БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УДК 004.934.1

ТКАЧЕНЯ
Андрей Владимирович

РАСПОЗНАВАНИЕ ЭМОЦИОНАЛЬНОЙ СЛИТНОЙ РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук
по специальности 05.13.01 - Системный анализ, управление и
обработка информации

Минск 2016

Работа выполнена в Белорусском государственном университете.

Научный руководитель - **Хейдоров Игорь Эдуардович,**
кандидат физико-математических наук, доцент, доцент кафедры радиофизики и цифровых медиатехнологий Белорусского государственного университета.

Официальные оппоненты: **Лобанов Борис Мефодьевич,**
доктор технических наук, профессор, главный научный сотрудник лаборатории распознавания и синтеза речи ГНУ «Объединенный институт проблем информатики НАН Беларуси»;

Скакун Виктор Васильевич,
кандидат физико-математических наук, доцент, заведующий кафедрой системного анализа и компьютерного моделирования Белорусского государственного университета.

Оппонирующая организация - **УО «Белорусский государственный университет информатики и радиоэлектроники».**

Защита состоится 10 июня 2016 г. в 14:00 на заседании совета по защите диссертаций Д 02.01.14 при Белорусском государственном университете по адресу: 220030, г. Минск, ул. Ленинградская, 8 (корпус юридического факультета), ауд. 407; телефон ученого секретаря 209-57-09.

С диссертацией можно ознакомиться в Фундаментальной библиотеке Белорусского государственного университета.

Автореферат разослан «_6_» мая 2016 г.

Ученый секретарь
совета по защите диссертаций
кандидат физико-математических наук
доцент Ю.И. Воротницкий

ВВЕДЕНИЕ

В условиях бурного развития информационных технологий в современном мире постоянно растет актуальность проблемы быстрого взаимодействия человека и компьютера посредством речи. В связи с этим продолжают интенсивно развиваться и совершенствоваться методы распознавания речи, и одной из наиболее актуальных задач является разработка систем распознавания спонтанной речи, для которой характерно слитное и эмоциональное произношение.

Выбор скрытых марковских моделей (далее – СММ) для решения задачи распознавания слитной речи обусловлен тем, что она позволяет создавать гибкую акустическую модель. СММ применяется для моделирования последовательности векторов признаков как кусочно-стационарного процесса, в котором каждый стационарный участок относится к определенному состоянию СММ. Такой подход обеспечивает моделирование динамической структуры речевой единицы, то есть решает проблему различной длительности сигнала, соответствующего одной и той же фонеме. Использование в качестве речевой единицы связанных трифонов позволяет хорошо моделировать слитную речь, а также помогает избежать нехватки обучающих данных и дает возможность добавлять новые слова, составляя их из имеющихся акустических моделей без необходимости переобучать СММ.

Речевой сигнал помимо лингвистической составляющей содержит дополнительную информацию о дикторе: возраст, пол, социальный статус, здоровье, эмоциональное состояние и др. Различия в этой дополнительной информации приводят к изменению акустических характеристик речи. В работах было показано, что точность систем распознавания речи снижается для всех эмоциональных состояний по сравнению с таковой для нейтральной речи.

Для увеличения точности распознавания эмоциональной речи в литературных источниках было предложено несколько подходов, которые можно разделить на 3 категории: использование инвариантных информативных признаков (далее – ИП), методы компенсации эмоций в ИП, методы адаптации моделей. Два последних метода требуют предварительного анализа базы эмоциональной речи, который необходим для моделирования статистических данных о каждой из эмоций с последующим их включением в систему распознавания речи. Недостатком этих двух методов является необходимость определения эмоций в распознаваемой речи для применения соответствующей компенсации эмоции или адаптированной модели. Таким образом, на точность распознавания речи будет сильно влиять надежность классификации эмоций.

Более общее решение проблемы распознавания эмоциональной речи заключается в улучшении этапа параметризации сигнала. Этот подход заключается в создании вектора признаков (далее – ВП), пространство которого

не изменяется в зависимости от эмоциональной окраски речи диктора, что, в целом, с учетом ошибок классификации эмоций в речевом сигнале, должно приводить к большему увеличению точности распознавания речи, чем применение методов компенсации эмоций или адаптации моделей.

В диссертации разрабатывается система распознавания эмоциональной слитной речи на основе скрытых марковских моделей, дополненная методиками декодирования спонтанной речи при помощи триггерной сети спутывания и комбинированной верификации слов распознанной слитной речи, а также алгоритмом интерактивной неконтролируемой адаптации СММ к распознаваемой эмоциональной слитной речи с механизмом обновления. Полученная система позволяет повысить точность распознавания слитной речи в целом, а разработанный алгоритм адаптации СММ наряду с использованием инвариантного к эмоциям ВП приводят к росту точности распознавания эмоциональной речи. При этом ВП предлагается формировать на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот для спектра, рассчитанного по параметрам линейного предсказания.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами и темами

Тема диссертации соответствует направлению 5 «Информационно-коммуникационные, авиационные и космические технологии и аппаратура», пункту 5.4 «математические и интеллектуальные методы, информационные технологии и системы распознавания и обработки образов, сигналов, речи и мультимедийной информации» Перечня приоритетных направлений фундаментальных и прикладных научных исследований Республики Беларусь на 2011–2015 годы, утверждённого Постановлением Совета Министров Республики Беларусь №585 от 19.04.2010.

Тема диссертации также подпадает под направление 7 «Информационно-коммуникационные и авиакосмические технологии», пункт 7.6 «технологии развития информационного общества» Приоритетных направлений научно-технической деятельности в Республике Беларусь на 2016–2020 годы, утверждённого Указом Президента Республики Беларусь № 166 от 22.04.2015.

Диссертационная работа выполнена в рамках:

- ГПНИ «Информатика и космос, научные основы и инструментальные средства информационных и космических технологий» на 2011-2015 годы, задание - 1.4.04 «Разработка новых методов акустического и семантического анализа и распознавания аудио сигналов», № гос. регистрации 20115604;

- ГПНИ «Информатика и космос, научное обеспечение безопасности и защиты от чрезвычайных ситуаций» на 2014-2015 годы, задание - 1.39 «Разработка методов анализа и распознавания эмоционально окрашенных звуковых сигналов для мониторинга акустической обстановки», № гос. регистрации 20142127.

Цель и задачи исследования

Целью работы является разработка алгоритмов, методик и программного комплекса, позволяющих повысить точность распознавания эмоциональной слитной речи на основе скрытых марковских моделей.

Для достижения поставленной цели требуется решить следующие задачи:

- изучить влияние эмоциональной окраски речи на характеристики речевого сигнала для задач классификации эмоций и распознавания речи;
- разработать методику формирования инвариантного к эмоциям ВП, позволяющего повысить точность распознавания эмоциональной речи;
- разработать методику генерации языковой модели эмоциональной слитной речи на основе сети спутывания и триггерной языковой модели;
- разработать методику верификации слов при помощи метода динамической трансформации шкалы времени, повышающую точность распознавания слитной речи;
- разработать алгоритм адаптации СММ к эмоционально окрашенной речи, позволяющий проводить интерактивное уточнение параметров СММ;
- осуществить программную реализацию предложенной системы распознавания эмоциональной слитной речи на основе СММ.

Объектом исследования является процесс распознавания эмоциональной слитной речи. Предмет исследования – математическая модель, методики и алгоритмы повышения точности распознавания эмоциональной слитной речи, а также принципы программной реализации распознавания эмоциональной слитной речи на основе скрытых марковских моделей.

Научная новизна

Научная новизна разработанной системы распознавания эмоциональной слитной речи на основе СММ состоит в том, что в ней предложены новые методики и алгоритм (формирование инвариантного к эмоциям вектора признаков речевого сигнала, декодирование спонтанной речи при помощи триггерной сети спутывания, комбинированная верификация слов распознанной слитной речи и интерактивная неконтролируемая адаптация СММ с механизмом обновления), которые позволяют обеспечить точность распознавания слитной эмоционально окрашенной речи в среднем на 14,5 % выше по сравнению с известными системами.

Положения, выносимые на защиту

1. Методика декодирования спонтанной речи на основе сети спутывания и триггерной языковой модели, позволяющая, за счет определения оптимального размера окна поиска и порога средней взаимосвязи между словами, выбирать репрезентативный набор триггерных пар и эффективно уточнять на их основе апостериорную вероятность распознанного слова.
2. Методика комбинированной верификации слов распознанной слитной речи, объединяющая статистический подход к распознаванию речи на основе скрытых марковских моделей (СММ) и динамический подход на основе алгоритма динамической трансформации шкалы времени, позволяющая повысить точность распознавания слитной речи.
3. Алгоритм интерактивной неконтролируемой адаптации СММ с механизмом обновления, который за счет формирования коэффициента доверия препятствует уточнению параметров СММ на ложно распознанных словах эмоциональной слитной речи.
4. Система распознавания эмоциональной слитной речи на основе скрытых марковских моделей, обеспечивающая повышение точности распознавания в среднем на 14,5 % по сравнению с известными системами.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Хейдорова И.Э. связан с постановкой цели и задач исследования. В публикациях с соавторами вклад соискателя определяется рамками излагаемых в диссертации результатов.

Апробация результатов диссертации и информация об использовании ее результатов

Результаты, полученные в ходе выполнения исследований, докладывались и обсуждались на: 10th International conference on pattern recognition and information processing, PRIP 2009 (Минск, Беларусь, 2009); 17th International conference on conceptual structures for extracting natural language semantics, SENSE’09 (Москва, Россия, 2009); международной научно-практической конференции «Информационные технологии, электронные приборы и системы» ITEDS 2010 (Минск, Беларусь, 2010); 2-ой международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» OSTIS-2012 (Минск, Беларусь, 2012); международной конференции «Диалог» Диалог-2012 (Бекасово, Россия, 2012); научной конференции студентов и аспирантов БГУ (Минск, Беларусь, 2009-2010).

Результаты диссертационной работы внедрены компанией ООО «Речевые технологии» в программном продукте «Система голосовой биометрии», компанией ООО «ИТТАС» в программном приложении «Менеджер сертификатов itKeyMng», компанией ЧУП «Сакрамент ИТ» в системе распознавания речи «Sakrament ASR Engine» и используются в учебном процессе кафедры радиофизики и цифровых медиатехнологий БГУ. Разработанный в ходе выполнения диссертации «Программно-методический комплекс распознавания эмоциональной слитной речи на основе скрытых марковских моделей» включен в Государственный регистр информационных ресурсов (регистрационное свидетельство № 1201505539 от 22.09.2015 г.).

Опубликование результатов диссертации

Основные научные результаты диссертации опубликованы в 19 научных работах, из которых:

9 статей в научных изданиях в соответствии с п. 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь (общим объемом 3,8 авторского листа);

3 статьи в российском научном журнале «Речевые технологии»;

7 статей в сборниках материалов научных конференций.

Структура и объем диссертации

Диссертационная работа состоит из перечня условных обозначений, введения, общей характеристики работы, четырех глав, заключения, библиографического списка и приложений. Полный объем работы составляет 128 страниц. Она содержит 20 рисунков на 10 страницах, 21 таблицу на 7 страницах и 3 приложения на 20 страницах. Библиографический список состоит из 215 наименований, включая собственные публикации автора.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Первая глава посвящена анализу современных подходов к распознаванию эмоциональной речи и систем классификации эмоций в речи. Эмоциональная речь характеризуется изменением пространства информативных признаков по сравнению с нейтральной речью в соответствии с выражаемой эмоцией. Данные изменения прослеживаются на длительном промежутке времени и носят квазистационарный характер. Этим обусловлена возможность классификации эмоционального состояния диктора по голосу, в то же время изменение пространства информативных признаков приводит к значительному снижению точности распознавания речи. Чтобы увеличить точность распознавания эмоциональной речи, в литературных источниках

рассматривается несколько подходов, которые можно разделить на 3 категории: использование инвариантных ИП, компенсация эмоций в ИП и адаптация моделей. Подход на основе использования инвариантных ИП не требует предварительного анализа распознаваемой речи, в свою очередь для применения подхода компенсации эмоций в ИП необходимо провести классификацию эмоций и сегментацию речевого сигнала на группы, подход на основе адаптации моделей требует только классификацию эмоций по голосу. Таким образом, на эффективность использования подходов компенсации эмоций в ИП и адаптации моделей влияет надежность предварительного анализа речевого сигнала. Средняя точность современных классификаторов эмоций в речи составляет 59,9 %, что ставит под сомнение эффективность их использования для предварительного анализа речевого сигнала в системах распознавания эмоциональной речи.

В современных работах, посвященных распознаванию эмоциональной речи, большое внимание уделяется методикам параметризации речевого сигнала и акустическим моделям. Тем не менее, ряд работ свидетельствует о важности языкового моделирования эмоциональной речи и его влиянии на точность распознавания. Также при разработке системы распознавании эмоциональной слитной речи следует уделить внимание дополнительным методикам, которые позволяют увеличить точность распознавания речи. Необходимость этого обусловлена тем, что точность, достигаемая за счет акустического и языкового моделирования, в большинстве случаев будет недостаточна для решения прикладных задач. Приводимые в литературных источниках данные показывают, что в настоящее время значение точности распознавания эмоциональной слитной речи лежит в пределах от 35 % до 65 %, что сильно уступает качеству распознавания нейтральной слитной речи, для которой указанная величина составляет 88 %.

Во **второй главе** реализован и исследован ряд классификаторов эмоций, наилучшим из которых оказался классификатор на основе метода опорных векторов и критерия Джини в качестве функции расстояния для снижения количества информативных признаков [3]. При этом средняя точность классификации составляет 83 %, что соответствует максимальным результатам, полученным на базе Emo-DB, для современных систем голосового распознавания эмоций. При этом разработанный классификатор не требует больших вычислительных ресурсов, что является весомым аргументом для его использования в подходах компенсации эмоций в ИП и адаптации моделей. Тем не менее, при анализе спонтанных эмоциональных высказываний точность классификации не превышает 70 %, что ставит под сомнение эффективность использования подходов компенсации эмоций в ИП и адаптации моделей для распознавания эмоциональной слитной речи, точность распознавания которой

будет напрямую зависеть от надежности классификации эмоций. Это свидетельствует о том, что на данный момент наиболее перспективным, в плане научных изысканий в сфере распознавания эмоциональной слитной речи, является подход на основе использования инвариантных ИП.

Исследование информативных признаков показало, что использование инвариантных к шуму кепстральных коэффициентов на основе быстрого преобразования Фурье, приводит к снижению точности распознавания эмоциональной речи. Это обусловлено тем, что для эмоциональной речи характерны изменения частоты основного тона голоса, которые приводят к смазыванию спектра, полученного на основе преобразования Фурье. В свою очередь, преимуществом линейного предсказания является то, что изменение частоты основного тона не вызывает смазывание спектра, что обуславливает высокую точность распознавания гласных звуков, а его недостаток заключается в отсутствии нулей в спектре, что приводит к спутыванию схожих согласных. Дальнейшие исследования показали, что использование экспоненциально-логарифмической шкалы частот, наряду с применением фильтрации ВП и кепстрального взвешивания, а также использованием информации о динамике речевого сигнала, позволяют повысить точность распознавания эмоциональной речи. В диссертации разработана методика формирования инвариантного к эмоциям вектора признаков на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот для спектра, рассчитанного по параметрам линейного предсказания [9] (далее - ЛПСКК), размерность его пространства признаков равна 39 (логарифм энергии фрейма + 12 кепстральных коэффициентов + 13 первых и 13 вторых производных).

Используя в качестве вектора признаков мел-частотные кепстральные коэффициенты (далее - МЧКК) и предложенный в диссертации ЛПСКК, а также систему распознавания речи на основе СММ со связанными трифонами и равновероятной биграммной языковой моделью, было исследовано влияние типа речевых сигналов, из которых состоит обучающая выборка, на точность распознавания речи. Полученные результаты (таблица 1) приводятся для случая распознавания только нейтральной «Нейтр.» или только эмоциональной «Эмоц.» речи. Эксперименты осуществлялись на базах эмоциональной слитной речи японского MULTEXT и русского ЭМО-РУСС языков.

Как видно из таблицы 1, точность распознавания нейтральной речи для МЧКК выше, чем для ЛПСКК, в то время как для эмоциональной речи - наоборот. Это обусловлено свойствами линейного предсказания, применение которого позволяет добиться увеличения точности распознавания гласных звуков с различной эмоциональной окраской, но, с другой стороны, приводит к спутыванию схожих согласных, что сказывается на точности распознавания нейтральной речи.

Таблица 1. - Зависимость точности распознавания слитной речи для ЛПСКК и МЧКК от типа речевых сигналов в обучающей выборке

Обучающая выборка	Вектор признаков			
	ЛПСКК		МЧКК	
	Нейтр., %	Эмоц., %	Нейтр., %	Эмоц., %
	MULTEXT			
Только нейтральная речь (Нейтр.)	71,5±1,0	44,3±1,9	**86,3**±0,8	13,8±3,1
Только эмоциональная речь (Эмоц.)	50,4±1,8	**58,2**±1,2	52,2±1,7	26,9±1,9
Нейтральная и эмоциональная речь 1 : 1	62,5±1,3	52,8±1,5	75,4±0,9	21,6±2,3
	ЭМО-РУСС			
Только нейтральная речь (Нейтр.)	70,7±1,0	43,5±2,0	**86,1**±0,9	14,6±3,0
Только эмоциональная речь (Эмоц.)	49,2±1,9	**56,1**±1,3	52,9±1,7	27,5±1,8
Нейтральная и эмоциональная речь 1 : 1	61,6±1,3	49,7±1,6	75,7±0,9	22,6±2,2

Эмоциональная слитная речь характеризуется большой акустической изменчивостью, особенно в местах слияния слов, что не позволяет добиться высокой точности распознавания речи только за счет акустического моделирования. Точность базовой системы распознавания речи на основе СММ равна 58,2 и 56,1 % для MULTEXT и ЭМО-РУСС соответственно (таблица 1), что свидетельствует о необходимости использования методик, повышающих точность распознавания эмоциональной слитной речи, таких как верификация результатов распознавания речи и адаптация СММ, а также языковой модели, которая учитывает лингвистические особенности эмоциональной речи.

В **третьей главе** рассмотрены методы генерации языковых моделей и декодирования речи на их основе. Для спонтанной речи характерна высокая степень корреляции отдельных слов, поскольку, чем сильнее слова связаны между собой, тем больше семантической информации они несут. Информацию о контекстных взаимосвязях между словами можно получить посредством анализа транскрипции обучающих данных или специализированных текстовых баз данных. В большинстве современных систем распознавания речи используются *n*-граммные языковые модели, способные задавать контекстную связь в рамках окна из *n* слов. На текущий момент наиболее часто используемое *n* для спонтанной речи равно двум, так называемые биграммные языковые модели, задающие контекстную связь между двумя стоящими рядом словами. Однако большое количество контекстных взаимосвязей встречается даже за пределом окна из 3 слов (триграммная языковая модель).

При этом создание *n*-граммной языковой модели в пределах окна из 10 слов, влечет за собой большие вычислительные затраты и приводит к генерации громоздких языковых моделей, которые занимают много места и значительно замедляют процесс декодирования речи. В этом случае, чтобы определить контекстную связь между словами для задачи декодирования спонтанной речи можно применить триггерную модель, которая, обладая преимуществами

n-граммных языковых моделей, лишена их недостатков. Проведенные исследования показывают, что наилучшие результаты декодирования достигаются при использовании сети спутывания, полученной из графа слов на основе алгоритма выравнивания по опорным точкам Тура, в которую была интегрирована триггерная языковая модель. В диссертации предложен алгоритм создания репрезентативного набора триггерных пар, состоящий в оптимизации порога минимальной взаимосвязи между словами и размера окна поиска для зависимой от расстояния между словами триггерной языковой модели.

Апостериорная вероятность для распознанного слова в сети спутывания с триггерной языковой моделью равна сумме акустической P_a и триггерной P_T апостериорной вероятности соответственно. Это обусловлено тем, что $P_T\left(w_k^m \mid O\right)$ задает только изменение вероятности появления слова w_k^m.

$$\hat{P}_{CN}\left(w_k^m \mid O\right) = P_a\left(w_k^m \mid O\right) + P_T\left(w_k^m \mid O\right).$$

Так как апостериорная вероятность P_{CN} всех слов в *k*-ом множестве спутывания $w_k^{M_k}$ в сумме должна быть равна единице, необходимо провести нормирование полученной величины $\hat{P}_{CN}$. Таким образом, базовая система распознавания речи на основе СММ дополняется разработанной методикой декодирования спонтанной речи при помощи триггерной сети спутывания [7]. Результаты экспериментов по оценке зависимости точности распознавания эмоциональной слитной речи от количества используемых в языковой модели триггерных пар для баз MULTEXT и ЭМО-РУСС приведены в таблице 2. Отметим, что случай для нуля триггерных пар, соответствует точности распознавания эмоциональной слитной речи без декодирования спонтанной речи при помощи триггерной сети спутывания.

Таблица 2. - Зависимость точности распознавания эмоциональной слитной речи на основе ЛПСКК от количества триггерных пар в языковой модели

Кол-во триггерных пар, шт.	Точность распознавания речи, %	
	MULTEXT	ЭМО-РУСС
0	58,2±1,3	56,1±1,3
150	58,5±1,3	56,3±1,3
300	61,2±1,2	58,9±1,3
500	63,1±1,2	60,7±1,2
750	**65,0**±1,1	**62,2**±1,2
1000	64,9±1,1	62,3±1,2
1500	64,1±1,2	61,6±1,3

Как видно из таблицы 2, существует предельное значение количества триггерных пар, превышение которого приводит к незначительному росту, а порой и к снижению точности распознавания речи, так как начинают

учитываться несущественные пары слов. Использование предложенной методики декодирования спонтанной речи при помощи триггерной сети спутывания приводит к увеличению точности распознавания эмоциональной слитной речи на 6,8 и 6,1 % для базы MULTEXT и ЭМО-РУСС соответственно.

В **четвертой главе** разработаны методики, обеспечивающие дальнейшее повышение точности распознавания эмоциональной слитной речи. Так, увеличить точность системы распознавания речи можно за счет верификации слов распознанной слитной речи. Для решения данной задачи хорошо себя зарекомендовали подходы на основе динамической трансформации шкалы времени (далее - ДТВ) [4].

Предложенный алгоритм верификации слов на основе ДТВ заключается в сравнении анализируемого слова (слово-кандидат) со всеми соответствующими ему образцами произношения (набор слов-образцов) из обучающей выборки [6]. Окончательный результат верификации получается посредством выбора медианного результата промежуточных сравнений с каждым из слов-образцов, что позволяет исключить чрезмерную зависимость полученного результата от конкретного слова-образца. Главное достоинство предложенного алгоритма верификации слов состоит в нормировании на длительность слова-кандидата, что позволяет уменьшить вычислительные затраты на определение величины сходства между анализируемыми словами.

Уточненная апостериорная вероятность распознанного слова определяется как линейная интерполяция апостериорной вероятности для триггерной сети спутывания P_{CN} и верификации на основе алгоритма ДТВ P_{DTW}:

$$P_k^m = \alpha P_{DTW} + (1-\alpha) P_{CN} \ .$$

Уточненная таким образом вероятность позволяет увеличить точность распознавания эмоциональной слитной речи. Это обусловлено тем, что P_{CN} (на основе СММ) хорошо описывает статистику акустических характеристик речи, тем самым, являясь менее восприимчивой к их вариации в анализируемых данных, в то же время приводит к плохому распознаванию схожих слов. Напротив, P_{DTW} (на основе алгоритма ДТВ) хорошо отражает изменения акустических характеристик в анализируемых данных, что дает возможность эффективно различать схожие слова, но порождает ошибки распознавания при вариации акустических характеристик речи.

В ходе экспериментов получена зависимость точности распознавания эмоциональной слитной речи от величины весового коэффициента α. Как видно из графика (рисунок 1), максимальная точность распознавания эмоциональной слитной речи достигается при величине весового коэффициента α равной 0,65 и 0,58 для MULTEXT и ЭМО-РУСС соответственно.

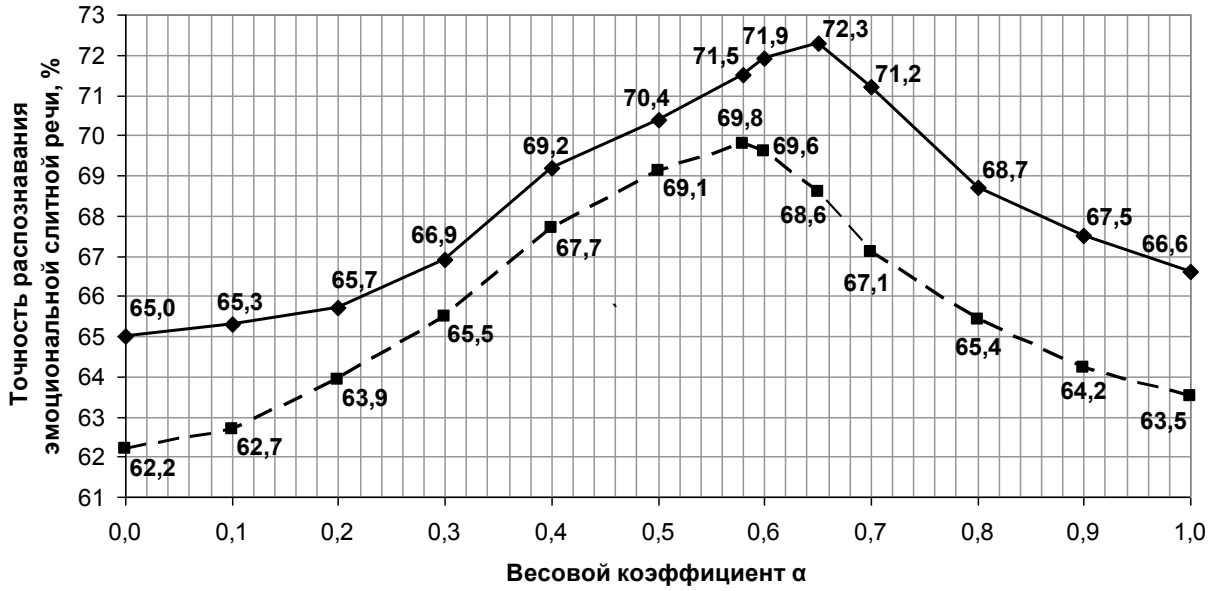

Рисунок 1. - Зависимость точности распознавания эмоциональной слитной речи от величины весового коэффициента α для баз MULTEXT (сплошная линия) и ЭМО-РУСС (пунктир)

При дальнейшем увеличении коэффициента α падение точности распознавания речи обусловлено ошибками, возникающими при определении границ слова-кандидата, и изменчивостью акустических характеристик эмоциональной слитной речи. Использование разработанной методики комбинированной верификации слов распознанной слитной речи [6] позволяет увеличить точность распознавания эмоциональной слитной речи на 7,3 и 7,6 % для базы MULTEXT и ЭМО-РУСС соответственно. Оценка доверительного интервала для случая максимальной точности распознавания равна 0,9 и 1,0 % для базы MULTEXT и ЭМО-РУСС соответственно. Так как транскрипция слов-образцов размечена вручную, полученная в ходе алгоритма ДТВ транскрипция по трифонам для слов-кандидатов может считаться вполне достоверной.

Полученные при верификации на основе алгоритма ДТВ транскрипции по трифонам для распознанных слов могут использоваться для адаптации СММ с гауссовым распределением значений параметров на основе квази-Байесовского обучения, алгоритм которого заключается в последовательном уточнении параметров скрытых марковских моделей. Для увеличения эффективности адаптации СММ предлагается использовать механизм забывания, а также разработанный механизм обновления [8], который за счет формирования коэффициента доверия препятствует уточнению параметров СММ на ложно распознанных словах эмоциональной слитной речи. Разработанный алгоритм адаптации СММ с механизмом обновления [8] позволяет проводить интерактивное неконтролируемое уточнение параметров СММ на тестовой выборке. Механизм обновления [8] характеризуется коэффициентом $0 \le \psi \le 1$,

который задает степень доверия к новым входным данным, чтобы не допускать адаптацию СММ на ложно распознанных словах. Коэффициент ψ предлагается определять исходя из формулы:

$$\psi = \begin{cases} 0, & \text{при} \quad P_k < 2BL - 1 \\ \dfrac{P_k - BL}{1 - BL} + 1, & \text{при} \quad 2BL - 1 \le P_k < BL \\ 1, & \text{при} \quad P_k \ge BL \end{cases},$$

где BL – это среднее значение апостериорной вероятности распознанного слова для корректно распознанных слов. BL определяется при обучении системы и при анализе тестовой выборки никак не меняется.

Дополнение системы распознавания слитной речи алгоритмом интерактивной неконтролируемой адаптации СММ с механизмом обновления [8] обеспечивает увеличение точности распознавания речи для последующего речевого сигнала, так как он уже будет анализироваться на уточненных СММ.

В таблице 3 показана зависимость точности распознавания эмоциональной слитной речи для разработанной системы (рисунок 2) от количества распознанных слов, используемых для адаптации СММ. Отметим, что случай для нуля распознанных слов, соответствует точности распознавания эмоциональной слитной речи без адаптации СММ на данных из тестируемой выборки. Как видно из таблицы 3, максимальная точность распознавания речи равная 77,4 и 74,6 % достигается при анализе 30 слов, что соответствует 20 секундам эмоциональной слитной речи.

Таблица 3. - Зависимость точности распознавания эмоциональной слитной речи от количества распознанных слов для баз MULTEXT и ЭМО-РУСС

Количество распознанных слов, шт.	Точность распознавания эмоциональной слитной речи, %	
	MULTEXT	ЭМО-РУСС
0	72,3±0,9	69,8±1,0
5	72,5±0,9	69,9±1,0
10	73,4±1,0	70,6±1,2
15	75,7±1,3	73,3±1,4
20	76,8±1,4	74,2±1,6
30	**77,4**±1,5	**74,6**±1,7
40	77,4±1,6	74,5±1,8
50	77,3±1,6	74,5±1,9

Как видим, разработанный алгоритм интерактивной неконтролируемой адаптации СММ к распознаваемой эмоциональной слитной речи с механизмом обновления [8] позволяет повысить точность распознавания эмоциональной слитной речи на 5,1 и 4,8 % для базы MULTEXT и ЭМО-РУСС соответственно.

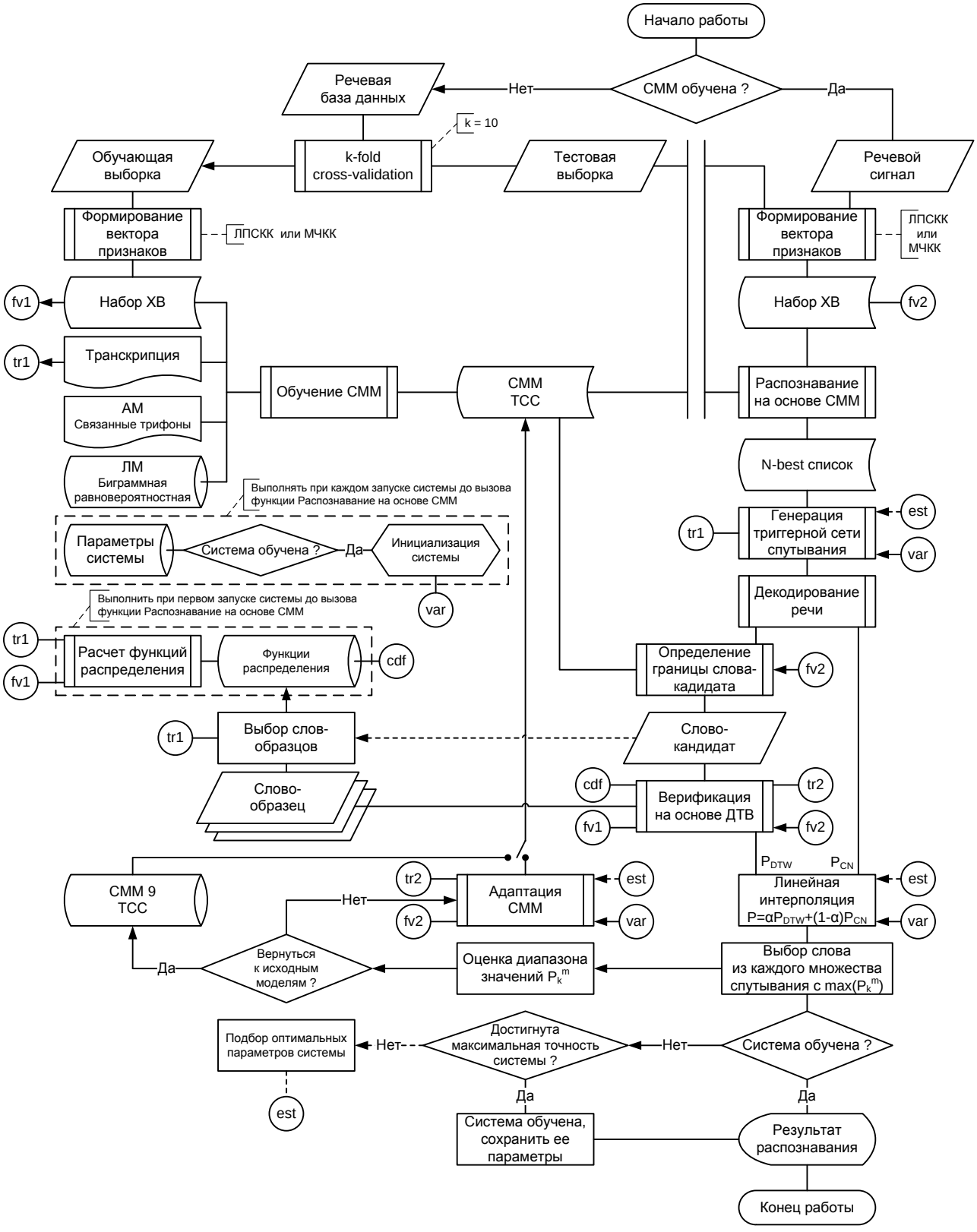

Рисунок 2. - Процедурная блок-схема системы распознавания эмоциональной слитной речи на основе СММ, включающая разработанные методики декодирования спонтанной речи при помощи триггерной сети спутывания и комбинированной верификации слов распознанной слитной речи, а также алгоритм интерактивной неконтролируемой адаптации СММ к распознаваемой эмоциональной слитной речи с механизмом обновления

Использование разработанных методик декодирования спонтанной речи [7] и комбинированной верификации слов [6], а также алгоритма интерактивной неконтролируемой адаптации СММ [8] позволяет повысить точность распознавания эмоциональной слитной речи на 19,2 % (с 58,2 % до 77,4 %) и на 18,5 % (с 56,1 % до 74,6 %) для базы MULTEXT и ЭМО-РУСС соответственно, что является важным практическим результатом.

На основе разработанной в ходе выполнения диссертации системы (рисунок 2) реализован программно-методический комплекс распознавания эмоциональной слитной речи на основе СММ (регистрационное свидетельство № 1201505539 от 22.09.2015 года). Программный комплекс создан в свободной кроссплатформенной среде разработки Code::Blocks версии 13.12 при использовании компилятора MinGW/GCC версии 4.7.1 на языке программирования C++. Для работы программно-методического комплекса персональный компьютер должен удовлетворять следующим техническим характеристикам: двухъядерный процессор Intel Core2 Duo @ 2.00GHz и 2 Гб оперативной памяти. Полученный программно-методический комплекс был интегрирован в пользовательский графический интерфейс «EmotionalASR», разработанный при помощи кроссплатформенной библиотеки Qt версии 5.3.1.

В ходе проведенных экспериментов, результаты которых представлены в таблице 4, было показано, что интернет сервис Google ASR, коммерческий программный продукт Phonexia LVCSR и библиотека функций с открытым исходным кодом CMU Sphinx имеют более низкую точность распознавания эмоциональной слитной речи.

Таблица 4. - Точность распознавания эмоциональной слитной речи для различных систем

Система	Точность распознавания речи, %	
	MULTEXT	ЭМО-РУСС
EmotionalASR (вектор признаков - ЛПСКК)	**77,4**±1,5	**74,6**±1,7
Hybrid ANN/HMM ASR with DEA	72,9±1,0	70,7±1,3
Google ASR	-	57,2±3,2
Phonexia LVCSR	-	59,8±2,9
CMU Sphinx	-	52,6±2,6

В свою очередь, система Hybrid ANN/HMM ASR with DEA показала неплохие результаты, что свидетельствует об эффективности использования гибридных алгоритмов распознавания речи, а также подхода адаптации моделей, на точность которого влияет надежность классификации эмоций в речевом сигнале. В целом, эксперименты на русскоязычной базе ЭМО-РУСС показывают, что точность распознавания эмоциональной слитной речи для

разработанной в диссертации системы в среднем на 14,5 % выше по сравнению с известными системами (таблица 4).

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Разработан классификатор эмоций на основе метода опорных векторов и критерия Джини в качестве функции расстояния для снижения количества информативных признаков [3, 11, 19], средняя точность которого составляет 83 %, что сопоставимо с лучшими известными классификаторами; при этом его вычислительные затраты на распознавание эмоций ниже.

2. Разработана методика формирования инвариантного к эмоциям вектора признаков на основе кепстральных коэффициентов [5, 12], определенных на экспоненциально-логарифмической шкале частот для спектра, рассчитанного по параметрам линейного предсказания [9]. Точность распознавания эмоциональной слитной речи для русскоязычной базы ЭМО-РУСС и предложенного вектора признаков в среднем на 6,8 % выше по сравнению с известными методиками параметризации речевого сигнала.

3. Разработана методика декодирования спонтанной речи на основе сети спутывания и триггерной языковой модели [7], позволяющая выбирать репрезентативный набор триггерных пар и уточнять на их основе апостериорную вероятность распознанного слова в сети спутывания. Экспериментально установлено, что триггерная языковая модель позволяет повысить точность декодирования спонтанной речи на основе сети спутывания [1, 2, 10, 14].

4. Разработана методика комбинированной верификации слов распознанной слитной речи на основе совместного использования статистического и динамического подходов к распознаванию речи [6]. Достоинством предложенной методики является определение границ слова в слитной речи и сравнение слов на основе алгоритма ДТВ с нормированием на длительность анализируемого слова [4, 12], что позволяет уменьшить вычислительные затраты на определение величины сходства между словами.

5. Разработан алгоритм интерактивной неконтролируемой адаптации СММ с механизмом обновления [8], который позволяет эффективно уточнять параметры скрытых марковских моделей на распознанных данных при помощи алгоритма Баума - Велша, адаптируя модели к эмоционально окрашенной речи и индивидуальным особенностям голоса диктора.

6. Реализована система распознавания эмоциональной слитной речи на основе скрытых марковских моделей, включающая разработанные методики декодирования спонтанной речи при помощи триггерной сети спутывания [7] и

комбинированной верификации слов распознанной слитной речи [6], а также алгоритм интерактивной неконтролируемой адаптации СММ к распознаваемой эмоциональной слитной речи с механизмом обновления [8]. Экспериментально установлено, что рост точности распознавания эмоциональной слитной речи обеспечивается за счет разработанных методик и алгоритма; точность распознавания эмоциональной слитной речи для русскоязычной базы ЭМО-РУСС в среднем на 14,5 % выше по сравнению с известными системами.

Рекомендации по практическому использованию результатов

Результаты, полученные в ходе выполнения диссертации, позволяют создавать эффективные программные системы распознавания эмоциональной слитной речи, на основе которых можно решать прикладные задачи.

Результаты диссертационной работы внедрены и используются:

- компанией ООО «Речевые технологии», являющейся резидентом Парка высоких технологий Республики Беларусь, в программном продукте «Система голосовой биометрии»;
- компанией ООО «ИТТАС» в программном приложении «Менеджер сертификатов itKeyMng»;
- компанией ЧУП «Сакрамент ИТ», являющейся резидентом Парка высоких технологий Республики Беларусь, в системе распознавания речи «Sakrament ASR Engine»;
- в учебном процессе кафедры радиофизики и цифровых медиатехнологий БГУ в лабораторном практикуме по специальному курсу «Статистическая радиофизика» в программе «EmotionalASR»;
- разработанный в ходе выполнения диссертации «Программно-методический комплекс распознавания эмоциональной слитной речи на основе скрытых марковских моделей» включен в Государственный регистр информационных ресурсов (регистрационное свидетельство № 1201505539 от 22.09.2015 г.).

Перспектива дальнейшего увеличения точности распознавания эмоциональной слитной речи заключается в развитии гибридных алгоритмов распознавания речи на основе скрытых марковских моделей и нейронных сетей с применением генетических алгоритмов, метода опорных векторов и алгоритма динамической трансформации шкалы времени.

Результаты диссертационной работы могут быть использованы:

- для создания средств речевого управления – программ, управляющих действиями компьютера или другого электронного устройства с помощью голосовых команд;
- для распознавания речи людей в стрессовых ситуациях, для которых характерно сильное изменение акустических характеристик голоса.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Статьи в научных изданиях в соответствии с п. 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь

1. Поиск ключевых слов в слитной речи на основе усовершенствованной меры достоверности / Цзинбинь Янь, Ши У, А. В. Ткаченя, И. Э. Хейдоров // Вестн. Белорус. гос. ун-та. Сер. 1, Физика. Математика. Информатика. – 2009. – № 3. – С. 44–48.
2. Янь, Цзинбинь. Исследование алгоритма поиска ключевых слов на основе сети спутывания / Цзинбинь Янь, А. М. Сорока, А. В. Ткаченя // Электроника Инфо. – 2009. – № 9 (69). – С. 70–73.
3. Классификация эмоционального состояния диктора с использованием метода опорных векторов и критерия Джини / А. В. Ткаченя, А. Г. Давыдов, В. В. Киселёв, М. В. Хитров // Изв. вузов. Приборостроение. – 2013. – Т. 56, № 2. – С. 61–66.
4. Киселёв, В. В. Алгоритм сравнения фонограмм в обучающих системах / В. В. Киселёв, А. Г. Давыдов, А. В. Ткаченя // Информатика. – 2013. – № 4 (40). – С. 30–35.
5. Киселёв, В. В. Разработка каналонезависимых информативных признаков / В. В. Киселёв, А. В. Ткаченя, М. В. Хитров // Изв. вузов. Приборостроение. – 2014. –Т. 57, № 2. – С. 12–17.
6. Ткаченя, А. В. Верификации результатов распознавания эмоциональной слитной речи / А. В. Ткаченя // Электроника Инфо. – 2014. – № 7 (109). – С. 32–34.
7. Ткаченя, А. В. Декодирование речи на основе триггерной сети спутывания / А. В. Ткаченя // Электроника Инфо. – 2014. – № 8 (110). – С. 20–23.
8. Ткаченя, А. В. Адаптация скрытых марковских моделей к распознаванию эмоционально окрашенной речи / А. В. Ткаченя // Информатика. – 2014. – № 3 (43). – С. 21–27.
9. Ткаченя, А. В. К вопросу об устойчивых к эмоциям информативных признаках для задачи распознавания речи / А. В. Ткаченя // Вестн. Белорус. гос. ун-та. Сер. 1, Физика. Математика. Информатика. – 2014. – № 3. – С. 56–61.

Статьи в других научных журналах

10. Янь, Цзинбинь. Исследование характеристик системы поиска ключевых слов на основе минимального интервала редактирования и мер доверительности / Цзинбинь Янь, И. Э. Хейдоров, А. В. Ткаченя // Речевые технологии. – 2008. – № 4. – С. 5–14.
11. Автоматическое распознавание эмоций по речи с использованием

метода опорных векторов и критерия Джина / М. В. Хитров, А. Г. Давыдов, В. В. Киселёв, Ю. Н. Ромашкин, А. В. Ткаченя // Речевые технологии. – 2012. – № 4. – С. 34–43.

12. Киселёв, В. В. Алгоритм сравнения фонограмм на основе каналонезависимых информативных признаков / В. В. Киселёв, А. В. Ткаченя // Речевые технологии. – 2013. – № 3–4. – С. 3–11.

Статьи в сборниках материалов научных конференций

13. Speech segmentation to phonemes based on hybrid hidden Markov models / Jingbin Yan, Shi Wu, I. E. Kheidorov, A. V. Tkachenia // Pattern Recognition and Information Processing (PRIP)'2009 : Proceedings of the 10th International Conference (19–21 May, 2009, Minsk, Belarus). – Minsk : Publ. center of BSU, 2009. – P. 192–194.

14. Tkachenia, A. V. Robust keyword search using subword lattice / A. V. Tkachenia, Jingbin Yan, A. A. Trus // Workshop at the 17th International Conference on Conceptual Structures (ICCS'09) : Proceedings of conceptual Structures for Extracting Natural language SEmantics (SENSE)'09 (26 July, 2009, Moscow, Russia). – Moscow : the Higher School of Economics, 2009. – P. 101–107.

15. Ткаченя, А. В. Разработка методов поиска ключевых слов в слитной речи на основе скрытых марковских моделей / А. В. Ткаченя // Сборник работ 66-й научной конференции студентов и аспирантов Белорусского государственного университета, 18–21 мая 2009 г., Минск. В 3 ч. / Белорус. гос. ун-т. – Минск : Изд. центр БГУ, 2010. – Ч. 1. – С. 187–191.

16. Tkachenia, A. V. Phoneme Segmentation System Based on Hybrid Hidden Markov Models and Wavelet Transforms / A. V. Tkachenia, A. M. Soroka // Информационные технологии, электронные приборы и системы (ITEDS'2010) : Материалы Международной научно-практической конференции, 6–7 апреля 2010 г., Минск / Белорусский государственный университет. – Минск : Национальная библиотека Беларуси, 2010. – С. 138–145.

17. Ткаченя, А. В. Разработка системы поиска ключевых слов в слитной речи на основе скрытых марковских моделей / А. В. Ткаченя // Сборник работ 67-й научной конференции студентов и аспирантов Белорусского государственного университета, 17–20 мая 2010 г., Минск. В 3 ч. / Белорус. гос. ун-т. – Минск : Изд. центр БГУ, 2011. – Ч. 1. – С. 159–162.

18. Киселёв, В. В. Система определения эмоционального состояния диктора по голосу / В. В. Киселёв, А. Г. Давыдов, А. В. Ткаченя // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2012) : материалы 2 Междунар. науч.-техн. конф. (Минск, 16–18 февраля 2012 г.) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2012. – С. 355–358.

19. Выбор оптимального набора информативных признаков для классификации эмоционального состояния диктора по голосу / А. Г. Давыдов, В. В. Киселёв, Д. С. Кочетков, А. В. Ткаченя // Компьютерная лингвистика и интеллектуальные технологии : По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). – М.: Изд-во РГГУ, 2012. – Вып. 11 (18), т. 1: Основная программа конференции. – С. 122–128.

РЭЗЮМЭ

Ткачэня Андрэй Уладзіміравіч

Распазнаванне эмацыйнай злітным гаворкі
на аснове схаваных маркаўскіх мадэляў

Ключавыя словы: маўленчы сігнал, эмацыйная гаворка, распазнанне злучнай гаворкі, схаваныя маркаўскія мадэлі (СММ), вектар прыкмет, дэкадаванне гаворкі, верыфікацыя слоў, адаптацыя СММ.

Мэта работы: распрацоўка алгарытмаў, методык і праграмнага комплексу, якія дазваляюць павысіць дакладнасць распазнання эмацыйнай злучнай гаворкі на аснове СММ.

Метады даследавання: цыфравая апрацоўка сігналаў, тэорыя акустыкі і псіхаакустыкі, матэматычнае мадэляванне, тэорыя верагоднасцяў і матэматычнай статыстыкі.

Атрыманыя вынікі і іх навізна. У дысертацыі прыведзены аналіз змены прасторы інфарматыўных прыкмет для эмацыйнай гаворкі і яго ўплыў на зніжэнне дакладнасці распазнання гаворкі. Распрацавана методыка фарміравання інварыянтнага да эмоцый вектару прыкмет на аснове кепстральных каэфіцыентаў, вызначаных на экспанентна-лагарыфмічнай шкале частот для спектру, вылічанага па параметрах лінейнага прадказання.

Распрацаваны методыкі дэкадавання спантаннай гаворкі на аснове сеткі зблытвання і трыггернай моўнай мадэлі, камбінаванай верыфікацыі слоў распазнанай злітнай гаворкі на аснове сумеснага выкарыстання статыстычных і дынамічных падыходаў да распазнання гаворкі, а таксама алгарытма інтэрактыўнай некантралюемай адаптацыі СММ да распазнаваемай эмацыйнай злітнай гаворкі з механізмам абнаўлення.

Створаная сістэма распазнання эмацыйнай злітнай гаворкі на аснове схаваных маркаўскіх мадэляў забяспечвае павышэнне дакладнасці распазнання ў сярэднім на 14,5 % у параўнанні з вядомымі сістэмамі.

Галіна ўжывання і рэкамендацыі па выкарыстанні. Атрыманыя вынікі маюць як тэарэтычнае значэнне, так і практычную накіраванасць. Распрацаваная сістэма распазнання эмацыйнай злітнай гаворкі можа выкарыстоўвацца ў складзе інтэрфейсаў ўзаемадзеяння з дапамогай гаворкі паміж чалавекам і кампутарам для распазнання эмацыйных частак маўленчага сігналу, сегментаваных у спантаннай гаворкі пры дапамозе дэтэктару эмоцый.

РЕЗЮМЕ

Ткаченя Андрей Владимирович

Распознавание эмоциональной слитной речи
на основе скрытых марковских моделей

Ключевые слова: речевой сигнал, эмоциональная речь, распознавание слитной речи, скрытые марковские модели (СММ), вектор признаков, декодирование речи, верификация слов, адаптация СММ.

Цель работы: разработка алгоритмов, методик и программного комплекса, позволяющих повысить точность распознавания эмоциональной слитной речи на основе скрытых марковских моделей.

Методы исследования: цифровая обработка сигналов, теория акустики и психоакустики, математическое моделирование, теория вероятностей и математической статистики.

Полученные результаты и их новизна. В диссертации приведен анализ изменения пространства информативных признаков для эмоциональной речи и его влияние на снижение точности распознавания речи. Разработана методика формирования инвариантного к эмоциям вектора признаков на основе кепстральных коэффициентов, определенных на экспоненциально-логарифмической шкале частот для спектра, рассчитанного по параметрам линейного предсказания.

Разработаны методики декодирования спонтанной речи на основе сети спутывания и триггерной языковой модели, комбинированной верификации слов распознанной слитной речи на основе совместного использования статистических и динамических подходов к распознаванию речи, а также алгоритм интерактивной неконтролируемой адаптации СММ к распознаваемой эмоциональной слитной речи с механизмом обновления.

Реализована система распознавания эмоциональной слитной речи на основе СММ, обеспечивающая повышение точности распознавания в среднем на 14,5 % по сравнению с известными системами.

Рекомендации по использованию и область применения. Полученные результаты имеют как теоретическое значение, так и практическую направленность. Разработанная система распознавания эмоциональной слитной речи может быть использована в составе интерфейсов взаимодействия посредством речи между человеком и компьютером для распознавания эмоциональных участков речевого сигнала, сегментированных в спонтанной речи при помощи детектора эмоций.

SUMMARY

Tkachenia Andrei Vladimirovich
Emotional continuous speech recognition
based on hidden Markov models

Keywords: speech signals, emotional speech, continuous speech recognition, hidden Markov models (HMM), feature vector, speech decoding, word verification, HMM adaptation.

The purpose of research: development of algorithms, methods and software package, which increase the accuracy of emotional continuous speech recognition based on HMM.

Research methods: digital signal processing, theory of acoustics and psychoacoustics, mathematical modeling, probability theory, mathematical statistics.

The obtained results and scientific novelty. The research provides an analysis of changes in the space of informative features for emotional speech and its impact on reducing the performance of speech recognition. The method of emotional invariant feature vector parameterization based on cepstral coefficients which are obtained on the basis of a linear prediction power spectrum defined on an ExpoLog frequency scale was released.

Methods of spontaneous speech decoding based on confusion network with trigger language model, composite verifying the words of recognized continuous speech based on the combining of statistical and dynamic speech recognition approaches, also on-line unsupervised adaptive learning of HMM to recognizing emotional continuous speech with updating mechanism were developed.

It was shown that the constructed emotional continuous speech recognition system based on hidden Markov models provides accuracy increasing by an average of 14,5 % compared with known up-to-date systems.

Application field and usage recommendations. The obtained results have both theoretical and practical. The developed emotional continuous speech recognition system can be used as a part of interaction speech interface between human and computer for recognizing emotional parts of speech signal segmented by emotion detector from spontaneous speech.