НАН Украины «Разработка методов компьютерного восприятия слитной речи на основе пофонемного распознавания речевых образов», шифр РСМ-2005, № 0105U001160; «Исследование проблем искусственного интеллекта по компьютерному распознаванию речи с учетом семантики произнесенного и использованием разработанного инверсионного грамматического словаря украинского языка», шифр РСМ-2008, № 0108U003014; «Разработка модуля пословной диктовки со словарем 100 тысяч словоформ для текстового редактора Word», шифр СМС_РІС 2013, №0113U0011327.

– в учебном процессе на кафедре программной инженерии Донецкого национального технического университета в курсе «Цифровая обработка сигналов и распознавание речи»;

– на предприятии ООО «Техно КМВ» приняты к использованию методы и алгоритмы, разработанные диссертации.

Основные положения и результаты диссертационной работы докладывались и обсуждались на следующих конференциях: IV МНК «Интеллектуальные и многопроцессорные системы – 2003» (Дивноморское, 2003); V, VII, VIII, XI, XII и XIII МНТК «Искусственный интеллект. Интеллектуальные и многопроцессорные системы – 2004» (Кацивели, 2004, 2006, 2007, 2010, 2012 и 2013 гг.); VI МНТК «Искусственный интеллект. Интеллектуальные и многопроцессорные системы – 2005» (Дивноморское, 2005); VIII Всероссийской конференции с международным участием «Новые информационные технологии в исследовании сложных структур», (Томск, 2010).

По теме диссертации опубликованы 17 статей, в том числе 2 статьи в изданиях, рекомендованных ВАК РФ, 2 работы в журнале, индексируемом базой данных SCOUPUS. Все результаты, представленные в диссертационной работе, получены автором лично. В совместных научных публикациях имеет место неделимое соавторство.

Структура и объём диссертации. Диссертационная работа состоит из введения, четырех разделов, заключения, списка литературы и приложений. Полный объем работы – 192 страницы, из них основной текст 143 страницы, 2 приложения на 28 страницах, список литературы на 21 странице, 188 источников, 48 рисунков и 7 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы цель и задачи, описаны методы исследования, научная новизна, практическая значимость, основные научные результаты, выносимые на защиту, достоверность и обоснованность научных положений диссертации, апробация работы.

В первом разделе проведен анализ научных источников, связанных с диссертацией. Рассмотрен объект исследования – речь, а также способы ее представления в цифровой форме. Приведена артикуляционная классификация звуков речи, поставлена задача распознавания речи и описана общая структура системы автоматического распознавания речи. Установлено, что для автоматической сегментации речевых данных используются скрытые марковские модели, нейросетевые модели, методы динамического программирования а также методы, основанные на анализе временных и частотных характеристик речи. Эти подходы имеют следующие недостатки: приводят к появлению дополнительных границ на участках, соответствующих одной фонеме, требуют длительной и трудоемкой процедуры обучения. Приведена классификация и описание подходов составления векторов-признаков, работающих как в частотной области (коэффициенты линейного предсказания, мел-кепстральные коэффициенты), так и во временной (частота проходов через ноль, кратковременная энергия). Среди методов получения признаков распознавания для использования в данной работе был выбран метод, основанный на вычислении относительных частот длин полных колебаний.

Сделан вывод об эффективности метода сопоставления с эталонами целых слов при помощи алгоритма динамического программирования для

распознавания отдельных слов с малым словарем (единицы и десятки слов). Он был взят за основу для создания метода анализа речевых данных с использованием дифонов, который может быть применен и к большим словарям, содержащим тысячи и десятки тысяч слов.

Определены основные направления исследований диссертационной работы:

Во втором разделе выполнена разработка алгоритма определения начальной и конечной точек речи и метода априорной сегментации речевых данных.

Предложен алгоритм определения начальной и конечной точек речи, использующий в качестве признака отношение $R = \frac{V}{C}$, где V – вариация, определяемая как

$$V = \sum_{i=1}^{N-1} |x_{i+1} - x_i|, \quad (1)$$

где x_i – значение *i*-го отсчета речевых данных;

N – количество отсчетов в блоке данных.

Величина C представляет собой количество «точек постоянства», то есть моментов времени, для которых в следующий момент значение данных остается неизменным:

$$C = \sum_{k=1}^{N-1} C_k \,,\; C_k = \begin{cases} 0, \textit{если} & x_k \neq x_{k+1} \\ 1, \textit{если} & x_k = x_{k+1} \end{cases}, \quad (2)$$

где x_i – значение *i*-го отсчета речевых данных;

N – количество отсчетов в блоке данных.

Отношение R вычисляется на отрезках (блоках) данных речи размером в 300 отсчетов (14 миллисекунд), которые поступают от устройства записи. Решение о начале речевого отрезка принимается на основе сравнения значения признака на текущем участке данных с пороговым значением $T_{нач.}$, которое рассчитывается по значениям величины R на первых 10 блоках данных:

$$T_{\text{нач.}} = \frac{3 \cdot \sum_{i=1}^{10} R_i}{10}.$$

Если этот порог превышен на 5 блоках подряд, то первый из этих блоков считается начальной точкой речи. Конечная точка речи определяется сравнением текущего значения признака с порогом $T_{кон.}=5 \cdot T_{нач.}$. Если значение признака ниже порога $T_{кон.}$ на 30 блоках данных подряд, то последний блок перед ними считается концом речевого отрезка.

Предлагаемый метод автоматической априорной сегментации состоит в разбиении речевых данных на участки, соответствующие гласным (сегмент «W»), голосовых согласным (сегмент «C»), глухим шипящим (сегмент «F») и глухим взрывным (паузообразным) фонемам (сегмент «P»). Предлагается метод сегментации, состоящий из 3-х этапов: сегментация на невокализованные и вокализованные участки, сегментация невокализованных участков на шипящие и

паузы, сегментация вокализованных участков на гласные и звонкие согласные (рисунок 1).

Для первого этапа в качестве языковых единиц взяты сегменты, формируемые по групповым признакам: «NV» – невокализованные сегменты (не содержащие голоса), «V» – вокализованные. «NV» может соответствовать одной глухой взрывной «P» или шипящей «F» фонеме, или включать возможные сочетания «FF», «PP», «FP», «PF», «FPF», «PFP» и т.д. из глухих взрывных и шипящих фонем. «V» может содержать одну гласную «W» или звонкую согласную фонему «C», или сочетание из этих фонем.

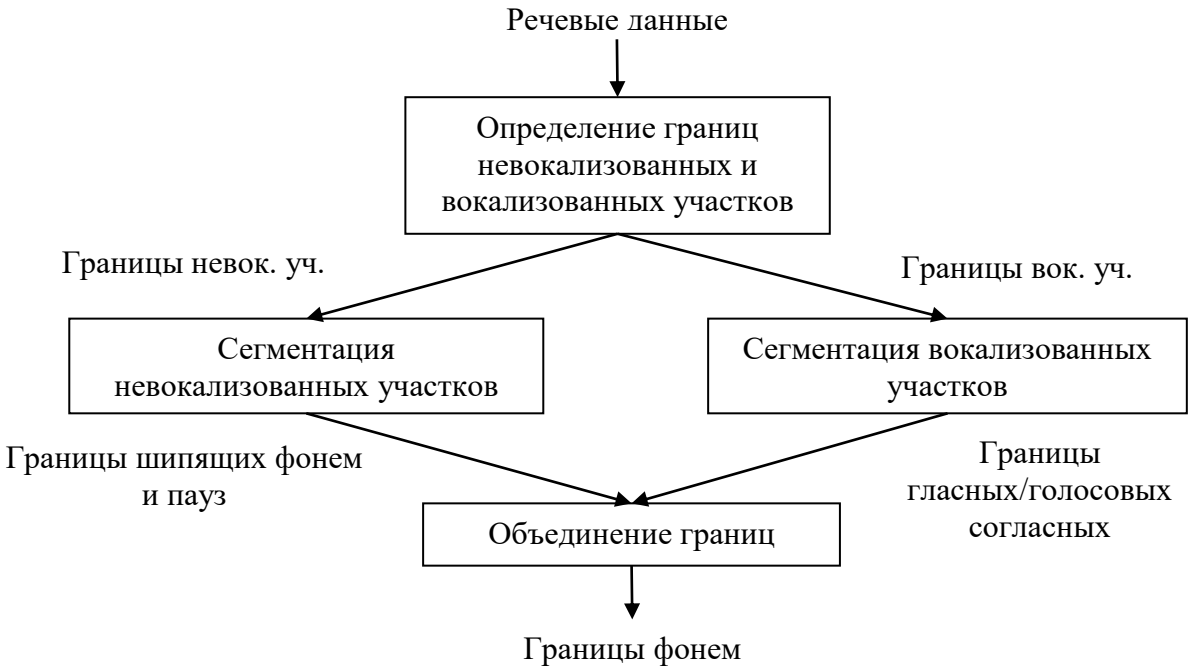

Рисунок 1 – Схема метода автоматической априорной сегментации речевых данных

На втором этапе осуществляется анализ параметров для разделения данных участка «NV» на возможные сочетания «PP», «FP», «PF», «FPF», «PFP» и т.д., т.е. проверка каждого из сегментов «NV» на принадлежность его к одному или нескольким классам фонем. В случае отнесения к разным классам проводится дополнительная сегментация и идентификация выделенных сегментов.

На третьем этапе анализируются данные сегментов «V» на возможные сочетания звонких согласных «C» и гласных «W». Для выделения невокализованных и вокализованных участков, речевые данные подвергаются обработке цифровым полосовым фильтром. Участки данных, соответствующие шипящим и аффрикатам превращаются в паузообразные с большим количеством точек постоянства (т.е. моментов дискретного времени, для которых в следующий момент значение речевых данных остается неизменным). Обработанные фильтром данные разбиваются на смежные неперекрывающиеся блоки по 256 отсчетов (12 мс). Такие участки данных с глухими согласными можно выделить, оценив разницу между количеством точек непостоянства и количеством точек

постоянства в каждом блоке: $D = 256 - 2 \cdot C$. Если в нескольких идущих подряд блоках эта разница отрицательна, то они относятся к невокализованному сегменту, в противном случае – к вокализованному.

Чтобы определить границы между сегментами, соответствующими шипящим и глухим взрывным фонемам, на невокализованных сегментах проводится дополнительная сегментация с использованием численного аналога полной вариации «с переменным верхним пределом»

$$V(0) = 0, \qquad V(n) = \sum_{i=0}^{n-1} |x_{i+1} - x_i|. \tag{3}$$

Так как для шипящих характерна высокая частота изменения речевых данных, то эта функция будет возрастать для шипящих быстрее, а для паузообразных медленнее. Чтобы использовать этот факт для классификации шипящих и пауз, определим также вспомогательную функцию $W(n) \overset{\text{mod } 256}{=} V(n)$, возрастающую вместе с V, однако «сбрасываемую» на 0 по достижении значения 255. В результате появляется массив чисел

$$N_1, N_2 - N_1, N_3 - N_2 \ldots . \tag{4}$$

Каждое число из массива (4) – это длина участка, на котором величина $W(n)$ возрастает от 0 до 255. На сегменте шипящей величина (3) быстро возрастает, поэтому участки возрастания величины $W(n)$ от 0 до 255 короткие, то есть числа (4) относительно малы. На сегменте паузы величина (3) растет медленно и поэтому числа (4) относительно велики. Для разделения сегментов шипящих и пауз вводится порог $T_{шип.}$ (в системе автора он равен 200). На выделенных вышеописанным методом невокализованных сегментах строится последовательность чисел (4). Те блоки данных, для которых числа (4) превышают $T_{шип.}$, относим к сегменту паузы (символ «*P*»), остальные – к сегменту шипящей (символ «*F*»). Получали границы сегментов шипящих и пауз.

Далее голосовые сегменты данных делятся на сегменты, соответствующие гласным и звонким согласным фонемам с использованием численного аналога полной вариации. Речевые данные разбиваются на последовательные блоки по 256 отсчетов (12 мс), и на каждом из них вычисляется значение вариации (1). От начала данных берется интервал 20 таких блоков или столько, сколько позволяет длина участка, и вычисляется среднее значение соответствующих величин (1), которое принимается за порог. Затем интервал, на котором выполняется описанная процедура, сдвигается вправо на один блок и процедура повторяется. Это происходит до тех пор, пока левый конец интервала не выйдет за пределы участка. В результате возникает последовательность строк следующего вида, показанного на рисунке 2.

Затем пересматриваются все строки полученной последовательности и создается новая символьная строка S. Если текущая i-я строка таблицы начинается и заканчивается одним и тем же символом («Н» или «В»), то на i-ю позицию в S записывается соответствующий символ. Иначе подсчитывается количество каждого из символов в этой строке. Если количество «В» превышает количество «Н» или равно ему, то в S на соответствующую позицию записывается

«В», иначе «Н». Границы сегментов определяются как номера блоков, где происходит смена символов «Н» на «В», или «В» на «Н». «В» – участок считается соответствующим сегменту гласной фонемы (около левой метки проставляется символ «W»). «Н» – участок считается соответствующим сегменту звонкой согласной фонемы (около левой метки проставляется символ «С»). Результатом являются номера блоков, соответствующих границам между сегментами.

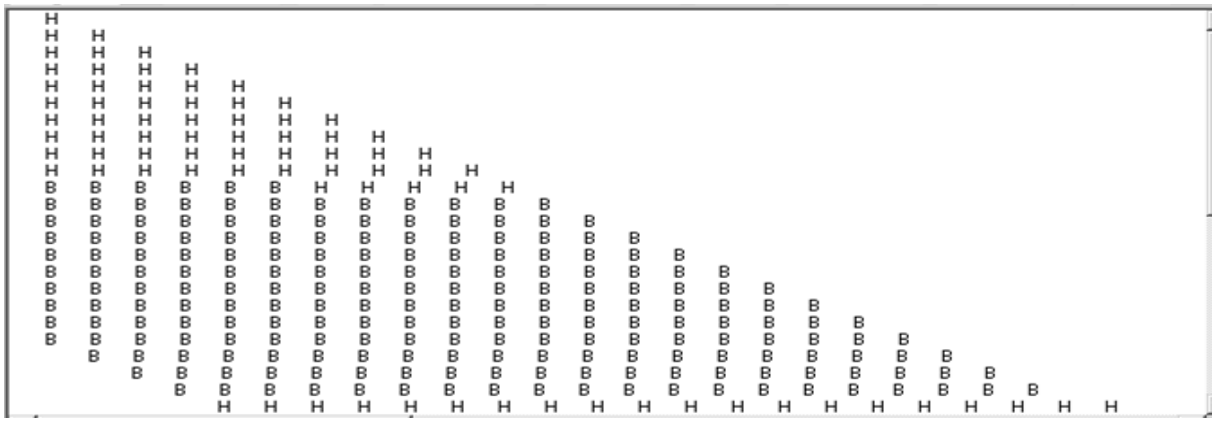

Рисунок 2 – Результат «В-Н» обработки речевых данных: «В» – значения выше среднего, «Н» – значение ниже среднего на интервале («В-Н» обработка)

В третьем разделе разработан метод анализа речевых данных на основе сравнения с эталонами с помощью алгоритма DTW, который заключается в использовании эталонов слов, автоматически синтезируемых из эталонов дифонов по транскрипции слова. Применяются вектора признаков, связанные с относительными частотами длин полных колебаний на блоках речевых данных размером в 368 отсчетов (17 миллисекунд).

Модификация классического метода распознавания с помощью алгоритма DTW заключается в использовании для распознавания эталонов слов, которые автоматически синтезируются из эталонов дифонов, полная база которых в объеме около 1700 эталонов создается для каждого диктора заранее. Схема метода показана на рисунке 3.

Под дифоном, который соответствует межфонемному переходу внутри слова, понимается участок стандартной длины: 3 блока в 368 отсчетов слева от границы между сегментами и 3 таких же блока справа от той же границы. Эталон дифона – набор из шести соответствующих векторов признаков. Кроме того, используется участок в 3 блока в начале речи и участок в 3 блока в конце речи, условно называемый соответственно начальным и конечным полудифоном (переход от молчания к речи и наоборот) (рисунок 4). Каждому эталону дифона присваивается свой уникальный идентификатор. Всем векторам признаков, входящим в эталоны, также присваивается идентификатор.

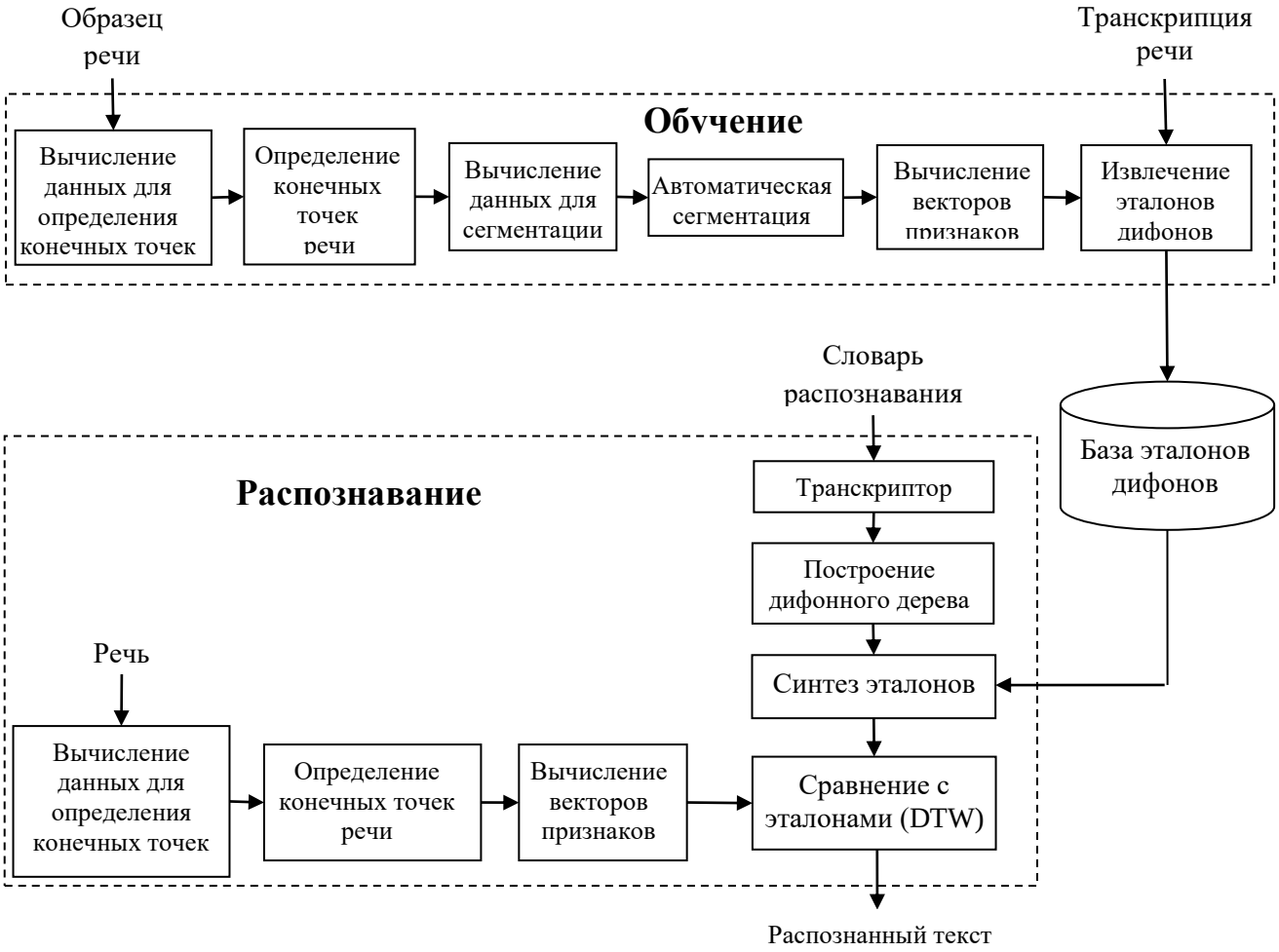

Рисунок 3 – Общая схема метода анализа устной речи с использованием дифонов на основе априорной сегментации

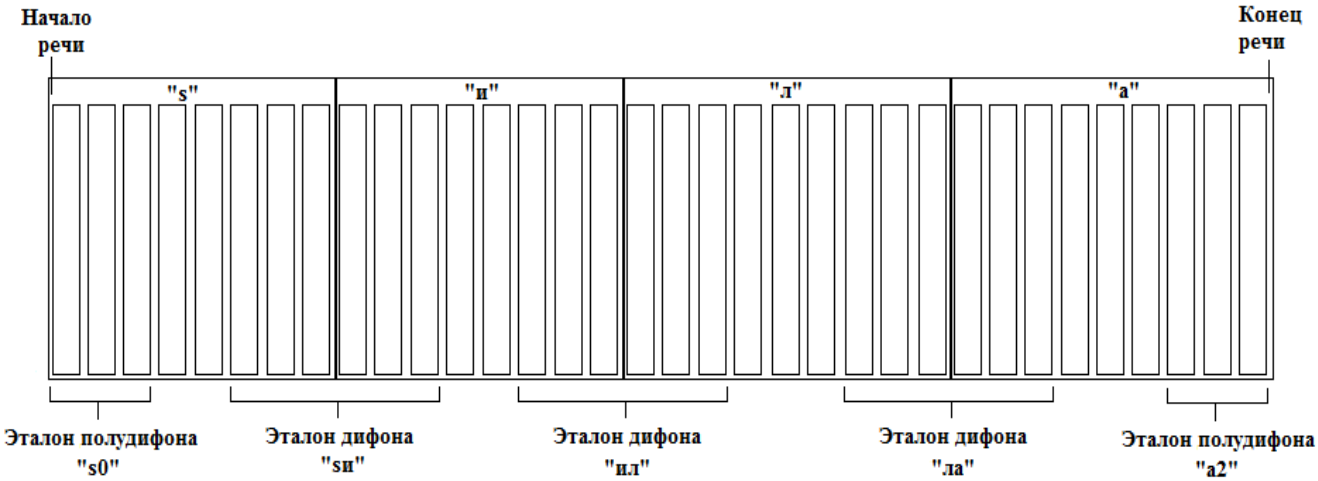

Рисунок 4 – Извлечение эталонов дифонов из речи (слово «сила»)

Очевидно, что распознавание слов или фраз с использованием дифонов требует предварительного создания их фонетической транскрипции.

В диссертации разработан транскриптор, который позволяет по мере накопления опыта модифицировать систему транскрипции путем простейших

изменений в управляющем файле с целью учета ранее не учтенных транскрипционных ситуаций. Управляющий файл содержит набор правил, каждое из которых записано в виде двух частей, соединенных знаком равенства. Слева стоят символы буквенной записи слова, справа – символы, которыми они заменяются в транскрипции. Символ «\» означает ударение. Машина, транскрибируя слово, последовательно ищет вхождения левой части очередного правила, и если такое обнаруживается, заменяет его правой частью.

В качестве транскрипционных знаков применены буквы русского алфавита. Исключение составляют символы «*w*», «*q*» для ударных [е], [я] соответственно. Твердые русские согласные транскрибируются также русскими буквами, а соответствующие мягкие согласные аналогичными латинскими буквами. Исключения: символом «*@*» обозначается мягкий [п], символом «*$*» – мягкий [ж], символом «*&*» – южнорусский (украинский) [г], значком + обозначается слитный звук [д'ж'] (звонкая параллель [ч]), символом «*» твёрдый [ч] (в слове *«лучше»*), символом «*%*» – слитный звук [дз] (звонкая параллель [ц]), символом «^» фрикативный звук, появляющийся следом за [й] в конце слова (факультативно). Знак «#» обозначает начало слова, если стоит перед любой буквой, и конец слова, если стоит после какой-либо буквы. Каждое слово из словаря распознавания автоматически транскрибируется, по транскрипции строится цепочка имен дифонов, например, *остановка → астанофка → а0-ас-ст-та-ан-но-оф-фк-ка-а2*.

Каждое из них заменяется эталоном соответствующего дифона. Полученная цепочка векторов образует эталон слова. Словарь эталонов слов реализован в виде дерева дифонов, использование которого существенно ускоряет процесс распознавания. Дифоны представлены в дереве своими идентификаторами.

Эталон каждого слова представляется в виде пути в этом дереве, от начального до конечного дифона цепочки. Если несколько путей имеют общую часть, то вычисления, заполняющие соответствующую часть DTW-матрицы, выполняются только один раз. Уровни дерева соответствуют позициям дифонов в слове (рисунок 5).

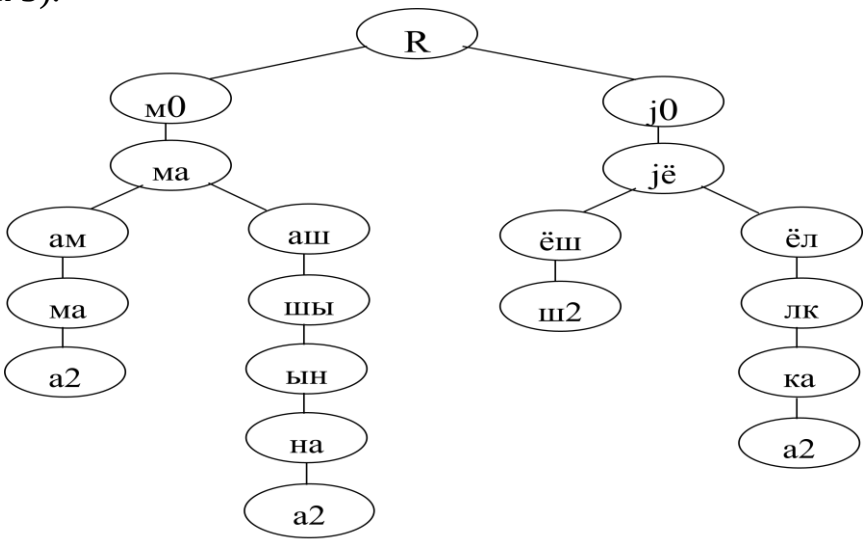

Рисунок 5 – Схема дерева синтеза эталонов для простого словаря

Каждый узел в рамках каждого уровня представляет собой идентификатор дифона, что находится в слове на соответствующей позиции. Узлы, соответствующие конечным дифонам слов, обозначаются как концы соответствующих слов (в узле записывается порядковый номер соответствующего слова в словаре). Если узел не конечный, то записывается значение «–1». Максимальная глубина дерева соответствует максимальной длине пути (выраженной в количестве дифонов) для соответствующего слова в словаре.

Процесс распознавания осуществляется следующим образом. Создается параметрическое представление речевых данных в виде набора n векторов признаков и строится таблица D расстояний этих векторов ко всем векторам признаков эталонов дифонов. Далее вычисляются DTW-расстояния от распознаваемого слова до всех эталонов слов путем рекурсивного обхода дерева эталонов «в глубину». Сначала просматривается корень дерева, а затем рекурсивно смежные узлы вглубь дерева, пока не достигнут узел, помеченный как конец слова. После того, как достигнут конец слова, происходит возврат назад вдоль пройденного пути пока не найден узел, у которого есть еще не просмотренный потомок. Затем движемся в новом обнаруженном направлении. Процесс завершается, когда просмотрены все узлы дерева.

При прохождении ветвей дерева, по идентификаторам дифонов строится цепочка соответствующих им идентификаторов векторов признаков, образующих эталон слова. При движении в глубину, в цепочку добавляются идентификаторы, соответствующие пройденным узлам, а при движении назад они удаляются из нее. Достигнув узла, являющегося концом очередного слова, вычисляется DTW-расстояние от построенной цепочки векторов (эталона данного слова) до распознаваемой последовательности векторов признаков. Расстояния между векторами берутся из таблицы D. Особенностью процесса вычисления расстояний является то, что матрица DTW не пересчитывается полностью, а обновляются только столбцы, соответствующие новым кодовым векторам, номера которых добавлены в цепочку после возврата назад по окончании предыдущего этапа.

Время работы данного алгоритма зависит от количества векторов признаков в исследуемом образце – n, количества векторов признаков в синтетическом эталоне – m и общего количества эталонов слов – v. Для классического алгоритма DTW временная сложность алгоритма может быть оценена как $O(n\sum_{i=1}^{v} m_i)$. За счет использования дерева сложность алгоритма будет иметь нелинейную зависимость от v и будет тем меньше, чем больше транскрипций в словаре имеют общую часть. Это позволяет достичь значительного выигрыша в скорости распознавания по сравнению со стандартным алгоритмом DTW.

Распознавание между собой словоформ одного и того же слова представляет более трудную задачу, чем распознавание словоформ различных слов. Это вызвано тем, что они, как правило, отличаются окончаниями, которые чаще всего безударны и редуцируются при произношении. С целью увеличения надежности распознавания словоформ предлагается начинать распознавание с распознавания окончаний. В пользу этого можно привести следующие качественные соображения. Две словоформы достаточно длинного слова имеют общую основу и, следовательно, имеют больше общего, чем различий, что может

служить источником ошибок. Если же ограничиться распознаванием одних окончаний, то их отличия относительно больше, чем отличия полных словоформ. Поэтому ошибки в их распознавании должны быть менее частыми. С другой стороны качество распознавания с помощью DTW возрастает при увеличении длины распознаваемых речевых отрезков. Поэтому целесообразно присоединять к окончанию часть суффикса, и работать с этими объектами, которые естественно назвать квазифлексиями. Соответственно оставшуюся часть слова будем называть квазиосновой.

Использование квазифлексий приводит также к сокращению размеров распознаваемых словарей. Квазифлексии, очевидно, являются общими для больших групп слов. Если имеется m квазиоснов и n квазифлексий, то их комбинации образуют $m \times n$ словоформ и, при распознавании словоформы как целого, размер словаря для распознавания составил бы $m \times n$ объектов. При распознавании же квазиосновы и квазифлексии отдельно, размер словаря составляет $m+n$ объектов. В результате время распознавания значительно сокращается, а надежность распознавания имеет тенденцию к увеличению.

Итак, для решения указанных проблем предлагается распознавать словоформы в два этапа: вначале распознавая изменяющуюся часть слова (квазифлексию), затем неизменяющуюся часть (квазиоснову) из множества, соответствующего распознанной квазифлексии. Введенное понятие квазиосновы родственно используемому в лингвистике понятию основы слова, которая при простейшем описании определяется как его неизменяемая часть (приставка+корень+суффикс), то есть является результатом отбрасывания окончания. Короткие словоформы (состоящие менее чем из 5 фонем) включаются в число квазиоснов целиком. Исходя из того, что русский язык является флективным языком, слова языка моделируются в виде комбинации постоянной и переменной составляющих:

$$x = c(x)\&f(x) \qquad (4)$$

где $c(x)$ – часть лексемы x, которая в процессе словоизменения остается неизменной (квазиоснова), $f(x)$ – ее переменная составляющая (квазифлексия), $\&$ – знак конкатенации.

Так как распознавание будет вестись с использованием эталонов дифонов, то для каждой квазиосновы и квазифлексии используется транскрипция и по ней создается цепочка соответствующих дифонов. Для распознавания применяется алгоритм на основе DTW, описанный выше. Определение квазифлексии производится по принципу минимума DTW-расстояния. Он заключается в последовательном распознавании с помощью алгоритма DTW заключительных частей речевых данных, начиная с двух конечных сегментов: вначале берется два последних сегмента, затем три, четыре и так далее до задаваемого заранее максимального количества фонетических сегментов. При этом запоминается минимальное значение DTW-расстояния среди всех эталонов и соответствующая этому эталону квазифлексия, и далее производится сравнение эталонов со следующим участком данных. Таким образом, получается список гипотетических квазифлексий и DTW-расстояний от их эталонов до рассматриваемых отрезков речевых данных. Из этого списка выбирается квазифлексия с наименьшим

расстоянием. Затем для выделенной таким образом квазифлексии происходит обращение к словарю соответствующих квазиоснов, и в пределах этого словаря осуществляется DTW-распознавание участка данных от начальной точки речи до начала участка, соответствующего распознанной квазифлексии.

В четвертом разделе разработано программное обеспечение для исследования предложенных методов сегментации и анализа речевых данных. Реализованы алгоритмы определения конечных точек речи, автоматической сегментации, автоматического извлечения эталонов дифонов и распознавания. На базе этих алгоритмов создана экспериментальная информационная технология распознавания изолированных слов для произвольного словаря большого объема (тысячи и десятки тысяч слов). Разработана программная реализация алгоритмов сегментации и классификации сегментов речевых данных, алгоритма автоматического построения транскрипции, а также алгоритма распознавания изолированных слов на основе сегментации и модифицированного алгоритма DTW с эталонами слов, автоматически синтезируемыми из эталонов дифонов. Разработаны структуры данных для представления базы эталонов дифонов.

С помощью разработанного программного обеспечения проведен анализ функционирования алгоритмов сегментации и анализа речевых данных. Проведено сравнение эффективности распознавания отдельно произносимых слов с методом на основе скрытых марковских моделей (в качестве тестовой платформы был использован HTK Toolkit) и коммерческой программой распознавания речи Voco, также работающей на основе статистических моделей. Для оценки качества распознавания речи использовался показатель процента корректно распознанных слов (WCR – Word Correctly Recognized).

В экспериментах участвовали 5 дикторов. Был сформирован словарь объемом 100 слов. Для каждого диктора был создан банк речевых данных – результатов произнесения слов словаря. Банк был записан в 5 версиях: одна версия предназначалась для режима обучения, остальные 4 – для режима тестирования. Запись производилась в условиях низкого уровня фонового шума. Параметры записи наборов слов: частота дискретизации – 22050 Гц; разрядность квантования – 8 бит; средняя длительность записанного слова – 2 с (включая окружающие слово паузы, длительностью не менее 0,3 с каждая).

В качестве тестовой платформы для метода скрытых марковских моделей был использован HTK Toolkit, чтобы построить базовую систему с MFCC коэффициентами. Для каждого диктора на обучающей выборке была обучена своя статистическая модель.

По результатам тестирования при распознавании отдельных слов с помощью DTW и синтетических эталонов (использования эталонов слов, синтезированных из эталонов дифонов), качество распознавания на тех же аудиоданных повышается на 3 – 18% по сравнению с распознаванием методом скрытых марковских моделей, и на 9 – 14% по сравнению с системой Voco (таблица 1).

Было проведено также исследование эффективности метода анализа речевых данных для распознавания изолированных слов на большом словаре. В процессе исследования из грамматического словаря русского языка А.А. Зализняка объемом около 100 тыс. (точная цифра: 94604) слов в начальных

формах случайным образом было отобрано 2 тысячи слов – тестовый словарь для распознавания. Далее из них случайным образом выбиралось 10 слов, которые произносил диктор. Этот последний этап повторялся 50 раз. Исследование показало высокую эффективность разработанного метода: доля корректно распознаваемых слов составляла не менее 90%.

Таблица 1 – Результаты сравнительного тестирования качества распознавания

Диктор	DTW+дифоны (WCR,%)	Классический DTW (WCR,%)	HTK Toolkit (WCR,%)	Voco (WCR,%)
1	98%	99%	95%	89%
2	98%	98%	93%	87%
3	96%	97%	92%	82%
4	96%	97%	78%	85%
5	97%	99%	89%	90%

Указанные результаты распознавания отдельных слов программой Voco объясняются тем, что она предназначена для распознавания слитной речи с использованием *n*-граммной языковой модели. Для отдельных слов языковая модель практически не работает, в этом случае система распознавания опирается только на встречаемость слова в тренировочных данных, что, по сути, не несет никакой полезной информации, и распознавание ведется только за счет статистической речевой модели. Таким образом, можно сделать вывод, что собственно распознавание отдельных слов в данной работе реализовано лучше.

Результатами экспериментов являются также время распознавания одного слова (таблица 2) и зависимость среднего времени распознавания от объема словаря (рисунок 6). Как видно из результатов, для стандартного метода DTW наблюдается существенный линейный рост времени распознавания одного слова в зависимости от количества слов в словаре. В то же время для метода на основе дифонов оно растет существенно медленнее, за счет использования дифонного дерева при распознавании. При этом для словаря в 100 слов модифицированный метод работает в 2 раза быстрее, а для словаря в 1000 слов – в 5 раз быстрее.

Таблица 2 – Характеристики среднего времени распознавания одного слова для словаря в 100 слов

Диктор	Классический DTW (среднее время распознавания,мс)	DTW+дифоны (среднее время распознавания,мс)
1	101	46
2	104	48
3	97	45
4	99	46
5	103	47

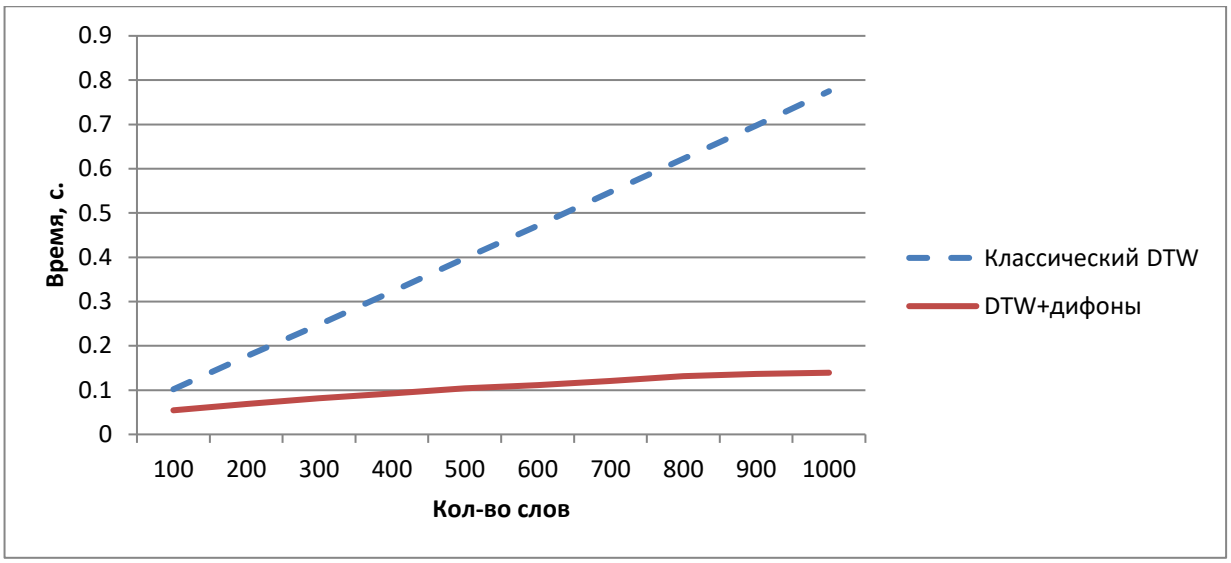

Рисунок 6 – Зависимость среднего времени распознавания слова от количества слов в словаре

Заключение содержит полученные в работе результаты.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

В диссертационной работе решена актуальная научно-техническая задача, которая состоит в развитии и усовершенствовании методов, средств и технологий анализа устной речи на основе априорной сегментации и алгоритма динамического программирования DTW, использующего эталоны слов, автоматически синтезируемые из эталонов дифонов. Это позволяет использовать метод сравнения эталонами в системах распознавания с большим объемом словаря.

Разработан метод анализа устной речи, отличающийся тем, что эталонные образы целых слов синтезируются из эталонов дифонов (сравнительно небольшого числа речевых данных, содержащих межфонемные переходы), что позволяет выполнять распознавание речи со словарями большого объема без предварительного создания голосовых эталонов всех слов; за счет использования дифонного дерева обеспечивается повышение скорости распознавания в 3-4 раза по сравнению с классическим методом на основе DTW.

Разработан метод автоматической априорной сегментации речевых данных, отличающийся тем, что для определения границ между фонемами в речи с заранее неизвестным фонемным составом применяется новый способ анализа структуры коротких участков речи, позволяющий с использованием численного аналога полной вариации и полной вариации с переменным верхним пределом автоматически определять межфонемные переходы в данных речи.

Разработан метод автоматического извлечения эталонов дифонов из речи, отличающийся использованием информации о границах между фонемами, полученной с помощью априорной сегментации, и позволяющий автоматизировать процесс создания базы дифонов при обучении системы распознавания.

Разработан автоматический фонетический транскриптор русского языка, позволяющий переводить любой текст в последовательность транскрипционных символов.

Разработано программное обеспечение на языке высокого уровня C++ в среде программирования Microsoft Visual Studio, реализующее алгоритмы сегментации, алгоритм автоматического построения транскрипции, а также алгоритм распознавания изолированных слов на основе сегментации и модифицированного алгоритма DTW с эталонами слов, автоматически синтезируемыми из эталонов дифонов.

Проведено исследование эффективности распознавания отдельных слов с использованием разработанного метода анализа речи. Результаты исследования показали, что при распознавании отдельных слов с помощью DTW и синтетических эталонов качество распознавания на тех же аудиоданных повышается на 3 – 18% по сравнению с распознаванием методом скрытых марковских моделей, и на 9 – 14% по сравнению с системой Voco. Исследование эффективности метода дифонного распознавания изолированных слов на большом словаре показало высокую эффективность разработанного метода: доля корректно распознаваемых слов составляла не менее 90%.

Разработанные методы и алгоритмы могут быть использованы при создании систем компьютерного распознавания речевых образов.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в изданиях, включенных в перечень ВАК

1. Бурибаева А.К. Сегментация и дифонное распознавание речевых сигналов / А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко, В.Ю. Шелепов. // Труды СПИИРАН. – 2013. – № 31. – С. 20 – 42.

2. Шелепов В.Ю. О распознавании сверхбольших словарей русских словоформ с использованием квазиоснов / В.Ю. Шелепов, А.В. Ниценко // Известия ЮФУ. Технические науки. – 2016. – № 4. – С. 82 – 92.

Статьи, опубликованные в журналах, включенных в список индексируемых базой данных SCOUPUS

3. Shelepov V.Ju. Recognition of the continuous-speech Russian phrases using their voiceless fragments / V.Ju.Shelepov, A.V.Nicenko //Eurasian Journal of Mathematical and Computer Applications – 2016. – Vol. 4. – Iss.4. – P.19-24.

4. Nitsenko A.V. A «by part» method of Russian word speech recognition / A.V. Nitsenko // Eurasian Journal of Mathematical and Computer Applications. – 2014. – Vol.1, Iss. 2 – P. 102 – 109.

Свидетельство о регистрации авторских прав

5. Герасимов И.Г., Ниценко А.В., Азаренко Д.С., Шелепов В.Ю. Компьютерная программа «Голосовой калькулятор» // Свидетельство о регистрации авторского права на произведение № 58222. Украина. Министерство образования и науки. Государственный департамент интеллектуальной собственности. – Дата регистрации 22.01.2015.

Прочие публикации

6. Шелепов В.Ю. К проблеме пофонемного распознавания / В.Ю.Шелепов, А.В. Ниценко // Искусственный интеллект. – 2005. – №4. – С.662 – 668.

7. Шелепов В.Ю. Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала, распознавания некоторых классов фонем / В.Ю.Шелепов, А.В. Ниценко // Искусственный интеллект. – 2007. – № 1. – С. 213 – 224.

8. Шелепов В.Ю. О распознавании фонем с помощью анализа речевого сигнала в частотной и временной областях. Приложение к распознаванию синтаксически связных фраз / В.Ю. Шелепов, А.В. Ниценко, А.В. Жук, Д.С.Азаренко // Речевые технологии. –2008. – №2.– С. 43 – 52.

9. Бекманова Г.Т. О некоторых вопросах, связанных с распознаванием казахской речи / Г.Т. Бекманова, А.В. Ниценко, А.А. Шарипбаев, В.Ю. Шелепов // Вестник Евразийского национального университета им. Л. Н. Гумилева. – Астана, 2009. – № 6 – С. 172 – 177.

10. Шелепов В.Ю. Построение системы голосового управления компьютером на примере задачи набора математических формул / В.Ю. Шелепов, А.В. Ниценко, А.В. Жук // Искусственный интеллект. – 2010. – № 4. – С.259 – 267.

11. Шелепов В.Ю. Новый подход к определению границ речевого сигнала. Проблемы конца сигнала / В.Ю. Шелепов, А.В. Ниценко // Речевые технологии– 2012.–№1 – С. 74 – 78.

12. Шелепов В.Ю. О распознавании речи на основе межфонемных переходов / В.Ю. Шелепов, Г.В. Дорохина, А.В. Ниценко // Искусственный интеллект. –2012. – №1 – С.132 – 139.

13. Шелепов В.Ю. К проблеме распознавания слитной речи / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – 2012. – №4 – С.272 – 281.

14. Шелепов В.Ю. О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – 2013. – №3 – С. 209 – 216.

15. Ниценко А.В. Сегментация и дифонное распознавание речевых сигналов / А.В. Ниценко, В.Ю. Шелепов // Материалы Международной молодежной научной школы «Системы и средства искусственного интеллекта ССИИ – 2013». – 2013. – С.163 – 166.

16. В.Ю. Шелепов. О возможностях алгоритма DTW при распознавании речевых сигналов / В.Ю. Шелепов, А.В. Ниценко // Информатика и кибернетика. – 2017. – №2(8). – С. 73-82.

17. А.В. Ниценко. Метод коррекции фонемной сегментации речи / А.В. Ниценко // Проблемы искусственного интеллекта. – 2017. – №1(4). – С.43-48.

18. В.Ю. Шелепов. Распознавание русских слитно произносимых фраз с некоторыми специальными словарями / В.Ю. Шелепов, А.В. Ниценко // Проблемы искусственного интеллекта. – 2017. – №2(5). – С. 27-31.

В совместных научных публикациях имеет место неделимое соавторство.

Соискатель Ниценко А.В.