Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

УДК 004.032.6; 004.383.3

ЗАХАРЬЕВ
Вадим Анатольевич

КОНВЕРСИЯ ГОЛОСА НА ОСНОВЕ МОДЕЛИ ГАУССОВЫХ СМЕСЕЙ В СИСТЕМАХ СИНТЕЗА РЕЧИ ПО ТЕКСТУ С НАСТРОЙКОЙ НА ГОЛОС ДИКТОРА

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

по специальности 05.13.17 – Теоретические основы информатики

Минск 2017

Работа выполнена в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники».

Научный руководитель	**Петровский Александр Александрович**, доктор технических наук, профессор, заведующий кафедрой электронных вычислительных средств учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»
Официальные оппоненты:	**Мурашко Игорь Александрович**, доктор технических наук, доцент, профессор кафедры информационных технологий учреждения образования «Гомельский государственный технический университет имени П.О. Сухого»
	Хейдоров Игорь Эдуардович, кандидат физико-математических наук, доцент, заведующий кафедрой радиофизики и цифровых медиа технологий Белорусского государственного университета
Оппонирующая организация	Учреждение образования «Белорусский государственный технологический университет»

Защита состоится 21 июня 2017 г. в 14.00 на заседании совета по защите диссертаций Д 02.15.04 при учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, г. Минск, ул. П. Бровки, 6, корп. 1, ауд. 232, тел. 293-89-89, e-mail: dissovet@bsuir.by.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

Автореферат разослан « » мая 2017 г.

Ученый секретарь
совета по защите диссертаций,
кандидат технических наук, доцент П. Ю. Бранцевич

КРАТКОЕ ВВЕДЕНИЕ

Разговорная речь является одной из наиболее естественных и эффективных форм передачи информации между людьми. Этот факт объясняет значительный интерес исследователей и инженеров к вопросам разработки речевых интерфейсов, а также их широкое распространение для обеспечения человеко-машинного взаимодействия в составе современных коммуникационных, мультимедийных и интеллектуальных систем. Блок вывода речевой информации в подобных системах зачастую реализован в виде синтезатора речи по тексту (СРТ).

На данном этапе развития речевых технологий ставится вопрос не только о создании СРТ с высокими уровнями разборчивости и натуральности синтезируемой речи, но и их адаптации под различные сценарии применения в случае конкретного пользователя. В связи с наметившейся устойчивой тенденцией к персонализации устройств и программ возникает потребность в создании мультиголосовых синтезаторов речи по тексту (МГСРТ), обладающих возможностью настройки на голос определенного целевого диктора. В существующих решениях данная функциональность достигается путем подготовки и смены речевых баз данных (БД) для каждого нового диктора либо же изменением характеристик голоса в речевом сигнале (РС) таким образом, что он становится не похож на первоначальный.

Особый интерес в настоящее время представляет подход к разработке МГСРТ, основанный на использовании технологии конверсии голоса, позволяющий реализовать процесс настройки на голос целевого диктора на базе имеющихся в синтезаторе акустических ресурсов и сравнительно небольшого количества обучающих данных. В контексте данного подхода значимыми являются направления исследований, касающиеся совершенствования способов синтеза речи (параметрические или компиляционные), которые позволяют осуществлять мультиголосовую генерацию речевого сигнала с максимально приближенными к естественным характеристиками голоса диктора. Следовательно, весьма актуальны вопросы, связанные с разработкой методов конверсии голоса, развитием способов обучения системы (текстозависимый или текстонезависимый), а также задача поиска архитектуры, позволяющей интегрировать блоки, реализующие процесс конверсии голоса в состав мультиголосового синтезатора речи по тесту.

Таким образом, решение данных вопросов в совокупности позволит создать систему мультиголосового синтеза речи по тексту с настройкой на голос диктора, характеризующуюся высокими значениями параметров разборчивости, натуральности и узнаваемости синтезируемой речи, а также возможностью обучения системы в текстонезависимом режиме.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами и темами

Диссертационная работа выполнена в соответствии с научно-техническими заданиями и планами работ кафедры «Электронные вычислительные средства», научно-исследовательской лаборатории 3.1 «Мультипроцессорные системы реального времени» учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» и проводилась в соответствии с государственными научными темами, выполненными в рамках бюджетного финансирования Республики Беларусь:

1. Конверсия голоса для синтезатора речи по тексту (№ 12-1099 Б) / Белорусский республиканский ФФИ, Министерство образования Республики Беларусь; рук. д-р техн. наук, проф. Б. М. Лобанов. – Минск, 2012. – № Ф12ОБ-053. – 01.07.2012 – 30.06.2014 гг.

2. Разработка алгоритмического и аппаратно-программного обеспечения обработки мультимедиа данных; рук. д-р техн. наук, проф. А. А. Петровский. – Минск, 2011. – ГБ № 11-2008. – 01.01.2011 – 31.12.2016 гг.

3. Разработка универсального аудиокодера на основе разреженной аппроксимации с оптимизированным словарем частотно-временных функций в качестве встраиваемой системы реального времени / Грант Министерства образования для аспирантов; рук. д-р техн. наук, проф. А. А. Петровский. – Минск, 2016. – ГБЦ № 16-3137. – 01.09.2016 – 31.12.2016 гг.

Цель и задачи исследования

Целью диссертационной работы является разработка методов конверсии голоса на основе модели гауссовых смесей для систем синтеза речи по тексту с настройкой на голос диктора.

Поставленная цель определяет следующие задачи исследования:

1. Выполнить анализ способов синтеза речи и методов настройки СРТ на голос диктора для создания МГСРТ.

2. Предложить архитектурные решения мультиголосового синтезатора речи по тексту, позволяющие включить в состав акустического процессора модули конверсии голоса, учесть особенности этапов обработки речевой информации согласно выбранным способам синтеза речи и конверсии голоса, а также форму представления лингвистических ресурсов СРТ в виде, наиболее подходящем для реализации алгоритмов конверсии и генерации речевого сигнала.

3. Разработать метод конверсии голоса, учитывающий дикторозависимые характеристики голоса в речевом сигнале.

4. Разработать метод настройки системы на голос целевого диктора, позволяющий осуществлять текстонезависимое обучение при ограниченном наборе используемых данных.

5. Реализовать программную систему мультиголосового синтеза речи по тексту, характеризующуюся высокими значениями параметров разборчивости, натуральности и узнаваемости синтезируемой речи.

6. Провести экспериментальные исследования и оценить эффективность предлагаемых решений, методов и алгоритмов.

Научная новизна

1. Интегрированная архитектура системы мультиголосового синтеза речи по тексту, которая позволяет включить модуль конверсии голоса в состав компиляционного синтезатора речи по тексту на уровне акустического процессора системы и использовать имеющиеся лингвистические ресурсы, содержащиеся в базе данных акустических фрагментов речевого сигнала синтезатора (в параметризированном виде), в процессе текстонезависимого обучения модуля конверсии голоса и синтеза речи. Предлагаемое решение позволяет уменьшить количество ошибок при конверсии тембральных и просодических характеристик голоса диктора на 11 и 13 % по сравнению с каскадной архитектурой.

2. Метод конверсии голоса на основе модели гауссовых смесей, позволяющий выполнить кластеризацию пространства векторов параметров, характеризующих голос диктора, и осуществить их трансформацию с использованием регрессионной функции конверсии специального вида, который позволил увеличить значение коэффициента детерминации статистической модели на 0,1 , а также оценки узнаваемости и естественности согласно шкале оценки мнений (*MOS*) на 0,45 и 0,34 единиц соответственно.

3. Алгоритм поиска параметров функции конверсии на основе алгоритма максимизации функции правдоподобия статистической модели (*EM*-алгоритма) и использования информационного критерия Акаике (учитывающего кроме значения функции правдоподобия еще и размерность модели в качестве штрафного коэффициента), который позволил осуществлять определение необходимого количества компонент смеси непосредственно в процессе обучения системы.

4. Метод текстонезависимого обучения системы конверсии голоса на базе скрытых марковских моделей лево-правого типа с непрерывным множеством алфавита наблюдений и определением параметров модели на основе модифицированного алгоритма Витерби, который предусматривает возможность перехода к текстонезависимому способу обучения МГСРТ с относительной величиной ошибки обучения, равной 15 % , по сравнению с текстозависимым вариантом.

Положения, выносимые на защиту

1. Интегрированная архитектура системы мультиголосового синтеза речи по тексту, включающая модуль конверсии голоса на уровне акустического процессора и параметризированную базу данных акустических фрагментов речевого сигнала компиляционного синтезатора речи по тексту.

2. Метод конверсии голоса на основе модели гауссовых смесей и регрессионной функции отображения специального вида.

3. Алгоритм поиска параметров функции конверсии на основе алгоритма максимизации значения логарифмической функции правдоподобия модели (*EM*-алгоритма) и использования информационного критерия Акаике.

4. Метод текстонезависимого обучения системы конверсии голоса на базе скрытых марковских моделей лево-правого типа с непрерывным множеством алфавита наблюдений и модифицированного алгоритма Витерби.

Личный вклад соискателя ученой степени

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя доктора технических наук, профессора А. А. Петровского связан с постановкой целей и задач исследований, определением возможных путей решения и обсуждением результатов исследований, проводимых автором. В публикациях с соавторами вклад соискателя определяется рамками излагаемых в диссертации результатов.

Апробация диссертации и информация об использовании ее результатов

Основные результаты диссертационной работы докладывались и обсуждались на 17 международных и республиканских научных конференциях: Международной научной конференции «Информационные технологии и системы» (*ITS*) – Минск, Беларусь, 2011, 2012, 2013, 2014, 2015, 2016; 12th International Conference «Pattern Recognition and Information Processing» (*PRIP*) – Minsk, Belarus, 2014; Inter International Conference «Automatic Processing of Natural-Language Electronic Texts» (*NOOJ*) –Minsk, Belarus, 2015; 47-й научной конференции аспирантов, магистрантов и студентов БГУИР – Минск, Беларусь, 2011; Международной научно-технической конференции «Наука образованию, производству, экономике – Минск, Беларусь, 2013; Международной научно-технической конференции, приуроченной к 50-летию МРТИ–БГУИР – Минск, Беларусь, 2014; Международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (*OSTIS*) – Минск, Беларусь, 2014, 2015, 2017; Международной

научно-технической конференции «Цифровая обработка сигналов и ее применение» (*DSPA*) – Москва, Россия, 2013, 2015; 9-й научной конференции «Информационные технологии в управлении» (*ITU*) – Санкт-Петербург, Россия, 2016.

Опубликование результатов диссертации

По материалам диссертации опубликованы 22 печатные работы, в том числе 5 статей в рецензируемых научных журналах, 14 статей в сборниках материалов научных конференций и симпозиумов, 3 тезиса докладов научных конференций. Результаты диссертационной работы включены в 2 отчета по НИР.

Общий объем публикаций по теме диссертации, соответствующий пункту 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь, составляет около 7,5 авторских листов.

Структура и объем диссертации

Диссертационная работа состоит из введения, общей характеристики работы, пяти глав, заключения, библиографического списка и четырех приложений. Общий объем диссертационной работы составляет 218 страниц, из них 102 страницы основного текста, 42 рисунка на 39 страницах, 39 таблиц на 33 страницах, библиография из 228 наименований, включая 22 публикации автора, на 19 страницах и четыре приложения на 25 страницах.

ОСНОВНАЯ ЧАСТЬ

Во **введении** обоснована актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, определена область, основные направления, цель и задачи исследования.

Первая глава посвящена анализу существующих современных методов и структурных решений синтезаторов речи по тексту, систем конверсии голоса и подходов к их обучению, а также моделей представления речевого сигнала в контексте возможного их применения для построения МГСРТ. В настоящий момент времени большинство систем мультиголосового синтеза речи построено на основе принципа «смены речевых баз» при смене диктора. Главным его недостатком является необходимость длительной и трудоемкой подготовки речевых баз, количество которых должно быть пропорционально количеству добавляемых дикторов. В диссертации предлагается использовать альтернативный подход на основе конверсии голоса.

Конверсия голоса (КГ) – это техника цифровой обработки речевого сигнала, которая позволяет осуществлять преобразование параметров речевого сигнала, характеризующих голос исходного диктора (ИД), в параметры целевого диктора (ЦД).

На вход системы, реализующей функцию КГ, поступает речевое сообщение, озвученное голосом ИД, на выходе системы получается то же сообщение, но озвученное голосом ЦД. В процессе обработки осуществляется трансформация тембральных и просодических характеристик ИД в ЦД согласно заданной функции конверсии. Тембральные характеристики голоса проявляются в РС через параметры спектральных огибающих, характерные для конкретных ИД и ЦД. Просодические характеристики определяются через параметры контура частоты основного тона (ЧОТ).

Предложена архитектура системы МГСРТ, которая подразумевает полную интеграцию двух типов систем путем включения соответствующих блоков КГ в состав СРТ на уровне акустического процессора, что позволяет учесть особенности этапов обработки речевой информации согласно выбранным способам синтеза речи (компиляционный) и конверсии голоса (модель гауссовых смесей), уменьшив количество ошибок по сравнению с каскадной архитектурой системы (рисунок 1).

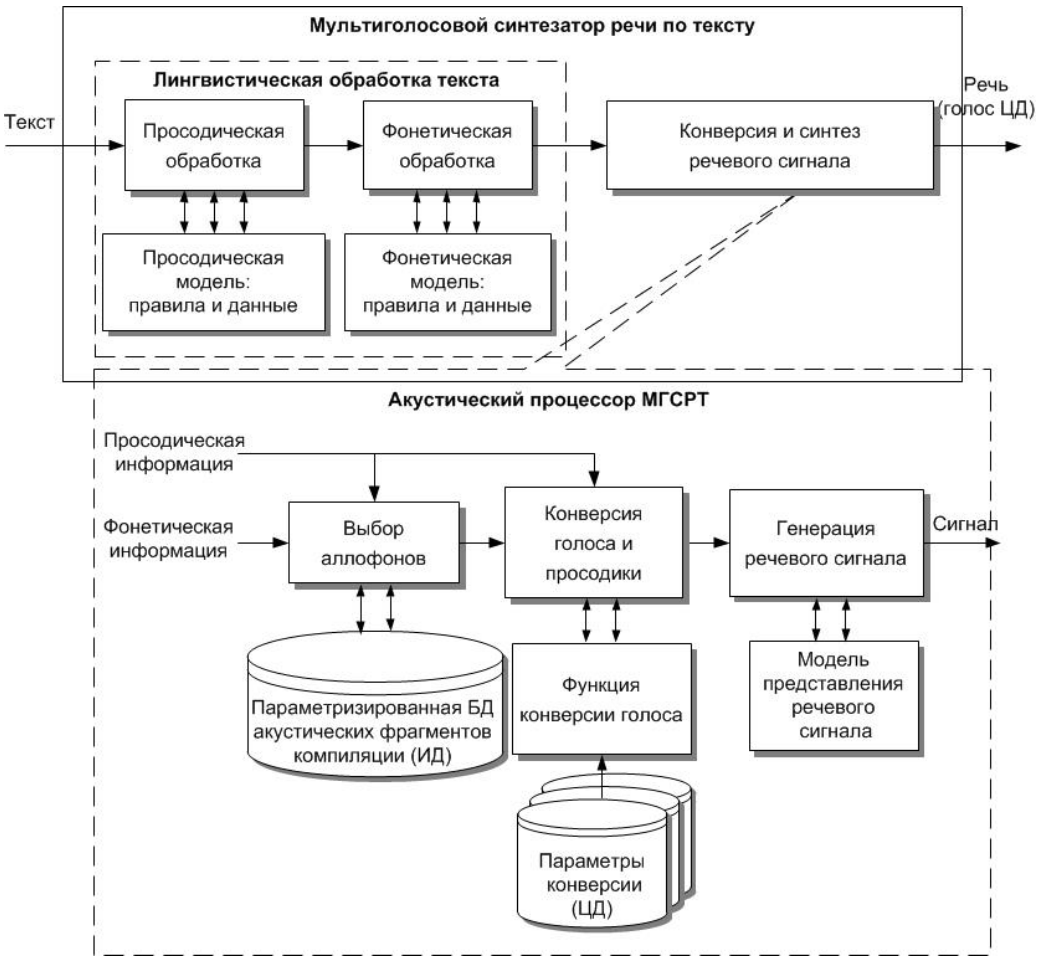


Рисунок 1. – Интегрированная архитектура МГСРТ

Модифицированный акустический процессор МГСРТ компиляционного типа в рамках интегрированной архитектуры включает блок выбора аллофонов, реализующий процедуру поиска элементов компиляции в параметризированной БД

аллофонов, блок конверсии голоса и просодики, выполняющий преобразование векторов параметров спектральных огибающих, и просодические модификации, связанные с изменением длительности элементов и контура ЧОТ, а также блок, осуществляющий окончательную генерацию речи на основе параметров выбранной модели представления речевого сигнала.

Особенность предлагаемой архитектуры МГСРТ компиляционного типа заключается в наличии параметризированной БД элементов компиляции. Каждый аллофон хранится не в виде записей отрезков речевой волны, а в виде набора параметров, определяемых выбранной моделью представления сигнала. В диссертационной работе предлагается использовать гибридную модель, позволяющую представить речевой сигнал в виде суммы периодической и апериодической составляющих. В качества инструмента анализа сигнала, реализующего идеи данной модели, был выбран метод представления и преобразования речи на основе адаптивной интерполяции взвешенного спектра (*Speech Transformation and Representation by Adaptive Interpolation of weighted Spectrum*), известный как *STRAIGHT*. Такой способ представления БД синтезатора на основе гибридной модели сигнала позволяет выполнять сложные манипуляции с единицами компиляции (например, изменение скорости частоты основного тона $f_0/\Delta t$ или длительности ΔT фрагментов компиляции в широком диапазоне значений), а также уменьшить амплитудные и фазовые искажения на стыках фрагментов.

Таким образом, при разработке интегрированной архитектуры МГСРТ были учтены следующие важные моменты: блок конверсии голоса функционирует внутри системы на уровне акустического процессора, этапы конверсии характеристик (тембральных и просодических) выполняются единым блоком, что позволяет изменять характеристики сигнала только один раз, а синтез акустических фрагментов по параметрам модели и их компиляция выполняются непосредственно после КГ, что позволяет сократить количество артефактов в результирующем РС.

Вторая глава посвящена исследованию существующих, а также разработке оригинального метода КГ для последующего применения в МГСРТ. В настоящий момент времени одним из самых распространенных и широко используемых подходов в КГ, доказавших свою эффективность, является КГ на основе модели гауссовых смесей (МГС). На этапе обучения МГСРТ, после того как выполнен анализ РС, а также проведена дополнительная параметризация коэффициентов модели, описывающих тембральные характеристики диктора, с целью уменьшения размерности, получаем две последовательности векторов параметров спектральных огибающих $\boldsymbol{X} = \{\boldsymbol{x}_i \mid \boldsymbol{x} \in \mathbb{R}^{d\times 1}\}$ исходного и $\boldsymbol{Y} = \{\boldsymbol{y}_i \mid \boldsymbol{y} \in \mathbb{R}^{d\times 1}\}$ целевого дикторов для каждого i-го фрейма сигнала, где $i = 1,2 \ldots T$; d – размерность вектора пара-

метров огибающей. Будем считать, что данные векторы расположены в пространстве параметров размерностью D и характеризуют особенности голоса конкретного диктора. В рамках МГС данное пространство векторов параметризиурется с помощью суммы плотностей вероятностей распределений Гаусса. Функция плотности вероятности МГС представляет собой взвешенную сумму Q гауссовых компонент и определяется следующим выражением:

$$p(\boldsymbol{x}) = \sum_{q=1}^{Q} \alpha_q \, N_q\left(\boldsymbol{x}\middle|\boldsymbol{\Theta}_q\right),$$

где $\boldsymbol{x} = [x_0, x_1, \dots x_{d-1}]^{\mathrm{T}}$– вектор параметров спектральной огибающей размерности d; $N_q\left(\boldsymbol{x}\middle|\boldsymbol{\Theta}_q\right)$ – плотность вероятности q-й компоненты смеси; α_q – весовой коэффициент q-й составляющей.

Каждая из компонент N_q представляет собой функцию плотности вероятности многомерного распределения Гаусса размерностью d:

$$N_q\left(\boldsymbol{x}\middle|\boldsymbol{\Theta}_q\right) = \frac{1}{(2\pi)^{p/2}\sqrt{\left|\boldsymbol{\Sigma}_q\right|}} \boldsymbol{e}^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_q})^T \boldsymbol{\Sigma}_q (\boldsymbol{x}-\boldsymbol{\mu_q})\right]},$$

где $\boldsymbol{\mu}_q$ – вектор математических ожиданий класса q размерностью d; $\boldsymbol{\Sigma}_q$–ковариационная матрица многомерного распределения Гаусса размерностью $d \times d$. Скалярные веса смесей α_q принимают значения больше нуля, $\alpha_q \geq 0, \forall q = 1, \dots, Q$, а их сумма равна единице: $\sum_{q=1}^{Q} \alpha_q = 1$. Таким образом, полное параметрическое представление МГС, описывающей пространство параметров диктора, включает в себя множество $\boldsymbol{\Theta} = \{\alpha_{\mathrm{q}}, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\}$ характеристик для $q = 1, \dots, Q$ компонент.

Центральной задачей конверсии голоса является поиск функции конверсии голоса, позволяющей выполнить оптимальное отображение вектора параметров исходного диктора на каждом i-м фрейме анализа в параметры целевого диктора. В качестве такого критерия оптимальности, как правило, выступает минимум расстояния между векторами в пространстве параметров. Для непосредственного выполнения процедуры отображения (конверсии векторов) в диссертационной работе за основу предлагается взять регрессионную функцию вида

$$F(\boldsymbol{x}) = \sum_{q=1}^{Q} h_q(\boldsymbol{x})[\boldsymbol{\nu}_q + \boldsymbol{\Gamma}_q \boldsymbol{\Sigma}_q^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_q)], \qquad (1)$$

где $h_q(\boldsymbol{x})$ – апостериорная вероятность того, что вектор $\boldsymbol{x}$ принадлежит q-й гауссовой компоненте. Параметры функции конверсии $\{\boldsymbol{\nu}_q, \boldsymbol{\Gamma}_q\}$ вычисляются с применением методов среднеквадратичной оптимизации с целью минимизации ошибки преобразования между сконвертированными и целевыми векторами:

$$E[\|\boldsymbol{y} - F(\boldsymbol{x})\|^2] \to \min.$$

Модель может быть расширена для случая совместного пространства параметров векторов исходного и целевого дикторов $\boldsymbol{z} = [\boldsymbol{x}^T, \boldsymbol{y}^T]^T$ путем использования в выражении (1) совместной плотности вероятности векторов $p(\boldsymbol{x}, \boldsymbol{y})$**.** В данном случае функция конверсии является зависимостью следующего вида:

$$F(\boldsymbol{x}) = \sum_{q=1}^{Q} h_q(\boldsymbol{x}) \, [\boldsymbol{\mu}_q^y + \boldsymbol{\Sigma}_q^{yx} \boldsymbol{\Sigma}_q^{xx^{-1}} (\boldsymbol{x} - \boldsymbol{\mu}_q^x)], \tag{2}$$

$$h_q(\boldsymbol{x}) = \frac{\alpha_q N(\boldsymbol{x} | \boldsymbol{\mu}_q^{xx}, \boldsymbol{\Sigma}_q^{xx})}{\sum_{j=1}^{Q} \alpha_j N(\boldsymbol{x} | \boldsymbol{\mu}_q^{xx}, \boldsymbol{\Sigma}_q^{xx})}, \boldsymbol{\Sigma}_q = \begin{bmatrix} \boldsymbol{\Sigma}_q^{xx} & \boldsymbol{\Sigma}_q^{xy} \\ \boldsymbol{\Sigma}_q^{yx} & \boldsymbol{\Sigma}_q^{yy} \end{bmatrix}, \quad \boldsymbol{\mu}_q = \begin{bmatrix} \boldsymbol{\mu}_q^x \\ \boldsymbol{\mu}_q^y \end{bmatrix},$$

где $\boldsymbol{\Sigma}_q$ и $\boldsymbol{\mu}_q$ – блочная ковариационная матрица и вектор математических ожиданий q-й компоненты смеси для совместной плотности вероятности $2d \times 2d$; $\boldsymbol{\Sigma}_q^{xy}$ –кросс-ковариационная матрица q-ой компоненты МГС для совместной плотности вероятности размерностью $d \times d$; $\boldsymbol{\Sigma}_q^{xx}$ и $\boldsymbol{\mu}_q^x$ –ковариационная матрица и вектор математических ожиданий q-й компоненты МГС ИД; $\boldsymbol{\Sigma}_q^{yy}$ и $\boldsymbol{\mu}_q^y$ – ковариационная матрица и вектор математических ожиданий q-й компоненты МГС ЦД.

Проблемой приведенных функций конверсии голоса (1) и (2) на основе МГС является то, что они в полной мере не учитывают наличия локальных особенностей в процессе изменения спеткральных огибающих во времени, поскольку в рамках данных подходов последовательность векторов обучения рассматривается как простой набор элементов, для которых корреляционные связи учитываются лишь для одной пары векторов на каждом i-м фрейме анализа. В результате чего увеличивается ошибка преобразования, что приводит к искажениям параметров огибающей на выходе функции конверсии, которые перцептуально воспринимаются как уменьшение сходства и ухудшение натуральности сконвертированной речи.

В диссертации предлагается развитие регрессионных моделей вида (1) и (2) за счет введения в функцию конверсии новых факторов, использующих контекстную информацию из соседних с i-м вектором параметров исходного и целевого дикторов. Поскольку последовательность векторов параметров речевого сигнала обладает свойствами марковского процесса, была выдвинута гипотеза о том, что параметры контекстных векторов также могут коррелировать с i-м вектором целевого диктора.

$$\widehat{\boldsymbol{y}}_i = F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{q=1}^{Q} h_q(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}, \boldsymbol{y}_{i-1}) \left[\boldsymbol{\nu}_q + \boldsymbol{\Phi}_q \overline{\boldsymbol{x}}_i^q + \boldsymbol{\Psi}_q \overline{\boldsymbol{y}}_{i-1}^q + \boldsymbol{\Omega}_q \overline{\boldsymbol{x}}_{i-1}^q \right], \tag{3}$$

где $\widehat{\boldsymbol{y}}_i$ – вектор параметров, полученный в результате конверсии на i-м фрейме речевого сигнала; $\overline{\boldsymbol{x}}_i^q$ и $\overline{\boldsymbol{x}}_{i-1}^q$ – векторы параметров, учитывающие ковариационную матрицу и математическое ожидание q-й компоненты МГС для i-го и $(i-1)$-го фрейма анализа для исходного диктора, такие, что $\overline{\boldsymbol{x}}_k^q = \boldsymbol{\Sigma}_q^{\boldsymbol{xx^{-1}}} (\boldsymbol{x}_k - \boldsymbol{\mu}_q^{\boldsymbol{x}})$, $k = \overline{i-1, i}$;

$\overline{\boldsymbol{y}}_{i-1}^{q} = \boldsymbol{\Sigma}_{q}^{\boldsymbol{yy^{-1}}}(\boldsymbol{y_{i-1}} - \boldsymbol{\mu}_{q}^{\boldsymbol{y}})$ – вектор параметров, учитывающий ковариационную матрицу q-й компоненты МГС для $(i-1)$-го фрейма ссигнала целевого диктора; $\overline{\boldsymbol{x}}_{i}^{q} = \boldsymbol{\Sigma}_{q}^{\boldsymbol{xx^{-1}}}(\boldsymbol{x_i} - \boldsymbol{\mu}_{q}^{\boldsymbol{x}})$ – вектор параметров q-й компоненты МГС ИД для i-го фрейма сигнала.

Для определения параметров регрессионной функции метод на основе совместной плотности вероятности напрямую применен быть не может. Однако при условии $E[\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2] \to 0$ показано, что осуществлять поиск коэффициентов данной модели возможно на основе метода наименьших квадратов. Тогда выражение (3) можно представить в матричном виде:

$$[\boldsymbol{P} \vdots \boldsymbol{B} \vdots \boldsymbol{C} \vdots \boldsymbol{D}\,] \cdot \begin{bmatrix} \boldsymbol{\nu} \\ \cdots \\ \boldsymbol{\Phi} \\ \cdots \\ \boldsymbol{\Psi} \\ \cdots \\ \boldsymbol{\Omega} \end{bmatrix} = \boldsymbol{y}, \tag{4}$$

где $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \vdots \cdots \vdots \boldsymbol{y_T}]^T$ – последовательность векторов параметров целевого диктора $\mathbf{y}_j \in \mathbb{R}^{1\times d}$; d – размерность вектора параметров; $\boldsymbol{\nu} = [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \vdots \cdots \vdots \boldsymbol{\nu}_Q\ ]^T$ – вектор математических ожиданий для каждой компоненты смеси, где $\boldsymbol{\nu}_j \in \mathbb{R}^{1\times d}$; $\boldsymbol{\Phi} = [\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}, \vdots \cdots \vdots \boldsymbol{\Phi_Q}\ ]^T$ – матрица регрессионных коэффициентов для всех компонент смеси при независимой переменной $\boldsymbol{x}_i$, где $\boldsymbol{\Phi}_j \in \mathbb{R}^{d\times d}$; $\boldsymbol{\Psi} = [\boldsymbol{\Psi_1}, \boldsymbol{\Psi_2}, \vdots \cdots \vdots \boldsymbol{\Psi_Q}\ ]^T$ – матрицы регрессионных коэффициентов при независимой переменной $\boldsymbol{y}_{i-1}$, $\boldsymbol{\Psi}_j \in \mathbb{R}^{d\times d}$; $\boldsymbol{\Omega} = [\boldsymbol{\Omega_1}, \boldsymbol{\Omega_2}, \vdots \cdots \vdots \boldsymbol{\Omega_Q}\ ]^T$ – матрицы регрессионных коэффициентов при независимой переменной $\boldsymbol{x}_{i+1}$, а $\boldsymbol{\Omega}_j \in \mathbb{R}^{d\times d}$. Матрицы $\{\boldsymbol{P}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\}$ размерностью $Q \times T-1$ представляют собой известные характеристики модели и определяются согласно следующим выражениям:

$$\boldsymbol{P} = \begin{bmatrix} p_1(1) & p_1(2) & \cdots & p_1(T-1) \\ p_2(1) & p_2(2) & \cdots & p_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1) & p_Q(2) & \cdots & p_Q(T-1) \end{bmatrix}^T,$$

$$\boldsymbol{B} = \begin{bmatrix} p_1(1)\boldsymbol{\Sigma}_1^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1^{\mathbf{x}}) & p_1(2)\boldsymbol{\Sigma}_1^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_1^{\boldsymbol{x}}) & \cdots & p_1(T-1)\boldsymbol{\Sigma}_1^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_{T-1} - \boldsymbol{\mu}_1^{\boldsymbol{x}}) \\ p_2(1)\boldsymbol{\Sigma}_2^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_2^{\boldsymbol{x}}) & p_2(2)\boldsymbol{\Sigma}_2^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2^{\boldsymbol{x}}) & \cdots & p_2(T-1)\boldsymbol{\Sigma}_2^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_{T-1} - \boldsymbol{\mu}_2^{\boldsymbol{x}}) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1)\boldsymbol{\Sigma}_Q^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_Q^{\boldsymbol{x}}) & p_Q(2)\boldsymbol{\Sigma}_Q^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_Q^{\boldsymbol{x}}) & \cdots & p_Q(T-1)\boldsymbol{\Sigma}_Q^{\boldsymbol{x^{-1}}}(\boldsymbol{x}_{T-1} - \boldsymbol{\mu}_Q^{\boldsymbol{x}}) \end{bmatrix}^T,$$

$$\boldsymbol{C} = \begin{bmatrix} p_1(1)\boldsymbol{\Sigma}_1^{y^{-1}}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1^y) & p_1(1)\boldsymbol{\Sigma}_1^{y^{-1}}(\boldsymbol{y}_2 - \boldsymbol{\mu}_1^y) & \cdots & p_1(T-1)\boldsymbol{\Sigma}_1^{y^{-1}}(\boldsymbol{y}_{T-1} - \boldsymbol{\mu}_1^y) \\ p_2(1)\boldsymbol{\Sigma}_2^{y^{-1}}(\boldsymbol{y}_1 - \boldsymbol{\mu}_2^y) & p_2(1)\boldsymbol{\Sigma}_2^{y^{-1}}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2^y) & \cdots & p_2(T-1)\boldsymbol{\Sigma}_2^{y^{-1}}(\boldsymbol{y}_{T-1} - \boldsymbol{\mu}_2^y) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1)\boldsymbol{\Sigma}_Q^{y^{-1}}(\boldsymbol{y}_1 - \boldsymbol{\mu}_Q^y) & p_Q(1)\boldsymbol{\Sigma}_Q^{y^{-1}}(\boldsymbol{y}_2 - \boldsymbol{\mu}_Q^y) & \cdots & p_Q(T-1)\boldsymbol{\Sigma}_Q^{y^{-1}}(\boldsymbol{y}_{T-1} - \boldsymbol{\mu}_Q^y) \end{bmatrix}^T,$$

$$\boldsymbol{D} = \begin{bmatrix} p_1(2)\boldsymbol{\Sigma}_1^{x^{-1}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1^x) & p_1(3)\boldsymbol{\Sigma}_1^{x^{-1}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_1^x) & \cdots & p_1(T)\boldsymbol{\Sigma}_1^{x^{-1}}(\boldsymbol{x}_T - \boldsymbol{\mu}_1^x) \\ p_2(2)\boldsymbol{\Sigma}_2^{x^{-1}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_2^x) & p_2(3)\boldsymbol{\Sigma}_2^{x^{-1}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2^x) & \cdots & p_2(T)\boldsymbol{\Sigma}_2^{x^{-1}}(\boldsymbol{x}_T - \boldsymbol{\mu}_2^x) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(2)\boldsymbol{\Sigma}_Q^{x^{-1}}(\boldsymbol{x}_1 - \boldsymbol{\mu}_Q^x) & p_Q(3)\boldsymbol{\Sigma}_Q^{x^{-1}}(\boldsymbol{x}_2 - \boldsymbol{\mu}_Q^x) & \cdots & p_Q(T)\boldsymbol{\Sigma}_Q^{x^{-1}}(\boldsymbol{x}_T - \boldsymbol{\mu}_Q^x) \end{bmatrix}^T,$$

Выполнив подстановку $\boldsymbol{A} = [\boldsymbol{P} \vdots \boldsymbol{B} \vdots \boldsymbol{C} \vdots \boldsymbol{D}\ ]$ и $\boldsymbol{\chi} = [\boldsymbol{\nu} \vdots \boldsymbol{\Phi} \vdots \boldsymbol{\Psi} \vdots \boldsymbol{\Omega}\ ]^T$, уравнение (4) можно привести к нормальной форме. Тогда процесс нахождения неизвестных параметров $\{\boldsymbol{\nu}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}\}$ формулируется как задача оптимизации. Для ее решения может быть использован метод наименьших квадратов. Представим выражение (4) в следующем виде:

$$\boldsymbol{A} \cdot \boldsymbol{\chi} = \boldsymbol{y} \Rightarrow\ \boldsymbol{A}^T\boldsymbol{A} \cdot \boldsymbol{\chi} = \boldsymbol{y}\boldsymbol{A}^T,$$

тогда решение ищется в виде

$$\boldsymbol{\chi_{opt}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{y}\boldsymbol{A}^T.$$

Сложности, которые могут возникнуть при решении данного уравнения связаны с возможной необходимостью инверсии плохо обусловленных матриц большой размерности, с тенденцией роста количества параметров системы. Общая размерность матрицы, требующей инверсии в правой части выражения, зависит от количества компонент смеси Q и размерности векторов параметров d и определяется как $(3 \times Q \times d + d)^2$. Решение данной проблемы возможно с использованием диагональных ковариационных матриц вместо их полных версий, а также с использованием декомпозиции на основе разложения Холецкого. При практической реализации данного метода был использован алгоритм решения уравнения на основе метода наименьших квадратов с применением методик, разработанных для систем линейных уравнений для матриц высокой размерности.

В **третьей главе** предложен метод текстонезависимого обучения МГСРТ. В процессе этапа обучения осуществляется настройка на голос целевого диктора путем определения параметров функции конверсии F . Метод включает в себя стадии анализа входной выборки обучающих сигналов, их временного «выравнивания» друг относительно друга, определения параметров функции конверсии. Обучение системы может быть текстозависимым и текстонезваисимым. В текстозависимом случае обучающие данные будут представлять собой базу фонограмм, полученных на основе одного и того же текстового корпуса (т. е. идентичных наборах фраз) для ИД

и ЦД. Недостатком данного подхода являются существенные трудозатраты на составление таких корпусов. Текстонезависимый подход наоборот подразумевает отсутствие заранее подготовленного общего текстового корпуса, по которому может быть записан набор обучающих фонограмм.

Метод текстонезависимого обучения в рамках интегрированной архитектуры МГСРТ, использующий скрытые марковские модели (СММ) и модифицированный алгоритм Витерби, представлен на рисунке 2.

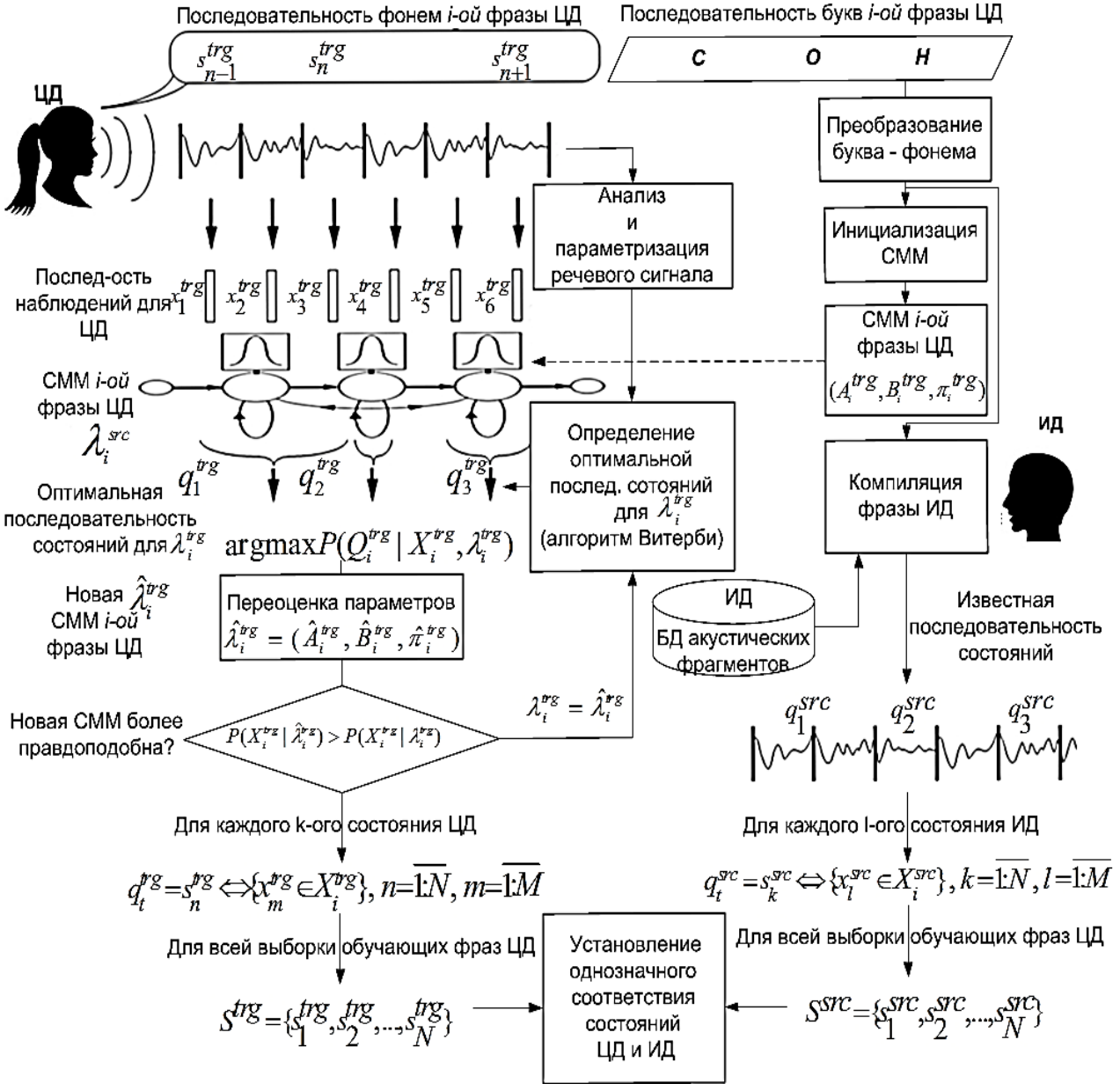


Рисунок 2. – Процесс текстонезависимого обучения

Данный метод основан на определении строгого соответствия фонемы набору векторов наблюдений. Необходимо найти оптимальную последовательность состояний с учетом всех возможных реализаций фонемы во фразах обучающей выборки. С

точки зрения СММ данный процесс заключается в том, чтобы связать оптимальную последовательность состояний с текущей последовательностью наблюдений для данной модели, что можно записать :

$$\begin{cases} \text{argmax}P(X^{src}, Q^{src}|\lambda^{src}), \ \lambda^{src} \in (A^{src}, B^{src}, \pi^{src}), \\ \text{argmax}P(X^{trg}, Q^{trg}|\lambda^{trg}), \ \lambda^{trg} \in (A^{trg}, B^{trg}, \pi^{trg}), \end{cases} \quad (5)$$

где X^{src} и X^{trg} – последовательности наблюдений СММ, которые соответствуют последовательностям векторов параметров спектральных огибающих для ИД и ЦД соответственно; Q^{src} и Q^{trg} – последовательности состояний СММ, которые соответствует появлению фонем в речевом сигнале фраз ИД и ЦД; λ^{src} и λ^{trg} – параметры СММ моделей для текущей фразы ИД и ЦД; $P(X^k, Q^k|\lambda^k)$ – вероятность соответствия текущей последовательности X^k наблюдений последовательности состояний Q^k для параметров модели λ^k для $k = \{src, trg\}$ ИД и ЦД соответственно.

Решение задачи (5) возможно с помощью использования итерационного алгоритма Витерби. Далее путем объединения статистик всех наблюдений по каждому из состояний для одного диктора по всем фразам обучающей выборки возможно найти соответствие векторов параметров, относящихся к определенной фонеме, благодаря равенству фонемных алфавитов с точки зрения состава его символов $S^{trg} = S^{src}$. Это позволяет сформировать совместную последовательность векторов параметров сигнала по фонетическому принципу для его последующей кластеризации и определения параметров функции конверсии.

Четвертая глава посвящена рассмотрению программной реализации мультиголосового синтезатора речи по тексту с использованием разработанного метода конверсии голоса на базе МГС и регрессионной функции отображения, в ней приведена структура МГСРТ, алгоритмы, указаны особенности имплементации акустического и просодического процессоров на основе разработанных методов.

Предложен алгоритм конверсии просодических характеристик голоса диктора, который позволяет выполнить параметризацию контура ЧОТ на основе определения особых точек контура, а затем найти его трансформацию с учетом этих особенностей, что повышает точность процедуры конверсии.

На этапах проектирования и реализации программной системы мультиголосового синтеза речи по тексту с применением технологии конверсии голоса был использован объектно-ориентированный подход через такие его механизмы, как абстрагирование, инкапсуляция, наследование и полиморфизм (посредством использования программных интерфейсов), что позволило значительно снизить степень связности между программными компонентами системы и предоставило возможность оперативной смены одного варианта реализации предлагаемых методов на другой в рамках интегрированной архитектуры системы.

В **пятой главе** представлены результаты экспериментальных исследований и оценки эффективности предлагаемых методов и алгоритмов, а также программной реализации МГСРТ. Один из экспериментов проводился с целью оценки качества работы МГСРТ путем установления степени соответствия характеристик голоса, полученного на выходе системы голосу ЦД. Для этого из специального фонетически сбалансированного текста сформированы один обучающий (40 предложений), а также пять тестовых (I, II, III, IV, V по 5 предложений) наборов фраз различного содержания. На основе данных наборов обучающих текстовых данных были записаны фонограммы для четырёх различных целевых дикторов: двух женщин (Ж1, Ж2) и двух мужчин (М1, М2). Средняя длительность фразы составила 7 – 10 с. Аудиофайлы были закодированы в формате «wav» с частотой дискретизации 16 кГц и разрядностью 16 бит. Метрика близости рассчитывалась как среднеквадратичная ошибка между векторами кепстральных коэффициентов исходного и сконвертированного фреймов сигнала ($\varepsilon_{ИД}$), и целевого и сконвертированного фреймов ($\varepsilon_{ЦД}$). Результирующие оценки приведены в таблице 1.

Таблица 1. – Результаты экспериментальной оценки качества работы МГСРТ

Диктор	**Набор тестовых фраз**									
	I		**II**		**III**		**IV**		**V**	
	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$	$\varepsilon_{ид}$	$\varepsilon_{цд}$
Ж1	0,89	0,62	0,82	0,76	0,90	0,67	0,87	0,63	0,80	0,59
Ж2	0,99	0,78	0,96	0,91	0,99	0,78	0,98	0,72	0,91	0,68
М1	0,75	0,56	0,71	0,57	0,78	0,57	0,78	0,52	0,70	0,50
М2	0,73	0,54	0,74	0,63	0,80	0,62	0,73	0,49	0,76	0,44
Среднее:	0,84	0,63	0,81	0,72	0,87	0,66	0,84	0,59	0,79	0,55

В ходе эксперимента было установлено, что общие среднеквадратичные ошибки по всем наборам фраз составляют $\varepsilon_{ИД} = 0{,}84$ для голоса ИД и синтезированного МГСРТ, а также $\varepsilon_{ЦД} = 0{,}63$ для синтезированного МГСРТ и голоса ЦД соответственно. По результатам экспериментов можно сделать вывод, что речевой сигнал, генерируемый МГСРТ, настроенной на диктора, оказывается в параметрическом представлении на 30 % ближе к речевому сигналу ЦД чем ИД, что свидетельствует о корректной работе системы с точки зрения объективных метрик оценки.

Целью второго эксперимента являлось определение эффективности осуществления процедуры конверсии на основе предложенного метода КГ. Осуществлялся сравнительный анализ предложенного метода конверсии голоса на основе МГС и регрессионной функции отображения (МГС*), выражение (3), искусственных нейронных сетей (ИНН), спектрального взвешивания (СВ), стандартного метода на базе МГС (МГС), выражение (1). Исходные данные эксперимента анало-

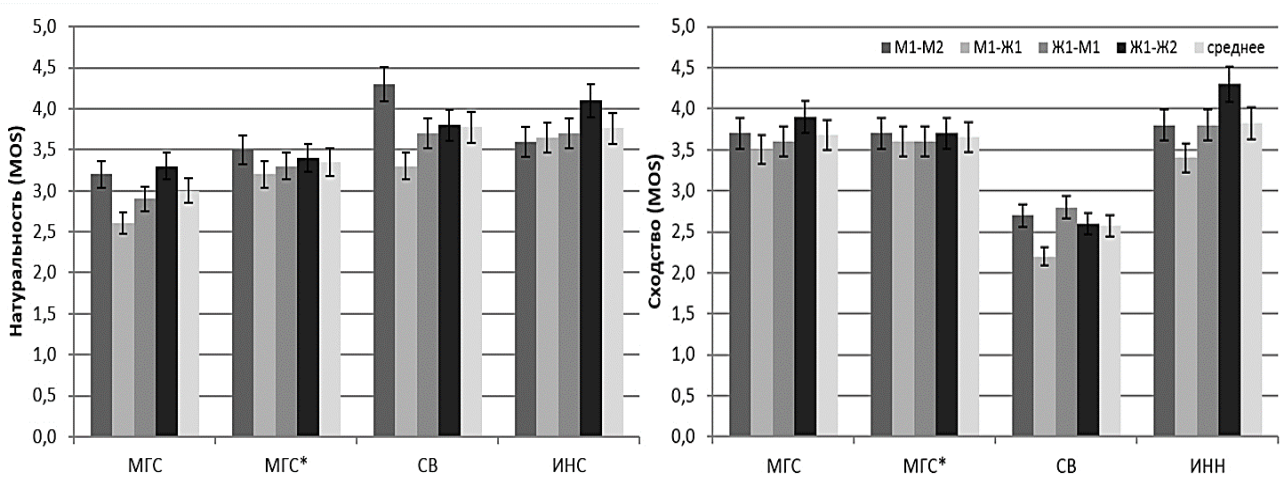
гичны данным, используемым в предыдущем эксперименте. Для оценки была использована субъективная метрика оценки мнений экспертов (*MOS*) по показателям натуральности и сходства речи. Результаты представлены в таблицах 2, 3 и на рисунке 3.

Таблица 2. – Оценки *MOS* для характеристики натуральности речи

Метод конверсии	**Конверсия голоса (ИД–ЦД)**				**Средняя оценка**
	М1–М2	**М1–Ж1**	**Ж1–М1**	**Ж1–Ж2**	
МГС	3,2	2,6	2,9	3,3	3,0
МГС*	**3,5**	**3,2**	**3,3**	**3,4**	**3,4**
СВ	4,3	3,3	3,7	3,8	3,8
ИНН	3,6	3,7	3,7	4,1	3,8

Таблица 3. – Оценки *MOS* для характеристики сходства речи

Метод конверсии	**Конверсия голоса (ИД–ЦД)**				**Средняя оценка**
	М1–М2	**М1–Ж1**	**Ж1–М1**	**Ж1–Ж2**	
МГС	3,7	3,5	3,6	3,9	3,7
МГС*	**3,7**	**3,6**	**3,6**	**3,7**	**3,7**
СВ	2,7	2,2	2,8	2,6	2,6
ИНН	3,8	3,4	3,8	4,3	3,8



а б

а – натуральности; б – сходства

Рисунок 3. – Результаты экспериментов по оценки характеристик

Анализ результатов экспериментов показывает, что предложенный метод МГС* позволяет добиться улучшения характеристик натуральности по сравнению с классическим методом конверсии на основе МГС в среднем на 10 % по параметрам натуральности и на 5 % по параметру сходства. По параметру натуральности предложенный метод уступает подходу на основе СВ и ИНН. Этот факт можно объяснить

тем, что перечисленные методы позволяют получить более выраженное, менее усредненное представление спектральной огибающей в результате конверсии, в то время как при использовании статистических методов КГ может наблюдаться эффект усреднения данных характеристик. По степени сходства предложенный метод МГС* превосходит стандартный МГС, а также метод на основе СВ и лишь незначительно уступает методу на основе ИНН при достижении большой простоты в обучении и меньших ресурсах затрачиваемых на обучение. Высокие характеристики последнего обосновываются использованием нелинейной функции отображения, однако требуют предварительной разработки архитектуры нейронной сети и выбора алгоритма ее обучения.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Предложена интегрированная архитектура системы мультиголосового синтеза речи по тексту, которая позволяет включить модуль конверсии голоса в состав компиляционного синтезатора речи по тексту на уровне акустического процессора системы и использовать имеющиеся лингвистические ресурсы, содержащиеся в базе данных акустических фрагментов речевого сигнала синтезатора (в параметризированном виде), в процессе текстонезависимого обучения модуля конверсии голоса и синтеза речи. Предлагаемое решение позволяет уменьшить количество ошибок при конверсии тембральных и просодических характеристик голоса диктора на 11 и 13 % по сравнению с каскадной архитектурой [2, 6, 12, 13, 20, 21].

2. Разработан метод конверсии голоса на основе модели гауссовых смесей, позволяющий выполнить мягкую кластеризацию пространства векторов параметров, характеризующих голос диктора, и осуществить их трансформацию с использованием регрессионной функции конверсии специального вида, который позволил увеличить значение коэффициента детерминации статистической модели на 0,1 , а также оценки узнаваемости и естественности согласно шкале оценки мнений (*MOS*) на 0,45 и 0,34 единиц соответственно [1, 4, 10, 11, 15, 19].

3. Разработан алгоритм поиска параметров функции конверсии на основе алгоритма максимизации функции правдоподобия статистической модели (*EM*-алгоритма) и использования информационного критерия Акаике (учитывающего кроме значения функции правдоподобия еще и размерность модели в качестве штрафного коэффициента), который позволил осуществлять определение необходимого количества компонент смеси непосредственно в процессе обучения системы. [3, 8, 14, 17, 22].

4. Метод текстонезависимого обучения системы конверсии голоса на базе скрытых марковских моделей лево-правого типа с непрерывным множеством алфавита наблюдений и определением параметров модели на основе модифицированного алгоритма Витерби, который дает возможность осуществления перехода к текстонезависимому способу обучения МГСРТ с относительной величиной ошибки обучения, равной 15 % , по сравнению с текстозависимым вариантом [5, 7, 9, 16, 18].

Рекомендации по практическому использованию результатов

Разработанные методы конверсии голоса на основе модели гауссовых смесей могут использоваться для создания мультиголосовых синтезаторов речи по тексту с настройкой на голос диктора в систем мультимедиа, интернет-сервисах передачи голосовых сообщений, программах переозвучивания аудиокниг голосом целевого диктора, детских развивающих игрушках с возможностью речевой обратной связи и настойкой на голос родителей или родственников и др. Представленные в работе методы и алгоритмы разработаны и внедрены на предприятиях Республики Беларусь в следующие программные средства:

1. Модуль мультиголосового синтеза речи по тексту на основе интегрированной архитектуры систем синтеза речи и конверсии голоса для системы синтеза речи «SpeechFLY Voices» – ООО «Речевые системы», Минск, Республика Беларусь.

2. Модуль оценки и коррекции просодических характеристик речи диктора для программного продукта «Speech trainer» («Речевой тренер»). – ООО «IT4YOU», Москва, Российская Федерация.

3. Модуль текстонезависимого обучения системы конверсии голоса на базе скрытых марковских моделей для мультиголосового синтезатора речи по тексту «Мультифон». – «ОИПИ НАН Беларуси», Минск, Республика Беларусь.

4. Программный комплекс, предназначенный для изучения принципов построения систем мультиголосового синтеза речи по тексту в рамках курса «Системы обработки мультимедиа данных» для студентов 4-го курса специальности 1 - 40 02 02 «Электронные вычислительные средства» – БГУИР, Минск, Республика Беларусь.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ УЧЕНОЙ СТЕПЕНИ

Статьи в рецензируемых научных журналах

1. Захарьев, В. А. Конверсия просодических характеристик диктора на основе методов параметризации контура частоты основного тона / В. А. Захарьев, А. А. Петровский // Доклады БГУИР. – 2013. – № 1. – С. 39–46.

2. Захарьев, В. А. Архитектура мультиголосового синтезатора речи по тексту / В. А. Захарьев, А. А. Петровский // Доклады БГУИР. – 2013. – № 7. – С. 57–64.

3. Захарьев, В. А. Конверсия голоса на основе множественной регрессионной функции отображения и методе спектрального взвешивания / В. А. Захарьев, А. А. Петровский // Речевые технологии. – 2014. – № 3. – С. 40–54.

4. Захарьев, В. А. Система синтеза речи по тексту с возможностью настройки на голос целевого диктора / В. А. Захарьев, А. А. Петровский, Б. М. Лобанов // Труды СПИИРАН. – 2014. – № 1(32). – С. 82–98.

5. Grapheme-to-phoneme and phoneme-to-grapheme conversion in Belarusian with NooJ for TTS and STT systems / V. Zahariev, S. Lysy, A. Hiuntar, Y. Hetsevich / Communications in Computer and Information Science book series. Springer. – 2016. – Vol. 607 – P. 137–150. doi:10.1007/978-3-319-42471-2_12

Статьи в сборниках материалов научных конференций

6. Захарьев, В. А. Анализ подходов конверсии голоса в системах мультимедиа / В. А. Захарьев // Информационные технологии и системы 2011 (ИТС 2011): материалы Международной конференции, Минск, 26 октября 2011 г. / БГУИР. – Минск, 2011. – С. 117–118.

7. Захарьев, В. А. Скрытые Марковские модели для решения задач текстонезависимого обучения в системе конверсии голоса / В. А. Захарьев // Информационные технологии и системы 2012 (ИТС 2012): материалы Международной конференции, Минск, 24 октября 2012 г. / БГУИР. – Минск, 2012. – С. 106–107.

8. Захарьев, В. А. Конверсия голоса на основе взвешенной деформации спектра / В. А. Захарьев, А. А. Петровский // Информационные технологии и системы 2013 (ИТС 2013): материалы Международной конференции, Минск, 23 октября 2013 г. / БГУИР. – Минск, 2013. – С. 108–109.

9. Захарьев, В. А. Текстонезависимое обучение в системе конверсии голоса на базе скрытых Марковских моделей и схемы преобразования буква-фонема / В. А. Захарьев, А. А. Петровский // Цифровая обработка сигналов и ее применение (DSPA 2013): материалы 15-й Международной научно-технической конференции, Москва, – 27 – 29 марта 2013 г. / ИПУ РАН. – М., 2013. – Т. 2. –С. 327–332.

10. Захарьев, В. А. Применение метода семантического дифференциала для оценки показателей качества конверсии голоса / В. А. Захарьев, А. А. Петровский // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS 2014): материалы IV Международной научно-технической конференции, Минск, 22 – 24 февраля 2014 г. / БГУИР. – Минск, 2014. – С. 453–456.

11. Захарьев, В. А. Конверсия голоса на основе множественной регрессионной функции отображения / В. А. Захарьев, А. А. Петровский // Международная научно-техническая конференция, приуроченная к 50-летию МРТИ–БГУИР: материалы конференции, 18 – 19 марта 2014 г. / БГУИР. – Минск, 2014. – С. 312–314.

12. Zahariev, V. Multivoice text to speech synthesis system / V. Zahariev, A. Petrovsky, B. Lobanov // 12th International Conference on Pattern Recognition and Information Processing (PRIP 2014): conference proceedings, Minsk, 28 – 30 May 2014 / UIIP NASB. – Minsk, 2014 – P. 320–324.

13. Захарьев, В. А. Аспекты практической реализации мультиголосового синтезатора речи по тексту / В. А. Захарьев, А. А. Петровский // Информационные технологии и системы 2014 (ИТС 2014): материалы Международной конференции. – Минск, – 29 октября 2014 г. / БГУИР. – Минск, 2014. – С. 94–95.

14. Адкрытыя кампаненты «www.corpus.by» для натуральнага маўленчага інтерфейсу / Ю. С. Гецевіч, Б. М. Лабанаў, С. И. Лысы, А. В. Гюнтар, Д. А. Дзенісюк В. А. Захар'еў // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS 2015): материалы V Международной научно-технической конференции, Минск,19 – 21 февраля 2015 г. / БГУИР. – Минск, 2015. – С. 499–506.

15. Захарьев, В. А. Применение мультирегрессионной модели для решения задач конверсии голоса / В. А. Захарьев, А. А. Петровский // Цифровая обработка сигналов и ее применение (DSPA 2015): материалы 17-й Международной научно-технической конференции, Москва 25 – 27 марта 2015 г. / ИПУ РАН. – М., 2015. – С. 231–234.

16. Захарьев, В. А. Конверсия голоса для систем мультиголосового синтеза речи по тексту / В. А. Захарьев, А. А. Петровский // Информационные технологии и системы 2015 (ИТС 2015): материалы Международной конференции, Минск, 28 октября 2015 г. / БГУИР. – Минск, 2015. – С. 92–93.

17. Захарьев, В. А. Система синтеза речи по тексту с возможностью настройки на голос целевого диктора / В. А. Захарьев, А. А. Петровский // Информационные технологии в управлении 2016 (ИТУ 2016): материалы 9-й Международной конференции, Санкт-Петербург, 5 октября 2016 г. / АО «Концерн «ЦНИИ «Электроприбор». – СПб, 2016. – С. 120–132.

18. Захарьев, В. А. Алгоритм текстонезависимого обучения для систем мультиголосового синтеза речи по тексту / В. А. Захарьев,А. А. Петровский // Информационные технологии и системы 2016 (ИТС 2016): материалы Международной конференции, Минск, 26 октября 2016 г. / БГУИР. – Минск, 2016 – С. 90–91.

19. Zahariev, V. Multivoice Text-to-Speech Synthesis for Natural-Language Interfaces of Intelligent Systems / V. Zahariev, A. Petrovsky // 7th International Conference

Open Semantic Technologies for Intelligent Systems (OSTIS 2017): conference proceedings, Minsk, 16 – 18 February 2017. / BSUIR. – Minsk, 2017. – P. 167–169.

Тезисы докладов в сборниках материалов научных конференций

20. Захарьев, В. А. Технология конверсии голоса для систем мультимедиа / В. А. Захарьев, А. А. Петровский // Информационные технологии и управление: материалы 47-й научной конференции аспирантов, магистрантов и студентов, Минск, 25 – 29 апреля 2011 г. / БГУИР. – Минск, 2011. – С. 23.

21. Захарьев, В. А. Методы параметризации речевого сигнала на основе анализа синхронизированного с частотой основного тона в системах конверсии голоса / В. А. Захарьев, А. А. Петровский // Наука – образованию, производству, экономике : материалы 11-й Международной научно-технической конференции, Минск, 15 мая 2013 г. / БНТУ. – Минск, 2013. – Т. 1. – С. 203–204.

22. Захарьев, В. А. Построение многоголосого синтезатора речи по тексту на базе системы текстонезависимой конверсии голоса / В. А. Захарьев, А. А. Петровский // Наука – образованию, производству, экономике : материалы 11-й Международной научно-технической конференции, Минск, 15 мая 2013 г. / БНТУ. – Минск, 2013. – Т. 1. – С. 204–205.

РЭЗЮМЭ

Захар’еў Вадзім Анатольевіч

КАНВЕРСІЯ ГОЛАСУ НА АСНОВЕ МАДЭЛІ ГАЎСАВЫХ СУМЯСЯЎ У СІСТЭМАХ СІНТЭЗУ МАЎЛЕННЯ ПА ТЭКСЦЕ З НАЛАДКАЙ НА ГОЛАС ДЫКТАРА

Ключавыя словы: мультыгаласавы сінтэз маўлення па тэксце, канверсія голасу, мадэль гаўсавых сумясяў, тэкстанезалежнае навучанне.

Мэта работы: распрацоўка метадаў канверсіі голасу для сістэм сінтэзу маўлення па тэксце з наладкай на голас дыктара, а таксама рэалізацыя праграмнай сістэмы мультыгаласавога сінтэзу маўлення па тэксце.

Атрыманыя вынікі і іх навізна: прапанавана інтэграваная архітэктура сістэмы мультыгаласавога сінтэзу маўлення па тэксце, асаблівасцю якой з'яўляецца ўключэнне ў склад акустычнага працэсара модуля канверсіі галасоў, што дазваляе ўлічваць асаблівасці этапаў апрацоўкі маўленчай інфармацыі як на стадыі сінтэзу, так і на стадыі канверсіі голасу, што дазваляе паменшыць колькасць памылак у параўнанні з каскаднай архітэктурай; распрацаваны метад і алгарытм канверсіі голасу, асаблівасцю якога з'яўляецца магчымасць параметрызацыі прасторы характарыстык голасу дыктара і яго кластэрызацыі на базе мадэлі гаўсавых сумесяў і рэгрэсійнай функцыі канверсіі, якія дазваляюць павялічыць значэнне каэфіцыента дэтэрмінацыі статыстычнай мадэлі, а таксама павялічыць адзнакі пазнавальнасці і натуральнасці сканвертаванага голасу; прапанаваны метад і алгарытм тэкстанезалежнага навучання, асаблівасцю якога з'яўляецца выкарыстанне схаваных маркаўскіх мадэляў і мадыфікаванага алгарытму Вітэрбі, які дазволіў ажыццявіць пераход да тэкстанезалежнага навучання сістэмы з адноснай велічыней памылак выраўноўвання, якая супаставіма з тэкстазалежным варыянтам навучання сістэмы мультыгаласавога сінтэзу маўлення па тэксце.

Рэкамендацыі па выкарыстанні і вобласць ужывання: распрацаваныя метады могуць быць выкарыстаны для стварэння мультыгаласавога сінтэзу маўлення па тэксце ў сістэмах мультымедыя рознага прызначэння, інтэрнэт-сэрвісах перадачы галасавых паведамленняў, праграмах пераагучвання электронных аўдыякніг і г.д. Вынікі ўкаранены ў арганізацыях і на прадпрыемствах Рэспублікі Беларусь (АІПІ НАН Беларусі, г. Мінск, ТАА «Маўленчыя сістэмы», г. Мінск), Расійскай Федэрацыі (ТАА «АйТиФо Ю», г. Масква), а таксама выкарыстоўваюцца ў навучальным працэсе БДУІР для спецыяльнасці 1-40 02 02 «Электронныя вылічальныя сродкі».

РЕЗЮМЕ

Захарьев Вадим Анатольевич

КОНВЕРСИЯ ГОЛОСА НА ОСНОВЕ МОДЕЛИ ГАУССОВЫХ СМЕСЕЙ В СИСТЕМАХ СИНТЕЗА РЕЧИ ПО ТЕКСТУ С НАСТРОЙКОЙ НА ГОЛОС ДИКТОРА

Ключевые слова: мультиголосовой синтез речи по тексту, конверсия голоса, модель гауссовых смесей, регрессионная функция, текстонезависимое обучение.

Цель работы: разработка методов конверсии голоса для систем синтеза речи по тексту с настройкой на голос диктора и реализация программной системы мультиголосового синтеза речи по тексту.

Полученные результаты и их новизна: предложена интегрированная архитектура системы мультиголосового синтеза речи по тексту, особенностью которой является включение в состав акустического процессора модуля конверсии голоса, что позволяет учесть особенности этапов обработки речевой информации как на стадии синтеза, так и на стадии конверсии голоса и вследствие этого уменьшить количество ошибок по сравнению с каскадной архитектурой; разработан метод и алгоритм конверсии голоса, особенностью которого является параметризация пространства характеристик голоса диктора и его кластеризация на базе модели гауссовых смесей и регрессионной функции конверсии, который позволил увеличить значение коэффициента детерминации статистической модели, а также увеличить оценки узнаваемости и естественности сконвертированного голоса; предложен метод и алгоритм текстонезависимого обучения, особенностью которого является использование скрытых марковских моделей и модифицированного алгоритма Витерби, который позволил осуществить переход к текстонезависимому обучению системы с относительной величиной ошибок выравнивания, сопоставимой с текстозависимым вариантом обучения мультиголосового синтезатора речи по тексту.

Рекомендации по использованию и область применения: разработанные методы могут быть использованы для создания мультиголосовых синтезаторов речи по тексту в системах мультимедиа различного назначения, интернет-сервисах передачи голосовых сообщений, программах переозвучивания аудиокниг и т. д. Результаты внедрены в организациях и на предприятиях Республики Беларусь (ОИПИ НАН Беларуси, г. Минск, ООО «Речевые системы», г. Минск), Российской Федерации (ООО «АйТиФо Ю», г. Москва), а также используются в учебном процессе БГУИР для специальности 1-40 02 02 «Электронные вычислительные средства».

SUMMARY

Zahariev Vadim Anatolievich

VOICE CONVERSION
BASED ON THE GAUSSIAN MIXTURE MODELS
FOR TEXT TO SPEECH SYNTHESIS SYSTEMS
WITH TUNING FOR THE SPEAKER VOICE

Key words: multivoice text to speech synthesis, voice conversion, Gaussian mixture models, regression conversion function, text-independent training.

The purpose of research is the development of methods based on the voice conversion for text-to-speech synthesis systems (TTS) with tuning for the speaker voice and software realization of the multivoice text-to-speech synthesis system (MVTTS).

The obtained results and their novelty: an integrated system architecture of multivoice text-to-speech synthesis system, which distinctive feature is the inclusion of voice conversion module into acoustic processor of TTS. It allows taking into account the peculiarities of speech information processing stages both at the steps of speech synthesis, and voice conversion. And, consequently, it allows to reduce the number of errors compared to the canonical cascade architecture. The method and the algorithm for the voice conversion, which is unique in parameterization and clustering of voice features space is based on Gaussian mixtures models and regression conversion function. It allows to increase the value of the determination coefficient of the statistical model, as well as to increase the recognition and evaluation of naturalness converted voice was developed. The key feature of the method and the algorithm of text independent learning is the usage of hidden Markov models and modified Viterbi algorithm, what allows to move on to the text-independent approach in system training with relative value alignment errors, comparable with the text-dependent learning approach for multivoice text to speech synthesis system.

Recommendations on the use and field of application: the developed methods can be used for creation of multivoice text to speech synthesis systems, multimedia systems, online services with vocalizing features of the text messages, scoring of electronic audiobooks and virtual assistants etc. Results are implemented in organizations and enterprises of the Republic of Belarus (UIIP NAS Belarus, Minsk; LLC «Speech systems», Minsk), Russian Federation (LLC «IT4YOU», Moscow), as well as used in the educational process in BSUIR for specialty 1-40 02 02 «Computer engineering».

Научное издание

Захарьев Вадим Анатольевич

КОНВЕРСИЯ ГОЛОСА НА ОСНОВЕ МОДЕЛИ ГАУССОВЫХ СМЕСЕЙ В СИСТЕМАХ СИНТЕЗА РЕЧИ ПО ТЕКСТУ С НАСТРОЙКОЙ НА ГОЛОС ДИКТОРА

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

по специальности 05.13.17 – Теоретические основы информатики

Подписано в печать | Формат 60x84 1/16 | Бумага офсетная.
Гарнитура «Таймс». | Отпечатано на ризографе. | Усл. печ. л.
Уч.-изд. л. | Тираж 60 экз. | Заказ

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/238 от 24.04.2014,
№ 2/113 от 07.04.2016, № 3/615 от 07.04.2014.
ЛП № 02330/264 от 14.04.2014.
220013, Минск, П. Бровки, 6