

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. М. В. ЛОМОНОСОВА
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи

Соломенник Анна Ивановна

**Оценка качества селективного синтеза речи:
методы и результаты**

Специальность 10.02.21 – «Прикладная и математическая лингвистика»

Диссертация на соискание учёной степени
кандидата филологических наук

Научный руководитель
доктор филологических наук,
ведущий научный сотрудник
Кривнова Ольга Фёдоровна

Москва – 2016

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	5
ГЛАВА 1. ИСТОРИЯ СИНТЕЗАТОРОВ РЕЧИ И СОВРЕМЕННЫЕ РАЗРАБОТКИ	10
1.1 Цели и задачи синтеза речи	10
1.2 История вопроса и типы синтезаторов речи	10
1.3 Перспективы дальнейшего развития синтеза речи.....	36
1.4 ВЫВОДЫ К ГЛАВЕ 1	37
ГЛАВА 2. СЕЛЕКТИВНЫЙ СИНТЕЗ РЕЧИ	38
2.1 Алгоритм Unit selection	38
2.2 Русскоязычные селективные синтезаторы	41
2.3 Структура современного селективного синтезатора речи типа «Текст–Речь»	43
2.4 ВЫВОДЫ К ГЛАВЕ 2	61
ГЛАВА 3. МЕТОДЫ ОЦЕНКИ КАЧЕСТВА СЕЛЕКТИВНОГО СИНТЕЗА РЕЧИ	62
3.1 Задачи и критерии оценки качества синтезированной речи.....	62
3.2 Методы оценки разборчивости речи.....	65
3.3 Методы оценки естественности речи	67
3.4 Факторы, влияющие на восприятие синтезированной речи человеком	71
3.5 Необходимость адаптации методов оценки к селективному синтезу речи ..	72
3.6 Структура и задачи системы оценки качества синтезированной речи.....	75
3.7 ВЫВОДЫ К ГЛАВЕ 3	84

ГЛАВА 4. РЕЗУЛЬТАТЫ ОЦЕНКИ КАЧЕСТВА СЕЛЕКТИВНОГО СИНТЕЗА РЕЧИ	85
4.1 Оценка степени влияния различных типов ошибок на качество синтезированной речи	86
4.2 Оценка лингвистической обработки	94
4.3 Оценка фонетической обработки	108
4.4 Оценка акустической обработки	120
4.5 Интегральная оценка качества синтезированной речи и оценка общего качества синтеза	121
4.7 ВЫВОДЫ К ГЛАВЕ 4	131
ЗАКЛЮЧЕНИЕ	133
СПИСОК СОКРАЩЕНИЙ	136
СПИСОК ЛИТЕРАТУРЫ	137
СПИСОК РИСУНКОВ	150
ПЕРЕЧЕНЬ ТАБЛИЦ	152
ПРИЛОЖЕНИЯ	154
Приложение А. Фонетически представительный текст	154
Приложение Б. Инструкция и пример анкеты в эксперименте по влиянию различных типов ошибок на качество синтезированной речи	157
Приложение В. Список предложений для тестирования графических сокращений	158
Приложение Г. Ошибки чтения сокращений: результаты тестирования	162
Приложение Д. Список предложений для тестирования озвучивания аббревиатур	164

Приложение Е. Ошибки озвучивания аббревиатур.....	169
Приложение Ж. Список предложений для тестирования расшифровки цифровых обозначений и выражений	171
Приложение И. Результаты тестирования расшифровки цифровых обозначений и выражений	174
Приложение К. Список предложений для тестирования чтения особых символьных обозначений.....	175
Приложение Л. Результаты тестирования чтения символьных обозначений ..	177
Приложение М. Список английских слов с их практической транскрипцией .	178
Приложение Н. Результаты тестирования по озвучиванию англоязычных вставок	180
Приложение П. Список предложений для тестирования правильности снятия омонимии.....	181
Приложение Р. Результаты тестирования правильности снятия омонимии.....	185
Приложение С. Список предложений для оценки правильности сегментной транскрипции	187
Приложение Т. Результаты оценки правильности сегментной транскрипции в синтезаторах речи	188
Приложение У. Набор фраз для тестирования интонации	190
Приложение Ф. Образец анкеты по оценке адекватности и естественности интонации синтезированной речи.....	193
Приложение Х. Образец бланка оценки качества речи	194

ВВЕДЕНИЕ

В течение нескольких последних десятилетий технология синтеза речи развивалась с большой скоростью и эффективностью. Главным достижением в этой области можно считать озвучивание произвольного текста, которое близко по качеству к естественному звучанию. Важно также и то, что это можно сделать в режиме реального времени даже на небольшом мобильном устройстве. Синтез используется в системах голосового самообслуживания банков, транспортных компаний, в проведении телефонных опросов. Синтезированными голосами «говорят» мобильные устройства, озвучиваются аудиокниги. Синтезированная речь всё более уверенно становится частью повседневной жизни человека и общества. В связи с этим проблема объективной оценки качества синтезированной речи и сравнения речевых синтезаторов между собой приобретает особую значимость.

Актуальность данной работы состоит в том, что именно селективный синтез речи, в англоязычных источниках называемый unit selection, в настоящее время является общепризнанным методом получения качественной синтезированной речи, наиболее близкой по звучанию к естественной. Этим обусловлено то, что при разработке большинства современных синтезаторов, особенно коммерческих приложений, используется именно этот метод. В этой связи при оценке качества синтезированной речи необходимо обратить особое внимание на специфические особенности звучания речи, связанные с его использованием.

В этой области речевых технологий существует множество исследований, однако в данный момент для современных русскоязычных синтезаторов (так называемых синтезаторов третьего поколения по классификации П. Тейлора [Taylor 2009]), способных генерировать речь, по звучанию очень близкую к естественной, **степень разработанности проблемы** оценки качества

синтезированной речи невелика: пока нет единой общепринятой системы оценки [Соломенник и др. 2013(а)]. Отдельные исследования либо несколько устарели, так как в них рассматриваются синтезаторы второго поколения (аллофонный и дифонный конкатенативный синтез), как в работе [Русанова 2004], либо не обладают достаточной полнотой, как, например, статья, посвящённая тестированию систем автоматического синтеза и распознавания речи в целях определения коммерческой целесообразности их использования [Корсакова, Засыпкина 2012]. Таким образом, очевидна необходимость разработки и описания новой системы оценки, учитывающей особенности именно современных методов синтеза речи.

Цель настоящего исследования – разработать комплекс методов оценки качества селективного синтеза речи на материале русского языка.

Задачи исследования:

1. Описать существующие проблемы и методы оценки качества синтеза речи с анализом и обобщением результатов предыдущих исследований.
2. Обосновать необходимость специального подхода к оценке селективного синтеза с учётом его характерных особенностей.
3. Предложить методы оценки синтеза речи, позволяющие объективно оценивать и сравнивать современные русскоязычные селективные синтезаторы речи.
4. Произвести комплексную оценку нескольких современных русскоязычных синтезаторов.

Научная новизна работы заключается в том, что впервые для русского языка были предложены и опробованы методы комплексной оценки современных селективных синтезаторов речи.

Теоретическая значимость заключается в анализе и выявлении характеристик селективного синтеза речи с точки зрения особенностей качества речи, порождаемой таким синтезатором.

Практическая значимость работы состоит в том, что появляется возможность использовать предложенные методы для объективной оценки и сравнения между собой современных синтезаторов речи высокого качества. На основании полученных результатов могут быть предложены методы улучшения качества синтезированной речи.

Предметом исследования данной работы являются методы комплексной оценки качества синтеза речи.

Объект исследования – искусственно порождённая речь, её характеристики с точки зрения восприятия слушающими, критерии качества синтезированной речи.

Материалом исследования является синтезированная речь, полученная с использованием нескольких современных русскоязычных селективных синтезаторов (Acapela, iSpeech, Ivona TTS, Mary TTS, Loquendo TTS, Nuance Vocalizer, VitalVoice TTS).

В ходе работы были использованы следующие **методы**: методы слухового, аудиторского и инструментального анализа фонограмм, статистические методы анализа результатов проведённых фонетических экспериментов.

На защиту выносятся следующие **положения**:

1. Для оценки селективного синтеза речи необходим специальный подход, учитывающий специфические особенности данной речевой технологии.

2. Предложенный в диссертации подход и его оценочные средства позволяют проводить комплексное диагностическое тестирование современных русскоязычных синтезаторов селективного типа и сравнивать их между собой на объективной основе.
3. Максимальное влияние на естественность звучания селективного синтеза речи для русского языка оказывают ошибки, связанные с выбором неправильного места ударения в словах и неадекватной интонацией.
4. Ошибки и недочеты в лингвистической обработке текста перед его фонетизацией являются основным источником качественных различий в работе современных селективных синтезаторов русской речи.

Достоверность результатов обеспечивается успешным практическим применением предложенной системы методов оценки в экспериментах по тестированию нескольких современных систем селективного синтеза речи.

Апробация работы. Основные положения диссертационной работы докладывались на научно-методических конференциях: «Международная конференция по компьютерной лингвистике «Диалог 2009» (Москва), «Международная конференция по компьютерной лингвистике «Диалог 2010» (Москва), «Международная конференция по компьютерной лингвистике «Диалог 2012» (Москва), «Конференция AINL 2013: Искусственный интеллект и естественный язык» (Санкт-Петербург), «Актуальные вопросы теоретической и прикладной фонетики: конференция к юбилею О. Ф. Кривновой» (Москва, 2013), «Международная конференция по компьютерной лингвистике «Диалог 2013» (Москва), «15th International Conference on Speech and Computer SPECOM 2013» (Чехия), «XXI Международная конференция студентов, аспирантов и молодых ученых «Ломоносов» (Москва, 2014), «2nd International Scientific Conference

«Contemporary Research in Phonetics and Phonology: Methods, Aspects and Problems» (Латвия, 2015).

По теме диссертации опубликовано 16 научных работ, из них 14 статей, 7 из которых опубликовано в журналах из перечня ВАК.

Структура диссертации. Диссертация изложена на 195 страницах и состоит из введения, четырех глав и заключения. Список литературы содержит 104 наименования. Работа иллюстрирована 21 рисунком и 22 таблицами. В 19 приложениях содержатся тестовые тексты и подробные результаты экспериментов.

ГЛАВА 1. ИСТОРИЯ СИНТЕЗАТОРОВ РЕЧИ И СОВРЕМЕННЫЕ РАЗРАБОТКИ

1.1 Цели и задачи синтеза речи

Искусственное создание человеческой речи издавна интересовало ученых и практических исследователей. Задачи, стоящие перед синтезаторами речи, с течением времени претерпевали существенные изменения. Синтез речи эволюционировал от простого получения речеподобных звуков и озвучивания ограниченного набора фраз до озвучивания произвольного текста с нужным интонационным оформлением и эмоциональной окраской, причем голосом, точно имитирующим речь определённого человека (так называемое «клонирование» речи по терминологии Б. М. Лобанова [Лобанов и др. 2001]). Если же рассматривать синтезатор не просто как инструмент произнесения определённых текстов, а как часть системы речевого интерфейса общения человека с ЭВМ [Потапова 1989], то перед ним в настоящее время встают и ещё более сложные задачи. Например, так называемый «реактивный» синтез [Moore, Nicolao 2011] должен отслеживать эффект, производимый им на слушающего, и в соответствии с этим изменять свои характеристики. Актуальной на сегодняшний день проблемой является также мультязычный адаптивный синтез речи, то есть синтез речи на разных языках голосом одного диктора, возможно даже и не владеющего целевым языком [Latorre 2006, Нао 2015]. С течением времени соответственно изменялись и требования к качеству синтеза речи [Соломенник 2013(b,d)]. Рассмотрим подробнее виды синтезаторов речи и историю их создания.

1.2 История вопроса и типы синтезаторов речи

Задача синтеза речи начала интересовать людей задолго до XX века. Существуют легенды о «говорящих головах», умевших отвечать на вопросы,

которые были созданы Гербертом Орильякским (ок. 946 – 1003), Альбертом Великим (1198 – 1280) и Роджером Бэконом (1214 – 1294) [Mattingly 1974]. Достоверная история создания машин, имитирующих человеческую речь, насчитывает уже более двух веков. С течением времени изменялись как и сами механизмы и принципы работы синтезирующих устройств, так и основные области интереса и задачи учёных, занимавшихся созданием и развитием синтеза речи, а, соответственно, и само понятие «качества» синтеза речи.

1.2.1 Первые механические синтезаторы

Первые синтезаторы, появившиеся во второй половине XVIII века, были механическими, они могли порождать отдельные звуки или небольшие фрагменты слитной человекоподобной речи и подобно музыкальным инструментам требовали участия оператора-исполнителя. Важным является то, что уже в них посредством различных механических приспособлений воспроизводились основные процессы, происходящие при производстве речи человеком.

В 1779 году Петербургская Академия наук объявила ежегодную премию за объяснение разницы между пятью гласными звуками и за создание устройства, их порождающего. Немецкий учёный Христиан Готлиб Кратценштейн (1723 – 1795), работавший в то время в Петербурге, предложил наилучшее решение. Он создал систему резонаторов, представленную на рис. 1, порождавших русские гласные при помощи пульсирующего воздушного потока. Воздушный поток порождался вибрирующими язычками, подобными голосовым связкам человека [Фланаган 1968].

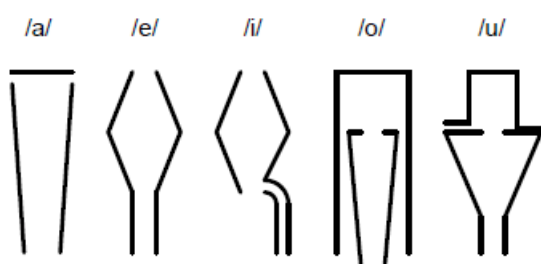


Рис. 1. Система резонаторов Кратценштейна [Lemetty 1999]

Ещё ранее и независимо от Кратценштейна над механической системой синтеза речи начал работать австрийский изобретатель Вольфганг фон Кемпелен (1734 – 1804), который представил результат своих трудов в 1791 году. Машина Кемпелена могла произносить различные звуки и их комбинации. В ней моделировалось продвижение струи воздуха через речевой тракт человека с модификациями, соответствующими реальному речепроизводству: имелись меха для подачи воздуха на язычок, пульсирующий воздушный поток возбуждал резонатор в форме резиновой камеры, управляемый рукой. Согласные, в том числе и носовые, получались с помощью четырёх каналов и свистков, зажимаемых пальцами [Фланаган 1968]. По утверждению самого Кемпелена, его машина производила 19 хорошо различимых согласных звуков [Кейтер 1985] и короткие фразы на нескольких языках [Mattingly 1974]. Для управления «говорящей машиной» требовался хорошо обученный оператор, а само порождение речи можно было сравнить с игрой на органе. Усовершенствованный вариант машины Кемпелена (рис. 2) был создан в 1837 году английским физиком Чарльзом Уитстоном (1802 – 1875). Также под впечатлением от машины Уитстона известный американский учёный и изобретатель Александр Грэм Бэлл (1847 – 1922) собрал собственную аналогичную модель [Кейтер 1985].

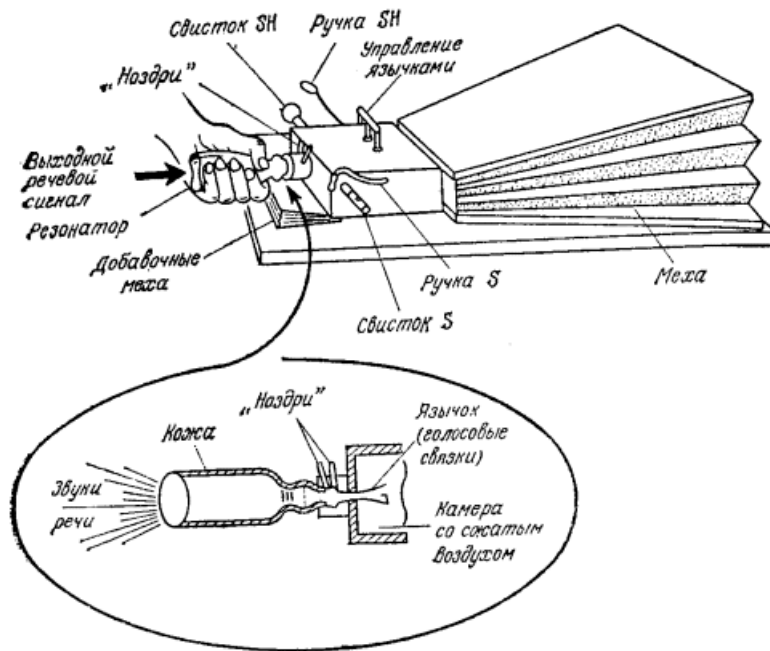


Рис. 2. Говорящая машина Кемпелена, построенная Уитстоном [Кейтер 1985]

В течение XIX века в технологии синтеза речи не было каких-либо революционных изменений. Известны исследования английского учёного Роберта Уиллиса (1800 – 1875), который подобно Кратценштейну экспериментировал с синтезом гласных звуков и установил связь между качеством гласных и геометрической формой речевого тракта. В своих работах 1828 года «О гласных звуках» и «О механизме гортани» Уиллис описал механизм производства гласных по аналогии порождением звуков при помощи органа.

В 1840 году Джозеф Фабер (ок. 1800 – ок. 1850) представил свою говорящую машину под названием «Эйфония», которая, по сообщениям современников, могла производить обычную и шёпотную речь, а также исполнять песни [Mattingly 1974].

В XX веке, несмотря на развитие электрических методов синтеза речи, разработка механических синтезаторов речи происходила вплоть до 60-х годов [Lemetty 1999]. Это было связано, с одной стороны, с малой доступностью

сложных электрических компонентов [Кейтер 1985], а с другой – с необходимостью имитации и измерения нелинейных эффектов в голосе, которые с трудом поддаются расчётам и не могут быть легко смоделированы с помощью линейных устройств [Фланаган 1968]. Среди наиболее известных устройств этого периода следует упомянуть механический синтезатор Р. Риша, продемонстрированный им в 1937 году (рис. 3). По структуре он практически повторял речевой тракт человека, был выполнен из резины и металла и управлялся клавишами, подобными клавишам музыкальной трубы [Кейтер 1985].

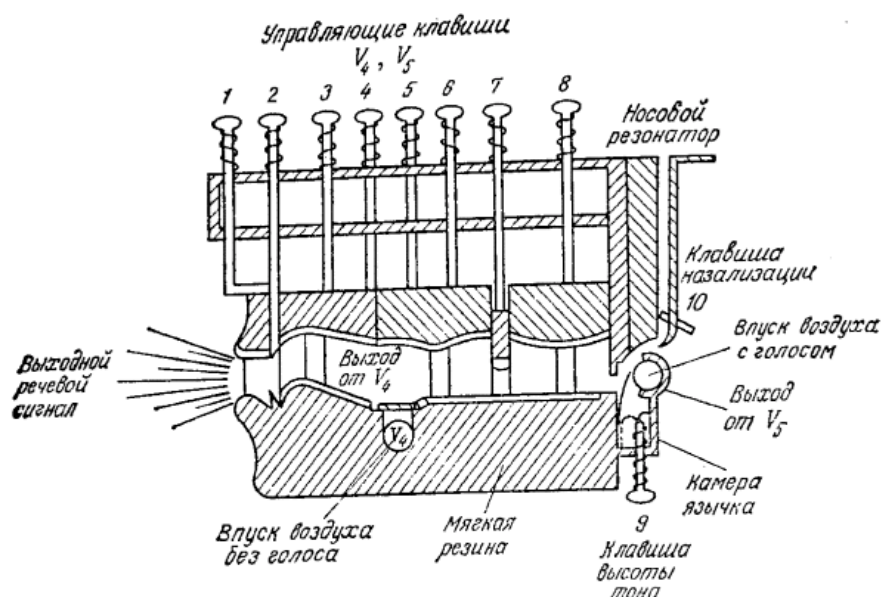


Рис. 3. Механический синтезатор Риша [Кейтер 1985]

Таким образом, общим методом создания механических синтезаторов являлась имитация или прямое моделирование речевого тракта человека. Основными рабочими компонентами таких моделей были: устройство для подачи воздуха (аналог лёгких), вибрирующая часть (аналог голосовых связок) и система резонаторов, в большей или меньшей степени воссоздававших форму речевого тракта человека. Механические синтезаторы стали прототипом современного артикуляционного синтеза речи. Основным показателем качества таких

синтезаторов было то, насколько похожие на человеческую речь звуки они могли издавать при помощи оператора, обученного управлению синтезатором.

1.2.2 Первые электрические синтезаторы

В XX веке, с освоением электрических устройств и появлением электроники, начались попытки построить синтезаторы речи в виде электрических аналогов речепроизводящей системы человека. Первый электрический синтезатор был создан Дж. Стюартом в 1922 году [Klatt 1987]. Его схема (рис. 4) включала электрический зуммер для моделирования голосовых связок и два индуктивно-ёмкостных резонатора для моделирования резонансов горла и ротовой полости [Кейтер 1985]. Таким способом генерировались первые две форманты (резонансные частоты речевого тракта), то есть устройство могло синтезировать только гласные звуки.

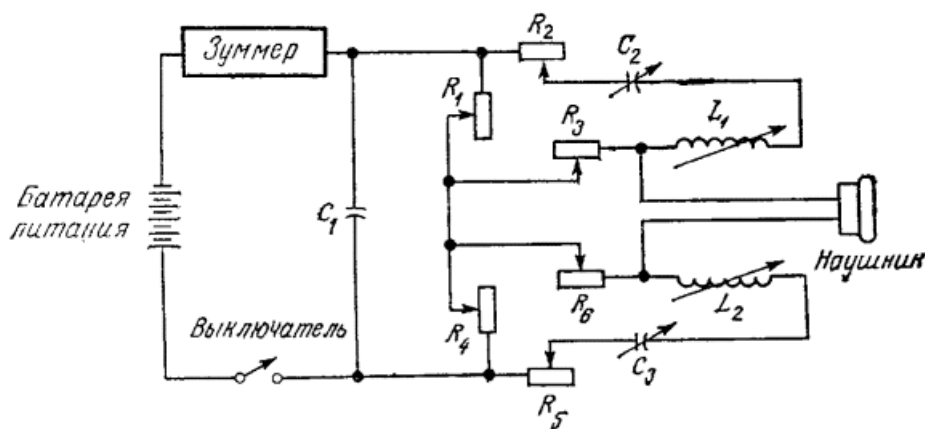


Рис. 4. Электрическая модель речевого тракта Стюарта [Кейтер 1985]

Аналогичный синтезатор, состоящий из четырёх подключенных параллельно резонаторов, возбуждаемых прерывателем тока, был создан немецким инженером К. В. Вагнером (1883 – 1953) в 1936 году [Фланаган 1968].

Следующий важный шаг в формировании технологии синтеза речи связан с развитием радиотехники, построением вокодеров (систем кодирования и

декодирования речи, в которых используются различные методы сжатия полосы частот для передачи сигналов, «voice coder») и ЭВМ [Обжелян, Трунин-Донской 1987].

Первым электрическим синтезатором, способным генерировать фрагменты *связной* речи, стал «водер» (Voder – Voice Operating Demonstrator), созданный американскими инженерами Г. Дадли (1896 – 1987), Р. Ришем и С. Уоткинсом. Водер был основан на вокодере, созданном в Bell Laboratories (США) в середине 30-х годов. От вокодера была взята синтезирующая часть, управлявшаяся вручную посредством тринадцати клавиш, ножной педали и переключателя источника шума на браслете (рис. 5) [Фланаган 1986].

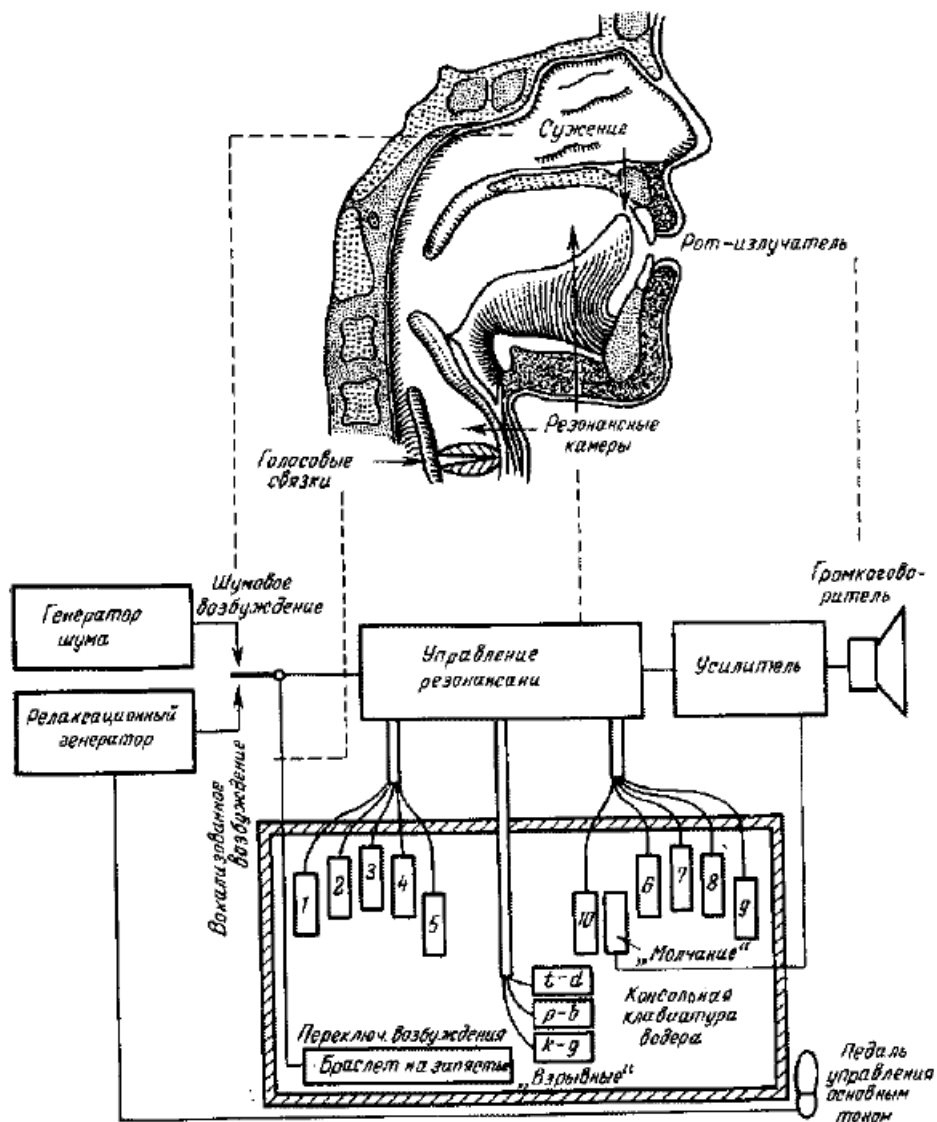


Рис. 5. Схема синтезатора «водер» [Фланаган 1986]

Водер синтезировал сигналы с заданным спектром посредством десяти включённых параллельно полосовых фильтров, охватывавших весь спектр речевых частот. Подготовка оператора для производства речи на водере занимала не менее года, однако в результате синтезируемая речь была вполне разборчива, что и спровоцировало новый интерес к синтезу речи после демонстрации водера на Всемирных выставках в Нью-Йорке в 1939 году и в Сан-Франциско в 1940 году.

В Советском Союзе производились попытки синтеза русской речи с использованием первых музыкальных электрических синтезаторов. «Вариофон» Е. А. Шолпо (1891 – 1951), сконструированный в 1931 году, представлял собой оптический синтезатор. Звук записывался на движущуюся плёнку с помощью вырезанных зубчатых дисков разной формы, изменявших очертания звуковой дорожки и трансмиссии, позволявшей синхронизировать контур и подачу плёнки. Первый электронный музыкальный синтезатор АНС был спроектирован Е. А. Мурзиным (1914 – 1970) в 1938 году и построен в 1958. АНС содержал 720 звуковых дорожек чистых тонов, которые можно было накладывать друг на друга. Клавиатуры не было, на стекле, покрытом специальной непрозрачной мастикой, прочерчивалась линия, через которую пускался световой луч на фотоэлементы синтезатора [Крейчи и др. 2013].

Важным этапом в развитии методов экспериментальных фонетических исследований и синтеза речи стала разработка звукового спектрографа в 1946 году. После этого появилась идея использования спектрограмм для управления синтезатором речи.

Для автоматического озвучивания речевых спектрограмм было создано несколько устройств. В устройстве Л. Шотта 1948 года использовался линейный источник света, расположенный вдоль оси частот спектрограммы и просвечивающий участки изображения с различной степенью прозрачности, а фотоэлементы, расположенные в ряд вплотную друг к другу по другую сторону спектрограммы, являлись источником управляющих сигналов для набора тех же полосовых фильтров, что и в водере. Дополнительные дорожки на спектрограмме управляли переключением тона и шума и несли информацию о частоте основного тона. Подобный метод использовался Дж. Борстом и Ф. Купером в устройстве «водек» (1957) [Фланаган 1968].

Наиболее известный «проигрыватель» спектрограмм, синтезатор Pattern Playback (рис. 6), был представлен американскими исследователями Ф. Купером, А. Либерманом и Дж. Борстом в 1951 году. Он состоял из оптической системы для динамической модуляции амплитуд гармоник основного тона в 120 Гц в зависимости от изображений на движущейся прозрачной ленте [Klatt 1987].

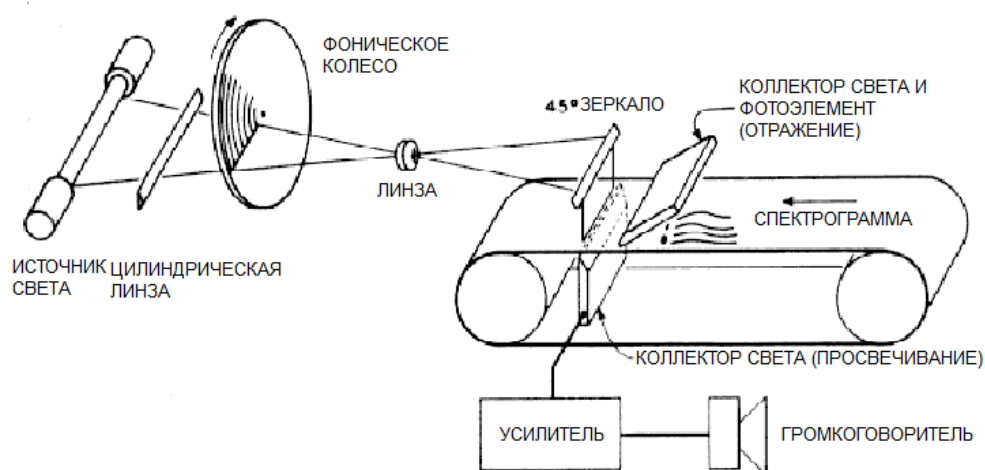


Рис. 6. Синтезатор Pattern Playback [Klatt 1987]

При помощи этого синтезатора, позволявшего производить монотонную, но разборчивую речь, проводились многочисленные эксперименты по оценке значимости для восприятия речи различных акустических характеристик путём упрощения и стилизации подаваемых на синтез фонограмм.

В первых электрических синтезаторах уже не моделируется напрямую речевой тракт человека. Вместо этого основным методом создания синтезированной речи является моделирование (или прямое считывание со спектрограммы) акустических характеристик речевого сигнала. Основными рабочими компонентами таких синтезаторов были устройства, генерирующие шум и периодический сигнал, и набор фильтров или резонаторов, усиливающих определённые частотные составляющие. Электрические синтезаторы стали прототипом современного компьютерного параметрического синтеза. Важной

характеристикой качества речи, производимой такими синтезаторами, становится её *разборчивость*.

Следующей важной вехой в истории синтеза речи стало развитие акустической теории речеобразования (1960), создавшей необходимую теоретическую базу для разработки основанных на ней формантных и артикуляционных синтезаторов, а также синтезаторов, использующих линейное предсказание. Эти три метода называют также технологиями синтеза *первого поколения* [Taylor 2009].

1.2.3 XX век: синтезаторы первого поколения

Синтезаторы первого поколения на основании используемых ими методов можно разделить на две большие группы: акустические и артикуляционные. К направлению акустического синтеза относится формантный синтез и синтез с использованием линейного предсказания. При создании акустических синтезаторов не ставится задача непосредственного отражения в синтезе процессов, связывающих артикуляцию с акустикой речевого сигнала, а вместо этого они воспроизводят в синтезируемом сигнале существенные для восприятия акустические характеристики естественной речи. В этом смысле акустический синтез является продолжением того направления, которое было начато созданием вокодеров и электрических параметрических синтезаторов разного типа [Кодзасов, Кривнова 2001].

1.2.3.1 Артикуляционный синтез

Артикуляционный (или артикуляторный) синтез в некоторой мере продолжил направление, заданное первыми механическими синтезаторами. В нём делается попытка синтезировать речевой сигнал на основе моделирования процесса речеобразования с учетом сведений об артикуляции, используемых для количественной оценки формы речевого тракта, его резонансных свойств и

характеристик звуковых источников. Затем на основе расчетных данных генерируется речевой сигнал [Кодзасов, Кривнова 2001]. В артикуляционной модели трубка, соответствующая речевому тракту, обычно разделяется на множество небольших секций, и таким образом она может быть представлена в качестве неоднородной электрической линии передачи [Фланаган 1968].

Первые электронные артикуляционные модели были статическими и требовали ручной настройки. Первый синтезатор американского исследователя Х. Данна 1950 года состоял из 25 одинаковых звеньев, между которыми для учёта влияния положения языка можно было ввести переменную индуктивность, а индуктивность на конце линии отражала влияние губ. Для произнесения вокализованных звуков синтезатор возбуждался пилообразным напряжением регулируемой частоты, а шумные звуки получались подключением белого шума к соответствующей точке линии [Фланаган 1968].

Первый артикуляционный синтезатор с динамическим контролем (рис. 7) DAVO (Dynamic Analog of the VOcal tract) был разработан Д. Розеном в 1958 году в Массачусетском технологическом институте. Он управлялся записанными на ленту контролирующими сигналами, созданными вручную [Lemetty 1999].

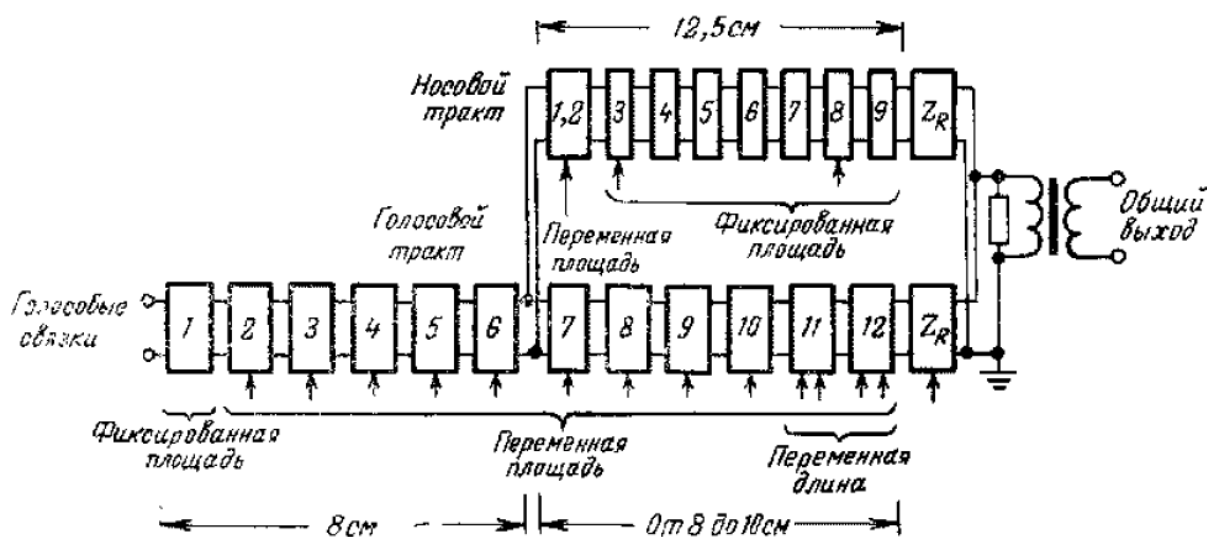


Рис. 7. Аналог речевого тракта с линией передачи, управляемый непрерывно [Фланаган 1968]

С течением времени артикуляционные синтезаторы развивались, в них вводилось дополнительное моделирование ослабления сигнала в речевом тракте, взаимодействия источника и фильтра, распространения сигнала от губ и, конечно, совершенствовалось моделирование голосового источника сигнала. Кроме этого, многие подходы включают моделирование движений и параметров мышц и управления моторикой. Однако из-за сложностей подобного моделирования в большинстве современных систем синтеза речи, позволяющих получать речь высокого качества, используются более «простые» подходы, а артикуляционный синтез чаще применяется в научных исследованиях в области артикуляционной фонетики и физиологии речи. Кроме этого, артикуляционный синтез непосредственно связан с областью аудиовизуального синтеза (или «говорящей головы»), задачей которого является построение визуальной модели головы и лица в процессе говорения [Taylor 2009].

1.2.3.2 Формантный синтез

Первым формантным синтезатором стал PAT (Parametric Artificial Talker) английского исследователя У. Лоуренса, представленный в 1953 году. Этот

синтезатор состоял из трёх электронных формантных резонаторов, соединённых параллельно, на вход которым подавался шум или гармонический сигнал. Синтезатор управлялся шестью временными функциями (три форманты, частота основного тона, амплитуда шума и амплитуда голосового источника), которые считывались с шаблонов, нарисованных на движущейся стеклянной дорожке [Klatt 1987]. Синтезатор Лоуренса был первым из параллельных формантных синтезаторов. Их главное преимущество состояло в относительной простоте управления. Вторым типом формантных синтезаторов стали каскадные синтезаторы (см. рис. 8), в которых формантные резонаторы были соединены последовательно, что позволило более точно моделировать передаточную функцию речевого тракта, но несколько усложнило их структуру [Klatt 1980].

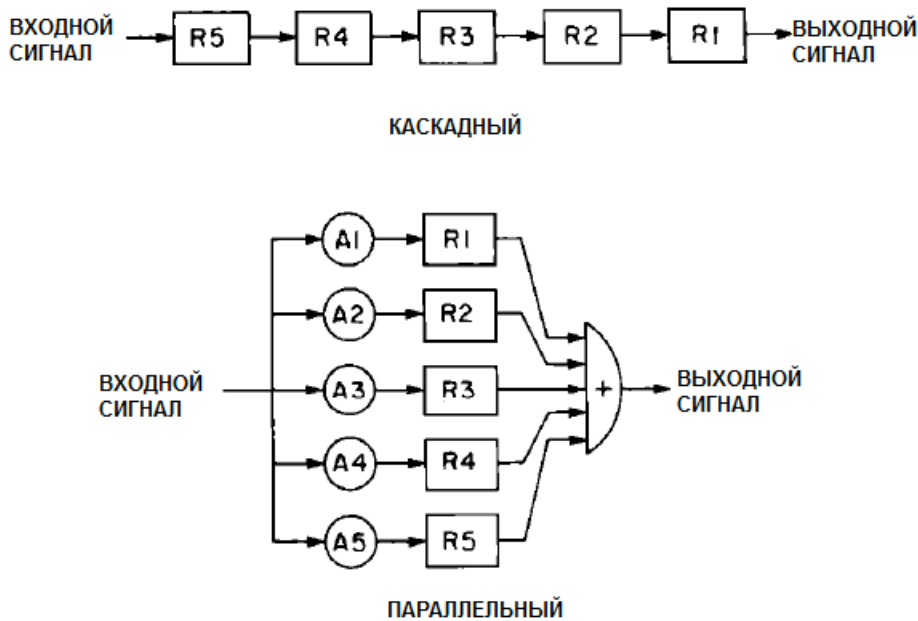


Рис. 8. Каскадный и параллельный синтезаторы. В параллельном синтезаторе амплитуда каждого формантного резонатора должна контролироваться отдельно. В каскадном – выходной сигнал каждого резонатора является входным сигналом следующего [Klatt 1980]

В том же 1953 году известный шведский исследователь речи, автор классической акустической модели речеобразования «источник-фильтр» Гуннар Фант продемонстрировал свой каскадный формантный синтезатор OVE I (Orator Verbis Electris). В нём частота двух нижних резонаторов контролировалась механической рукой, а амплитуда и частота основного тона определялись ручными потенциометрами [Klatt 1987].

В дальнейшем оба типа синтезаторов усложнялись и совершенствовались, позволяя каждой новой версии звучать всё ближе к естественной речи. В 1973 году английскому исследователю Дж. Холмсу удалось вручную настроить на своём синтезаторе (рис. 9) озвучивание предложения «I enjoy the simple life» так хорошо, что обычный слушатель не мог отличить его от произнесения того же текста человеком [Lemmetty 1999]. Однако оставалась существенная проблема с автоматическим контролем работы синтезатора, который не мог пока приблизиться к ручной настройке произнесения.

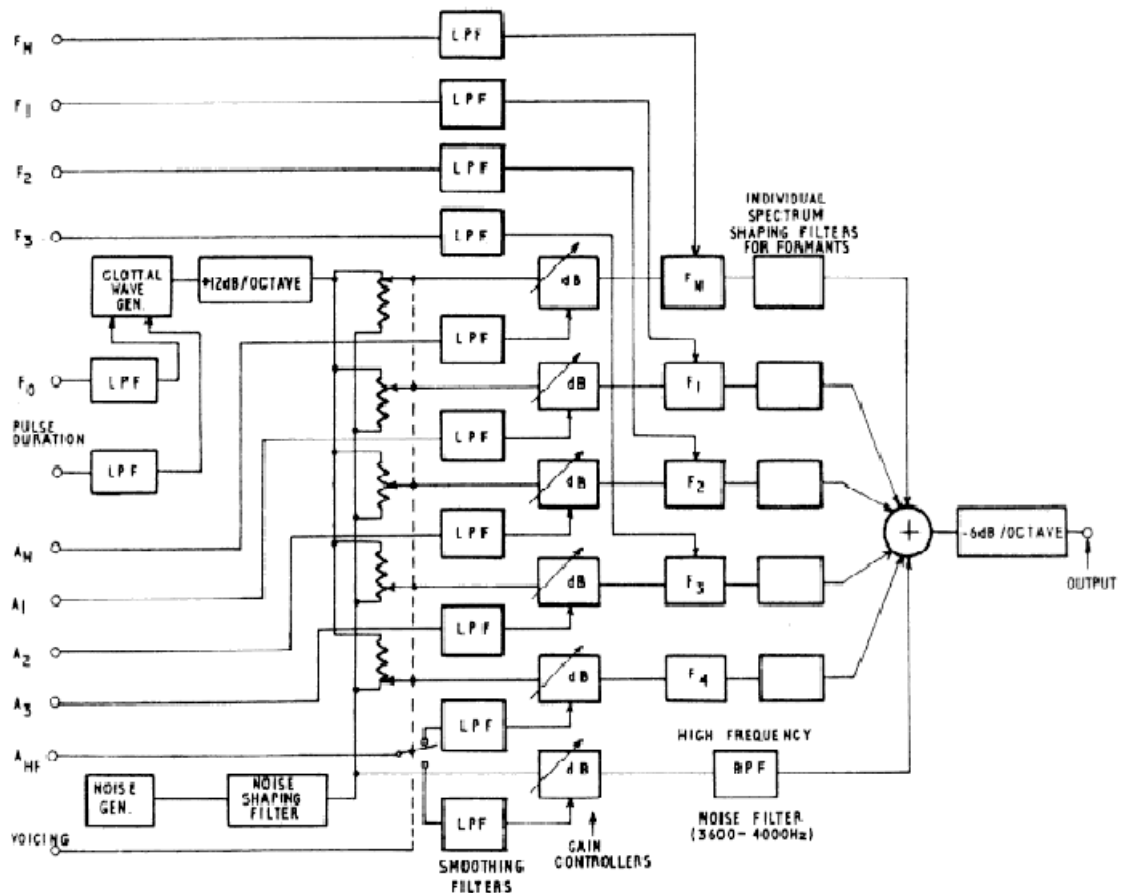


Рис. 9. Формантный синтезатор Холмса, состоящий из резонаторов для четырёх формант и носовой форманты, каждый из которых возбуждался вариативной смесью сигналов шумового и голосового источников [Klatt 1987]

С развитием компьютерной техники и появлением вычислительных машин в середине 50-х годов электрические аналоговые синтезаторы стали постепенно замещаться компьютерными программами или специально сконструированной цифровой аппаратурой, позволявшими работать с цифровым представлением речевого сигнала.

В 1972 году американский исследователь Д. Клатт предложил вариант гибридного формантного синтезатора, в котором сонорные и шумные звуки синтезировались каскадными и параллельными формантными резонаторами соответственно. Публикация исходного кода программы на языке Фортран в 1980

году позволила учёным в различных лабораториях оценить работу этого синтезатора, а также помогла в проведении перцептивных экспериментов [Klatt 1987].

Первая модель формантного синтезатора русской речи «Фонемофон-1» (рис. 10) была разработана в Минске в начале 70-х годов, а в его последующих версиях удалось добиться синтеза русской речи по произвольному тексту весьма высокого качества [Лобанов, Цирульник 2008].



Рис. 10. Внешний вид синтезатора «Фонемофон-1»

1.2.3.3 Синтезаторы, использующие линейное предсказание

Метод линейного предсказания позволяет использовать при синтезе искусственной речи параметры передаточной функции речевого тракта и является своеобразной альтернативой описанному выше аппаратно-цифровому формантному синтезу. Первые эксперименты с кодированием речи при помощи коэффициентов линейного предсказания (КЛП) были проведены в середине 60-х

годов. Эта технология впервые была использована в недорогих устройствах типа TI Speak'n'Spell (1980) [Lemmetty 1999].

Для синтеза речевого сигнала в КЛП-синтезаторе используются следующие изменяющиеся во времени параметры: период основного тона, средняя громкость звука, признак тон-шум и определённые заранее числовые коэффициенты линейного предсказания, связанные с передаточной функцией речевого тракта. При этом качество синтезированной речи зависит от числа коэффициентов, точности их вычисления, а также от того, насколько хорошо моделируются источники возбуждения [Обжелян, Трунин-Донской 1987]. В общем виде простейший цифровой КЛП-синтезатор имеет достаточно сложную структурную схему, представленную на рис. 11.

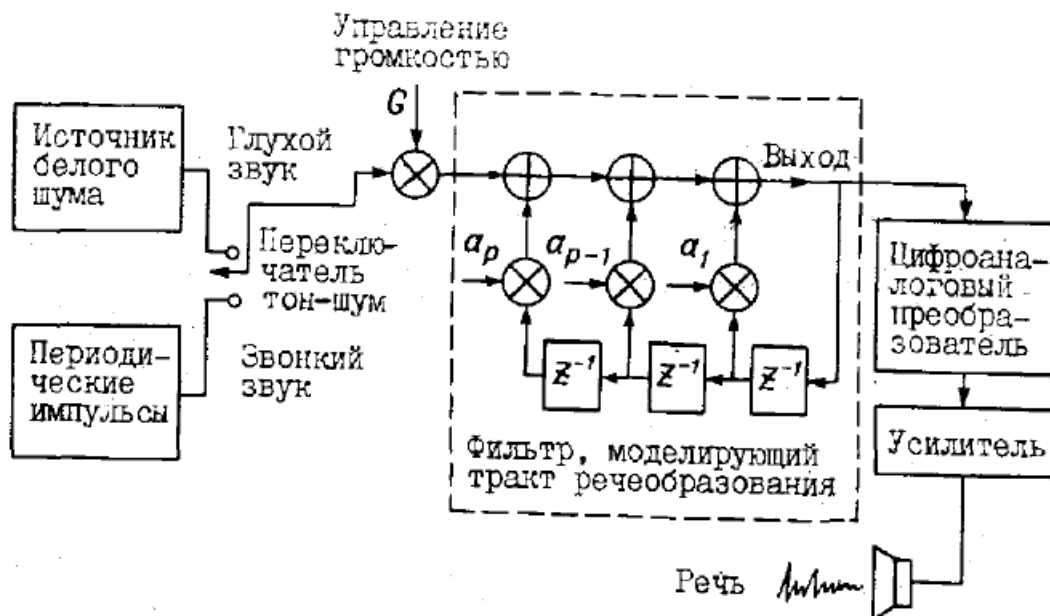


Рис. 11. Структурная схема КЛП-синтезатора [Обжелян, Трунин-Донской 1987]

Для работы КЛП-синтезатора по оцифрованной речи вычисляются требуемые параметры, а далее все необходимые единицы синтеза (слова или более короткие единицы) записываются в цифровом параметризованном виде в память и затем при синтезе извлекаются и соединяются, или конкатенируются в определённом порядке, затем преобразуются в звуковую волну и озвучиваются с использованием цифро-аналогового преобразователя. Таким образом, модель линейного предсказания косвенно способствовала развитию технологии конкатенативного синтеза речи.

Синтезаторы первого аппаратно-компьютерного поколения обычно требовали детального фонетико-акустического описания того, что должно быть произнесено, и не включали какого-либо автоматического способа получения подобного описания для произвольного сообщения или текста. Критериями качества оставались разборчивость речи и её сходство с живой человеческой речью.

1.2.4 XX век: синтезаторы второго поколения

В середине 60-х годов, в связи с продолжающимся развитием компьютерной техники и возросшими потребностями общества, перед разработчиками автоматического синтеза речи была поставлена более сложная задача озвучивания любого сообщения, вводимого в компьютер в текстовом виде и неизвестного заранее системе синтеза. Это привело к развитию синтезаторов типа «Текст–Речь» (Text-to-Speech или сокращённо TTS). В идеале такие устройства должны имитировать деятельность человека, который читает письменное сообщение или текст любой степени сложности [Кодзасов, Кривнова 2001]. Поэтому в синтезаторах такого типа (синтезаторах речи в современном понимании этого термина) появился блок лингвистической обработки, независимый от акустического блока и метода генерации речевого сигнала (рис. 12), тогда как

самые ранние синтезаторы и синтезаторы первого поколения были ориентированы в основном или полностью на модельную разработку акустического блока, то есть только на задачу генерации речевого сигнала, при сильном ограничении набора озвучиваемых речевых единиц и сообщений.

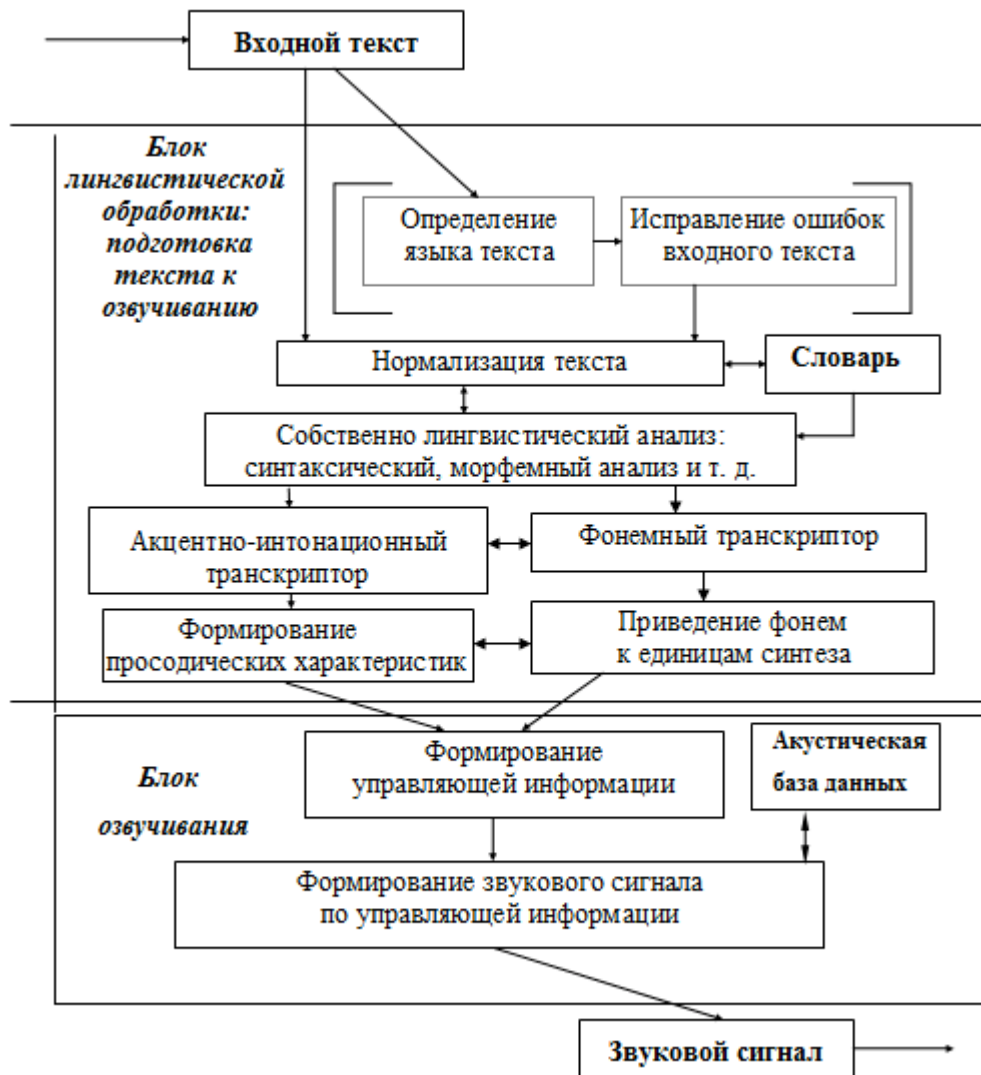


Рис. 12. Прототипическая схема системы синтеза речи [Кривнова 1998]

Первая полноценная система синтеза «Текст–Речь» для английского языка была создана в 1968 году в Японии исследователем Норико Умеда и его коллегами. Она была основана на артикуляционной модели акустического блока.

Лингвистический анализ текста и расстановка пауз производились при помощи сложных правил. По свидетельству специалистов, речь, производимая этой системой, была разборчивой, но довольно монотонной [Klatt 1987].

В дальнейшем алгоритмы лингвистической предобработки текста усложнялись благодаря увеличению скорости компьютерного анализа данных и объёма памяти для хранения вспомогательной лингвистической информации (различных словарей, речевых баз, моделей и т. п.). Это позволяло более точно представлять необходимые для акустического синтеза детальные фонетические описания: фонетическую транскрипцию и просодические характеристики сегментных единиц, получаемые на основе количественных интонационно-просодических моделей (длительность, частоту основного тона и интенсивность).

Следует подчеркнуть, что собственно фонетические описания, полученные на выходе лингвистического блока синтезатора, должны быть далее преобразованы в процессе синтеза во входные акустические данные, необходимые для блока генерации речевого сигнала (например, частоты формант), что может быть сделано двумя способами, в зависимости от используемого в этом блоке метода синтеза речевого сигнала: либо с помощью особых правил, либо посредством измерения (или «копирования») этих характеристик, полученных для отдельных звуков или целых фраз естественной речи. Копирование характеристик является наиболее простым и эффективным методом получения качественной (то есть разборчивой и естественной) синтезированной речи. Так называемый *ресинтез*, то есть подача на вход синтезатора акустических характеристик естественной речи, является также надёжным способом понять, насколько хорошо работает его акустический компонент.

1.2.4.1 Конкатенативный синтез

Конкатенативный синтез, называемый также техникой *второго* компьютерного поколения [Taylor 2009], смог появиться благодаря тому, что перед создателями систем синтеза уже не стояли такие жёсткие ограничения по объему доступной компьютерной памяти (как в 70-е и 80-е годы) и появилась возможность хранить большие массивы речевых данных. В отличие от систем первого поколения в них не используется классическая акустическая модель «источник-фильтр». Вместо этого в памяти компьютера хранятся фрагменты реальных акустических сигналов (либо в виде оцифрованных фрагментов звуковой волны, либо в параметризованной форме, полученной в результате акустического анализа исходных «живых» образцов), выделенных из речи определённого «диктора-донора». Из них путём «склейки» (или конкатенации) и создается первичная основа синтезируемого речевого сигнала. В дальнейшем эта основа подвергается модификации по правилам, функция которых состоит в том, чтобы придать склеенным фрагментам сигнала нужные просодические характеристики (или же модификация может производиться до склейки элементов) [Кодзасов, Кривнова 2001].

Различные системы конкатенативного синтеза используют в качестве базовых элементов для конкатенации звуковые единицы различного размера: фрагменты фонемной размерности: акустические аллофоны, полуслоги, слоги и образцы смешанных типов. Наиболее часто в таких системах используются дифоны – отрезки, начинающиеся в середине одного звука и заканчивающиеся в середине следующего. Дифоны как оптимальная единица для учёта акустических эффектов коартикуляции были впервые предложены американским исследователем Дж. Петерсоном с коллегами в 1958 году [Klatt 1987].

На качество речи, производимой конкатенативным синтезатором, влияет как качество и количество самих единиц для конкатенации (степень покрытия всех необходимых сегментных единиц), так и используемые алгоритмы просодической модификации речевого сигнала. Наиболее широко используемым методом модификации речи во временной и частотной области является алгоритм PSOLA (Pitch Synchronous Overlap and Add), разработанный в 1985 году, и его последующие варианты [Lemmetty 1999].

По современным меркам объём речевой базы для конкатенативного синтеза речи второго поколения является относительно небольшим, что позволяет построить синтезатор высокого качества довольно быстро. Однако главным недостатком систем такого типа является то, что они, в отличие от, например, формантного синтеза по правилам, не обладают достаточной гибкостью в изменении тембра голоса, так как для этого необходимо создавать новую базу акустических образцов для другого диктора-донора [Кривнова 1998].

Именно на этапе развития синтезаторов типа «Текст–Речь» стало возможным говорить об отдельной оценке различных модулей синтезатора: обработки текста, качестве просодического оформления синтезированной речи и т. п.

1.2.5 XX век: синтезаторы третьего поколения

К третьему поколению технологий автоматического синтеза речи обычно относят синтез на основе скрытых Марковских моделей и селективный синтез речи [Taylor 2009]. Их общей чертой является использование больших объёмов речевых данных, а также высокая естественность синтезированной речи.

1.2.5.1 Селективный синтез

В настоящее время доминирующей технологией автоматического синтеза речи является так называемый селективный синтез, оценка которого является

задачей данной работы. Он позволяет получать синтезированную речь, которая по своим характеристикам наиболее приближена к естественной [Taylor 2009]. Селективный синтез речи является разновидностью конкатенативного синтеза, то есть при генерации речевого сигнала используются заранее сделанные звукозаписи естественной речи. Первыми системами селективного синтеза стали n-Talk (1992) [Sagisaka et al. 1992] и CHATR (1994) [Black, Taylor 1994], а в 1996 году известные специалисты по синтезу речи А. Хант и А. Блэк предложили алгоритм выбора оптимальной последовательности единиц для конкатенации, который стал классическим [Hunt, Black 1996]. Подробнее метод селективного синтеза будет описан в разделе 2.1.

1.2.5.2 Статистический параметрический синтез

Статистический параметрический синтез, так же как и конкатенативный, является методом, основанным не на правилах, а на имеющихся акустических данных. Однако в отличие от конкатенативного метода, при котором необходимые для синтеза параметры речевого сигнала уже присутствуют в самих единицах конкатенации, сохраняемых в памяти компьютера, в статистическом параметрическом синтезе делается попытка машинного обучения синтезатора на имеющихся речевых данных с целью получения модели соответствия характеристик речи, поступающих на вход акустического блока, физическим параметрам целевых звуковых единиц. Полученная акустическая модель даёт два преимущества: уменьшение объема памяти для хранения модели вместо самой речевой базы и возможность её параметрической модификации, например, быстрого изменения тембра голоса [Taylor 2009].

Наиболее распространённой техникой в данном направлении синтеза является метод, основанный на использовании скрытых Марковских моделей (СММ, англ. HMM – hidden Markov models). Акустические параметры,

сгенерированные СММ, используются далее для управления вокодером. Это означает, что для порождения речевого сигнала используются в конечном итоге параметры речевого тракта и параметры возбуждения [Tokuda et al. 2013]. Скрытые Марковские модели звуковых единиц применялись в системах распознавания речи с конца 70-х годов и применяются до сих пор. Работу над автоматическими системами синтеза речи, основанными на СММ, начали в 1995 году японские учёные К. Токуда и его коллеги [Tokuda et al. 1995]. Использование статистического подхода для синтеза речи стало возможным благодаря возросшему быстродействию вычислительных машин и увеличению объёмов носителей информации для хранения больших речевых баз, необходимых для обучения статистических акустических моделей.

Приведём общую схему системы синтеза речи, основанной на СММ (рис. 13).

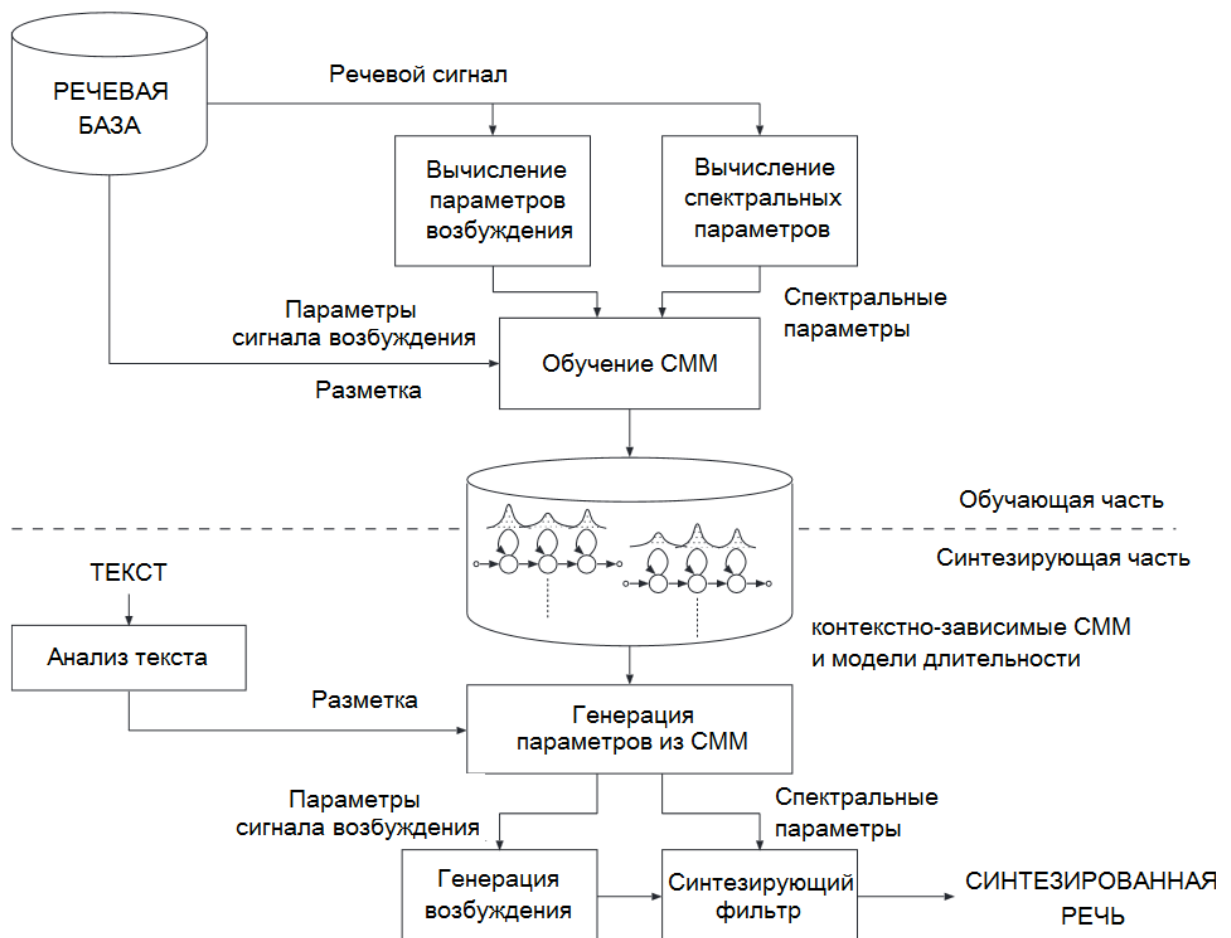


Рис. 13. Общая схема системы синтеза речи, основанной на СММ [Black et al. 2007]

Статистический параметрический синтез обеспечивает несколько меньшую естественность звучания синтезированной речи по сравнению с селективным синтезом из-за так называемого эффекта роботизированности речи. Однако, несмотря на это, данный подход обладает несколькими несомненными преимуществами:

1. Он позволяет легко модифицировать характеристики голоса с помощью адаптации или интерполяции акустических моделей дикторов, в то время как селективный алгоритм позволяет получить только такую синтезированную

речь, стиль и голос которой не отличается от стиля и голоса используемой речевой базы.

2. В речевом сигнале, порождаемом статистическим параметрическим синтезатором, отсутствуют резкие, не обусловленные контекстом перепады по частоте и энергии, обычно присущие конкатенативному синтезу. Кроме этого, при селективном синтезе качество речи может существенно ухудшиться в случае отсутствия подходящего звукового элемента в базе данных. При использовании акустических моделей звуковые элементы, отсутствующие в обучающей выборке, синтезируются на основе средних значений, максимально приближенных к требуемым, благодаря применению технологии кластеризации контекстов, основанной на деревьях решений. Это позволяет добиться хорошей разборчивости синтезированной речи в условиях ограниченного количества звуковых единиц в различных контекстах.
3. Статистический подход при использовании современных алгоритмов машинного обучения систем для построения акустических звуковых моделей позволяет разрабатывать новый голос синтезатора за гораздо меньший промежуток времени и требует значительно меньше памяти для хранения речевой базы синтеза.

1.3 Перспективы дальнейшего развития синтеза речи

Основными направлениями современных исследований в области автоматического синтеза речи являются аудиовизуальный синтез, синтез экспрессивной и эмоциональной речи, а также объединение двух подходов к синтезу речи третьего поколения: селективного синтеза и синтеза на основе скрытых Марковских моделей [Taylor 2009] – так называемый гибридный синтез. Наиболее часто в гибридном синтезе от селективного синтеза берётся та часть,

которая отвечает за подбор и соединение элементов конкатенации из речевой базы, поскольку соединяются элементы естественной речи – тем самым и синтезированная речь звучит более естественно. Значения физических параметров (длительности, энергии и ЧОТ звука), необходимые для оптимального селективного подбора элементов конкатенации, определяются не по созданным вручную правилам, а порождаются автоматически моделями, обученными на той же речевой базе, что позволяет быстро настроить просодическое оформление речи на нового диктора [Chistikov et al. 2013; Чистиков и др. 2013].

1.4 ВЫВОДЫ К ГЛАВЕ 1

В первой главе настоящей работы описана теоретическая база, лежащая в основе технологии синтеза речи, а также история создания синтезаторов речи и требований к их качеству. Обсуждается современное состояние данной речевой технологии, перспективы его развития и изменения в оценках качества синтезированной речи. Можно сделать вывод, что в уже довольно длительной истории технологии синтеза речи значительно менялись приоритеты и направления исследований. Это связано и с целями, которые ставились перед синтезаторами: от демонстрации возможности получения звуков, подобных человеческой речи, и моделирования процессов речеобразования до получения разборчивого, а затем и естественного, выразительного чтения компьютером произвольного текста. Нельзя не отметить также, что история и успехи разработок в области синтеза речи тесно связаны с развитием техники, в частности компьютерной, и других научных дисциплин: физики (механики, электродинамики, акустики), математики (статистики), информатики, физиологии, психологии и, конечно же, лингвистики (фонетики, автоматической обработки естественного языка).

ГЛАВА 2. СЕЛЕКТИВНЫЙ СИНТЕЗ РЕЧИ

В настоящее время системы, основанные на так называемой технологии unit selection или, иными словами, технологии селективного синтеза речи, обеспечивают получение наиболее естественной синтезированной речи. В данной главе рассматриваются принципиальные особенности и процесс выбора звуковых элементов из речевой базы синтезатора, приводится краткий обзор современных русскоязычных систем синтеза, а также подробно описывается общая структура современного селективного синтезатора речи.

2.1 Алгоритм Unit selection

Как было отмечено выше, селективный синтез речи является разновидностью конкатенативного синтеза, то есть при генерации речевого сигнала используются заранее полученные звукозаписи естественной речи. В отличие от более ранних аллофонных или дифонных синтезаторов речи, порождающих итоговый речевой сигнал из отдельных и специально подготовленных звуковых единиц, выделенных из небольшого и тщательно подобранного набора озвученных слов, при селективном синтезе для каждой целевой единицы синтеза производится выбор наиболее подходящего кандидата из множества вариантов, взятых из озвученных диктором предложений естественного языка. Для этого записываются специальные речевые базы, размер которых может достигать нескольких десятков часов звучания. [Black 2002]. В процессе акустического синтеза алгоритм селекции (выбора) строит оптимальную последовательность звуковых единиц, выбранных из речевой базы, (рис. 14), учитывая одновременно и то, насколько кандидат подходит под описание необходимых характеристик целевого звука (стоимость замены), и то, насколько хорошо выбранные элементы будут конкатенироваться с соседними (стоимость

связи). При этом с учетом указанных стоимостей из базы в качестве оптимальных могут быть выбраны не отдельные звуки, а их цепочки или даже целые предложения. Такой подход позволяет минимизировать необходимость модификаций речевого сигнала (или даже полностью от них отказаться), что повышает естественность синтезируемой речи.

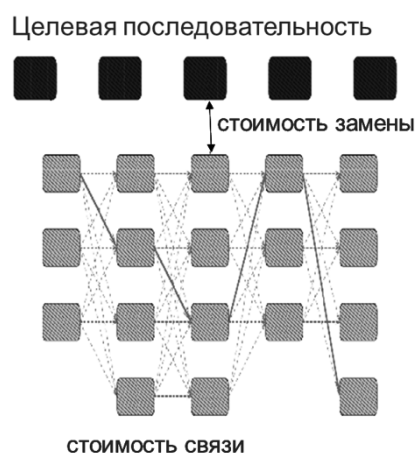


Рис. 14. Выбор целевой звуковой последовательности при селективном синтезе речи

В то же время именно из-за тенденции к минимизации акустических модификаций одной из особенностей селективного синтеза является возможность частичного или даже полного несоответствия характеристик выбираемых единиц-кандидатов целевым характеристикам, необходимым для синтеза. Поэтому при тестировании селективного синтеза особенно важным является отдельное тестирование лингвистической обработки текста в целях его дальнейшего озвучивания и собственно акустического модуля синтеза выходного речевого сигнала.

При селективном синтезе происходит выбор групп наиболее подходящих звуковых элементов из базы синтезатора на основе значений акустических характеристик, вычисленных для каждого аллофона [Clark et al. 2007]. Для того чтобы определить, насколько тот или иной элемент базы подходит для синтеза

целевой звуковой единицы, вводятся понятия стоимости замены (target cost) и стоимости связи (join/concatenation cost).

Стоимость замены для элемента из базы u_i по отношению к целевому элементу t_i вычисляется по формуле:

$$C^t(u_i, t_i) = \sum_{k=1}^p w_k^t C_k^t(u_i, t_i), \quad (1)$$

где:

C_k^t — расстояние между k -ыми характеристиками элементов (способ его вычисления зависит от конкретной характеристики)

w_k^t — вес k -ой характеристики (может подбираться опытным путём или настраиваться автоматически).

Иными словами, стоимость замены равна взвешенной сумме различий в признаках между целевым элементом и конкретным элементом речевой базы. В качестве признаков могут выступать любые просодические и лингвистические характеристики элементов. Как правило, используется следующая информация: частота основного тона (ЧОТ), длительность, контекст, позиция элемента в слове, слове, количество ударных слогов во фразе и др.

Выбранные из базы элементы должны не только мало отличаться от целевых, но и хорошо соединяться (конкатенироваться) друг с другом. Функция стоимости связи двух элементов в цепочке может быть определена как взвешенная сумма различий в признаках между двумя последовательно выбранными кандидатами:

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^q w_k^c C_k^c(u_{i-1}, u_i), \quad (2)$$

где:

C_k^c — расстояние между k-ми характеристиками элементов,

w_k^c — вес для k-ой характеристики.

Общая стоимость связи для целой последовательности из n элементов равна сумме введенных выше стоимостей:

$$C(u, t) = \sum_{i=1}^n C^t(u_i, t_i) + \sum_{i=2}^n C^c(u_{i-1}, u_i) \quad (3)$$

Задача алгоритма unit selection состоит в том, чтобы выбрать такое множество элементов $u_1, u_2, \dots, u_n$, которое бы минимизировало общую стоимость полученной цепочки согласно формуле (3).

В различных селективных синтезаторах могут варьироваться минимальные звуковые элементы базы: это могут быть аллофоны, дифоны, полуфоны, слоги и т. д. Также в разных селективных синтезаторах может использоваться различный набор характеристик, по которым подбираются эти элементы, способы и алгоритмы настройки весов для них [Vepa 2004; Vepa, King 2004].

На заключительном этапе синтеза происходит объединение выбранной последовательности элементов в звуковой поток, представляющий собой синтезированную речь. Структура современного селективного синтезатора типа «Текст–Речь» описывается подробно в разделе 2.3.

2.2 Русскоязычные селективные синтезаторы

Селективные синтезаторы для русского языка начали разрабатываться в 2000-х годах. В настоящий момент существует довольно много разнообразных синтезированных коммерческих и бесплатных синтезированных голосов, работающих с различными операционными системами. Приведём таблицу (табл. 1)

основных характеристик наиболее известных разработок, большая часть которых (кроме Mary TTS) является коммерческими системами, в которых для прослушивания и оценки синтеза речи доступны только онлайн демо-версии с различными ограничениями.

Таблица 1. Основные селективные синтезаторы для русского языка

Система синтеза	Доступные голоса	Онлайн-демо	Комментарии
Acapela Group (Бельгия, Франция, Швеция)	Alyona	http://www.acapela-group.com/voices/demo/ (не более 300 символов)	—
Apple (США)	Siri	Нет демо для Windows, работает под iOS как голос персонального помощника	—
Google Translate (США)	Женский голос	https://translate.google.com/#ru	Озвучивание переводимого текста. Чтение очень замедлено, слышна модификация речи
iSpeech (США)	Женский голос; мужской голос	http://www.ispeech.org/text.to.speech (не более 150 символов)	Есть возможность регулирования скорости произнесения (3 варианта)
Ivona TTS (Польша, США)	Tatyana; Maxim	http://www.ivona.com/ (не более 250 символов)	—
Mary TTS (Германия, бесплатное открытое ПО)	Мужской голос	http://mary.dfki.de:59125/	Есть возможность включения/отключения просодической модификации. Не читает числовые записи и иноязычные вставки
Microsoft (США)	Irina	Экранный диктор в ОС Windows	—
Nuance Loquendo (Италия, США)	Olga; Dmitri	http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/inbound-solutions/loquendo-small-business-bundle/interactive-tts-demo/index.htm (не более 500 символов, фоновая музыка)	—

Nuance Vocalizer (США)	Katya; Milena; Yuri	http://www.nuance.com/landing-pages/playground/Vocalizer_Demo2/vocalizer_modal.html?demo=true	–
ReadSpeaker (Нидерланды, Бельгия и др.)	Женский голос	http://www.readspeaker.com/ (не более 250 символов; фоновая музыка; добавление фразы про демо)	–
Svox (Швейцария, США)	Katja; Yuri	http://svoxmobilevoices.wordpress.com/demos/ (только образцы синтеза) Нет интерактивного демо (для Windows), работает под ОС Android	–
Tingwo (Швейцария)	Женский голос; Мужской голос	http://www.tingwo.co/en/interactive-tts-text-to-speech-demo (не более 200 символов; фоновая музыка)	–
VitalVoice TTS (ООО «ЦРТ») (Россия, С.-Петербург)	Юлия; Владимир; Анна; Виктория; Александр; Мария; Лидия	http://www.speechpro.ru/technologies/synthesis (не более 200 символов); прослушивание через «голосовые открытки» http://cards.voicefabric.ru/ (фоновая музыка)	–

2.3 Структура современного селективного синтезатора речи типа

«Текст–Речь»

Как будет показано ниже в главе 3, оценка качества работы синтезаторов речи часто выполняется для отдельных этапов преобразования текста в речь (иными словами, модулей синтезатора). Таким образом, необходимо дать краткое описание примерной структуры селективных синтезаторов.

Ниже на рис. 15 приводится примерная общая схема современного селективного синтезатора речи на русском языке. Конкретные реализации системы могут различаться, но, тем не менее, основные компоненты в том или ином виде должны присутствовать в любой системе, при этом они могут объединяться в различные модули, взаимодействующие друг с другом. Компоненты синтезатора,

отвечающие за преобразование текста в речь, которые работают до акустического процессора (лингвистическая обработка текста) не являются специфичными именно для селективного синтеза речи, а используются при синтезе речи любого типа (в синтезаторах типа «Текст–Речь»). Для удобства описания разделим блок лингвистической обработки на три части: 1 – собственно лингвистическая обработка (нормализация текста, расстановка ударений); 2 – просодическая обработка (определение локализации пауз и типов интонационных конструкций); 3 – фонетическая обработка (построение сегментной транскрипции и задание просодических параметров для интонации).



Рис. 15. Общая структура селективного синтезатора речи

В общем случае на вход программы синтеза поступает текст без каких-либо специальных помет, и синтезатор должен прочитать его. При этом синтезированная речь должна быть подобна чтению текста, а не спонтанной речи.

Многие современные синтезаторы позволяют обрабатывать специальные пользовательские обозначения ударений в словах, пауз и других элементов, а также вручную пополнять словари и т. п. Существует международный стандарт SSML (Speech Synthesis Markup Language¹), используемый для передачи управляющей информации синтезатору речи (выбор голоса, расстановка и длительность пауз, регулировка темпа произнесения и т. п.), который поддерживается большинством современных систем синтеза речи.

2.3.1 Блок лингвистической обработки

Текст, который должен озвучить синтезатор, может быть по-разному отформатирован, содержать различные специальные символы, формулы, рисунки и т. п. Всё это должно определённым образом учитываться при чтении текста. Мультиязычный синтезатор речи может также включать в себя модуль определения языка текста, также в этот этап может быть включено автоматическое исправление ошибок во входном тексте.

В первую очередь синтезатор речи должен определить границы слов, выделить знаки препинания, найти границы предложений, и, при необходимости, разделить текст на абзацы. Деление текста на предложения может быть затруднено из-за употребления различных сокращений с точкой или прямой речи. На этапе нормализации текста выполняется также расшифровка специальных обозначений: цифр, специальных символов и т. п.

Одной из главных проблем при обработке текста для дальнейшего озвучивания (при выборе места ударения в слове, формы числительного при расшифровке и т. п.) является неоднозначность, для русского языка возникающая довольно часто и по разным причинам.

¹ <http://www.w3.org/TR/2007/WD-speech-synthesis11-20070110/>

На этапе нормализации текста и его лингвистической обработки обычно используются словари и наборы правил. Могут также применяться и различные статистические методы, однако при этом необходимы обучающие текстовые корпуса значительного объёма, содержащие различные обозначения и их расшифровки, снятую омонимию и т. п. Полный синтаксический и семантический анализ текста является весьма трудоёмкой задачей, но для задач синтеза речи хорошо подходит и частичный синтаксический анализ, при котором анализируется ближайший контекст конкретного проблемного слова. Однако при этом могут возникать сложные случаи, для анализа которых требуется синтаксическая и семантическая обработка всего предложения.

2.3.1.1 Расшифровка сокращений и специальных знаков

Приведём общую классификацию нестандартных текстовых элементов, требующих расшифровки, созданную О. Д. Черепановой для русского языка [Черепанова 2015]:

Таблица 2. Общая классификация нестандартных текстовых элементов в русских текстах

Название класса	Контекстная зависимость	Примеры
1. Аббревиатуры		
1.1. Сложносокращенные слова	-	исполком; завлаб
1.2. Графические сокращения	+/-	филол.; т.е.; оз.; 60 км/ч; 20 кг; пр-т
1.3. Инициальные аббревиатуры		
<i>1.3.1. буквенные</i>	-	МГУ; СНГ
<i>1.3.2. звуковые</i>	-	МГИМО; ГУМ

1.3.3. <i>буквенно-звуковые</i>	-	ЦСКА; ГИБДД
1.4. Смешанные сокращения	-	БелАЗ
2. Объекты, содержащие цифры		
2.1. количественные числительные	+	12 домов
2.2. порядковые числительные	+	12-ый том; Иван IV; 1. ...; 2. ...
2.3. номера (не телефонные)	-	ауд. 956
2.4. телефонные номера	-	8 (495) 123 45 67; 123-45-67
2.5. адреса	-	д.1, к.2, кв.123; д.2/3
2.6. индексы	-	123456
2.7. указание времени	+	в 13:45; к 13:45
2.8. даты	+/-	2.05.06; 02/05; 2 мая 2006г.
2.9. годы	+	2001г.; 2010/11 гг.; 60-е
2.10. денежные суммы	+	\$ 1,5; 1.20 руб; 12 тыс. €
2.11. проценты	+	29,99%; 50%
2.12. дроби	+	1/4 финала; 2/3 опрошенных
2.13. номера серии	-	Android 2.3, § 1.2.3.
2.14. собирательные числительные	-	5-ро друзей; 2-е суток
2.15. мультипликативные числительные	+	3-ной подогрев; 4-ной сальхов
2.16. мультипликативные конструкции «число (р. п.) + прилагательное»	-	11-метровый; 4-кратный; 2,0-литровый
2.17. счетные числительные	-	16-ричный режим; 2-ичная

		нумерация
3. Спецсимволы		
<i>3.1. Контекстнозависимые</i>		
3.1.1. Символы валют	+/-	\$, €, ¥
3.1.2. Символы «§»; «№»; «%»	+	§12, №12, 12%
3.1.3. физические величины («°», «"»;	+/-	12°C; 12° по Цельсию; экран 6".
<i>3.2. Не зависящие от контекста</i>		
3.2.1. математические символы	-	+; -; ±; ≤; <; >; *; /; ^; ²; =; ...
3.2.2. символы валют + «тыс.», «млн», «трлн» и проч.	-	\$ 15 тыс.; 15 млрд руб.
3.2.3. Символы «&»; «#»; «@»; «~»; «x»	-	Маркетинг&Реклама; #100; 4x3; ~25%
3.2.4. сноски	-	[1], (1), ² , *, **
4. Смешанные объекты	+/-	№38-ФЗ;
5. Иностранные слова (на латинице)	-	Windows; microSD; MP3-плеер

При озвучивании нестандартных слов и обозначений может возникать целый ряд проблем. Многие сокращения являются неоднозначными (например, «м.» может обозначать «метр» или «метро», «г.» – «город», «год», «гражданин» и т. д.).

Для некоторых текстовых элементов (слова, написанные латиницей, чтение обозначений времени, телефонов) возможно несколько вариантов прочтения. При этом большинство вставок на латинице в современных русскоязычных текстах являются словами английского языка. Проблеме озвучивания английских слов и

словосочетаний на основе практической транскрипции посвящена вышеупомянутая работа [Черепанова 2015].

Расшифровка цифровых записей может осуществляться в несколько этапов:

1. Выделение специальных форматов (дата, время, телефон и т. п.).
2. Определение разряда числительного (количественное или порядковое).
3. Определение формы числительного (падеж, род).

Синтезатор также должен уметь правильно интерпретировать римские цифры (они должны быть переведены в арабские).

2.3.1.2 Расстановка ударений

Для выбора места словесного ударения в русских словах используется словарь. Выбор места ударения для несловарных слов может осуществляться при помощи набора правил или статистических методов.

Одной из основных проблем на данном этапе является выбор места ударения в омонимах, различающиеся произношением (омографах). Такие слова могут различаться местом ударения и/или наличием букв «ё»/«е», подробно проблема ё-омографов описана в статье [Лобанов 2009].

Омографы могут иметь одинаковые грамматические признаки («замОк» – «зАмок») либо различаться грамматическими характеристиками, ср.:

- Омонимичные формы внутри одной парадигмы (например, род. п. ед. ч. – им. п. мн. ч.: «облакА» – «Облака», «странЫ» – «стрАны» и т. п.).
- Омонимичные формы разных парадигм (например, существительное-инфинитив: «вестИ» – «вЕсти», «пропАсть» – «прОпасть»).

Омографы могут существенно различаться по частотности («ухА» – «Уха», сорокА – сорОка, кредИт – крЕдит, моЮ – мОю и т. п.), что важно учитывать при выборе нужного варианта.

В общем случае разрешение омонимии требует более глубокого анализа контекста. Он может производиться как на уровне индивидуальных слов (анализ слов, стоящих непосредственно рядом с текущим: «скрыто за семью замками»; поиск ключевых слов в том же предложении: «Дверь была заперта на необычный замок»), так и на уровне классов словоформ – при помощи анализа грамматического окружения и поиска согласованных слов в предложении. При этом могут использоваться грамматические правила, увеличивающие вес словоформы в зависимости от ее окружения.

Примеры реализации лингвистической обработки текстов в русскоязычных синтезаторах речи приведены в работах [Хомицевич и др. 2013; Гецэвіч 2012].

2.3.2 Блок просодической обработки

На этом этапе производится определение границ синтагм, локализации и длительности пауз, выбор интонационного типа фраз и места фразового ударения в них.

2.3.2.1 Определение локализации и длительности пауз

При определении места пауз нельзя полагаться только на знаки препинания: пауз может быть больше, а запятые, например, не всегда сигнализируют о паузе (к примеру, в предложениях с вводными словами). Существующие способы выбора места пауз можно разделить на следующие группы [Кривнова, Чардин 1999; Khomitsevich, Chistikov 2013]:

1. Определение мест пауз и границ синтагм по правилам (например, определение мест пауз по знакам препинания с заданными исключениями и определение

возможных мест пауз на длительных участках текста без знаков препинания). Этот метод является довольно трудоёмким, но в целом для русского языка может давать хорошие результаты. При таком подходе сложно учесть и новые случаи, которые ещё не предусмотрены правилами.

2. Определение места пауз при помощи полного синтаксического анализа предложений (здесь также требуются правила).
3. Определение места пауз при помощи статистических методов. При этом могут возникать грубые ошибки, нужен большой корпус для тренировки статистических моделей.
4. Статистические методы, дополненные и ограниченные правилами: например, статистика используется для отрезков без знаков препинания, а между определёнными словами паузы запрещаются правилами.

Длительности пауз могут задаваться правилами или при помощи статистики в зависимости от типа паузы (с учётом значений длительностей пауз у конкретного диктора).

2.3.2.2 Определение интонационного типа синтагм и места фразового ударения

На данном шаге может производиться интонационная транскрипция: определение интонационного типа синтагм и места фразового и эмфатического ударения. В зависимости от принятой системы интонационной транскрипции правила могут быть более или менее сложными, но, в общем случае, они основываются на анализе знаков препинания (наиболее простой вариант) или использовании полного/частичного синтаксического и семантического анализа предложения. Могут быть также использованы различные статистические методы, для обучения которых нужна текстовая база, заранее размеченная интонационной транскрипцией.

2.3.3 Блок сегментной фонетической обработки

На данном этапе выполняется построение сегментной транскрипции текстовых фрагментов (преобразование «буква–фонема») и задание физических параметров (частота основного тона, громкость, длительность) звуковых элементов.

2.3.3.1 Транскрибирование текста (сегментный компонент)

Для русского языка в большинстве систем синтеза речи буквенно-звуковое преобразование делается при помощи контекстных правил [Кузнецов, Отт 1989; Кривнова и др. 2001(а); Богданова 2001; Лобанов, Цирульник 2008], так как оно даёт возможность успешной автоматической обработки большинства слов с очень небольшим количеством исключений. Однако возможен также подход, предполагающий самообучение системы транскрибирования на основе данных. В рамках этого подхода также существует множество направлений (обзор см. в [Потапова 2015; Широкова 2014]).

При этом набор звуковых единиц, используемый при транскрибировании текстов, обычно основывается на стандартном для данного языка инвентаре фонем с возможным добавлением звуков, не имеющих статуса фонем, но отражающих речевую реальность (например, в заимствованиях). Для синтеза речи может также потребоваться более детальная фонетическая транскрипция. Правила транскрипции могут быть разделены на блоки, выполняемые последовательно, при этом более частные правила должны выполняться до общих. Для качественного буквенно-звукового преобразования необходимо также использовать различные словари исключений, задающие нестандартные произнесения слов, которые невозможно получить при помощи канонических правил (ср., например, фонетику заимствований).

2.3.3.2 *Определение просодических характеристик звуковых единиц*

После получения фонетической транскрипции текста, подаваемого на акустический блок синтезатора, нужно также задать необходимые просодические характеристики для каждой звуковой единицы текста, а именно:

- длительность;
- высоту голоса (ЧОТ);
- громкость (интенсивность).

Существуют различные методы формирования интонационно-просодического контура фразы в системах синтеза речи. В первую очередь их можно подразделить на две большие группы: ресинтез просодических характеристик синтезируемой фразы по готовому образцу (образцам) и параметризация на основе фонетико-акустических реализационных правил (по просодическим правилам) [Кривнова 2000].

К первой группе (помимо простого «копирования» интонационного контура известных фраз, заранее прочитанных диктором) также относятся системы, основанные на так называемых методах стилизации тонального контура – акустических или перцептивных.

К наиболее известным методам генерации интонационного контура (наиболее интересной и сложной задачей является генерация тонального контура фразы, т. е. изменения значений ЧОТ для звуков синтезированной речи) для произвольных предложений относятся следующие модели [Лобанов, Цирульник 2008]:

1. Автосегментная модель [Silverman et al. 1992].
2. IPO-модель [‘t Hart et al. 1991].

3. Суперпозиционная модель (модель Фуджисаки) [Fujisaki 1983].
4. Непрерывная параметрическая модель [Taylor 2000].
5. Модель портретов акцентных единиц (ПАЕ-модель) [Лобанов 1991; Lobanov 1987].

Все указанные методы требуют настройки под характеристики речи определённого диктора, чтобы учесть характерный темп речи, диапазон голоса и другие индивидуальные особенности, что особенно важно для селективного синтеза, так как необходимые для синтеза звуковые единицы с заданными интонационными параметрами должны присутствовать в самой речевой базе в достаточном количестве. Настройка параметров интонационных моделей может производиться вручную или при помощи статистических методов с использованием размеченной речевой базы.

2.3.5 Блок акустической обработки

На этом этапе выполняется выбор элементов из речевой базы синтезатора (Unit selection) и, при необходимости, модификация полученного речевого сигнала.

2.3.5.1 Выбор звуковых единиц из речевой базы

В селективном синтезе происходит выбор и конкатенация звуковых элементов из предварительно записанной речевой базы на основе значений акустических характеристик, вычисленных для каждого звукового элемента на предыдущем этапе, с учётом их весовых коэффициентов и сочетаемости выбранных элементов (подробное описание алгоритма Unit selection приведено выше в п. 2.1). При этом качество синтеза в большой степени зависит от характеристик используемой речевой базы (качества записи, постоянства темпа и тембра голоса и т. п.) и её объёма (далее в главе 4 описывается эксперимент по оценке синтеза с использованием речевых баз различного объёма). Имеющийся опыт показывает, что качественный синтез возможен только на основе

представительного, сбалансированного и корректно размеченного речевого корпуса.

С ростом объема корпуса в полной мере может быть учтена темповая и интонационная вариативность речи диктора. Иными словами, чем больше корпус, тем больше вероятность того, что в нем найдется элемент в необходимом контексте, с необходимой длительностью и контуром частоты основного тона (ЧОТ). Как следствие, в синтезированной речи будет меньше искажений от вынужденной модификации сигнала, а значит повысится её естественность.

В целом, использование корректно размеченного, сбалансированного корпуса является необходимым условием для получения синтезируемой речи высокого качества. Известно, что качество селективного синтеза не является постоянной величиной и зависит от озвучиваемого текста. Это свойство заложено в самой технологии. Действительно, когда выходной сигнал синтезатора составлен из оригинальных (немодифицированных) крупных фрагментов непрерывной речи, то качество речи практически совпадает с естественной. С другой стороны, когда требуется синтезировать речь по произвольному тексту, фрагменты которого представлены в корпусе лишь отдельными аллофонами, качество синтеза решающим образом определяется точностью разметки речевой базы. Ошибки в разметке обычно приводят к тому, что никакие дальнейшие усилия, связанные с модификацией речевого сигнала, не в состоянии сделать синтезированную речь близкой к естественной [Tatham, Morton 2005: 99].

Обычно объём речевых корпусов для селективного синтеза речи варьируется от 2 до 15 часов речи на одного диктора. При создании речевого корпуса могут использоваться специально подготовленные фонетически представительные тексты, списки частотных слов, числа, аббревиатуры, тексты с учётом области использования синтезированной речи (ср., например, обеспечение покупки

железнодорожных билетов, объявления остановок в маршрутном транспорте и т. п.) [Solomennik, Chistikov 2012]. Подбор оптимального текстового материала для речевого корпуса является отдельной интересной задачей [Кривнова и др. 2001(b); Chevelu et al. 2008; van Santen, Buchsbaum 1997]. Что касается записи речевого материала, то предпочтителен контроль записи диктора в студии для обеспечения единообразного нейтрального состояния голоса (для реализации эмоционального селективного синтеза нужны соответствующие базы эмоциональной речи диктора-донора). Процедуру подготовки речевой базы для синтезатора можно автоматизировать [Соломенник и др. 2013(b)]. На рис. 16 приводится примерная схема подготовки речевой базы для селективного синтезатора речи [Продан и др. 2010].

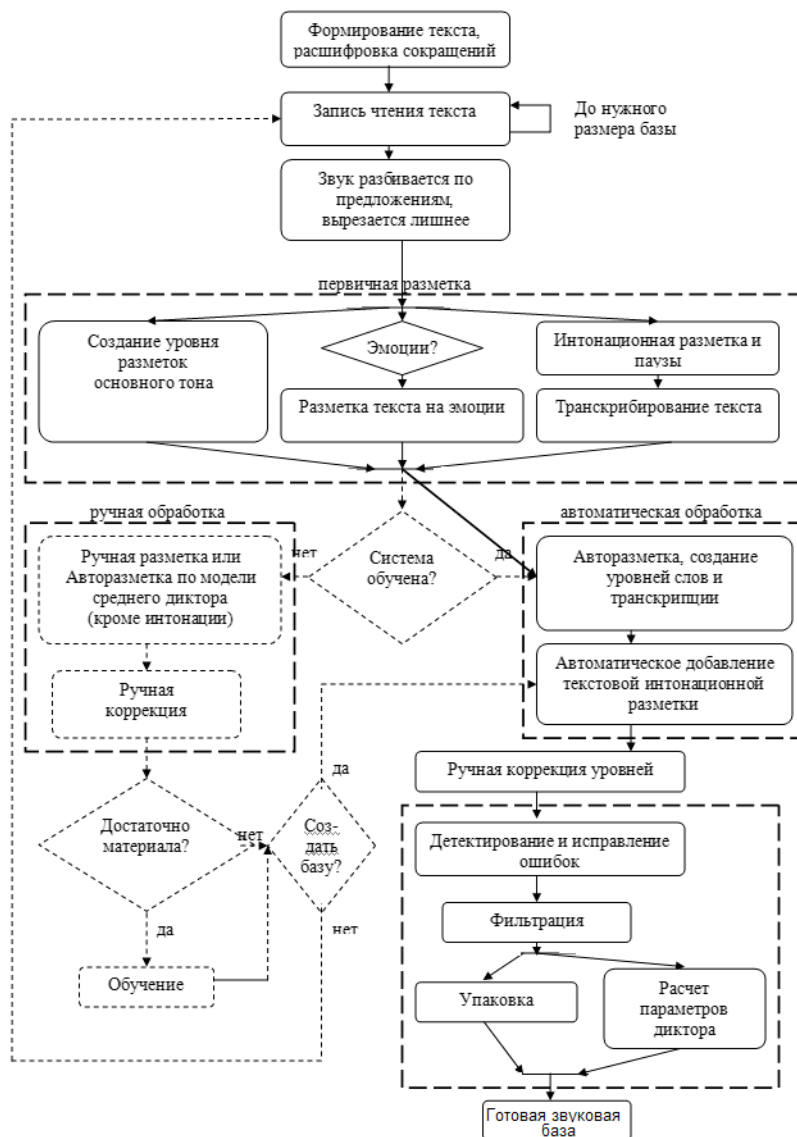


Рис. 16. Алгоритм создания речевой базы для селективного синтезатора

Разметка баз может выполняться вручную, что требует больших временных затрат, или автоматически с возможной последующей ручной корректировкой. Хороший результат даёт автоматическая разметка, подстроенная под определённого диктора, то есть обученная на части материала, размеченной вручную.

2.3.5.2 Модификация речевого сигнала

Просодические характеристики аллофонов, выбранных из речевой базы для конкатенации, во многих случаях не полностью соответствуют характеристикам, заданным на этапе фонетической обработки текста, поэтому может потребоваться дополнительная просодическая модификация выбранных из базы звуковых элементов. При этом (в отличие от аллофонного или дифонного компилятивного синтеза) в селективном синтезе модификация звуковых элементов сводится к минимуму. Значительная модификация (особенно частоты основного тона) уменьшает естественность речи, поэтому она используется только в отдельных случаях.

Существует множество алгоритмов просодической модификации речевого сигнала, приведём краткое описание двух наиболее часто используемых в системах синтеза речи по тексту.

Наиболее известным и часто используемым алгоритмом, обеспечивающим хорошее качество сигнала, является алгоритм TD-PSOLA (Time-Domain Pitch-Synchronous-Overlap-Add) [Moulines, Verhelst 1995]. Данный алгоритм работает с периодами основного тона сигнала. Сигнал разбивается на фрагменты, взвешенные окном Хеннинга, захватывающим два соседних периода с перекрытием в один период. Затем эти фрагменты перекомбинируются при помощи перемещения их центров и наложением с добавлением перекрывающихся частей (overlap and add), при этом в нужном направлении изменяется ЧОТ речевого сигнала. Иллюстрация применения данного алгоритма приводится на рис. 17.

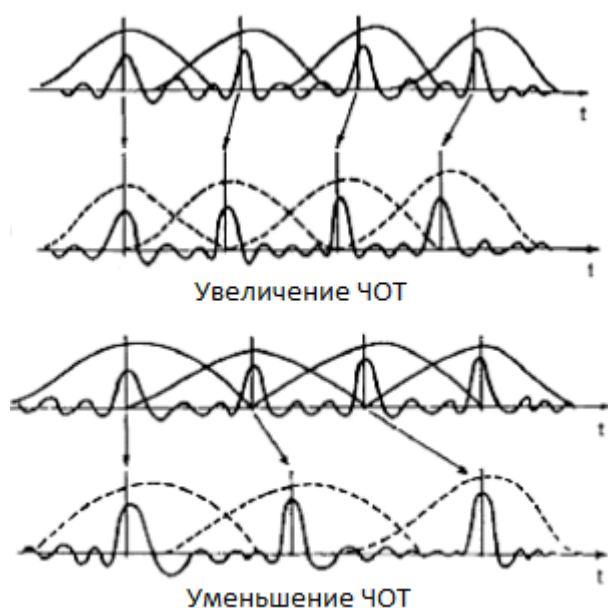


Рис. 17. Иллюстрация применения алгоритма TD-PSOLA

При использовании модели «гармоники плюс шум» (Harmonic plus Noise Model) [Stilianou 2001] исходный сигнал представляется как сумма периодической и стохастической (гармонической и шумовой) функций. При этом невокализованные участки речевого сигнала состоят только из шумовой компоненты. Недостатком этого метода является высокая вычислительная сложность его программной реализации.

После модификации звуковых элементов для приведения их к необходимым характеристикам синтеза результирующий сигнал может (при необходимости) быть подвергнут дальнейшей обработке: фильтрации, изменению частоты дискретизации, могут добавляться эффекты реверберации. В демо-версиях коммерческих синтезаторов на результирующий речевой сигнал часто накладывается музыка, что, с одной стороны, не даёт напрямую бесплатно использовать синтезированные фразы, а с другой – скрывает мелкие дефекты, которые могут возникать при конкатенации звуковых элементов и их модификации.

2.4 ВЫВОДЫ К ГЛАВЕ 2

В настоящее время имеется довольно большое количество разнообразных русскоязычных селективных синтезаторов. При проведении их сравнительных оценок следует учитывать как особенности технологии селективного синтеза речи, так и общую структуру селективных синтезаторов и качество решения задач, выполняемых на каждом этапе преобразования «Текст–Речь».

ГЛАВА 3. МЕТОДЫ ОЦЕНКИ КАЧЕСТВА СЕЛЕКТИВНОГО СИНТЕЗА РЕЧИ

В данной главе рассматриваются основные требования к качеству синтезированной речи и существующие методы ее оценки, обсуждается общая структура системы оценки качества селективного синтеза речи и особенности применяемых в ней методов, обосновывается также необходимость адаптации методов оценки к особенностям селективного синтеза речи.

3.1 Задачи и критерии оценки качества синтезированной речи

В государственном стандарте ГОСТ ISO 9000-2011 «Системы менеджмента качества. Основные положения и словарь» качество определяется как степень соответствия совокупности присущих характеристик требованиям [ГОСТ ISO 9000-2011: 10]. Синтез речи может рассматриваться как программный продукт, к которому предъявляются определённые требования в определённых условиях и для определённой задачи. Такой задачей может быть и озвучивание произвольного текста. Задача оценки качества синтезированной речи тесно переплетается с тестированием синтеза речи как компьютерной программы (или аппаратно-программного комплекса).

Среди задач, для решения которых может применяться система оценки качества синтезированной речи можно назвать следующие:

- **Тестирование системы синтеза в процессе её разработки.** Главная задача такого тестирования – последующее улучшение параметров оцениваемой системы. В этом случае к системе оценки предъявляются следующие требования: она должна быть автоматической, т. е. функционировать без участия человека; иметь достаточно высокое быстродействие; иметь возможность

оценивать синтезированную речь как на соответствие голосу конкретного диктора, так и на соответствие общим параметрам речевого сигнала. Для оценочного анализа доступны результаты всех этапов синтеза, и проверка может осуществляться с использованием промежуточной информации, генерируемой системой в явном виде.

- **Оценка собственной системы синтеза речи в сравнении с конкурентами.**

Основной задачей в этом случае является сравнение разных систем синтеза с разными голосами. Для этого может применяться как автоматическая дикторонезависимая оценка [Falk, Möller 2008], так и оценка экспертов. В данном тестировании может быть затруднен доступ к результатам синтеза: для коммерческих приложений обычно доступны только интерактивные демо-версии, при помощи которых можно получить образцы синтезированной речи довольно низкого качества, с фоновой музыкой или другими ограничениями в целях защиты от коммерческого использования, или же доступны только заранее подготовленные примеры звучания. Для корректного сравнения результатов работы синтезаторов необходимо использовать их полнофункциональные версии.

- **Участие в конкурсах, проводимых независимыми компаниями.**

В таких условиях система оценки может быть не автоматической, но автоматизированной. Для оценки синтезаторов в этом случае могут привлекаться большие человеческие ресурсы (например, заинтересованные пользователи интернета). При этом, хотя внутренняя структура систем синтеза и останется закрытой, имеется возможность получения промежуточных результатов работы системы в унифицированном виде, при заинтересованности в этом самих участников конкурса. Системы синтеза могут тестироваться на одной и той же речевой базе, с использованием которой строится синтезированный голос. Может применяться также эталонная оценка при сравнении с диктором-донором.

Две последние стратегии оценочного анализа приближают нас к созданию некоего стандарта в оценке различных систем синтеза речи по тексту. При этом в фокусе нашего внимания в настоящей работе будут особенности структуры и функционирования синтезаторов речи на русском языке.

Основными общепринятыми мерами качества синтезированной речи являются оценки её **разборчивости** и **естественности** [Фролова 1989; Dutoit 1997: 195]. Государственный стандарт России [ГОСТ 1995] вводит следующие определения:

Разборчивость речи – относительное количество (в процентах) правильно принятых элементов (слогов, слов, фраз) артикуляционных таблиц.

Качество речи – величина, характеризующая субъективную оценку звучания речи в испытуемом тракте:

- по сравнению со звучанием в контрольном тракте (принятом за пять баллов);
- по сравнению со звучанием речи в другом тракте (в процентах предпочтения).

Исходя из сказанного, можно считать, что под «качеством» синтезированной речи понимается её **естественность**, то есть величина, характеризующая субъективную оценку звучания синтезированной речи по сравнению со звучанием естественной речи.

Методы оценки качества синтеза можно разделить на две большие группы: субъективные и инструментальные. В отдельную промежуточную группу можно также выделить те методы, которые требуют участия человека для детектирования наличия/отсутствия ошибки (например, неправильного ударения), а не субъективной оценки речи по определённому параметру (например, естественности) [van Santen 1998; Campbell 2007]. Такие методы обычно

используются для тестирования отдельных модулей синтезатора: нормализации текста, расшифровки сокращений, чтения иноязычных вставок, расстановки ударений, фонетической обработки и т. п.

Также можно провести аналогию с терминами, принятыми в области тестирования программного обеспечения: так называемым тестированием «прозрачного ящика», когда тестирующий знает, каким образом устроена система, и строит свои тесты на основании этих знаний (тестирование и оценка работы отдельных модулей синтезатора, например, лингвистической обработки) и тестированием «чёрного ящика», то есть системы синтеза как некой целостной процедуры, преобразующий текст в звучащую речь [Van Heuven, Van Bezooijen 1995].

Далее рассмотрим существующие методы оценки качества синтезированной речи, их преимущества и недостатки для оценки современных селективных синтезаторов речи.

3.2 Методы оценки разборчивости речи

Долгое время основным критерием качества работы синтезаторов была разборчивость синтезированной речи. Лишь к концу 80-х годов с повышением ее разборчивости стала более актуальной проблема естественности синтезированной речи, так как изменились ожидания и требования к качеству речи, поскольку стало понятно, что даже полностью разборчивая, но «роботизированная» или «механическая» речь синтезаторов сложна для восприятия и не может считаться высококачественной.

Можно с определённой долей уверенности утверждать, что проблема разборчивости речи для современных синтезаторов в целом решена [Taylor 2009: 474]. Это означает, что, несмотря на возможную неразборчивость отдельных

ошибочно синтезированных слов или неправильно расшифрованных сокращений или аббревиатур, общий смысл синтезированных предложений и текстов остаётся понятным.

Существует ряд тестов для проверки разборчивости речи, порождаемой системами синтеза. Как было сказано выше, разборчивость определяется относительным количеством правильно распознанных элементов речи. Можно выделить несколько типов проверок в зависимости от длины речевых отрезков, подаваемых для тестирования, и задач, которые ставятся перед испытуемыми. Различают звуковую (фонемную), слоговую, словесную и фразовую разборчивость. Приведём данные (рис. 18) по взаимной зависимости различных видов разборчивости в естественной речи.

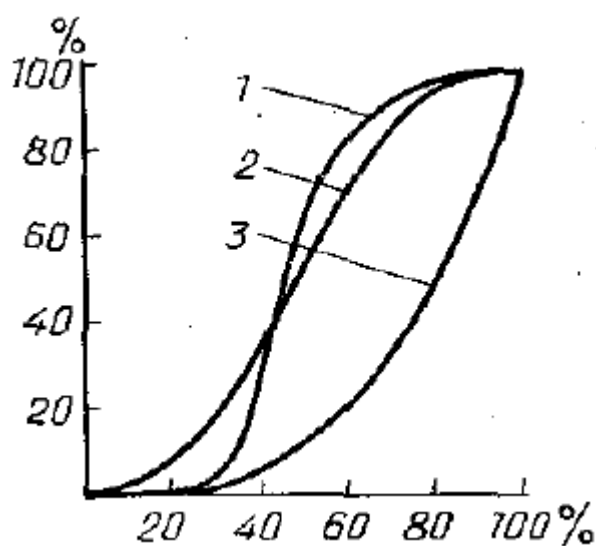


Рис. 18. Зависимость слоговой (3), словесной (2) и фразовой (1) разборчивости от фонемной [Покровский 1976]

Стоит отметить, что для синтезированной речи сами кривые могут заметно отличаться от соответствующих показателей для естественной речи, хотя несомненная взаимная зависимость приведенных показателей существует.

При оценке восприятия отдельных слов в предложении, то есть словесной разборчивости, обычно используются семантически непредсказуемые предложения [Benoit et al. 1992], слова в которых не связаны между собой по смыслу.

Звуковая разборчивость оценивается в диагностическом тесте на рифму и в модифицированном диагностическом тесте на рифму [House et al. 1965], где синтезируются цепочки слов, отличающиеся начальным или конечным согласными. Существуют также тесты на распознавание звукового кластера [Jekosch 1992]: слова с определённой слоговой структурой генерируются автоматически. Это позволяет оценивать разборчивость не только согласных звуков, а слов целиком, причём то, что заранее заданный список слов не используется, позволяет исключить возможность подстройки синтезаторов под конкретный набор слов. Возможность подстройки синтезаторов под заранее известный набор тестовых данных следует учитывать при подготовке тестов для независимого сравнения систем синтеза между собой.

Результаты отечественных разработок речевых тестов на разборчивость частично приводятся в [Речевые тесты 1986]. Слоговая разборчивость проверяется по специально составленным слоговым таблицам. В настоящее время проверка разборчивости озвучивания слогов для синтезаторов речи часто делается по ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости» [ГОСТ 1995], рекомендованном и для оценки синтезаторов речи, где приводятся специальные артикуляционные слоговые таблицы для русского языка.

3.3 Методы оценки естественности речи

К субъективным методам, позволяющим оценить степень естественности речи с точки зрения человека, её воспринимающего, относятся разного рода тесты

и опросники, заполняемые экспертами-специалистами, либо наивными слушателями, носителями синтезируемого языка. Опросники обычно основываются на рекомендации Международного союза электросвязи (International Telecommunication Union) P.85 ITU-T «Метод субъективной оценки качества речи устройств речевого вывода» [ITU-T Rec. 1994]. В них используется так называемая MOS-оценка (Mean Opinion Score или «метод мнений»), производимая по пятибалльной шкале по нескольким категориям: общее впечатление, слуховое усилие, естественность, понимание смысла сообщения, темп, разборчивость, приятность голоса. Приемлемость голоса (для определенных задач) оценивается в этой рекомендации по двухбалльной шкале.

Проведение подобных тестов является довольно трудоёмкой задачей, и для того, чтобы ускорить процесс оценки и сделать его более доступным для тестирования системы в процессе её разработки, начали создаваться различные инструментальные (или объективные) методы оценки качества синтеза. Такие методы основываются как на автоматическом сравнении синтезированной речи (с использованием различных мер близости) с «живой» речью того же диктора (диктора-донора) [Vepa et al. 2002; Stylianou, Syrdal 2001], так и на построении дикторонезависимых моделей естественной речи и различных методах оценки того, насколько синтезированная речь к ним приближена [Falk, Möller 2008]. При этом предполагается, что в естественной речи невозможны резкие скачки в частоте основного тона, энергии или спектральных составляющих, характерные для систем конкатенативного синтеза. Система оценки обучается на базе субъективных оценок образцов синтезированной речи, данных испытуемыми. Адекватность инструментальной оценки качества синтезированной речи измеряется также с той точки зрения, насколько она совпадает с субъективными оценками испытуемых. К объективным методам оценки можно отнести и систему [Norrenbrock et al. 2012(a)], позволяющую инструментально оценивать качество просодической

обработки на основании 18 акустических характеристик ЧОТ и длительностей гласных и согласных звуков.

Чисто инструментальные методы интегральной оценки для сравнения существующих речевых синтезаторов широкого применения пока не имеют и в основном используются в процессе их разработки и для автоматизации их настройки, однако последние исследования говорят о том, что задача адекватного автоматического вычисления субъективной оценки качества синтеза вполне осуществима [Norrenbrock et al. 2012(b)].

Существуют исследования по использованию систем автоматического распознавания речи для оценки качества синтезированной речи [Bachan 2012]. При этом показано, что «живая» человеческая речь распознаётся лучше всего, и более качественной считается синтезированная речь, на которой система распознавания даёт меньший процент ошибок.

В настоящий момент международные «соревнования» синтезаторов Blizzard Challenge² являются своеобразным эталоном по оценке качества систем синтеза речи. Их задачей является сравнение результатов работы различных систем синтеза, причем синтезированные голоса создаются на основе одних и тех же звуковых баз, предоставляемых организаторами перед началом соревнований. Затем, по прошествии времени, отведенного на создание синтезатора, участникам выдается набор тестовых текстов, которые они должны озвучить и предоставить организаторам для оценки. В 2010 году соревнования проводились для корпусов речи на английском и китайском языках. В 2012 году в центре внимания был синтез аудиокниг и методы оценки качества синтезированной речи, в 2013 – синтез по большим объёмам необработанных речевых данных (до 300 часов) и синтез

² <http://www.festvox.org/blizzard/>

речи на одном из индийских языков. В 2014 – 2015 годах – также синтез речи на нескольких индийских языках и синтез на материале двухчасовой базы англоязычных аудиокниг (только 2015) [Blizzard Challenge 2015].

По образцу Blizzard Challenge для испаноязычных синтезаторов было организовано отдельное тестирование Albayzin [Méndez et al. 2010]. Для французского языка существует стандартизованный набор тестов для синтеза речи на французском языке, разработанный в ходе национального проекта EvaSy («Evaluation of speech synthesis systems» – оценка систем синтеза речи) [EVALDA – EVASY 2006].

Для оценки русскоязычных синтезаторов чаще всего используется ГОСТ Р 50840-95. Наряду с этим используются различные тесты отдельных составляющих компонентов (модулей лингвистической обработки, модификации, полноты и качества речевой базы) [Бабкин, Захаров 1999; Гецэвіч 2012], но единого стандарта оценки, рассчитанного на современные синтезаторы, пока нет. В работе [Русанова 2004] была сделана попытка предложить единую комплексную оценку систем синтеза русской речи, однако она не имеет широкой известности и практически не применяется.

Особо следует упомянуть ещё две задачи, связанные с качеством синтезируемой речи. Это **оценка узнаваемости** речи исходного диктора, для которой могут применяться методы идентификации говорящего по голосу, и **оценка выразительности и эмоциональности** синтезированной речи. Вторая задача скорее относится к будущим разработкам, так как в настоящий момент для русского языка большинство селективных синтезаторов генерирует только нейтральную речь.

3.4 Факторы, влияющие на восприятие синтезированной речи человеком

При подготовке и проведении тестов для оценки синтезированной речи следует учитывать множество факторов, влияющих на восприятие речи человеком и выносимые им оценочные суждения. Приведём основные факторы, которые могут повлиять на восприятие и понимание синтезированной речи [Сорокин 1992; Санникова 2008]. К ним относятся:

1. Конкретные условия, связанные с выполняемой задачей.
2. Ограничения, присущие системе обработки информации, присущей человеку.
3. Опыт и тренировка слушателя.
4. Лингвистическая структура сообщения.
5. Качество записи речевого сигнала и условия восприятия (громкость, шум, реверберация, посторонние разговоры и т. п.).

Существуют специальные тесты, направленные на оценку понимания синтезированной речи. Например, в работе [Sydeserff et al. 1992] для того чтобы оценить, насколько хорошо воспринимается и понимается синтезированная речь, испытуемым предлагалось прослушать синтезированное предложение или несколько предложений, а затем ответить на вопросы по их содержанию. В [Санникова 2008] предлагается математическая модель для количественного оценивания понимания синтезированной речи.

Все перечисленные факторы учитывались нами при разработке методов и проведении экспериментов по оценке качества синтезированной речи, однако подробно проблема понимания синтезированной речи в данной работе рассматриваться не будет, так как выходит за рамки темы диссертации.

Отдельно следует отметить, что фактор «тренированности» аудиторов, участвующих в экспериментах по оценке качества синтезированной речи, в нашей

работе решался (в отличие от рекомендаций, приведённых в [Речевые тесты 1986]) в пользу того, чтобы не проводить с аудиторами каких-либо специальных предварительных тренировок по работе с синтезированной речью. Такое решение было принято по нескольким причинам: во-первых, это больше соответствует реальным условиям использования синтеза речи и, во-вторых, уровень естественности речи, производимой современными селективными синтезаторами, столь близок к естественной, что обычно для тренировки аудиторов перед экспериментом было достаточно прослушать один-два примера тестовых фраз. При этом полученные нами данные подтверждают факт влияния «тренированности» на оценки аудиторов: в целом, люди, совсем не имевшие опыта прослушивания синтезированной речи (по работе или в качестве озвучивания аудиокниг, в телефонных системах голосового самообслуживания и т. п.) были склонны оценивать качество любой синтезированной речи как более низкое, чем имевшие подобный опыт (подробнее см. главу 4).

3.5 Необходимость адаптации методов оценки к селективному синтезу речи

При оценке качества синтезированной речи необходимо обратить пристальное внимание на особенности звучания речи, связанные с использованием селективного метода [Соломенник 2013(с)], и соответствующим образом дополнить и изменить известные методы оценки синтезированной речи, а также при необходимости разработать новые.

Как было отмечено в главе 2, селективный синтез базируется на конкатенативном методе озвучивания, то есть при синтезе речевого сигнала используются заранее сделанные звукозаписи естественной речи. В отличие от более ранних аллофонных или дифонных синтезаторов речи, порождающих речевой сигнал из отдельных и специально подготовленных звуковых единиц, выделенных из небольшого и тщательно подобранного набора слов, в селективном

синтезе для каждой целевой единицы синтеза производится выбор наиболее подходящего кандидата из множества вариантов, взятых из естественно озвученных предложений базового языка. Для этого записываются специальные речевые базы, размер которых может составлять до нескольких десятков часов звучащей речи [Black 2002]. Алгоритм селективного синтеза строит оптимальную последовательность звуковых единиц, учитывая одновременно и то, насколько кандидаты подходят под описание необходимых характеристик целевых звуков, и то, насколько хорошо выбранные элементы будут конкатенироваться с соседними. При этом из базы могут быть выбраны не отдельные звуки, а их цепочки или даже целые предложения.

Таким образом, речь, порождаемая селективным синтезатором, имеет свои особенности. Это, в первую очередь, неравномерность распределения мест с неудачным звучанием: нередко отдельная фраза или её часть звучит гораздо естественнее остальных, а при стыковке «гладких» участков появляются помехи. Указанные особенности связаны с самим алгоритмом выбора звуковых единиц. Каждый раз при синтезе принимается (обычно на основании учёта штрафов) некое компромиссное решение относительно того, несоответствие каких характеристик выбираемых звуковых единиц характеристикам, требуемым системой, будет более критичным и насколько критичным будет несовпадение спектральных характеристик у соседних элементов, иными словами гладкость стыковки соседних единиц. При этом следует учитывать тип самих конкатенируемых звуков и место конкатенации (граница/середина звука) [Syrdal, Conkie 2005].

Кроме того, разработчики часто стараются минимизировать или вовсе устранить использование просодической модификации звуков-кандидатов (изменение их исходной длительности и ЧОТ) под требуемые значения, что может приводить к непредсказуемости просодического оформления фраз. Неестественное

звучание отрезка синтезированной речи может возникнуть из-за отсутствия нужной целевой единицы в речевой базе, при этом такая единица может искусственно конструироваться различными способами (собираться из меньших единиц, искусственно генерироваться), заменяться близкой ей по определённым правилам или вовсе пропускаться. Все вышеперечисленные возможности необходимо учитывать при разработке и составлении тестов для оценки качества селективного синтеза.

Следует также отметить, что для селективного синтеза невозможно, как, например, для простого аллофонного или дифонного конкатенативного синтезатора, составить тест, содержащий все или большинство элементов его речевой базы для тестирования их звучания, так как сегментные единицы языка (фонемы и их аллофоны) в базе будут представлены не одним, а, возможно, сотнями или тысячами вариантов. При этом объём материала для тестирования должен быть достаточно большим и разнообразным, включать в себя различные темы и жанры. В то же время это ни в коем случае не исключает и использования специально сконструированных текстов, например, на сложные с фонетической точки зрения сочетания звуков. Если синтезатор предполагается использовать для какой-то специфической задачи (например, чтения аудиокниг, озвучивания действий пользователя ПК или разговора с «искусственным» оператором по телефону), тесты обязательно должны быть составлены с учётом такого сценария использования.

При тестировании селективного синтеза особенно важным является раздельное тестирование лингвистической обработки текста для озвучивания и собственно акустического модуля синтеза сигнала, так как особенности алгоритма селективного синтеза часто предполагают возможность частичного или даже

полного несоответствия характеристик выбираемых единиц-кандидатов характеристикам, предсказанным системой на этапе лингвистической обработки.

Кроме этого, как и для любого другого вида синтеза, следует иметь в виду эффект привыкания пользователя к синтезированной речи. Возможно, при приближении искусственной речи к естественной может появляться и обратный эффект: одни и те же ошибки (например, ошибки в лингвистической обработке или неудачный подбор звуковой единицы) в речи, близкой к естественной, субъективно могут восприниматься как более грубые, чем аналогичные недочёты в речи, явно звучащей механически и роботизировано. То есть с повышением качества речи могут повыситься ожидания и требования к ней.

Из всего вышеизложенного следует, что основные адаптивные изменения при ориентированной оценке селективных синтезаторов речи должны коснуться как общей интегральной оценки качества синтеза речи, так и оценки акустической и фонетической обработки. В остальных тестах следует разграничивать причины возникновения ошибок в лингвистической обработке, связанные с работой лингвистического процессора, и ошибки, появившиеся вследствие неудачно подобранных звуковых элементов для конкатенации. Например, из-за неправильной длительности выбранных гласных звуков ударение может смещаться на другой слог.

3.6 Структура и задачи системы оценки качества синтезированной речи

Системы синтеза речи по произвольному тексту состоят из нескольких последовательных, частично независимых друг от друга модулей: модуля текстовой (или лингвистической) обработки, модуля просодической обработки и акустического модуля, который может включать в себя компонент просодической модификации звука (подробнее см. раздел 2.3). Очевидно, что при комплексной оценке качества синтезированной речи должны учитываться ошибки и неточности,

возникающие на различных этапах получения синтезированной речи [Соломенник 2013(a)].

3.6.1 Оценка лингвистической обработки

В процессе лингвистической обработки в системах синтеза речи обычно решаются следующие задачи:

1. Токенизация текста – выделение предложений в тексте и разбиение их на отдельные слова; разметка текста на буквы, специальные символы, цифры и знаки пунктуации;
2. Нормализация текста – расшифровка тех обозначений, которые не могут быть произнесены в исходном виде: сокращения, аббревиатуры, цифровые обозначения, номера телефонов, даты, время, римские цифры и т. п.;
3. Определение места ударения и морфограмматических характеристик слов в предложении, для этого обычно используется словарь и/или набор правил или статистические модели;
4. Снятие омонимии (в частности, омографии), т. е. выбор одной из нескольких словоформ, соответствующих тому или иному орфографическому слову текста. Эти словоформы могут отличаться ударением, наличием буквы «ё», фонемной транскрипцией (например, /t/ - /t'/: «ошибка в тесте»/«сосиска в тесте») или грамматическими характеристиками.

Тестовый материал для объективной оценки качества лингвистической обработки для каждого из оцениваемых параметров может быть получен различными способами:

- специально сконструирован лингвистами;
- автоматически сгенерирован текст с заданными параметрами;

- использован фрагмент готового текстового корпуса, на материале которого вычисляется процент ошибок по каждому из параметров [Sproat et al. 2001].

Оценки могут быть получены автоматически, если для сравнения доступен нормализованный и размеченный системой синтеза текст или же (вероятно, с определённой долей ошибок) оценка может делаться по итоговому речевому сигналу при помощи системы распознавания речи.

Однако для каждой из возможных задач текстовой обработки существуют свои нюансы.

- **Правильность выделения предложений** должна проверяться с использованием проблемных текстовых фрагментов, в которых присутствуют сокращения, имена собственные, т. е. набор специальным образом подобранных неоднозначных для системы обработки текста случаев, сконструированных вручную или отобранных из текстовых корпусов.
- При тестировании **расшифровки сокращений** в тестовых фразах должны обязательно присутствовать наиболее частотные сокращения, то же относится и к **чтению аббревиатур и специальных знаков** («%», «*», «#» и т. п.).
- Более развёрнутыми должны быть тесты для проверки правильности **чтения цифровых последовательностей** (чисел, дат, времени, телефонов и т. д.) и **снятия омонимии** (желательно включать в тест оба варианта прочтения омографа, но в то же время избегать неоднозначных предложений, из которых неясно, какой вариант должен быть правильным).
- При проверке **правильности указания места словесного ударения** желательно добавить в тестовый материал наиболее частотные имена

собственные: фамилии, имена, географические названия, особенно, если предполагается использование системы синтеза в соответствующей области.

Реже проверяется правильность работы при решении следующих задач (не во всех синтезаторах имеются специальные алгоритмы обработки таких случаев):

- **обработка иноязычных вставок** (транслитерация/практическая транскрипция): для русского языка в первую очередь важно чтение вставок на латинице и частотных английских слов;
- **транскрипция незнакомых слов и собственных имен**: для русского языка в первую очередь важно правильное предсказание места ударения);
- **автоматическое исправление орфографических ошибок и опечаток**.

При проверке прохождения тестов синтезатором следует учитывать, что в некоторых случаях может быть несколько допустимых вариантов прочтения, а также то, что один или несколько из допустимых вариантов прочтения могут быть более предпочтительными, чем остальные. Например, при чтении обозначений времени («12.10» допустимы варианты «двенадцать часов десять минут» и «двенадцать десять»), аналогично при озвучивании иноязычных вставок (подробное обсуждение предпочтительного варианта чтения англоязычных вставок приводится в работе [Черепанова 2015]).

3.6.2 Оценка фонетической обработки

Фонетическую обработку текста можно разделить на сегментную (построение сегментной транскрипции) и супraseгментную (паузация и выбор интонационных маркеров). Это задача фонетизации текста, которую надо отличать от последующей просодической параметризации и акустического озвучивания.

На этом этапе происходит построение сегментной транскрипции по правилам или по произносительному словарю (в зависимости от языка, для

русского языка обычно используются правила типа «буква–фонема» и набор словарей с исключениями). Таким образом, при оценке данного модуля необходимо проверить корректность транскрипции либо автоматически при наличии данных от синтезатора, либо при детектировании ошибок, обусловленных неправильной транскрипцией, на слух. Наиболее проблемными местами являются различные слова-исключения из правил канонического произношения и соответствующие транскрипции (современные иноязычные заимствования, слова разного типа из орфоэпического словаря), а также другие фонетические особенности русского языка: двойные согласные, позиционная мягкость согласных, дополнительное ударение в композитах, редуцированные безударные гласные, ассимиляция на стыках слов и т. п.

К просодической обработке на этапе фонетизации относятся те компоненты, которые определяют интонационное оформление текста (супрасегментная транскрипция). Сюда относится деление текста на просодические единицы – синтагмы и интонационные фразы, определение длительности пауз между ними и выбор типа (модели) интонационного оформления для каждой из синтагм и фраз. Затем на этапе просодической параметризации происходит вычисление соответствующих физических параметров: длительности, частоты основного тона (ЧОТ), энергии. Деление на синтагмы и выбор интонации может осуществляться как по правилам, так и на основе статистики, причем в конкретных системах синтеза операция выбора интонационного типа может быть пропущена, и система сразу, на основе имеющихся статистических данных, может переходить к предсказанию необходимых просодических параметров звуков в данной конкретной синтагме текста.

На данном этапе объективная количественная оценка работы синтезатора представляет собой уже менее тривиальную задачу, поскольку все вышеуказанные

функциональные признаки и физические параметры в естественной речи могут варьироваться от диктора к диктору и от произнесения к произнесению. Например, для оценки правильности **расстановки пауз** могут отдельно учитываться места, где пауза необходима, возможна (с разной степенью предпочтения) и недопустима [Хомицевич, Соломенник 2010].

При оценке точности определения локализации пауз могут использоваться следующие показатели, вычисленные по эталонному образцу или по частотности на многих дикторских прочтениях:

- Точность – доля предсказанных пауз, совпадающих с паузами в естественной речи в эталонном образце или по частотности на многих дикторских прочтениях.
- Полнота – доля пауз в естественной речи, правильно предсказанных программой синтеза речи.
- F-мера = $(2 * \text{полнота} * \text{точность}) / (\text{полнота} + \text{точность})$.
- Вставка паузы – «лишние» паузы, вставленные синтезатором.
- Пропуск паузы – паузы, не отмеченные синтезатором.
- Правильность связей – правильность определения наличия/отсутствия паузы на словоразделе для каждой пары слов в тексте, т. е. $((\text{количество слов} - 1) - \text{количество лишних пауз} - \text{количество пропущенных пауз}) / (\text{количество слов} - 1)$.

Выбор **интонационного оформления** (выбор типа интонационной модели и места ее интонационного центра) может быть оценен таким же образом (по набору допустимых вариантов), либо при сравнении с естественной речью (по эталонному образцу или по частотности на многих дикторских прочтениях). При оценке интонационного оформления предложений можно также отдельно учитывать правильность распознавания смысла или коммуникативного типа сообщения в зависимости от выбора интонационного оформления и естественность интонации.

К ошибкам модуля фонетической обработки относятся и недостатки реализации **темпа и громкости речи** (например, неожиданное ускорение или замедление произнесения), **ритма** (неадекватность акцентного контура, степени словесной и слоговой выделенности, проглатывание или затягивание отдельных звуков в общем потоке речи), которые могут оцениваться параллельно с тестированием интонационного оформления.

Отдельную и пока нерешённую задачу представляет собой тестирование **выразительности и эмоциональности** (уместности эмоций) синтезированной речи. Для проведения такой оценки предлагается включать в опросник соответствующие критерии оценки одновременно с субъективной оценкой общего качества синтезированной речи.

3.6.3 Оценка акустической обработки

К акустическому модулю относится выбор и объединение звуковых элементов из речевой базы в звуковые последовательности, а также просодическая модификация звуков при необходимости.

Возможные проблемы в работе акустического модуля существенно зависят от технологии синтеза: например, в формантном, аллофонном или дифонном конкатенативном синтезе второго поколения это может быть общая заметная неестественность (роботизированность) звучания одновременно с неудачными отдельными звуками и щелчки на звуковых стыках и подобные проблемы; в селективном синтезе – слышимые стыки между звуками, паразитические призвуки, несоответствие интонационного оформления ожидаемому, причем ошибки обычно неравномерно распределены по тексту; в синтезе, основанном на статистическом моделировании (СММ), – роботизированность всей речи или звуков определенного типа, в то время как резких «скачков» тона или энергии, как в селективном синтезе, обычно не наблюдается. Если в системе используется

значительный процент модификации звуков, в синтезированной речи появляются заметные призвуки и эффект роботизированности.

Субъективная оценка качества работы акустического модуля может быть в определенной степени объединена с общей субъективной оценкой качества синтеза (или оценкой общего впечатления) при помощи опросников, поскольку этот компонент является последним, формирующим выходной сигнал на основании работы предыдущих модулей.

Аудиторы могут указывать наличие и степень различных **искажений** (замена нужного звука неправильным, странное неестественное звучание отдельных звуков, неестественно громкое звучание некоторых звуков, неестественно тихое звучание некоторых звуков, неестественные скачки громкости и тона голоса, наличие посторонних призвуков, щелчков, «бульканья») в синтезированной речи.

Для оценки правильности **подбора элементов** из базы (в селективном синтезе) и работы акустического модуля в целом часто используется так называемый ресинтез, то есть синтез фраз, уже имеющихся в речевой базе, из которой подбираются элементы для конкатенации. В идеале они должны звучать одинаково. Также предлагается производить оценку качества просодической **модификации звука**, если в синтезаторе существует опция изменения скорости чтения и высоты голоса.

3.6.4 Интегральная оценка качества синтезированной речи

В интегральной оценке качества синтезированной речи необходимо учесть результаты всех тестов для отдельных компонентов синтезатора, а также степень их влияния на общее качество синтеза по мнению auditors. Испытуемые могут оценивать **естественность** и **приятность** голоса (обычно по пятибалльной шкале), трудность **понимания** (в тестах на понимание auditors обычно просят после

прослушивания ответить на вопросы по тексту или же написать его краткое содержание) и **приемлемость** голоса для конкретной задачи (например, в системе заказа железнодорожных билетов).

При этом следует учитывать конкретную задачу, стоящую перед программой синтеза речи (иными словами, сценарий его использования). Например, для озвучивания научных статей или новостных сообщений будет крайне важна правильность чтения различных сокращений и цифровых обозначений, в то время как для озвучивания ограниченного набора ответов системы голосового самообслуживания возможна предварительная подготовка текста до его подачи в синтезатор (с выполнением необходимых расшифровок). Решающим при оценке качества как совокупности характеристик, соответствующим определённым требованиям, является также тембр и приятность голоса по мнению auditors: для телефонных опросов тембр должен быть спокойным, приятным и доброжелательным, а для оповещения о какой-либо опасности – наоборот, резким и напряжённым. Экранный диктор может озвучивать действия пользователя в быстром темпе, а диктовка номера карты по телефону должна быть медленной и разборчивой. Поэтому вне конкретной задачи сложно говорить об интегральной оценке качества синтезаторов речи.

При оценке конкретной системы необходимо учитывать также ее интерактивную направленность, т. е. то, допускает ли она ручную корректировку ошибок пользователем: например, использование пользовательского словаря, где можно задать ударения в словах или расшифровки определённых сокращений, возможность регулирования длительности, энергии и ЧОТ отдельных участков речевого сигнала (например, при помощи тегов, вставляемых в текст), изменения места фразового ударения, интонации, исключения неудачных звуковых элементов из подбираемых вариантов и т. п.

Следует отметить, что сравнение различных систем синтеза должно проводиться максимально независимыми, незаинтересованными экспертами. Наличие открытых, находящихся в свободном доступе русскоязычных звуковых корпусов с достаточно большими объёмами записей одного и того же диктора (чтение текста) также облегчило бы сравнение синтезаторов между собой. Важно также иметь большие объёмы тестовых данных для исключения возможности ручной подстройки синтезаторов под определённый известный тест, так как стандартные тесты с заданным речевым материалом могут обеспечить лишь проверку «минимальной приемлемости» [van Santen 1998: 243] (то есть обеспечить не выбор лучшей системы, а отбрасывание худших синтезаторов, показывающих неприемлемый результат даже на «простых» тестах). Но даже задача создания корпуса, позволяющего произвести минимальную, но всестороннюю диагностику ошибок разных уровней, в настоящий момент не решена и является актуальной для русского языка.

3.7 ВЫВОДЫ К ГЛАВЕ 3

Система оценки селективных синтезаторов должна учитывать качество решения задач, возникающих на каждом этапе преобразования «Текст–Речь», то есть быть позволять производить помодульную диагностическую оценку качества. При этом должны учитываться особенности качества речи, порождаемой селективными синтезаторами. В данной главе были описаны основные задачи, способы и методы оценки качества синтезированной речи. Приведена структура системы оценки качества работы систем селективного синтеза речи. Даются рекомендации по составлению и проведению сравнительного и диагностического тестирования синтезаторов.

ГЛАВА 4. РЕЗУЛЬТАТЫ ОЦЕНКИ КАЧЕСТВА СЕЛЕКТИВНОГО СИНТЕЗА РЕЧИ

Целью экспериментов, описание и результаты которых приводятся в данной главе, является проверка пригодности предлагаемых методов оценки синтезированной речи для селективного синтеза, а также выявление основных слабых мест современных селективных синтезаторов речи: тех ошибок и недочётов, которые заставляют пользователей считать речь, производимую синтезатором, менее естественной.

Основное внимание уделено тем видам оценок, которые являются наиболее специфичными и важными для селективных синтезаторов: это оценки работы блоков акустической и просодической обработки. Приводятся также данные о влиянии разных типов ошибок на оценку естественности звучания синтезированной речи. С учетом анализа полученных результатов предлагаются новые методы оценки интонационного оформления синтезированной речи и её общего качества. Для других типов оценок предложены диагностические тесты, а также приведены примеры текстов, которые могут быть использованы для таких тестов.

При проведении обсуждаемых экспериментов материалом послужила синтезированная речь, полученная с использованием различных современных русскоязычных синтезаторов речи. Большинство из них являются коммерческими программами, что накладывает ограничения на длительность и качество тестируемых речевых записей. Этот же фактор заставил отказаться от тестирования разборчивости синтезированной речи. Поскольку, например, чтение слоговых таблиц синтезаторами требует подстройки самого синтезатора на особый режим чтения: чтение каждого слога как отдельной синтагмы, таким образом,

чтобы гласный звук был ударным, отключения модулей расшифровки сокращений и исправления ошибок и т. д., что не доступно для стороннего исследователя. Для тех типов тестов, которые в данный момент провести не представляется возможным, приводятся общие рекомендации по их составлению и проведению.

Для тестирования в общем случае были выбраны синтезаторы, имеющие демо-версии без фоновой музыки, которая может искажать восприятие синтезированной речи. Предпочтение отдавалось женским голосам, так как, во-первых, они чаще используются в голосовых системах обслуживания и в качестве экранных дикторов при общении пользователя с ЭВМ, а, во-вторых, из-за более высокой частоты основного тона (а, следовательно, меньшей длительности его периода), в связи с чем реализация качественного речевого синтеза женским голосом является более сложной задачей по сравнению с мужским, в частности с точки зрения модификации тональных характеристик речевого сигнала.

Общий объём протестированных звукозаписей составил более 4 часов речи (от 2 до 60 минут речи на синтезатор в зависимости от числа тестов, в которых данный синтезатор был использован).

4.1 Оценка степени влияния различных типов ошибок на качество синтезированной речи

При вычислении общей оценки качества системы синтеза речи следует учитывать влияние различных видов ошибок на естественность и разборчивость порождаемой речи. Приведём описание и результаты проведенного эксперимента по оценке синтезированной речи с учётом различных типов возможных ошибок синтезатора. Подробно эксперимент описывается также в статье [Соломенник 2015].

В работе [Санникова 2008: 6] показано, что восприятие синтезированной речи человеком подобно восприятию речи в шуме, то есть требует большей концентрации и умственных усилий. Однако у синтезированной речи есть свои особенности, в этом случае ошибки могут возникать на различных этапах обработки текста и различным образом влиять на её восприятие, разборчивость и естественность речи. Это могут быть ошибки в лингвистической обработке, ошибки просодического или акустического модуля и т. д.

В описываемом ниже эксперименте делается попытка оценить, какие ошибки наиболее распространены в современных селективных синтезаторах речи высокого качества и какие из них вызывают наибольшие проблемы при восприятии синтезированной речи, заставляя слушающих оценивать её как менее естественную.

Остановимся подробно на нескольких исследованиях, наиболее близких к рассматриваемой нами задаче. В работе [van Santen 1998: 241] описывается эксперимент по оценке синтезированной речи (на английском языке) с разграничением разных типов ошибок, приблизительно соответствующих отдельным компонентам синтезатора. При чтении коротких фраз основными оказались ошибки, связанные с акустическим блоком синтезатора, а именно: общее качество голоса и его прерывистость, наиболее редкими – ошибки лингвистической обработки озвучиваемого текста (неверное произнесение, пропуск слов или букв, неправильное место ударения и т. д.), промежуточное положение заняла категория «плохой ритм», отвечающая за длительности звуков. В работе [Русанова 2004: 83–84] при оценке трёх русскоязычных (не селективных) синтезаторов максимальный вес по влиянию на интегральную оценку правильности и естественности синтезированной речи получили ошибки, связанные с неверной постановкой ударения в словах. Ошибки в расстановке пауз

в предложении, интонационном выделении, темпе речи были отмечены экспертами только при прослушивании технического текста с более сложной структурой предложений. Более позднее исследование [Корсакова, Засыпкина 2012: 92–98] для четырёх систем синтеза речи (названия систем в работе не приводятся) на русском и английском языках показало, что основными типами ошибок оказались: смещение словесного ударения, выпадение звука, замена звука и неровная интонация, однако данные о влиянии различных типов ошибок на оценку естественности речи в этой работе не приводятся. Различия в результатах указанных исследований могут быть связаны с тем, что эксперименты проводились с разными поколениями синтезаторов.

В проведённом нами эксперименте для оценки качества и естественности русской синтезированной речи были выбраны два голоса современных синтезаторов последнего поколения: «Tatyana» (Ivona TTS) польской компании Ivona и «Анна» (VitalVoice TTS) петербургского ООО «ЦРТ». На основе анализа предыдущих исследований были выделены следующие категории возможных ошибок:

- 1) неверное место словесного ударения;
- 2) неверное произнесение слова (замена/выпадение/добавление лишнего звука);
- 3) неправильные паузы (отсутствие/лишние, слишком короткие/длинные);
- 4) плохой темп/ритм;
- 5) неровная/неверная интонация;
- 6) нарушения плавности речи (дефекты в речевом сигнале): прерывистость, скачки, «бульканье», стук и т. п.;
- 7) общее качество голоса;
- 8) иное.

В качестве тестового материала был использован фонетически представительный текст [Смирнова, Хитров 2013], включающий в себя описательную и диалоговую части, что позволило лучше оценить адекватность интонационного оформления синтезированной речи. Полный текст приводится в Приложении А. Текст состоял из 76 предложений (всего 532 слова), воспроизводившихся последовательно сначала одним синтезированным голосом, а затем другим, причём голоса подавались части испытуемых в разной последовательности, однако значимых различий в оценке из-за изменения порядка прослушивания не было. Каждое предложение повторялось два или (при необходимости) большее количество раз. Испытуемым было предложено оценить по пятибалльной шкале естественность звучания каждого предложения и при оценке ниже четырёх указать категории ошибок. В оценке участвовало 11 человек в возрасте от 18 до 40 лет, не занимавшихся профессионально синтезом речи. Общая длительность звукозаписи составила около 5 минут для каждого голоса.

Образец инструкции и ответного протокола приводится в Приложении Б. Краткие результаты эксперимента приводятся ниже в таблице 3.

Таблица 3. Количество ошибок разных типов, выделенных испытуемыми, и средняя оценка естественности речи (в скобках указано стандартное отклонение). Приводятся значения, усредненные по всем предложениям

Голос	Ударение	Произнесение	Паузы	Темп/ритм	Интонация	Плавность	Общ. кач. голоса	Иное	Средн. оценка
Tatyana	10,6 %	21,9 %	3,8 %	3,2 %	41,5 %	3,9 %	14,2 %	14,1 %	3,9 (0,3)
Анна	1,7 %	10,4 %	6,2 %	5,6 %	40,8 %	32,3 %	16,3 %	0,0 %	4,1 (0,4)

Общее количество ошибок, отмеченных аудиторами, оказалось примерно одинаковым (среднее количество ошибок, отмеченных аудиторами для голоса Tatyana – 89,5, а для голоса Анна – 88,6). Оба образца синтезированной речи

(голоса) также получили примерно одинаковую среднюю оценку естественности: 3,9 и 4,1 соответственно. Для сравнения оценок естественности был использован статистический анализ. Поскольку для измерения переменных используется порядковая шкала и неизвестны параметры распределения исследуемой выборки, был применён U-критерий Манна-Уитни (его значение для сравниваемых оценок равно 34). Критическое значение U-критерия Манна-Уитни при заданной численности сравниваемых групп составляет 30. $34 > 30$, следовательно, различия средних оценок данных синтезаторов статистически не значимы ($p > 0,05$).

Частотность ошибок разного типа для тестируемых синтезаторов различается. При этом, несмотря на большую вариативность ответов испытуемых (см. ниже рис. 19) можно выделить определённые закономерности, касающиеся влияния ошибок разного типа на общую оценку естественности звучания. Подробные результаты приводятся в таблице 4. Для тех типов ошибок, которые встречались в предложениях только одновременно с другими типами ошибок, данные о максимально возможных оценках не приводятся. Корреляцию ошибок и влияние на оценку естественности звучания одновременно нескольких типов ошибок следует исследовать отдельно. Необходимо уточнить, что высокий процент ошибок восьмого типа («иное») для голоса Tatyana отмечался испытуемыми с комментариями «акцент» в тех случаях, когда синтезированный голос неверно произносил отдельные звуки (неправильное смягчение, неполная редукция безударных гласных и т. п.), что можно было бы отнести и к типу два («неверное произнесение»). В будущих экспериментах описание этой категории ошибок должно быть уточнено.

Таблица 4. Максимально возможные оценки естественности речи при наличии ошибки определённого типа

Тип ошибки	Ударение	Произнесение	Интонация	Плавность
Макс. оценка естественности	3,5	4,0	4,4	3,9

Наиболее частотными для обеих тестируемых систем оказались ошибки и дефекты в интонации, однако следует отметить, что эта группа явно неоднородна: некоторые интонационные ошибки резко снижают оценку естественности фразы (единственная ошибка в интонации может снизить естественность до 3,2), а другие на нее практически не влияют. Это, скорее всего, связано с тем, что неточности в обработке интонации могут быть двух типов: неправильный выбор интонационного оформления фразы (например, повествовательная интонационная модель вместо вопросительной) или дефекты, состоящие в несколько неестественном движении ЧОТ. По всей видимости, эти интонационные неточности нужно оценивать отдельно. Подробнее (и для большего количества сравниваемых синтезаторов) этот вопрос рассматривается в разделе 4.3.3, где используется двухуровневая система оценки естественности интонации. Ошибки других типов отмечались значительно реже и не более чем половиной испытуемых.

Вариативность ответов испытуемых оказалась довольно высокой (см. рис. 19), особенно это касалось ошибок, связанных с акустическим модулем и варианта ответа «Иное». Возможно, на оценку общего качества голоса в большой степени влияет субъективное восприятие его приятности.

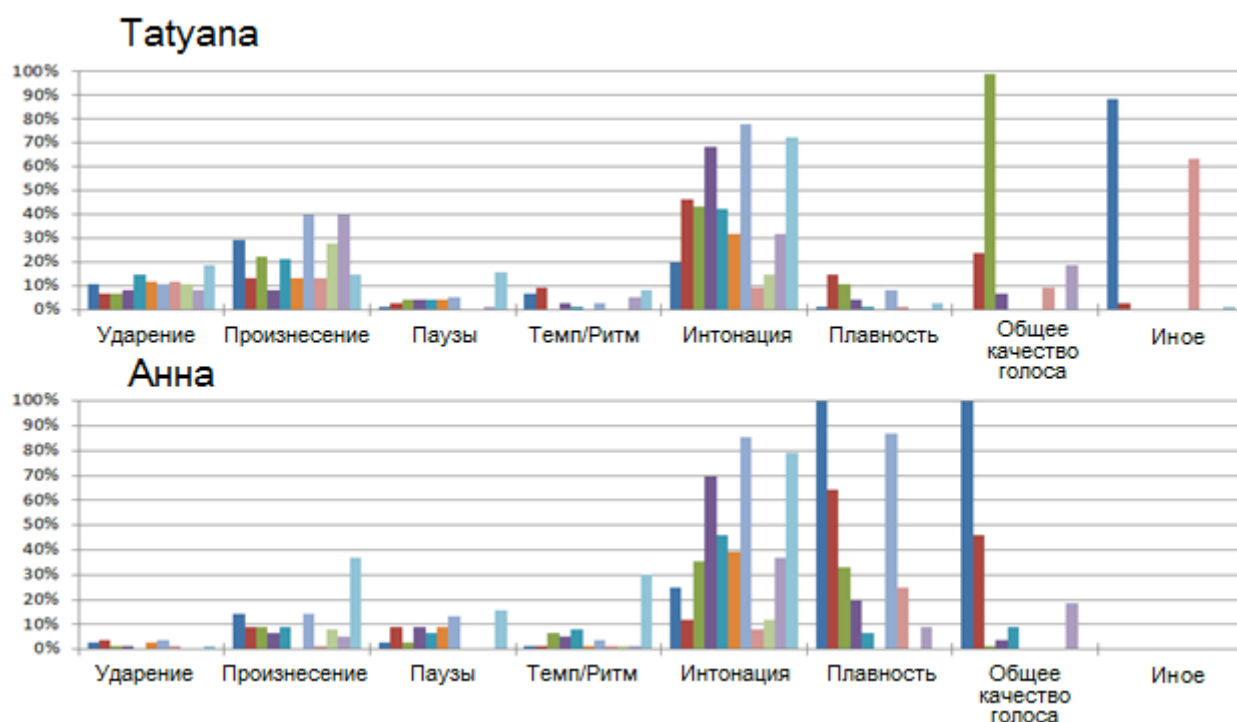


Рис. 19. Вариативность оценок аудиторов (процент предложений, в которых отмечены ошибки указанного типа) для разных синтезированных голосов. Цветами показаны проценты ошибок, указанные каждым испытуемым

Высокую согласованность при определении типа ошибки испытуемые показали только для наиболее частых и явных ошибок (например, ошибок в месте ударения). Данные по ошибкам, отмеченным более чем половиной и более чем тремя четвертями испытуемых, приводятся в таблице 5 и таблице 6 соответственно. Количество ошибок, отмеченных всеми испытуемыми одновременно, является незначительным. Следует упомянуть, что для голоса «Анна» большей частью испытуемых (более 75 процентов) отмечались только ошибки в интонационном оформлении предложений, в то время как для голоса «Tatyana» большинством были отмечены также ошибки в словесном ударении и произнесении.

Таблица 5. Данные по ошибкам, отмеченным большинством испытуемых для голоса «Tatyana»

Отметили ошибку, %	Ударение	Произнесение	Паузы	Темп/ритм	Интонация	Плавность	Общ. кач. голоса	Иное	Всего
Более 50	6	10	-	-	23	-	1	-	40
Более 75	4	2	-	-	5	-	-	-	11

Таблица 6. Данные по ошибкам, отмеченным большинством испытуемых для голоса «Анна»

Отметили ошибку, %	Ударение	Произнесение	Паузы	Темп/ритм	Интонация	Плавность	Общ. кач. голоса	Иное	Всего
Более 50	1	3	-	-	22	4	-	-	30
Более 75	-	-	-	-	4	-	-	-	4

Из проведенного теста можно сделать вывод, что ошибки в интонационном оформлении синтезированной речи являются главной проблемой современных русскоязычных селективных синтезаторов, что как раз может быть связано с алгоритмом селективного синтеза речи, используемым в современных системах, когда из звукового корпуса выбираются звуковые элементы, не точно соответствующие требованиям к их просодическим характеристикам. Также следует отметить, что тестируемые синтезаторы отличаются по качеству лингвистической и акустической обработки, причём несколько больший вес имеют ошибки, связанные с неправильной постановкой словесного ударения и неверной транскрипцией.

Некоторые возможные ошибки в синтезированной речи не были охвачены данным экспериментом. Для оценки пауз между предложениями нужен дополнительный эксперимент. В описанный тест не была включена также оценка слушающими возможных ошибок, связанных с неправильным чтением числительных, записанных цифрами, аббревиатур, телефонов, неправильной

расшифровкой сокращений, неправильным делением на предложения и т. п., то есть ошибок, возникающих при работе модуля нормализации текста. Такие ошибки влияют не только на естественность, но и на понимание синтезированной речи.

4.2 Оценка лингвистической обработки

Для оценки лингвистической обработки в общем случае были выбраны синтезаторы, имеющие демо-версии без фоновой музыки (Acapela, iSpeech, Ivona TTS, Mary TTS, Nuance Vocalizer, VitalVoice TTS). Оценка проводилась при помощи слухового анализа лично автором диссертации. Общая длительность протестированных звукозаписей на каждый синтезированный голос составила около 40 минут.

Следует отдельно отметить, что открытый для свободного использования синтезатор Mary TTS имеет несколько худшее качество, чем коммерческие продукты, что касается и лингвистической обработки: не читаются цифры и символы латиницы, инициальные аббревиатуры читаются как обычные слова. Поэтому для этого синтезатора данные по экспериментам приводятся только в тестах на правильность чтения омографов и правильность сегментной транскрипции. При составлении тестов предложения, содержащие анализируемые слова или обозначения, в общем случае случайным образом выбирались из Национального корпуса русского языка [Национальный корпус русского языка]. При оценке качества различных синтезаторов следует иметь возможность изменять набор тестовых данных для того, чтобы исключить возможность настройки синтезатора на конкретный ограниченный список тестовых предложений.

4.2.1 Оценка правильности выделения предложений

В тестовом тексте из п. 4.1 при слуховом анализе не было обнаружено ни одной ошибки в выделении предложений обоими (Ivona TTS, VitalVoice TTS) синтезированными голосами. Большинство подобных ошибок возникает в сложных случаях при чтении технических или новостных текстов, где встречаются сокращения с точкой, или же в случаях неправильной трактовки специальных символов (знак абзаца, перенос строки и т. п.), а также при чтении коротких фраз без точки, начинающихся с новой строки. Поэтому в общем случае правильное деление на предложения зависит от того, насколько хорошо исследуемый синтезатор справляется с расшифровкой графических сокращений с точкой (см. п. 4.2.2.1).

4.2.2 Оценка чтения аббревиатур

Тест на чтение аббревиатур был разделён на две части. Это связано с тем, что чтение графических сокращений зависит от контекста, что необходимо учитывать при анализе результатов.

4.2.2.1 Оценка расшифровки графических сокращений

Для эксперимента были выбраны 50 предложений из Национального корпуса русского языка [Национальный корпус русского языка], содержащие различные частотные графические сокращения. Список предложений приводится в Приложении В.

В тестовом тексте содержится 87 сокращений. При оценке повторные сокращения учитывались как различные ошибки, поскольку их расшифровка зависит от места в предложении. Если для конкретного сокращения допустимыми являются несколько вариантов прочтения, все они считались правильными. Ошибкой считалась неправильная форма расшифрованных слов.

Исследование проводилось с использованием демо-версий пяти голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Alyona (Acapela Group).
5. Женский голос синтезатора iSpeech.

Результаты теста представлены в таблице 7. Подробный список ошибок по предложениям дается в Приложении Г.

Таблица 7. Процент ошибочного чтения графических сокращений различными синтезаторами

Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Alyona Acapela	Female iSpeech
Процент ошибок	59,8	20,7	67,8	48,3	72,4

Из полученных данных можно сделать вывод о том, что с чтением сокращений полностью не справляется ни один из тестируемых синтезаторов. Наиболее хорошие результаты показала система VitalVoice (для которой русский язык является основным разрабатываемым языком), также более половины правильно озвученных сокращений у синтезатора компании Acapela. Худший результат у системы iSpeech, которая правильно озвучила только четверть тестовых предложений. Практически все синтезаторы правильно озвучивают только самые частотные сокращения типа «и т. д.», часто само слово расшифровывается правильно, но имеет неверную форму. Особо следует отметить, что у всех тестируемых синтезаторов были уникальные ошибки, что говорит о том, что они используют различные правила расшифровки сокращений.

Для оценки статистической значимости различий в работе синтезаторов и достаточности объёма тестового материала был применен критерий хи-квадрат К. Пирсона (поскольку необходимо было сравнить качественные показатели более двух систем). Исходными данными послужило количество ошибок, выделенных для каждого синтезированного голоса. Число степеней свободы равно 4. Значение критерия χ^2 составляет 59,691. Критическое значение χ^2 при уровне значимости $p < 0,01$ составляет 13,277. Различия между синтезаторами в количестве ошибок при чтении графических сокращений статистически значимы при уровне значимости $p < 0,01$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов.

4.2.2.2 Оценка чтения инициальных аббревиатур, сложносокращённых слов и смешанных сокращений

Для эксперимента были выбраны 50 предложений из Национального корпуса русского языка [Национальный корпус русского языка], содержащих частотные аббревиатуры (в том числе сложносокращённые слова и инициальные аббревиатуры на латинице). Список предложений приводится в Приложении Д.

В тестовом тексте содержится 91 различная аббревиатура. При оценке повторные аббревиатуры не учитывались, поскольку их расшифровка не зависела от места в предложении. При этом в спорных случаях при не очень хорошем качестве сегментных звуковых единиц аббревиатура прослушивалась и оценивалась отдельно вне контекста. Если для конкретной аббревиатуры допустимыми являются несколько вариантов прочтения, все они считались правильными.

Исследование проводилось методом слухового анализа автором диссертации с использованием демо-версий пяти голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Alyona (Acapela Group).
5. Женский голос синтезатора iSpeech.

Результаты теста представлены в Таблице 8. Подробный список ошибок по предложениям приводится в Приложении Е.

Таблица 8. Процент ошибочного чтения аббревиатур различными синтезаторами

Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Alyona Acapela	Female iSpeech
Процент ошибок	25,3	1,1	22	17,6	26,4

Здесь, как и в предыдущем тесте, лучший показатель у системы VitalVoice – практически стопроцентное правильное озвучивание аббревиатур (только одна ошибка при чтении аббревиатуры «МИМО»). В целом, задача озвучивания инициальных аббревиатур оказалась более простой, чем расшифровка графических сокращений с точкой, и максимальный процент ошибок составил чуть более четверти всех аббревиатур для синтезатора iSpeech.

Для оценки статистической значимости различий в работе синтезаторов был применен критерий хи-квадрат К. Пирсона (поскольку необходимо было сравнить качественные показатели более двух систем). Число степеней свободы равно 4. Значение критерия χ^2 составляет 25,609. Критическое значение χ^2 при уровне значимости $p < 0,01$ составляет 13,277. Различия в количестве ошибок при чтении графических аббревиатур между синтезаторами статистически значимы при уровне значимости $p < 0,01$, следовательно, выбранное количество тестовых

предложений достаточно для выявления различий в работе синтезаторов при чтении аббревиатур.

4.2.3 Оценка чтения цифровых обозначений

Для эксперимента были выбраны 25 предложений из Национального корпуса русского языка [Национальный корпус русского языка], содержащих цифровые обозначения (в том числе римские цифры, дроби, даты, обозначения времени и т. д.). Список предложений приводится в Приложении Ж.

В тестовом тексте содержится 63 различных цифровых обозначения. Если для конкретного цифрового обозначения допустимыми являлись несколько вариантов прочтения, все они считались правильными. В тех случаях, когда ошибки синтеза при чтении цифровых обозначений были связаны с неверной расшифровкой графических сокращений или специальных символов (например, «%»)), на вход синтезатора подавалось предложение с расшифрованными сокращениями и символами.

Исследование проводилось с использованием демо-версий пяти голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Alyona (Acapela Group).
5. Женский голос синтезатора iSpeech.

Результаты теста представлены в таблице 9. Подробный список ошибок по предложениям дается в Приложении И.

Таблица 9. Процент ошибочного чтения цифровых обозначений различными синтезаторами

Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Alyona Acapela	Female iSpeech
Процент ошибок	38,1	17,5	36,5	38,1	41,3

При расшифровке цифровых обозначений ни один синтезатор не смог правильно озвучить простые дроби («1/2», «1/3» и т. п.). Также большинство синтезаторов неправильно озвучивает время в формате чч.мм (как десятичную дробь), и при расшифровке не склоняются большие числа, записанные в нестандартной форме («5 млн. 800 тысячам»).

Для оценки статистической значимости различий в работе синтезаторов применялся, как и в прочих случаях (см. выше) критерий хи-квадрат К. Пирсона (поскольку сравнивались качественные показатели более двух систем). Число степеней свободы равно 4. Значение критерия χ^2 составляет 10,229. Критическое значение χ^2 при уровне значимости $p < 0,05$ составляет 9,488. Различия в количестве ошибок между синтезаторами при чтении цифровых обозначений статистически значимы при уровне значимости $p < 0,05$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов при чтении аббревиатур.

4.2.4 Оценка чтения специальных символов

Для эксперимента были выбраны 10 предложений из Национального корпуса русского языка [Национальный корпус русского языка], содержащих специальные символы, 10 предложений были сконструированы вручную, а также взяты из других интернет-источников. Список предложений приводится в Приложении К.

В тестовом тексте содержится 21 специальный символ («+», «%», «&» и т. п.). При этом не учитывались как разные повторяющиеся символы в сходном контексте. Если для конкретного символьного обозначения допустимыми являлись несколько вариантов прочтения, все считались правильными, в том числе пропуск символа при чтении для таких символов как «/», «™» и «®». Дополнительно для тестирования было взято одно предложение, содержащее сокращённые названия химических элементов («Fe», «Mg» и т. п.), которое не было правильно озвучено ни одним из исследуемых синтезаторов речи.

Исследование проводилось с использованием демо-версий пяти голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Alyona (Acapela Group).
5. Женский голос синтезатора iSpeech.

Результаты теста представлены в таблице 10. Подробный список ошибок по предложениям приведен в Приложении Л.

Таблица 10. Процент ошибочного чтения символьных обозначений различными синтезаторами

Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Alyona Acapela	Female iSpeech
Процент ошибок	57,1	19,1	28,6	4,8	66,7

Абсолютным лидером при чтении символьных обозначений стал синтезатор компании Acapela, не справившийся только с наиболее сложным предложением,

содержащим обозначение соотношения «в соотношении 1:1». Это предложение было правильно озвучено программой Ivona TTS. Демо-версия синтезатора Nuance Vocalizer целиком не читала предложения, содержащие символ «%». Вторым сложным случаем, озвученным только синтезатором Asapela, стало обозначение размера «406×40 мм».

Для оценки статистической значимости различий в работе синтезаторов был, как и ранее, применен критерий хи-квадрат К. Пирсона (поскольку сравнивались качественные показатели более двух систем). Число степеней свободы равно 4. Значение критерия χ^2 составляет в этом случае 24,873. Критическое значение χ^2 при уровне значимости $p < 0,01$ равно 13,277. Различия между синтезаторами в количестве ошибок при чтении символьных обозначений статистически значимы при уровне значимости $p < 0,01$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов.

4.2.5 Оценка чтения иностранных слов на латинице

Подробная оценка правильности чтения иноязычных вставок не входит в задачи данного исследования. Для получения представления о том, насколько с этой задачей справляются различные русскоязычные селективные синтезаторы, был взят небольшой список английских слов, использованный О. Д. Черепановой в работе [Черепанова 2015: 46] при оценке качества синтеза англоязычных вставок. При проведении эксперимента учитывались два показателя:

1. Процент точного соответствия практической англо-русской транскрипции, используемой в работе [Черепанова 2015].
2. Процент допустимого (с точки зрения автора диссертации) чтения таких слов, незначительно снижающего их восприятие (например, передача фонемы [w] через /y/ или изменение места ударения, соответствующее ударению в английском языке

для слова «Manchester»). Этот показатель также включал в себя и все точные соответствия.

Всего было озвучено 27 английских слов и словосочетаний следующими голосами систем синтеза:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Alyona (Acapela Group).
5. Женский голос синтезатора iSpeech.

Список слов с практической транскрипцией приводится в Приложении М. Результаты эксперимента приведены в таблице 11, подробные результаты – в Приложении Н.

Таблица 11. Проценты ошибочного и неточного чтения англоязычных вставок различными синтезаторами

Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Alyona Acapela	Female iSpeech
Процент явных ошибок	25,9	7,4	74,1	0	85,2
Процент несоответствий	44,4	22,2	92,6	66,7	88,9

Наилучший результат по чтению английских слов показала система Acapela, слова, озвученные данной системой, звучали допустимо (см. выше) (основными отличиями от «идеальной» транскрипции было чтение [w] через /y/, изменение места ударения на соответствующее место ударения в английском языке, чтение Samsung как «сэмсанг» и т. п.). В то же время у голоса «Юлия» наибольший процент точных совпадений с практической транскрипцией, предлагаемой

О. Д. Черепановой (более 75 %). На качество озвучивания англоязычных вставок синтезатором Ivona TTS в первую очередь влияет передача английской буквы «с» через русскую «ч» («блачк анд уйте» (black and white), «астропысичал» (astrophysical) и т. п.). В синтезаторе iSpeech буква «h» передаётся всегда через букву «х», независимо от контекста («партнерсхип» – partnership).

Для оценки статистической значимости различий в работе синтезаторов также был применен критерий хи-квадрат К. Пирсона (поскольку необходимо было сравнить качественные показатели более двух систем). Число степеней свободы равно 4. Значение критерия χ^2 составляет 69,001. Критическое значение χ^2 при уровне значимости $p < 0,01$ составляет 13,277. Различия между синтезаторами в количестве ошибок при чтении английских слов (для явных ошибок) статистически значимы при уровне значимости $p < 0,01$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов.

4.2.6 Оценка правильности определения места ударения

Для оценки правильности места словесного ударения были использованы два текста: фонетически представительный текст из эксперимента в п. 4.1, а также набор предложений, содержащих омографы как наиболее частые случаи ошибок в ударении, (см. п. 4.2.6.1).

По результатам анализа данных из эксперимента по влиянию различных типов ошибок можно сделать несколько выводов. Во-первых, аудиторы довольно точно отмечают наличие ошибок в месте ударения. Во-вторых, качество определения места ударения для тестируемых голосов различно (см. таблицу 12 ниже). Для голоса «Анна» (VitalVoice TTS) такая ошибка была всего одна – в слове «сЮжет», по-видимому, она связана с неудачным подбором звуковых единиц для конкатенации, а не с ошибкой лингвистической обработки, что было проверено

синтезом тестового текста другим голосом того же синтезатора (Юлия), а также изменением контекста слова. У голоса «Tatyana» (Ivona TTS) ошибок в ударении несколько больше, а именно семь. Из них пять были допущены в словах-омографах, также неправильно было прочитано слово «все» (как «всё»), что тоже связано с проблемой омографии. Также следует отметить две ошибки с лишними ударениями: «Вот тАк встреча!» и «рыба пО-бенгальски». Неправильно были прочитаны также несколько слов с буквой «ё», написанные через «е» (подробнее в разделе по оценке транскрипции), возможно, с какими-то проблемами с обработкой буквы «ё» связана и ошибка в слове «плЫвешь». В слове «попрощЕ» ошибка в ударении, также как и у голоса «Анна», скорее всего, связана с неудачным подбором звукового элемента. Таким образом, проверка правильности снятия омонимии для слов-омографов является одним из основных показателей качества лингвистической обработки текста по определению места словесного ударения.

Таблица 12. Ошибки различных синтезаторов в ударениях при чтении фонетически представительного текста (приведены неправильные ударения, в том числе в омографах)

Голос	Список ошибочных слов
Анна	сЮжет
Tatyana	самОй гОры городА дрУжны плЫвешь гОду видУ

Правильность определения места ударения в именах собственных в данной диссертационной работе не проверялась, так как выбор тестовых материалов в таком случае очень сильно зависит от конкретной ситуации использования

синтезатора речи. Однако следует подчеркнуть, что, наряду с ошибками в слова-омографах, ошибки чтения имён собственных являются наиболее частотными для всех протестированных синтезаторов.

4.2.6.1 Оценка правильности снятия омонимии

Для эксперимента были выбраны 50 предложений из Национального корпуса русского языка [Национальный корпус русского языка], содержащих частотные слова-омографы. Список предложений приводится в Приложении П.

В тестовом тексте содержится 25 пар исходных омографов, по которым производился поиск и дополнительно 11 омографов, также содержащихся в выбранных предложениях. В тех случаях, когда ошибки синтеза при чтении цифровых обозначений были связаны с неверной расшифровкой графических сокращений или специальных символов (например, «%»»), на вход синтезатора подавалось предложение с расшифрованными сокращениями и символами.

Исследование проводилось с использованием демо-версий шести голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Синтезатор Mary TTS (мужской голос)
5. Alyona (Acapela Group).
6. Женский голос синтезатора iSpeech.

Результаты теста представлены в таблице 13. Подробный список ошибок по предложениям дается в Приложении Р.

Таблица 13. Процент ошибочного чтения слов-омографов различными синтезаторами

Голос	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Male Mary TTS	Alyona Acapela	Female iSpeech
Процент ошибок	39,3	1,6	42,6	54,1	34,4	21,3

Синтезатор VitalVoice допустил при выборе места ударения всего одну ошибку, в то время как в синтезаторе Mary TTS имеется явная проблема с выбором места ударения не только в словах-омографах, что, по всей видимости, связано с малым размером словаря.

Для оценки статистической значимости различий в работе голосов/синтезаторов был применен критерий хи-квадрат К. Пирсона (поскольку сравнивались качественные показатели более двух систем). Число степеней свободы равно 5. Значение критерия χ^2 составляет 41,478. Критическое значение χ^2 при уровне значимости $p < 0,01$ равно 15,086. Различия между синтезаторами в количестве ошибок при чтении омографов статистически значимы при уровне значимости $p < 0,01$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов.

Выводы. Общие результаты исследований, описанных в данном разделе, являются в некоторой степени ожидаемыми. При выполнении задач лингвистической обработки текста, связанных непосредственно с русским языком, лучшие показатели у системы синтеза VitalVoice TTS, разрабатываемой в России, в первую очередь для русского языка. Однако при чтении иноязычных вставок и специальных символов самые хорошие результаты показывает синтезатор компании Acapela Group, что, по всей видимости, связано с тем, что соответствующие модули могли быть встроены в русскоязычный синтез из более

разработанных языков. Некоммерческий голос системы Mary TTS ожидаемо показал самый плохой результат, не справившись с большинством задач. Среди остальных коммерческих систем показатели качества варьируются в зависимости от конкретной задачи.

4.3 Оценка фонетической обработки

В данном разделе блоки просодической и фонетической обработки объединены в один, так как при исключительно аудитивной оценке интонации звучащей синтезированной речи без доступа к результатам работы соответствующих модулей невозможно протестировать правильность интонационной транскрипции независимо от результирующих физических характеристик речи.

4.3.1 Оценка сегментной транскрипции

Для русского языка, в котором для большей части слов соответствие «буква–фонема» может быть определено ограниченным набором правил чтения, можно выделить несколько зон потенциальных ошибок в сегментной транскрипции для синтезаторов речи (не считая ошибок в определении места ударения):

1. Чтение слов с буквой «ё» через «е» (например, «нем» вместо «нём»).
2. Смягчение согласных в заимствованных словах перед буквой «е» (например, «синтЕз»).
3. Чтение безударных «о» и «э» как «а» или «и»/«ы» соответственно в определённых контекстах и группах слов (например, «какаО»).
4. Озвучивание произносимых согласных (например, «соЛнце», «чуВство»).
5. Отсутствие ассимиляции по месту или способу образования согласных (например, «выеЗжаю»).
6. Ошибки, связанные с кратким или ошибочно долгим чтением двойных согласных (например, «класС»).

7. Неверное оглушение/озвончение согласных (например, оглушение предлогов перед словом с начальной буквой «в» – «беЗ ветки»)
8. Иные ошибки (например, чтение окончаний прилагательных «-оГо» или чтение слов с этим сочетанием не в окончании через «ово», например, в слове «мноГо») и т. д.

Для проверки правильности сегментной транскрипции из национального корпуса русского языка были выбраны 14 предложений, содержащих потенциально ошибочные слова. Одно предложение для контрастной проверки чтения двойных согласных было сконструировано вручную. Полный список предложений приводится в Приложении С. Общая длительность звукозаписей для каждого голоса составила около 2 минут речи.

Исследование проводилось лично автором диссертации методом слухового анализа с использованием демо-версий шести голосов различных систем синтеза речи:

1. Katya (Nuance Vocalizer).
2. Юлия (VitalVoice TTS).
3. Tatyana (Ivona TTS).
4. Синтезатор Mary TTS (мужской голос)
5. Alyona (Acapela Group).
6. Женский голос синтезатора iSpeech.

Результаты тестирования приведены в таблице 14. Всего при озвучивании тестовых предложений было обнаружено 16 различных ошибок в транскрипции. Подробный список ошибок приводится в Приложении Т.

Таблица 14. Количество ошибок в сегментной транскрипции у различных синтезаторов

Голос/ Синтезатор	Katya Nuance	Юлия VitalVoice	Tatyana Ivona	Male Mary TTS	Alyona Acapela	Female iSpeech
Количество ошибок	8	2	6	9	2	9

Наиболее сложным случаем оказалось чтение слова «нестрого», которое большинство синтезаторов озвучили как «нестроВо». Также следует отметить, что в синтезаторе Nuance Vocalizer существуют явные проблемы с чтением двойных согласных и буквы «ё», не обозначенной в тексте в явном виде. Синтезатором Mary TTS был допущен ряд ошибок сверх ожидаемых (что может быть связано не только с ошибками в транскрипторе, но и с некачественной разметкой речевой базы).

Для оценки статистической значимости различий в работе синтезаторов был применен критерий хи-квадрат К. Пирсона (поскольку необходимо было сравнить качественные показатели более двух систем). Число степеней свободы равно 5. Значение критерия χ^2 составляет 14,4. Критическое значение χ^2 при уровне значимости $p < 0,05$ составляет 11,07. Различия между синтезаторами в количестве ошибок в сегментной транскрипции статистически значимы при уровне значимости $p < 0,05$, следовательно, выбранное количество тестовых предложений достаточно для выявления различий в работе синтезаторов.

4.3.2 Оценка паузирования в синтезированной речи

Оценка правильности определения локализации пауз затруднена тем, что для получения корректных результатов требуется озвученный синтезаторами текст достаточной длины, что трудновыполнимо при наличии лишь демо-версий, так как в них существует ограничение на количество символов (в большинстве

синтезаторов – не более 300 символов). В озвученном двумя синтезированными голосами фонетически представительном тексте (см. п. 4.1) не было выявлено явно ошибочных лишних или недостающих пауз. Следует отметить, что при проведении эксперимента по влиянию различных типов ошибок на оценку естественности синтеза (см. п. 4.1), в некоторых предложениях несколько auditors (не более четырёх) отметили наличие ошибок в паузировании (в эту категорию входило не только наличие/отсутствие паузы, но также и её неверная длительность). Рассмотрим случаи, где трое или более auditors отметили ошибки такого рода, подробнее.

Для голоса «Анна» таких предложений с потенциальными ошибками в паузации было больше (далее в предложениях символ «//» обозначает паузы в чтении, сделанные обоими синтезаторами, /*T/ – только «Tatyana»):

1. *Прямо передо мной внизу – // пестрый узор из крыш городских домов, // а вдалеке, // на краю горизонта, // тянется серебряная цепь снеговых вершин...*
(Анна – 3 auditors отметили ошибку)
2. *Минуты две-три ищу в шкафу /*T/ электрическую схему, // привезенную французским коллегой.* (Tatyana – 3 auditors отметили ошибку)
3. *Останавливаюсь в начале мостика у ограды, // чтобы полюбоваться рельефными склонами горных хребтов, // всматриваюсь в речную рябь.*
(Анна – 3 auditors отметили ошибку)
4. *Мы были дружны, // но потом, // мало-помалу, // связь наша оборвалась.*
(Анна – 4 auditors отметили ошибку)
5. *- Следующим пунктом Уфа.* (Анна, Tatyana – 3 auditors отметили ошибку)
6. *А ты, // значит, /*T/ теперь здесь живешь? // Давно?* (Анна – 4 auditors, Tatyana – 3 auditors отметили ошибку)

В первом предложении, в котором были отмечены ошибки только для голоса Анна, возможной причиной при совпадении мест пауз послужили более длительные вторая и четвёртая паузы (более 300 мс) по сравнению со вторым голосом Tatyana, где самая длинная вторая пауза составляет примерно 200 мс, несмотря на более медленный общий темп чтения (см. рис. 20).

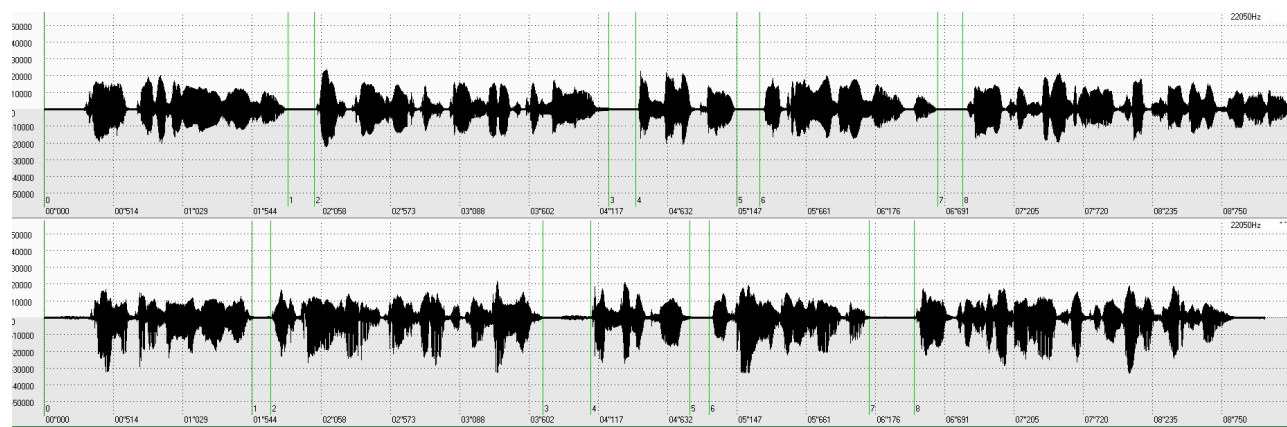


Рис. 20. Синхронизированные осциллограммы звукозаписей синтезированного предложения «Прямо передо мной внизу – // пестрый узор из крыши городских домов, // а вдалеке, // на краю горизонта, // тянется серебряная цепь снеговых вершин...» для голосов Tatyana (вверху) и Анна (внизу) с отмеченными паузами

Во втором предложении у голоса «Tatyana» по сравнению с голосом «Анна» в чтении есть дополнительная пауза на отрезке предложения между знаками препинания.

В третьем предложении возможной причиной ошибки, отмеченной для голоса «Анна», при совпадении пауз у обоих голосов послужило резкое падение тона и долгий гласный в слове «Остана\вливаюсь», возможно, выбранный в речевой базе из конечного во фразе слова с фразовым ударением, так что, несмотря на отсутствие паузы, на этом месте слышна граница синтагм.

В четвёртом предложении в голосе «Анна» пауза после слова «потом» также является слишком длинной – более 400 мс.

В предложении «Следующим пунктом Уфа» паузы не поставили оба синтезатора. По всей видимости, ожидаемой для трёх участников эксперимента была пауза перед словом «Уфа».

В последнем предложении лишними, по мнению auditors, являются паузы, выделяющие вводное слово «значит».

Можно сделать вывод, что большая длина пауз в голосе «Анна» могла повлечь за собой оценку auditors таких пауз как ошибочных. Также следует отметить, что «сомнения» auditors по поводу выделения паузами вводных слов согласуются с данными, приведёнными в исследовании [Кривнова, Чардин 1999].

4.3.3 Оценка интонации синтезированной речи

В настоящем разделе предлагается двухэтапный подход к диагностическому тестированию качества интонации русскоязычной синтезированной речи, также описанный в статье [Solomennik, Cherentsova 2013]. На основе прослушивания и оценки тестовых фраз испытуемыми были вычислены интегральные коэффициенты интонационной адекватности и естественности. Тестирование было проведено для нескольких голосов русскоязычных синтезаторов речи (см. ниже). В тесте от auditors требовалось принять лингвистическое решение относительно коммуникативного типа воспринимаемого высказывания. Одновременно оценивалась естественность интонации по трехбалльной шкале. По результатам эксперимента предлагаются рекомендации, позволяющие улучшить просодическую обработку в синтезаторах.

Будем называть интонацию синтезированной речи адекватной, если при прослушивании возможно восстановить пунктуацию исходного текста в той же степени, насколько это возможно для естественной речи при чтении текста.

При оценке рассматриваются перцептивно детектируемые ошибки, неточности и несоответствия в синтезированной речи в широком смысле, позволяющие производить улучшение просодического компонента системы синтеза с учетом слабых мест в определении интонационной структуры синтезируемых высказываний.

Текстовый материал для исследования выбирался с использованием существующих знаний о различных типах интонационных конструкций русского языка и их частотности.

Критерий адекватности коммуникативного типа высказываний основан на использовании оппозиции между интонацией завершенности (понижающийся тон «\»), незавершенности (умеренный подъем тона «/-») и вопросительной (значительный подъем тона «/») интонацией. Даже адекватная интонация в синтезированной речи может звучать неестественно и раздражающе и тем самым нарушать ее комфортное восприятие: на корректное понимание сообщения может влиять отклонение того, что было услышано, от ожиданий слушающего. Например, нейтральный тон эмоционального восклицания или агрессивного императива (произнесённого ровно или с гортанной смычкой как в русском «Какая приятная неожиданность!») может восприниматься как ироничный, а дружелюбная ровная манера произнесения угрозы или предупреждения, такого как «Опасная зона! Немедленно покиньте территорию!» может привести к непониманию или игнорированию предупреждения).

Для того чтобы получить представительный набор фраз для тестирования интонации использовался размеченный речевой корпус, состоящий из записей различных по стилю текстов, прочитанных восемью дикторами, четырьмя мужчинами и четырьмя женщинами, около десяти часов речи на диктора. Корпус был записан на кафедре фонетики СПбГУ для проекта по созданию синтеза

русской речи VitalVoice (ООО «ЦРТ»). Два текста («Обмен» Ю. Трифонова и «Поздний ребёнок» А. Алексина) были прочитаны всеми восемью дикторами. В процессе разметки корпуса использовалось 13 типов интонации [Вольская, Скредин 2009], соответствующих расширенному набору семи интонационных конструкций (ИК), обычно используемых для описания русской интонации [Брызгунова 1980]. На основе частотности различных типов интонации в корпусе была определена доля фраз с определённой интонацией в тестовом наборе.

Для эксперимента были подобраны нейтральные фразы с минимумом лексических или синтаксических маркеров коммуникативного типа, т. е. потенциально многозначные, допускающие несколько значений (как минимум два), если они представлялись вне контекста и без знаков препинания. Несколько таких маркеров, однако, были представлены в тестовых предложениях. Это в основном союзы и вопросительные слова, отмечающие вопросы с вопросительными словами («Где вы родились?») и восклицания («Какой день чудесный!»), а также повелительные глаголы и наречия, указывающие на императив («Сиди дома!»). Несмотря на менее значительный вклад интонации в корректное понимание таких фраз, они были добавлены в набор, чтобы провести различие между независимыми фразами, произнесёнными отдельно вне контекста, и незавершёнными фразами той же структуры, которые являются фрагментами, выделенными из соответствующего предложения. Например, вопрос с вопросительным словом в прямой речи будет иметь более интенсивное ударение на вопросительном слове и иметь динамичную основную частоту, в то время как тот же фрагмент, являющийся косвенной речью, будет произноситься невыразительно или даже нейтральным тоном.

Для прослушивания аудиторам были предложены 40 тестовых фраз, при этом в них присутствовало следующее количество синтагм с требуемой

интонацией: завершённости – 27, незавершённости – 18, специальных вопросов – 9, общих вопросов – 20, восклицаний – 6, побудительных предложений – 2. Эти фразы были просинтезированы несколькими русскоязычными голосами разных синтезаторов:

1. Alyona (Acapela TTS),
2. Milena (Nance Vocalizer)
3. Olga (Loquendo TTS),
4. Юлия и Владимир (VitalVoice TTS).

Было также записано естественное чтение того же набора фраз русскоязычным диктором-женщиной 28 лет. Набор фраз приводится в Приложении У. Общая длительность звукозаписей для каждого голоса составила около 2 минут речи.

В ходе перцептивного теста аудиторов просили поставить во фразах знаки препинания. Кроме принятия лингвистического решения о коммуникативном типе каждого высказывания, испытуемые должны были оценить его естественность (по двухбалльной шкале). Для окончательной оценки результатов использовалась трёхбалльная шкала (0-2): 2 соответствовало правильной интерпретации и выбору знака пунктуации и оценке фразы как естественной, 1 – правильной пунктуации, но оценке интонации фраз как неестественной и 0 ставился тем фразам, где знаки пунктуации были поставлены неправильно. Образец анкеты приводится в Приложении Ф.

В эксперименте принимали участие 10 испытуемых в возрасте от 23 до 50 лет, среди которых четыре мужчины и шесть женщин, для которых русский язык является родным.

Для того чтобы оценить качество интонации в синтезированной речи, были подсчитаны два показателя: общая правильность (естественность) интонации (процент фраз с интонацией, корректно проинтерпретированной пунктуационно и отмеченной как естественная) и её адекватность (процент фраз с корректно проинтерпретированной интонацией). Средневзвешенное значение использовалось для подсчёта этих показателей по всем типам интонации, в качестве весов использовалось распределение значений частотности каждой интонационной структуры.

Таблица 15. Показатели общей правильности и адекватности интонации для синтезированных голосов (среднеквадратичное отклонение даётся в скобках). Минимальные значения отмечены жирным шрифтом

		Завершён- ность	Специальный вопрос	Общий вопрос	Восклицание	Незавершён- ность	Всего
Alyona	Ест.	0,72 (0,24)	0,14 (0,12)	0,24 (0,17)	0,19 (0,16)	0,49 (0,17)	0,54 (0,13)
	Адекв.	0,74 (0,10)	0,45 (0,10)	0,31 (0,09)	0,21 (0,14)	0,77 (0,08)	0,67 (0,07)
Milena	Ест.	0,83 (0,14)	0,10 (0,09)	0,08 (0,13)	0,19 (0,14)	0,54 (0,08)	0,60 (0,07)
	Адекв.	0,84 (0,15)	0,26 (0,27)	0,21 (0,24)	0,26 (0,13)	0,62 (0,38)	0,66 (0,08)
Olga	Ест.	0,51 (0,25)	0,30 (0,19)	0,11 (0,11)	0,19 (0,19)	0,50 (0,16)	0,45 (0,16)
	Адекв.	0,66 (0,24)	0,45 (0,12)	0,21 (0,24)	0,24 (0,17)	0,75 (0,17)	0,65 (0,16)
Юлия	Ест.	0,88 (0,11)	0,25 (0,12)	0,25 (0,22)	0,34 (0,18)	0,50 (0,16)	0,64 (0,07)
	Адекв.	0,88 (0,11)	0,33 (0,11)	0,59 (0,15)	0,41 (0,19)	0,70 (0,17)	0,72 (0,08)
Владимир	Ест.	0,86 (0,13)	0,24 (0,17)	0,38 (0,30)	0,25 (0,15)	0,61 (0,22)	0,66 (0,15)
	Адекв.	0,87 (0,10)	0,39 (0,19)	0,57 (0,25)	0,25 (0,15)	0,74 (0,19)	0,72 (0,09)
Ест. речь	Ест.	0,97 (0,02)	1,00 (0,00)	0,78 (0,16)	0,89 (0,10)	0,94 (0,08)	0,92 (0,03)
	Адекв.	0,99 (0,00)	1,00 (0,00)	0,78 (0,16)	0,90 (0,09)	0,94 (0,08)	0,93 (0,02)

Таблица 16. Показатели общей правильности и адекватности интонации для синтезированных голосов, нормализованные относительно естественной речи. Минимальные значения отмечены жирным шрифтом

		Завершён- ность	Специальны й вопрос	Общий вопрос	Восклицание	Незавершён- ность	Всего
Alyona	Ест.	0,74	0,14	0,31	0,21	0,52	0,59
	Адекв.	0,75	0,45	0,40	0,23	0,82	0,72
Milena	Ест.	0,85	0,10	0,10	0,21	0,58	0,65
	Адекв.	0,85	0,26	0,27	0,29	0,66	0,71

Olga	Ест.	0,52	0,30	0,14	0,21	0,53	0,49
	Адекв.	0,67	0,45	0,27	0,27	0,80	0,70
Юлия	Ест.	0,90	0,25	0,32	0,38	0,53	0,70
	Адекв.	0,89	0,33	0,75	0,45	0,74	0,77
Владимир	Ест.	0,88	0,24	0,49	0,28	0,65	0,72
	Адекв.	0,88	0,39	0,73	0,28	0,79	0,77

В результате общая естественность интонации синтезированных голосов, участвовавших в тестировании, нормированная на естественный голос, изменяется от 49 % для голоса «Olga» до 72 % («Владимир»).

Показатель точности обнаруживает меньшую междикторскую вариативность, приближаясь к 70 процентам для всех синтезированных голосов. Восприятие голоса «Olga» (Loquendo) могло быть несколько нарушено фоновой музыкой в демо-версии синтезатора.

Различия в работе синтезаторов, подсчитанные для среднего количества фраз с точной/естественной интонацией, не являются статистически значимыми на уровне значимости $p < 0,05$. Был использован критерий χ^2 , число степеней свободы равно 4, значение критерия χ^2 для фраз с естественной интонацией составляет 4,598, с адекватной – 1,066. Критическое значение χ^2 при уровне значимости $p < 0,05$ составляет 9,488.

Показатели, подсчитанные отдельно для каждого интонационного типа, позволяют найти слабые места каждой системы синтеза в отношении определённых интонационных конструкций. Самый низкий показатель точности был получен для вопросов с вопросительными словами и восклицаний.

Можно соотнести полученные данные с результатами эксперимента по влиянию ошибок на естественность речи (п. 4.1) в части средней оценки процента ошибок в интонации (100 – показатель точности). При этом данные по

эксперименту из п. 4.1 совпадают со средним процентом «интонационно неточных» предложений данного эксперимента (42 %).

Таблица 17. Процент интонационных ошибок по данным двух экспериментов

Эксперимент из п. 4.3.3 (оценка интонации)	
Alyona (Acapela TTS)	46 %
Milena (Nuance TTS)	40 %
Olga (Loquendo TTS)	55 %
Юлия (VitalVoice TTS)	36 %
Владимир (VitalVoice TTS)	34 %
Эксперимент из п. 4.1 (влияние ошибок на естественность речи)	
Tatyana (Ivona TTS)	42 %
Анна (VitalVoice TTS)	41 %

Выводы. В данном разделе были рассмотрены результаты оценки фонетической обработки, применяемой в синтезаторах речи. Наибольшую правильность сегментной транскрипции по результатам тестов показали синтезаторы Acapela TTS и VitalVoice TTS. Доля правильно определенных локализаций пауз для протестированных синтезаторов при чтении художественного текста приближается к 100 %, в то время как правильность

мелодического оформления синтезированной речи в среднем для протестированных синтезаторов составляет 58 %, что связано в первую очередь с тем, что эксперименты проводились именно с селективными синтезаторами (ЧОТ синтезированной речи в которых может не полностью соответствовать смоделированной). Для точной оценки правильности интонационной транскрипции необходимы промежуточные данные синтезаторов, недоступные при тестировании коммерческих систем с использованием демо-версий. Также следует отметить, что, в отличие от оценки лингвистической обработки, нельзя обобщать данные, полученные для одного голоса, на всю систему синтеза. Например, для различных голосов синтезатора VitalVoiceTTS процент фраз с ошибками варьируется от 34 до 41, что может объясняться как различиями в размерах речевых баз для конкретных голосов, так и особенностями чтения «диктора-донора».

4.4 Оценка акустической обработки

В целом, оценка работы акустического модуля совпадает с оценкой общего качества синтеза (см. п. 4.6), однако при оценке селективного синтеза речи есть некоторые дополнительные возможности оценки отдельных модулей, которые приводятся ниже.

4.4.1 Оценка модуля выбора звуковых единиц из речевой базы

Одним из основных способов оценки качества селективного модуля и работы селективного синтезатора в целом является **ресинтез речевых образцов**, то есть синтез того же текста, который был озвучен диктором при записи речевой базы. При этом в идеальном случае звуковые элементы должны выбираться в той же последовательности и синтезированная речь должна быть неотличима от оригинала. Однако на практике достижение такого результата осложняется множеством факторов. При синтезе обычно моделируется максимально

нейтральный, «усреднённый», нормативный тип речи, как с точки зрения интонации и расстановки пауз, так и с точки зрения сегментной транскрипции. Даже если характерные особенности речи диктора специальным образом учитываются, на уровне моделируемых физических характеристик звуковых единиц совпадение может быть неполным. Поиск причин таких несовпадений – лучший способ отладки программы синтеза речи. Однако в общем случае такой тип тестирования недоступен, так как текст, начитываемый диктором при записи базы, неизвестен сторонним пользователям системы.

4.4.2 Оценка модуля модификации

При оценке модуля модификации можно использовать SSML-теги (см. п. 2.3), различные регуляторы громкости, высоты голоса, темпа, тембра, уровня реверберации и т. п., обычно недоступные в демо-версиях. Следует проверять, насколько точно изменяются физические характеристики (например, увеличение темпа в полтора раза), а также то, насколько применение модификации снижает естественность синтезированного речевого сигнала, так как для селективного синтеза речи такое снижение может быть очень резким.

4.5 Интегральная оценка качества синтезированной речи и оценка общего качества синтеза

При интегральной оценке синтеза речи ключевым фактором при составлении теста является сценарий использования программы синтеза речи, что было показано в главе 3 (п. 3.6.4). Поскольку в настоящем исследовании не ставилась задача оценки синтезаторов речи в какой-то определённой ситуации и из-за ограничений, накладываемых использованием демо-версий синтезаторов, при проведении испытаний мы остановимся только на основных показателях качества речи: разборчивости и естественности.

4.5.1 Оценка разборчивости

В ситуации прослушивания синтезированной речи с использованием качественной аппаратуры воспроизведения аудиторы без каких-либо нарушений слуха и восприятия речи (вследствие различных факторов), оценивают словесную разборчивость (при отсутствии грубых ошибок лингвистической и акустической обработки) практически всегда как стопроцентную. Повышенные требования к разборчивости синтезированной речи могут накладывать различные сценарии её использования (шум в самолёте, помехи в телефонном канале и т. п.), поэтому в настоящем исследовании разборчивость синтезированной речи никак специально не тестировалась. При прослушивании синтезированной речи и выполнении тестовых заданий аудиторами не были отмечены какие-либо нарушения разборчивости тестируемой речи.

Оценка разборчивости речи селективных синтезаторов с использованием слоговых таблиц (по ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости») также представляется затруднительной из-за того, что синтезатор должен выдерживать единый нейтральный темп, громкость речи, паузы между слогами и не выделять отдельных звуков. Иначе говоря, для TTS-синтезаторов необходимо создание некоего специального режима чтения слоговых таблиц, в котором будет отключена опция чтения аббревиатур, каждый слог будет ударным и прочитываться как отдельная синтагма без вариативности интонации и т. п., при этом при чтении сочетаний слов и более длинных фраз для селективного синтеза могут подбираться другие звуковые элементы и итоговая разборчивость может напрямую не зависеть от слоговой.

4.5.2 Оценка естественности

Оценка естественности звучания синтезированной речи является ключевой при разработке и сравнении синтезаторов. В качестве примера приведём эксперимент по оценке общей естественности синтезированной речи при решении вспомогательной задачи выбора оптимального объёма речевого корпуса для селективного синтезатора и метода просодической обработки (двух важных факторов, влияющих на качество селективного синтеза), подробно описанный в работе [Solomennik, Chistikov 2013]. Исследование проводилось с системой синтеза VitalVoice TTS ООО «ЦРТ» [Oparin, Talanov 2007].

Приведём краткое описание тестируемых подходов к моделированию интонации.

Первый подход основывался на правилах и состоял из двух шагов. На первом шаге определялся выбор типа интонации фразы (синтагмы) и слова, на которое падает фразовое ударение (ИЦ) в зависимости от пунктуации и наличия во фразе особых слов-интонационных ключей (вопросительных, союзов и др.). Это делалось при помощи правил, составленных вручную. К этому моменту границы синтагм уже были определены [Хомицевич, Соломенник 2010]. В системе интонационной транскрипции имелось шесть типов интонации, надёжно определяемых по тексту: завершённость, незавершённость, общий и специальный вопрос и также два типа восклицаний (сокращённый набор типов из [Вольская, Скрелин 2009]). На втором шаге (после фонетической транскрипции) каждому аллофону озвучиваемого отрезка текста присваивались определенные значения ЧОТ, длительности и энергии [Вольская и др. 2005]. Эти параметры являлись дикторозависимыми и изменялись, прежде всего, в зависимости от выбранного типа интонации и места интонационного центра (ИЦ). Для длинных и коротких фраз использовались различные параметры. Для ЧОТ задавалась характеристика

деклинации (на основе средней ЧОТ) и величина отклонения от неё. Длительность и энергия также задавались в зависимости от места аллофона во фразе и его ударности как отклонения от средних. Значения указанных просодических параметров для конкретного диктора настраивались вручную с использованием статистики их распределения в речевой базе диктора-донора. Такой подход является довольно трудоёмким.

Второй подход (гибридный) объединяет методы СММ и селективного синтеза, он подробно описан в статье [Chistikov, Korolkov 2012]. Лингвистические и просодические признаки для каждого аллофона, используемые для обучения СММ и последующей генерации необходимых физических характеристик, приводятся в таблице 18. Параметры речевого сигнала получают из СММ, векторы наблюдения которых состоят из мел-кепстральных коэффициентов (MFCC – mel-frequency cepstral coefficients), характеристик ЧОТ и длительности аллофонов. Речевой сигнал порождается путём компиляции выбранных звуковых единиц из речевой базы.

Таблица 18. Признаки, используемые для обучения СММ

Аллофонные признаки	
Аллофон перед предыдущим	Аллофон, следующий за следующим
Предыдущий аллофон	Позиция аллофона от начала слога
Текущий аллофон	Позиция аллофона от конца слога
Следующий аллофон	
Слоговые признаки	
Тип предыдущего слога	Позиция слога от конца слова
Тип текущего слога	Позиция слога от начала синтагмы
Тип следующего слога	Позиция слога от конца синтагмы
Количество аллофонов в предыдущем слоге	Количество ударных слогов перед текущим слогом в синтагме
Количество аллофонов в текущем слоге	Количество ударных слогов после текущего слога в синтагме
Количество аллофонов в следующем слоге	Гласный звук в текущем слоге
Позиция слога от начала слова	
Словные признаки	
Часть речи предыдущего слова	Количество слогов в текущем слове
Часть речи текущего слова	Количество слогов в следующем слове
Часть речи следующего слова	Позиция слова в текущей синтагме от начала
Количество слогов в предыдущем слове	Позиция слова в текущей синтагме от конца
Синтагматические признаки	
Количество слогов в текущей синтагме	Тип пунктуационного знака в конце синтагмы
Количество слов в текущей синтагме	

Фонетическая и лингвистическая информация для обучения параметров извлекается из размеченного речевого корпуса [Продан и др. 2009].

В описываемом эксперименте оценка качества производилась в соответствии с рекомендациями ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости».

Оценивался синтезированный голос «Юлия» (VitalVoiceTTS). Для оценки были предложены следующие варианты синтеза:

1. С интонацией, основанной на правилах, на маленьком речевом корпусе объёмом в 20 минут звучащей речи (вся разметка данного корпуса была откорректирована вручную).
2. С интонацией, основанной на правилах, на речевом корпусе, объёмом в два с половиной часа записей (разметка была откорректирована вручную).
3. С интонацией на основе СММ на том же корпусе речи (2,5 часа, ручная коррекция).
4. С интонацией на основе правил на относительно большом (6 часов) корпусе речи, размеченном автоматически (без ручной корректировки меток).

В перцептивном эксперименте принимали участие 17 человек, 8 женщин и 9 мужчин в возрасте от 20 до 55 лет. Среди них 11 были тем или иным образом знакомы с синтезированной речью, в то время как 6 человек не имели опыта или имели очень маленький опыт прослушивания синтезированной речи.

Аудиторам было дано задание оценить естественность четырёх синтезированных и одного естественного варианта чтения (тем же диктором) семи тестовых фраз:

- (1) *Если хочешь быть здоров, советует Татьяна Илье, чисть зубы пастой «Жемчуг»!*
- (2) *Вчера на московском заводе малолитражных автомобилей состоялось собрание молодежи и комсомольцев.*
- (3) *В клумбах сочинской здравницы «Пуща», сообщает нам автоинспектор, обожгли шихту.*
- (4) *Тропический какаду — это крупный попугай? Ты не злословишь?*

(5) *Актеры и актрисы драматического театра часто покупают в этой аптеке антибиотики.*

(6) *Нам с вами сидеть и обсуждать эти слухи некогда!*

(7) *Так ты считаешь, что техникой мы обеспечены на весь сезон?*

Оценки могли изменяться от 0 до 5 с шагом в 0,1, описание соответствия оценок качеству [ГОСТ 1995] приводится в таблице 19:

Таблица 19. Оценки и их количественное значение

Характеристика речи	баллы
Естественное звучание речи, отдельные малозаметные искажения помехового типа. Искажения типа дребезжание, хрип отсутствуют. Высокая узнаваемость	> 4,5
Некоторое нарушение естественности и узнаваемости, слабое присутствие одного вида искажений (картавость, гнусавость, дребезжание, хрип и др.)	3,6 – 4,5
Заметное нарушение естественности и ухудшение узнаваемости, присутствие нескольких видов искажений (картавость, гнусавость, дребезжание, хрип и др.)	2,6 – 3,5
Постоянное присутствие искажений типа картавость, гнусавость, дребезжание, хрип и др. Существенное искажение естественности и ухудшение узнаваемости	1,7 – 2,5
Сильные механические искажения типа картавость, гнусавость, дребезжание, хрип и др.; механический голос. Наблюдается потеря естественности и узнаваемости	< 1,7

Пять вариантов произнесения каждой фразы (см. выше) были даны аудиторам в произвольном порядке с возможностью прослушать при необходимости каждую фразу несколько раз. Испытуемые должны были занести свои количественные оценки в специальные бланки. Образец бланка представлен в Приложении X.

Полученные оценки приведены в таблице 20:

Таблица 20. Результаты оценки естественности звучания различных образцов синтезированной речи

Тип синтеза	Средняя оценка	Стандартное отклонение
20-ти минутный речевой корпус	3,6	0,9
Интонация по правилам (корпус 2,5 часа)	4,1	0,7
Интонация на основе СММ (корпус 2,5 часа)	4,3	0,6
Автоматически размеченная база (корпус 6 часов)	3,7	0,8
Естественная речь	4,9	0,1

Если исключить результаты двух испытуемых, показавших более чем двадцатипроцентное отклонение от средних значений и нормализовать оценки на оценку естественной речи (как рекомендует ГОСТ), для интонации на основе правил и гибридного подхода могут быть получены средние оценки естественности речи 4,4 и 4,5 соответственно. Все типы синтеза, таким образом, попадают в один и тот же I класс качества (оценки от 3,6 до 4,5 по ГОСТу, см. таблицу 21).

Таблица 21. Классы качества речи по ГОСТ Р 50840-95

Класс качества	Характеристика класса качества речи	Норма качества, баллы
Высший	Естественное звучание речи, отдельные малозаметные искажения помехового типа. Искажения типа дребезжание, хрип отсутствуют. Высокая узнаваемость	> 4,5
I	Некоторое нарушение естественности и узнаваемости, слабое присутствие одного вида искажений (картавость, гнусавость, дребезжание, хрип и др.)	3,6 – 4,5
II	Заметное нарушение естественности и ухудшение узнаваемости, присутствие нескольких видов искажений (картавость, гнусавость, дребезжание, хрип и др.)	2,6 – 3,5

III	Постоянное присутствие искажений типа картавость, гнусавость, дребезжание, хрип и др. Существенное искажение естественности и ухудшение узнаваемости	1,7 – 2,5
IV	Сильные механические искажения типа картавость, гнусавость, дребезжание, хрип и др.; механический голос. Наблюдается потеря естественности и узнаваемости	< 1,7

Следует отметить, что прослеживается явная связь между оценками и «знакомством» испытуемого с синтезированной речью. Это можно увидеть на рис. 21 с диаграммой, где буквой «а» обозначены «наивные» слушатели, а «b» – слушатели, знакомые с синтезом (оценки усреднены для всех типов синтеза):

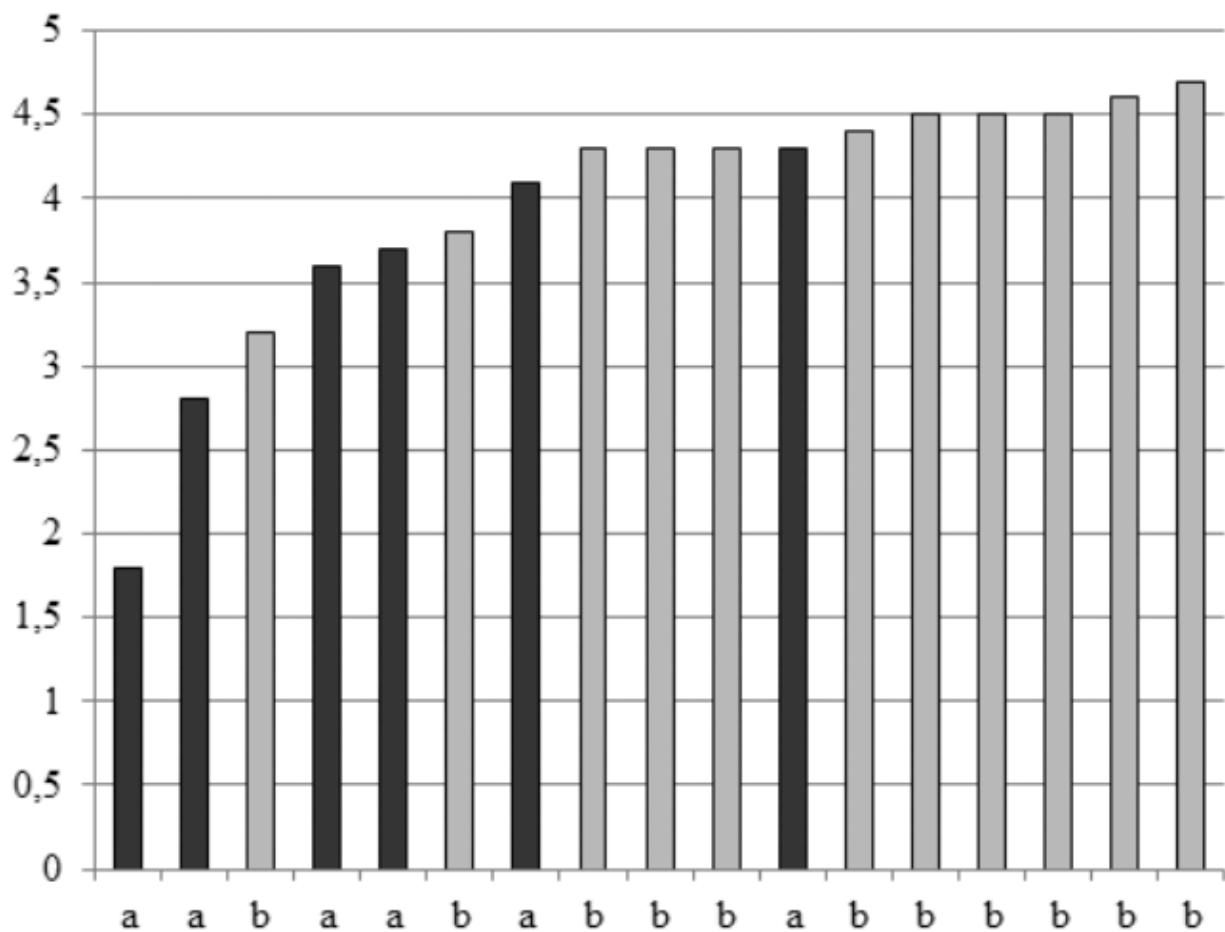


Рис. 21. Средние оценки различных типов синтезированной речи по отношению к «знакомству» испытуемых с синтезом речи

Можно видеть, что испытуемые, не привыкшие к синтезированной речи, склонны давать более низкие оценки естественности синтезированной речи, чем остальные. Чтобы проверить, является ли эта закономерность статистически значимой, был использован U-критерий Манна-Уитни (равен 7,5). Критическое значение U-критерия Манна-Уитни при заданной численности сравниваемых групп составляет 13, следовательно, различия в оценках в сравниваемых группах статистически значимы ($p < 0,05$).

Полученные результаты показывают, что метод оценки качества речи по ГОСТу может использоваться для грубых диагностических тестов синтезатора на

естественность (в режиме сравнений). При этом сам эксперимент показал, что использование гибридного подхода, комбинирующего СММ и селективный синтез речи, позволяет получать более естественно звучащую синтезированную речь. По результатам эксперимента генерация просодических параметров на основе СММ была включена в систему подготовки нового голоса для протестированного синтезатора [Продан и др. 2010]. Другой важный результат состоит в том, что даже небольшой, но фонетически сбалансированный речевой корпус [Solomennik, Chistikov 2012; Чистиков и др. 2014] может обеспечить приемлемое качество синтезированной речи.

4.7 ВЫВОДЫ К ГЛАВЕ 4

Произведенное в настоящей работе тестирование (см. итоговую таблицу 22) показало, что на основании диагностических тестов отдельных модулей синтезатора нельзя определить лучшую систему синтеза речи, так как для различных задач использования синтезаторов, критичными могут оказаться различные показатели, даже те, которые не обсуждались в данной главе (например, степень устойчивости синтеза речи к шумовым помехам или качество речи в телефонном канале). Даже система Mary TTS, показавшая худший результат практически по всем тестам, обладает одним несомненным преимуществом – открытым доступом к исходному коду программы и, следовательно, возможностью улучшения и настройки синтезатора под определённую задачу. В то же время проведённые тесты указывают на слабые стороны, которые требуют той или иной доработки для различных систем синтеза и могут существенно улучшить их качество.

Таблица 22. Сводная таблица результатов проведённых тестов по успешности выполнения системами синтеза речи различных задач

[illegible]

ЗАКЛЮЧЕНИЕ

В рамках данного диссертационного исследования ставились следующие задачи:

1. Описать существующие проблемы и методы оценки качества синтеза речи, обобщить результаты предыдущих исследований, посвящённых методам оценки качества синтезированной речи.
2. Обосновать необходимость специального подхода к оценке технологии селективного синтеза речи на основании его характерных особенностей.
3. Предложить методы оценки синтеза речи, позволяющие объективно оценивать и сравнивать современные русскоязычные селективные синтезаторы.
4. Произвести комплексную оценку качества работы нескольких современных русскоязычных синтезаторов селективного типа.

В **первой** главе диссертации изложена теоретическая база исследования: история создания и типы синтезаторов речи, качество порождаемой ими речи.

Во **второй** главе приводится подробное описание технологии селективного синтеза речи и даётся аналитический обзор современных русскоязычных синтезаторов.

В **третьей, центральной,** главе описываются существующие методы и способы оценки качества синтезированной речи, преимущества и недостатки синтезаторов разного типа, предлагается структура системы для оценки качества селективного синтеза речи, даётся обоснование необходимости разработки соответствующих оценочных методов.

В **четвёртой, экспериментальной,** главе описаны эксперименты и тесты, проведённые в диссертационном исследовании по оценке качества нескольких современных систем селективного синтеза речи на русском языке, даётся анализ полученных результатов.

В диссертации получены следующие основные **результаты:**

1. Произведено описание и анализ существующих методов оценки синтезированной речи.
2. Обоснована необходимость специального подхода к оценке селективного синтеза речи.
3. Предложена система методов оценки, адаптированных для селективного синтеза речи.
4. Подготовлены тестовые тексты и опросники для проведения комплексной оценки различных модулей синтезаторов речи.
5. Произведено комплексное тестирование русскоязычных селективных синтезаторов и получены оценки качества речи, синтезируемой с их использованием.

Дальнейшие исследования по данной тематике наиболее актуальны в следующих направлениях: оценка качества статистического параметрического синтеза речи, инструментальная автоматизированная оценка качества синтезаторов речи, оценка выразительности и эмоциональности синтезированной речи, оценка точности воспроизведения особенностей речи конкретного диктора-донора. Полезным направлением могло бы стать также проведение независимого конкурса синтезаторов речи на материале русского языка, что, к сожалению, трудно осуществить, так как большинство разработчиков современных синтезаторов речи высокого качества являются зарубежными компаниями, для которых разработка и

совершенствование русскоязычного синтеза не всегда является приоритетной задачей.

СПИСОК СОКРАЩЕНИЙ

ИК интонационная конструкция

ИЦ интонационный центр

КЛП коэффициенты линейного предсказания

СММ скрытая Марковская модель

ЧОТ частота основного тона

HMM Hidden Markov Model

MFCC Mel-Frequency Cepstral Coefficients

PSOLA Pitch Synchronous Overlap and Add

SSML Speech Synthesis Markup Language

TD-PSOLA Time Domain Pitch Synchronous Overlap-Add

TTS Text-to-Speech

СПИСОК ЛИТЕРАТУРЫ

1. Бабкин А. В., Захаров Л. М. Оценка качества системы синтеза речи, разработанного в МГУ // Труды Международного семинара «Диалог'99». — Таруса, 1999. — С. 12–25.
2. Богданова Н. В. Живые фонетические процессы русской речи: пособие по спецкурсу. — СПб., 2001. — С. 286.
3. Брызгунова Е. А. Интонация // Русская грамматика. М., 1980. — Т. 1. — С. 96–122.
4. Вольская Н. Б., Скредин П. А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Труды третьего междисциплинарного семинара «Анализ разговорной русской речи (АРЗ-2009)». — СПб., 2009. — С. 28–40.
5. Вольская Н. Б., Скредин П. А., Таланов А. О. Автоматическое моделирование просодического оформления фразы // Интегральное моделирование звуковой формы естественных языков. — СПб., 2005. — С. 64–74.
6. Гецавіч Ю. С. Алгарытмы лінгвістычнай апрацоўкі тэкстаў для сінтэзу маўлення на беларускай і рускай мовах. Дысертацыя на атрыманне навуковай ступені кандыдата тэхнічных навук. — Мінск. 2012. — 191 с.
7. ГОСТ ISO 9000–2011 Системы менеджмента качества. Основные положения и словарь. — М.: Стандартинформ, 2012. — 32 с.
8. ГОСТ Р 50840–95 Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. — М.: Госстандарт России, 1995 — 234 с.

9. Кейтер Дж. Компьютеры — синтезаторы речи. — М.: Мир, 1985. — 237 с.
10. Кодзасов С. В., Кривнова О. Ф. Общая фонетика. — М., 2001. — 592 с.
11. Корсакова Н. С., Засыпкина К. А. Методика тестирования систем автоматического синтеза и распознавания речи в целях определения коммерческой целесообразности их использования // Речевые технологии. — 2012. — Вып. 3. — С. 92–98.
12. Крейчи С. А., Кривнова О. Ф., Егоров А. М. Фотооптический синтезатор АНС и его роль в развитии методов синтеза речи // журнал «Речевые технологии». — 2013. — №1. — С. 58-71.
13. Кривнова О. Ф. Автоматический синтез русской речи по произвольному тексту (вторая версия с женским голосом) // Труды Международного семинара "Диалог'98". — Таруса, 1998. — С. 498–511.
14. Кривнова О. Ф. Генерация тонального контура фразы в системах автоматического синтеза речи // Труды Международного семинара по компьютерной лингвистике и ее приложениям "Диалог'2000". — М., 2000. — С. 211–220.
15. Кривнова О. Ф., Захаров Л. М., Строкин Г. С. Мультифункциональный автоматический транскриптор русских текстов // Труды Международного конгресса исследователей русского языка. — М., 2001(а). — С. 408–409.
16. Кривнова О. Ф., Захаров Л. М., Строкин Г. С. Подбор текстового материала и статистический инструментарий для создания речевых корпусов // Сборник трудов XI сессии Российского акустического общества. Том 3. Акустика речи. Медицинская и биологическая акустика. — М.: ГЕОС, 2001(б). — С. 87–92.

17. Кривнова О. Ф., Чардин И. С. Паузирование при автоматическом синтезе речи // Теория и практика речевых исследований (АРСО-99). Материалы конференции. — 1999. — С. 87–103.
18. Кузнецов В. Б., Отт А. В. Автоматический синтез речи. Алгоритм преобразования «буква-звук» и управление длительностью речевых сегментов. — Таллинн, 1989. — 135 с.
19. Лобанов Б. М. и др. Синтезатор персонализированной речи по тексту “ЛобаноФон-2000” // Тр. Международной конференции, посвящённой 100-летию российской экспериментальной фонетики. — СПб, 2001 — С. 101–104.
20. Лобанов Б. М. Микроволновой синтез речи по тексту // Анализ и синтез речи: сб. науч. трудов / научн. ред. Б. М. Лобанов. — Мн.: Институт технической кибернетики АН БССР, 1991. — С. 57–73.
21. Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и клонирование речи. — Минск, «Белорусская Наука», 2008. — 316 с.
22. Лобанов Б. М. Проблема разрешения «Ё»-омографов при синтезе речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». — Вып. 8 (15). — М.: РГГУ, 2009. — С. 291–298.
23. Национальный корпус русского языка [Электронный ресурс]. — Режим доступа: <http://www.ruscorpora.ru/>.
24. Обжелян Н. К., Трунин-Донской В. И. Машины, которые говорят и слушают. — Кишинев, 1987. — 175 с.

25. Покровский Н. Б. Расчет и измерение разборчивости речи. — М.: Связьиздат, 1976. — 391 с.
26. Потапова Р. К. Речевое управление роботом. — М.: Радио и связь, 1989. — 248 с.
27. Потапова Р. К. Речь: коммуникация, информация, кибернетика. — М., 2015. — 600 с.
28. Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О. Особенности использования многоуровневой разметки звукового корпуса // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». — Вып. 8 (15). — М.: РГГУ, 2009. — С. 415–419.
29. Продан А. И., Таланов А. О., Чистиков П. Г. Система подготовки нового голоса для системы синтеза «VitalVoice» // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 9 (16). — М.: Изд-во РГГУ, 2010. — С. 394–399.
30. Речевые тесты и их применение / Отв. ред. Л. В. Златоустова. — М.: МГУ, 1986. — 99 с.
31. Русанова О. А. Исследование и разработка методов анализа и оценки качества синтезированной устной речи: дис. ... канд. техн. наук: 05.13.17 / Русанова Ольга Александровна. — Красноярск, 2004. — 107 с.
32. Санникова Е. Г. Восприятие и понимание естественной и синтезированной речи: автореф. дис. ... канд. псих. наук: 19.00.01 / Санникова Елена Георгиевна. — Екатеринбург, 2008. — 23 с.

33. Смирнова Н. С., Хитров М. В. Фонетически представительный текст для фундаментальных и прикладных исследований русской речи // Изв. вузов. Приборостроение. — 2013. — Вып. 2. — С. 5–10.
34. Соломенник А. И. Структура системы оценки качества синтезированной русской речи // Структурная и прикладная лингвистика. — Вып. 10. — СПб, 2013(a). — С. 251–266.
35. Соломенник А. И. Технология синтеза речи: история и методология исследований // Вестник Московского университета. Сер. 9. Филология. — 2013(b). — № 6. — С. 149–162.
36. Соломенник А. И. Зависимость естественности звучания синтезированной речи от наличия ошибок различных типов // Актуальные проблемы филологической науки: взгляд нового поколения. Доклады участников XX–XXI Международных конференций студентов, аспирантов и молодых ученых «Ломоносов». Секция «Филология». Вып. 6. — Изд. Московского университета, 2015. — С. 475–480.
37. Соломенник А. И. Особенности оценки качества селективного синтеза речи. Актуальные вопросы теоретической и прикладной фонетики. Сборник статей к юбилею О. Ф. Кривновой / Под ред. А. В. Архипова, И. М. Кобозевой, Кс. П. Семёновой. — М.: ООО «Буки-Веди», 2013(c). — С. 336–341.
38. Соломенник А. И. Технология синтеза речи в историко-методологическом аспекте. Речевые технологии. — №1, — 2013(d). — С. 42–57.
39. Соломенник А. И., Таланов А. О., Соломенник М. В., Хомицевич О. Г., Чистиков П. Г. Оценка качества синтезированной речи: проблемы и

- решения // Изв. вузов. Приборостроение. Тематический выпуск "Речевые информационные системы". — 2013(a). — №2. — С. 38–42.
40. Соломенник А. И., Чистиков П. Г., Рыбин С. В., Томашенко Н. А. Автоматизация процедуры подготовки нового голоса для системы синтеза русской речи // Изв. вузов. Приборостроение. — 2013(b). — №2. — С. 29–32.
 41. Сорокин В. Н. Синтез речи. — М., Наука, 1992. — 392 с.
 42. Фланаган Дж. Анализ, синтез и восприятие речи. — М.: Связь, 1968. — 396 с.
 43. Фролова И. Г. Естественность и разборчивость как параметры оценки качества синтезированной речи // Экспериментальная фонетика. — М.: Изд-во МГУ, 1989. — С. 95–102.
 44. Хомицевич О. Г., Рыбин С. В., Аничкин И. М. Использование лингвистического анализа для нормализации текста и снятия омонимии в системе синтеза русской речи // Изв. вузов. Приборостроение. Тематический выпуск "Речевые информационные системы". — 2013. — №2. — С. 42–46.
 45. Хомицевич О. Г., Соломенник М. В. Автоматическая расстановка пауз в системе синтеза русской речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 9 (16). — М.: Изд-во РГГУ, 2010. — С. 531–537.
 46. Черепанова О. Д. Нормализация текста в системе русскоязычного синтеза «Текст-Речь»: озвучивание английских слов и словосочетаний на основе практической транскрипции. Дипломная работа. — Москва, 2015. — 80 с.

47. Чистиков П. Г., Корольков Е. А., Таланов А. О., Соломенник А. И. Гибридная технология синтеза русской речи на основе скрытых Марковских моделей и алгоритма Unit Selection // Изв. вузов. Приборостроение. — 2013. — №2. — С. 33–38.
48. Чистиков П. Г., Таланов А. О., Захаров Д. С., Соломенник А. И. Технология синтеза естественной речи с использованием базы данных небольшого объема // Научно-технический вестник информационных технологий, механики и оптики. — №4 (91) — 2014. — С. 83–97.
49. Широкова А. М. Буквенно-звуковое преобразование в системах автоматической обработки речи // Структурная и прикладная лингвистика. — СПб., 2014. — Вып. 10. — С. 279–302.
50. Bachan J., Kuczmarski T., Francuzik P. Evaluation of synthetic speech using automatic speech recognition // Proc. XIV International PhD Workshop OWD. — 2012. — P. 500–505.
51. Benoit C., Grice M., Hazan V. The SUS test: a method for the assessment of text to speech synthesis intelligibility using Semantically Unpredictable Sentences // Speech Communication. — 1996. — Vol. 18. — P. 381–392.
52. Black A. W., Taylor P. CHATR: A Generic Speech Synthesis System // COLING-94. — Japan, 1994. — P. 983–986.
53. Black A. W. Perfect Synthesis for all of the people all of the time // Keynote, IEEE TTS Workshop. — Santa Monica, CA, 2002. — P. 167–170.
54. Black A. W., Zen H., Tokuda K. Statistical parametric speech synthesis // Proc. ICASSP 2007. — 2007. — P. 1229–1232.

55. Blizzard Challenge 2015 [Электронный ресурс]. — 2015. — Режим доступа: http://www.synsig.org/index.php/Blizzard_Challenge_2015.
56. Campbell N. Evaluation of Speech Synthesis // Dybkjær L., Hamsen H. and Minker W. (Eds.) Evaluation of text and speech systems. — Springer: The Netherlands, 2007. — P. 29–64.
57. Chevelu J., Barbot N., Boeffard O., Delhay A. Comparing set-covering strategies for optimal corpus design // Proceedings of the 6th International Language Resources and Evaluation. — 2008. — P. 969–974.
58. Chistikov P., Korolkov E. Data-driven speech parameter generation for Russian text-to-speech system // Proceedings of the Dialogue-2012 International Conference. — Bekasovo, 2012. — No 11 (18). — P. 103–111.
59. Chistikov P. G., Korolkov E. A., Talanov A. O. Combining HMM and Unit Selection technologies to increase naturalness of synthesized speech // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — М.: Изд-во РГГУ, 2013. — Вып. 12 (19). — Т. 2. — С. 2–10.
60. Clark R. A. G., Richmond K., King S. Multisyn: Open-domain unit selection for the Festival speech synthesis system // Speech Communication. — 2007. — Vol. 49. — Issue 4. — P. 317–330.
61. Dutoit T. An Introduction to Text-to-Speech Synthesis. — Dordrecht–Boston–London, 1997. — 286 p.
62. EVALDA – EVASY. Evaluation des Synthétiseurs de parole en français [Электронный ресурс]. — 2006. — Режим доступа: http://www.technolanguage.net/article.php3?id_article=202.

63. Falk T. H., Möller S. Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems // IEEE Signal Processing Letters. — 2008. — Vol. 15. — P. 781–784.
64. Fujisaki H. Dynamic characteristics of voice fundamental frequency in speech and singing // Production of Speech. — N. Y. 1983. — P. 39–55.
65. Hao W., Soong, F. K., Meng H. A spectral space warping approach to cross-lingual voice transformation in HMM-based TTS // ICASSP. — 2015. — P. 4874–4878.
66. House A., Williams C., Hecker M., Kryter K. Articulation-testing methods: Consonantal differentiation with a closed-response set // The Journal of the Acoustical Society of America 37. — 1965. — P. 158–166.
67. Hunt A., Black A. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database // Proceedings of ICASSP 96. — 1996. — P. 373–376.
68. Jekosch U. The Cluster-Identification Test // Proceedings of ICSLP 92 (1). — 1992. — P. 205–208.
69. Khomitsevich O. G., Chistikov P. G. Using statistical methods for prosodic boundary detection and break duration prediction in a Russian TTS system // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — М.: Изд-во РГГУ, 2013. — Вып. 12 (19). — Т. 2, с. 11–19.
70. Klatt D. Review of Text-to-Speech Conversion for English // JASA. — 1987. — Vol. 82 (3). — P. 737–793.

71. Klatt D. H. Software for a cascade/parallel formant synthesizer // JASA. — 1980. — Vol. 67. — P. 971–995.
72. Latorre J. A. Study on Speaker-Adaptable Multilingual Synthesis. — Tokyo Institute of Technology, 2006. — 121 p.
73. Lemmetty, S. Review of Speech Synthesis Technology. Master's Thesis, Helsinki University of Technology. — 1999. — 104 p.
74. Lobanov B. The phonemophon text-to-speech system // International Congress of Phonetic Sciences: proc. Of the 11-th section ICPhS'87, Tallin, USSR, 6-10 August 1987. — Tallin, 1987. — Vol. 1. — P. 120–124.
75. Mattingly I. G. Speech Synthesis for Phonetic and Phonological Models // Current Trends in Linguistics, edited by T. S. Sebeok. — Mouton, the Netherlands, 1974. — Vol. 12. — P. 2451–2487.
76. Méndez F. et al. The Albayzín 2010 Text-to-Speech Evaluation // Proc. Fala-2010. — 2010. — P. 317–340.
77. Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, ITU-T Rec. P.85. — Int. Telecom. Union, 1994. — 13 p.
78. Moore R. K., Nicolao M. Reactive Speech Synthesis: Actively Managing Phonetic Contrast along an H&H Continuum // ICPhS 2011. — Hong Kong, China, 2011. — P. 1422–1425.
79. Moulines E., Verhelst W. Time-domain and frequency-domain techniques for prosodic modification of speech in Speech Coding and Synthesis // IEEE. — Netherland, 1995. — P. 519–555.

80. Norrenbrock C. R. Hinterleitner F., Heute U., Moller S. Towards Perceptual Quality Modeling of Synthesized Audiobooks – Blizzard Challenge 2012 [Электронный ресурс]. — 2012 (b). — Режим доступа: http://festvox.org/blizzard/bc2012/Norrenbrock_etal_Blizzard_workshop_2012_final.pdf
81. Norrenbrock C. R., Hinterleitner F., Heute U., Moller S. Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals // Signal Processing Letters, IEEE. — 2012(a). — P. 255–258.
82. Oparin I., Talanov A. (2007), Outline of a New Hybrid Russian TTS System // Proceedings of the 12th International conference on Speech and Computer, SPECOM 2007. — Moscow, 2007. — P. 603–608.
83. Sagisaka Y. et al. ATR – n-Talk speech synthesis system // Proceedings of ICSLP92. — Banff, Canada, 1992. — P. 483–486.
84. Silverman K. et al. TOBI: a standard for labelling English prosody // Spoken Language Processing: proceedings of 2-nd International conference ICSLP'92. Alberta, Canada, 13-16 October 1992. — Alberta, 1992 — P. 867–870.
85. Solomennik A. I., Cherentsova A. E. A Method for Auditory Evaluation of Synthesized Speech Intonation // Miloš Železný et al. (Eds.): SPECOM 2013, Lecture Notes in Artificial Intelligence 8113. — Springer, 2013. — P. 9–16.
86. Solomennik A. I., Chistikov P. G. Evaluation of naturalness of synthesized speech with different prosodic models // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — М.: Изд-во РГГУ, 2013. — Вып. 12 (19). — Т. 2. — С. 31–38.

87. Solomennik A., Chistikov P. Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system // Proceedings of the Dialogue-2012 International Conference. — Bekasovo, 2012. — No 11 (18). — P. 607–615.
88. Sproat R., Black A. W., Chen S., Kumar S., Ostendorf M., Richards C. Normalization of non-standard words // Computer Speech and Language. — 2001. — Vol. 15. — P. 287–333.
89. Stilianou Y. Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis // IEEE transactions on speech and audio processing. — 2001. — Vol. 9. — No. 1. — P. 21–29.
90. Stylianou Y., Syrdal A. Perceptual and objective detection of discontinuities in concatenative speech synthesis // Proc. Int. Conf. Acoustics, Speech, and Signal Processing. — 2001. — P. 837–840.
91. Sydeserff H. A., Caley R. J., Isard S. D. et al. Evaluation of speech synthesis techniques in a comprehension task // Speech Communication 11, 2–3. — 1992. — P. 89–194.
92. Syrdal A. K., Conkie A. D. Perceptually-based data-driven join costs: Comparing join types // In Proceedings of Eurospeech, Interspeech. — 2005. — P. 2813–2816.
93. 't Hart J., Collier R., Cohen A. A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody. — Cambridge, 1991. — 212 p.
94. Tatham M, Morton K. Developments in Speech Synthesis // John Wiley & Sons Ltd. — 2005. — 342 p.

95. Taylor P. Analysis and synthesis of intonation using the tilt model // J. Acoust. Soc. America. — 2000. V. 107. — № 3, p. 1697–1714.
96. Taylor P. Text-to-Speech Synthesis. Cambridge University Press, 2009. 474 p.
97. Tokuda K., Masuko T., Yamada T. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features // Proceedings of Eurospeech-1995. — 1995. — P. 757–760.
98. Tokuda K., Nankaku Y., Toda T., Zen H., Yamagishi J., Oura K. Speech Synthesis Based on Hidden Markov Models // Proceedings of the IEEE. — 2013. — Vol. 101. — No. 5. — P. 1234–1252.
99. Van Heuven V. J., Van Bezooijen R. Quality Evaluation of Synthesized Speech. // Speech Coding and Synthesis. — New York, 1995. — P. 723–725.
100. Van Santen J. Evaluation // Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. — Kluwer, Dordrecht, 1998. — P. 229–244.
101. Van Santen J. P. H., Buchsbaum A. L. Methods for optimal text selection. Proc. of Eurospeech. — Rhodes, Greece, 1997. — P. 553–556.
102. Vepa J. Join Cost for Unit Selection Speech Synthesis. — University of Edinburgh, 2004. — 219 p.
103. Vepa J., King S. Subjective evaluation of join cost functions used in unit selection speech synthesis // In Proceedings of the International Conference on Speech and Language Processing 2004. — Jeju, Korea, 2004. — P.1181–1184.
104. Vepa J., King S., Taylor P. Objective distance measures for spectral discontinuities in concatenative speech synthesis // Proc. Int. Conf. Spoken Language Processing. — 2002. — P. 2605–2608.

СПИСОК РИСУНКОВ

Рис. 1. Система резонаторов Кратценштейна	12
Рис. 2. Говорящая машина Кемпелена, построенная Уитстоном	13
Рис. 3. Механический синтезатор Риша	14
Рис. 4. Электрическая модель речевого тракта Стюарта	15
Рис. 5. Схема синтезатора «водер»	17
Рис. 6. Синтезатор Pattern Playback	19
Рис. 7. Аналог речевого тракта с линией передачи, управляемый непрерывно	22
Рис. 8. Каскадный и параллельный синтезаторы. В параллельном синтезаторе амплитуда каждого формантного резонатора должна контролироваться отдельно. В каскадном – выходной сигнал каждого резонатора является входным сигналом следующего	23
Рис. 9. Формантный синтезатор Холмса, состоящий из резонаторов для четырёх формант и носовой форманты, каждый из которых возбуждался вариативной смесью сигналов шумового и голосового источников	25
Рис. 10. Внешний вид синтезатора «Фонемофон-1»	26
Рис. 11. Структурная схема КЛП-синтезатора.....	27
Рис. 12. Прототипическая схема системы синтеза речи	29
Рис. 13. Общая схема системы синтеза речи, основанной на СММ.....	35
Рис. 14. Выбор целевой звуковой последовательности при селективном синтезе речи.....	39

Рис. 15. Общая структура селективного синтезатора речи.....	45
Рис. 16. Алгоритм создания речевой базы для селективного синтезатора	58
Рис. 17. Иллюстрация применения алгоритма TD-PSOLA.....	60
Рис. 18. Зависимость слоговой (3), словной (2) и фразовой (1) разборчивости от фонемной	66
Рис. 19. Вариативность оценок auditors (процент предложений, в которых отмечены ошибки указанного типа) для разных синтезированных голосов. Цветами показаны проценты ошибок, указанные каждым испытуемым	92
Рис. 20. Синхронизированные осциллограммы звукозаписей синтезированного предложения «Прямо передо мной внизу – // пестрый узор из крыш городских домов, // а вдалеке, // на краю горизонта, // тянется серебряная цепь снеговых вершин...» для голосов Tatyana (вверху) и Анна (внизу) с отмеченными паузами	112
Рис. 21. Средние оценки различных типов синтезированной речи по отношению к «знакомству» испытуемых с синтезом речи	130

ПЕРЕЧЕНЬ ТАБЛИЦ

Таблица 1. Основные селективные синтезаторы для русского языка	42
Таблица 2. Общая классификация нестандартных текстовых элементов в русских текстах.....	47
Таблица 3. Количество ошибок разных типов, выделенных испытуемыми, и средняя оценка естественности речи (в скобках указано стандартное отклонение). Приводятся значения, усредненные по всем предложениям	89
Таблица 4. Максимально возможные оценки естественности речи при наличии ошибки определённого типа	91
Таблица 5. Данные по ошибкам, отмеченным большинством испытуемых для голоса «Tatyana»	93
Таблица 6. Данные по ошибкам, отмеченным большинством испытуемых для голоса «Анна»	93
Таблица 7. Процент ошибочного чтения графических сокращений различными синтезаторами	96
Таблица 8. Процент ошибочного чтения аббревиатур различными синтезаторами	98
Таблица 9. Процент ошибочного чтения цифровых обозначений различными синтезаторами	100
Таблица 10. Процент ошибочного чтения символьных обозначений различными синтезаторами	101
Таблица 11. Проценты ошибочного и неточного чтения англоязычных вставок различными синтезаторами	103

Таблица 12. Ошибки различных синтезаторов в ударениях при чтении фонетически представительного текста (приведены неправильные ударения, в том числе в омографах)	105
Таблица 13. Процент ошибочного чтения слов-омографов различными синтезаторами	107
Таблица 14. Количество ошибок в сегментной транскрипции у различных синтезаторов	110
Таблица 15. Показатели общей правильности и адекватности интонации для синтезированных голосов (среднеквадратичное отклонение даётся в скобках). Минимальные значения отмечены жирным шрифтом	117
Таблица 16. Показатели общей правильности и адекватности интонации для синтезированных голосов, нормализованные относительно естественной речи. Минимальные значения отмечены жирным шрифтом	117
Таблица 17. Процент интонационных ошибок по данным двух экспериментов .	119
Таблица 18. Признаки, используемые для обучения СММ.....	125
Таблица 19. Оценки и их количественное значение	127
Таблица 20. Результаты оценки естественности звучания различных образцов синтезированной речи	128
Таблица 21. Классы качества речи по ГОСТ Р 50840-95	128
Таблица 22. Сводная таблица результатов проведённых тестов по успешности выполнения системами синтеза речи различных задач	132

ПРИЛОЖЕНИЯ

Приложение А. Фонетически представительный текст

Дом, в котором я живу, расположен на окраине маленького городка, у самой подошвы горы. Здесь мягкий климат и редко идут дожди. Ночью небосвод бывает так густо усеян звездами, что кажется, будто все миллиарды их из нашей галактики разбросаны вверху над моей головой. Летним утром, как только я открываю окно, моя большая комната наполняется запахом цветов. Ветки черешен смотрят мне в окна, и легкий теплый ветер усыпает мой письменный стол белыми лепестками.

Я слушаю щебет птиц. Вот с искрометным задором пропел зяблик. Где-то дятел устраивает дупло. А это черные дрозды — поют не хуже соловьев. Прямо передо мной внизу — пестрый узор из крыш городских домов, а вдалеке, на краю горизонта, тянется серебряная цепь снеговых вершин... Весело жить в такой земле! Отрадное чувство разливается в жилах: вокруг величественные горы, воздух чистый и свежий, солнце яркое, небо синее — чего еще желать?.. Бьют настенные часы над камином: пять, шесть, семь, восемь, девять... Нужно торопиться в бюро. Минуты две-три ищу в шкафу электрическую схему, привезенную французским коллегой. Наконец заглядываю в портфель, нахожу ее внутри и вкладываю в книгу. После этого плотно закрываю жалюзи, однако сквозь щелки все равно пробивается солнечный свет. Выхожу на крыльцо и запираю ключом дверь.

Спустившись ниже к центру города, иду бульваром. Часть дороги проходит по пешеходному мосту через реку. Гибкие стальные тросы держат невысокий мост. Они привязаны к специальным тяжелым якорям, врытым в землю. Останавливаюсь в начале мостика у ограды, чтобы полюбоваться рельефными склонами горных хребтов, всматриваюсь в речную рябь. Под мостом с шумом плещутся мелкие рыбешки, возмущая водную гладь. Откуда-то доносится музыка: ноктюрн Шопена — позывные местной радиостанции.

Вдруг позади себя я слышу: „Сережа, неужели это ты? Вот так встреча!“. Я узнаю этот низкий голос. Оборачиваюсь — так и есть: Андрей Сафонов! Очень радостно видеть его вновь. С Андреем мы знакомы с конца восьмидесятых — служили вместе в армии. Его, энергичного и общительного, всегда на помощь готового прийти, любили все. Меня покорили его честность и недюжинная сила. Мы были дружны, но потом, мало-помалу, связь наша оборвалась.

— Здравствуй, Андрюша! Как ты тут оказался?

— Командировка в архив: предлагают снять сюжет про судоверфь. Вчера приехали — сегодня уезжаем. Вернее, улетаем — к четверем в аэропорт.

— Так скоро? И куда?

— Следующим пунктом Уфа. Прямой рейс. А ты, значит, теперь здесь живешь? Давно?

— Два года будет в феврале. Обменял свою городскую квартиру на бревенчатую избу.

— Серьезно? Не жалеешь?

В глазах моего приятеля мелькнул веселый огонек.

— Нисколько. Отдыхаю от километровых пробок, сутолоки и пыли.

— По-прежнему плывешь против течения? Счастливый ты человек, Сергей.

— Ладно, расскажи лучше о себе. Мы тысячу лет не виделись. Как жизнь? Как семья?

— Все у нас хорошо, все здоровы. Мы с женой работаем, дочь гимназию заканчивает.— Ну а Федор как? Учится?

— Он в этом году поступил в медицинский.

— Какой молодец! Поздравляю вас!

— Спасибо.

— Кстати, ты позавтракал?

— Немного кофе выпил в гостинице. С удовольствием бы съел что-нибудь.

— Недалеко отсюда есть кафе. Мы привыкли там есть. Пойдем, провожу тебя. Вполне приличный сервис, разнообразное меню. Одно из их „фирменных“ блюд — рыба по-бенгальски. Рекомендую: вкус необыкновенный — для настоящих гурманов.

— Звучит слишком изысканно. Попроше ничего нет?

— Как насчет яичницы с грибами?

— В самый раз. А ты торопишься? Может быть, составишь мне компанию?

За разговорами и воспоминаниями незаметно пролетел завтрак. Приближалось время сказать „до свидания“. Мы расстались в твердом намерении больше не терять друг друга из виду.

Приложение Б. Инструкция и пример анкеты в эксперименте по влиянию различных типов ошибок на качество синтезированной речи

ГОЛОС 1

Оцените, пожалуйста, естественность синтезированной речи по пятибалльной шкале (1 – плохо, 5 – отлично). При оценке 1 – 4 отметьте, пожалуйста, тип ошибки (или несколько типов) описанием (например, «уд», «инт»...) или номером. Место ошибки можно подчеркнуть в тексте.

1. Неверное ударение
2. Неверное произнесение: замена/выпадение/добавление лишнего звука
3. Неправильные паузы (отсутствие/лишние, слишком короткие/длинные)
4. Плохой темп/ритм
5. Неровная/неверная интонация
6. Прерывистость, скачки, «бульканье», стук и т. п.
7. Общее качество голоса
8. Иное

Предложение	Оценка	Тип и место ошибки
Дом, в котором я живу, расположен на окраине маленького городка, у самой подошвы горы.		
Здесь мягкий климат и редко идут дожди.		
Ночью небосвод бывает так густо усеян звездами, что кажется, будто все миллиарды их из нашей галактики разбросаны над моей головой.		
Летним утром, как только я открываю окно, моя большая комната наполняется запахом цветов.		
Ветки черешен смотрят мне в окна, и легкий теплый ветер усыпает мой письменный стол белыми лепестками.		
Я слушаю щебет птиц.		

Приложение В. Список предложений для тестирования графических сокращений

1. Клепиковский район, г. Спас-Клепики, ул. Московская, д. 23. [Приглашение // «Народное творчество», 2004]
2. В соответствии со **ст. 7** ФЗ № 115-ФЗ от 07.08.2001 "О противодействии легализации (отмыванию) доходов» [Разъяснение (2004)]
3. Кроме того, по слухам, М. Ходорковский обещал выделить **10 млн. долл.** на выборные нужды союза СПС и "Яблока". [Александр Колесниченко. СПС - «Яблоко»: Когда пожертвуют Чубайсом? (2003) // «Аргументы и факты», 2003.01.29]
4. Вот он описывает Йер: "Городок, окружённый, **т. е.** с трех сторон (четвёртая— море) тремя цепями гор. [Вадим Крейд. Георгий Иванов в Йере // «Звезда», 2003]
5. К нелепым, но шумным показательным процессам и **т. д.** и **т. п.** [Александр Рыклин. Пусть игра стоит свеч (2003) // «Еженедельный журнал», 2003.04.08]
6. Он то утверждает, что у них там всё хорошо, а через неделю говорит, что всё совсем плохо (**см.** таблицу на **стр. 23**). [Александр Рыклин. Спор правительствующих субъектов (2003) // «Еженедельный журнал», 2003.03.24]
7. В 2000-м **и. о.** президента Путин назначает Беспалова **и. о.** представителя президента в Санкт-Петербурге. [Александр Садчиков. Партийный дестабилизатор. Александр Беспалов привел «Единую Россию» к краху (2003) // «Известия», 2003.02.24]
8. В 1994-1995 **гг.** начался массированный накат на мэра Петербурга Анатолия Собчака. [Александр Садчиков. Партийный дестабилизатор. Александр Беспалов привел «Единую Россию» к краху (2003) // «Известия», 2003.02.24]
9. Он был первым в спринтерской гонке на **10 км** в немецком местечке Оберхоф. [Павел Абаренов. Информ пробежка (2002) // «Вечерняя Москва», 2002.01.10]
10. Открывали фестиваль три спектакля, поставленные Андреем Гончаровым в Театре **им.** Маяковского. [Дина Абрамова. Мы говорим вам «До свидания!» (2002) // «Вечерняя Москва», 2002.02.07]
11. Частично слова **г-на** Хужамова подтверждаются статистикой департамента страхнадзора. [Таисия Мартынова. Стань сильнее или умри (2002) // «Финансовая Россия», 2002.09.19]

12. Как был, так и остался бензиновый двигатель 1,3 л (82 л. с.). [Андрей Патрикеев. Высокая мода (2002) // «Автопилот», 2002.07.15]
13. На самом деле даже в пятом сете ему удавались титанические подачи, скорость которых достигала 212 км/ч. [Роман Средиземский. Мертвее всех мертвых. Марат Сафин: «Я больше не мог. Это было ужасно» (2002) // «Известия», 2002.08.28]
14. Формализация мышления от античных времён до эпохи кибернетики. — 2-е изд., перераб. и доп. — М.: Знание, 1985. [В. А. Успенский. Витгенштейн и основания математики (2002)]
15. Ещё в 1989 г. руководство депо Челябинск предлагало увеличить до ст. Петропавловск плечо обслуживания пассажирских поездов, проходящих по ст. Курган. [Диктует экономика (2001) // «Локомотив», 2001.07.26]
16. Всего на площади более 7000 км кв. изучены инфразвуковые поля и дан прогноз нефтегазоносности 140 ловушек структурного и неструктурного типов. [Инфразвуковая разведка на нефть и газ АНЧАР — новая российская геофизическая технология. Результаты // «Геоинформатика», 2001]
17. Помехи наблюдают на всех участках, где для питания рельсовых цепей используется ток промышленной частоты 50 Гц. [Как защитить устройства АЛСН от помех (2001) // «Локомотив», 2001.06.28]
18. Температура 8 град., непонятно, почему стоящие рядом люди зябко кутаются в куртки и пальто. [Об Изумрудном острове, кельтах, Св. Патрике, погоде и современном образовании (2001) // «Туризм и образование», 2001.03.15]
19. Соковыжималка в сборе 1 шт. [Инструкция к электросоковыжималке (1981)]
20. Основные модели, на которые делает ставку компания, — Mini One (двигатель 1,6 л, 90 л. с.) и Mini Cooper (1,6 л, 115 л. с.) [Официальные дилеры (2002) // «Автопилот», 2002.05.15]
21. Для известкования огорода или газона потребуется 2,5-7 кг золы на 10 м², а под каждый куст или дерево — 250-450 г. [Сад на кислых почвах (2003) // «Сад своими руками», 2003.09.15]
22. Поэтому оценивая размер входного текста надо исходить из нескольких МБ (уже сейчас маленькая коллекция текстов Диалога 2000-2001 составляет 3,5 МБ). [Подборка кратких электронных сообщений одного автора (2003)]
23. Греф пишет, что они насчитали от 3 до 5 трлн. функций министерств. [Определить присутствие государства в экономике (2003) // «Время МН», 2003.08.07]
24. 25-220 гг. н. э. [История восточной философии (1998)]

25. Наш адрес: 119034, Москва, Турчанинов **пер.**, 6. [Информация // «Народное творчество», 2003]
26. Стоимость тура определена в условных единицах (1 **у. е.** = 1 USD). [Договор о туристических услугах (2003)]
27. Мифология и повседневность. **Вып. 2.** [Елена Душечкина. Дед Мороз и Снегурочка // «Отечественные записки», 2003]
28. **Пер.** с французского Е. В. Морозовой. [Эдуард Вирапян. Разорванный маршрут // «Знание -- сила», 2003]
29. Цена флакончика 0,5 л – 77 **руб.** [Татьяна Булгакова. Цветочная «косметичка» (2003) // «Сад своими руками», 2003.01.15]
30. Ленин предложил платить чиновникам "не выше платы рабочего" (ПСС, т. 33, с. 109). [Б. Варецкий. Стыдные уроки барства. Власть и бедность (2003) // «Советская Россия», 2003.08.21]
31. Оснащён же ноутбук отнюдь не "по-женски": процессор Pentium М с частотой 1,1 **ГГц**, оперативная память до 1 **Гбайт**, жёсткий диск 60 или 80 **Гбайт**, встроенный Ethernet-адаптер 10/100 **Мбит/с**, адаптер беспроводной сети **и т. д.** [Ноутбук для леди // «Computerworld», 2004]
32. Для приготовления мясных шницелей "Классических" (350 г, 206 **ккал**) от ООО "МЛМ-РА" из г. Раменское тебе потребуется всего несколько минут. [К нашему любимому празднику: все угощение -- на стол! // «Даша», 2004]
33. Театр "Мурзилки", **пос.** Лаишево, Татарстан. [Классика на школьной сцене // «Народное творчество», 2004]
34. 141980 Дубна, Московская **обл.**, Россия. [Магнитные фазовые переходы в допированных железом манганитах Pr_{0.7}Ca_{0.3}Mn_{1-y}FeyO₃ при высоких давлениях // «Физика твёрдого тела», 2004]
35. Голышмановский **р-н**, д. Усть-Ламенка. [Свадьба тюменских старожилов // «Народное творчество», 2004]
36. 512 **Мбайт** оперативной памяти. [Терабайт за полчаса // «Computerworld», 2004]
37. Наиболее важным водотоком на территории Раменского района является **р.** Москва. [Экологическое состояние р. Москвы на территории Раменского района Московской области // «Геоинформатика», 2004]
38. Долина **р.** Родинг, горы Дун в **Нов.** Зеландии. [А. А. Антонов. Минералогия родингитов Баженовского гипербазитового массива (2003)]
39. Именно таковы бывшие "корейцы", ныне почти русские КИА К 2700 II, К 3000 S и К 3600 II грузоподъёмностью 1,0; 1,5 и 2,5 **т.** соответственно. [Анатолий

Карпенков, Юрий Нечетов. Балтийские головастики (2003) // «За рулем», 2003.05.15]

40.Использованные карандаши и бумага оплачиваются по цене 125 **руб** карандаш и 20 **руб** 2 листа бумаги — а всего 145 **руб**. [С. Артюхов. ...И паек в натуре (2002) // «Сочи», 2002.08.22]

41.Рост самой Юхсии — всего 1,53 **м**, а весит она 55 **кг**. [Ольга Бойцова. «Плейбой» чуть не раздел депутата (2002) // «Вечерняя Москва», 2002.02.07]

42.В июле-ноябре 2000 **г.** в Доминиканской Республике возникла вспышка заболеваний, поразившая 19 человек в возрасте от 9 **мес** до 21 года. [Ликвидация полиомиелита и роль вакцинных вирусов в этом процессе (2002) // «Вопросы вирусологии», 2002.10.14]

43.Ветер северный, 4-9 **м/с**. В отдельных районах небольшие осадки.

44.Есть сирконстанты, которые сочетаются с глаголами, обозначающими **т. н.** конкретные ситуации. [В. С. Храковский. Понятие сирконстанта и его статус (1999)]

45.Домашний адрес: **г.** Владимир, ул. Пушкина, д. 6, **кв.** 18. [Автобиография (1999)]

46.Взыскано по тарифу: 45 **руб** 67 **коп.** [Доверенность на внесение сведений (1998)]

47.Удостоверение водолаза рекордсмена, выдано **тов.** Чертанову. [Екатерина Боярина. Подъем! (1997) // «Столица», 1997.01.06]

48.Он приказал заложить лошадей и поехал внезапно ревизовать **С.-** Петербургскую таможеню. [Ю. Н. Тынянов. Малолетний Витушишников (1933)]

49.Входят Дашенька, Мозговой, шафера, кавалеры, барышни и **проч.** [А. П. Чехов. Свадьба (1889)]

50.**Т. о.**, срок исковой давности истек в 2009 году. [коллективный. Срок исковой давности (2011-2012)]

Приложение Г. Ошибки чтения сокращений: результаты тестирования

Голос	Katya		Юлия		Tatyana		Alyona		iSpeech	
Номер предложения	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок
1	2	ул. пропуск д.			1	гэ	2	г. д.	3	
2	1	станцией			1	статья	1	ст	1	
3	1	долл			1	долл				
4	1	те								
5	2	итд итп								
6	1	см			2	эсэм строение			2	
7	2	иа	1	исполняющей обязательно	7	ио	2	ио	2	
8	1	гОда					1	год	1	
9										
10					1	им	1	им	1	
11							1	гэна	1	
12	1	литрас			2	эль эль эс	1	литра эс	1	литра с
13					1	вэ час				
14	4	изд., перераб. и доп. — М.	3	перераб и дополнительный эм	4	изд., перераб. и доп. — М.	4	издательство, перераб. и доп. — М.	4	
15	2	станция	1	по станцию	2	станция	2	ст ст	3	гОда ст ст
16	1	кв			1	кв	1	километров квартира	1	
17					1	гц				
18	1	град	1	град	1	град	1	град	1	
19					1	шатэ	1	штук	1	
20	2	литровс			4	эль эль эс	2	литров эс два раза	2	
21	1	г							1	к
22	1	эмбэ	1	эмбэ	1	эмбэ			1	
23					1	трлн	1	триллион А	1	
24	2	гнэ			1	гэгэ			2	к н э
25	1	пёр			1	пёр			1	
26	1	уе			1	уе	1	уе	1	
27	1	вып	1	вып	1	вып	1	вып	1	
28	1	пёр			1	пёр	1	переулок	1	
29					1	эль				

30	1	с	2	телефон эс	1	тэ	2	т с	2	
31	2	мбитс итд			4	ггц гбайт – в секунду	4	гбайт ггц мбитс	4	гбайт ггц мбитс
32	2	г город	1	ккал	2	килокало рия город	1	г	2	кал к
33	1	пос							1	
34	1	обл							1	
35	1	д	1	д	1	д	1	д	2	
36					1	мбайт	1	мбайт	1	
37	1	р	1	р	1	эр	1	р	1	
38	2	р нов	2	р нов	2	эр нов	2	р нов	2	
39					1	тэ				
40									1	рУблей, рубль
41									1	метр
42	1	мес			2	год мес	1	мес	2	годА мес
43	1	мс			1	метр слеш эс			1	мс
44	1	тн							1	
45	3	ул пропуск д кв			1	г	1	д	4	
46	1	коп			1	копейка				
47	1	тов	1	тов	1	товарищ	1	товарище ство	1	
48	1	эс	1	с	1	с	1	с	1	
49	1	проч					1	проч	1	
50	1	то	1	то	1	т о	1	то	1	

Приложение Д. Список предложений для тестирования озвучивания аббревиатур

1. Ожидается, что к 2012 г. сумма денежных переводов в указанный регион превысит 70 млрд. долларов **США**. [Денежные переводы мигрантов -- фактор инновационного развития мировой финансовой инфраструктуры // «Вопросы статистики», 2004]
2. Он выступил перед студентами и преподавателями **МГИМО** с лекцией о состоянии и перспективах развития отношений между Россией и США в области энергетики. [День за днем // «Дипломатический вестник», 2004] [омонимия снята]
3. Ротационная **БД** в **ЛВС** типа «звезда». [А. Б. Барский. Применение SPMD-технологии при построении сетевых баз данных с циркулирующей информацией // «Информационные технологии», 2004]
4. С 1971 по 1977 гг. работал инженером **НПО "Вектор"**. [Новости ВКО // «Воздушно-космическая оборона», 2001]
5. В качестве полузаказных **СБИС** на основе **БМК** возьмём серию К 1574, производимую на **НПО "Интеграл"**. [П. Н. Бибило. Логическое перепроектирование схем, реализованных на FPGA, в схемы на базовых матричных кристаллах // «Информационные технологии», 2004]
6. Так что не надо замазывать влияние **КПРФ**. [Вячеслав Костиков. Тень КПРФ становится гуще (2003) // «Аргументы и факты», 2003.01.29]
7. Такую сверхзадачу поставили перед собой в прямом радиозэфире "Эха Москвы" один из собственников **ЮКОСа** Василий Шахновский и "политкорректно" пособлявшие ему на разные лады, обозреватели А. Венедиктов и М. Ганопольский. [Иван-дурак и миллиардер Шахновский (2003) // «Завтра», 2003.08.22]
8. В мае 1933 года министр **МВД**, правая рука Ф. Рузвельта, неукротимый Гарольд Икес доложил президенту о "сплошной деморализации" нефтяной промышленности. [Иван-дурак и миллиардер Шахновский (2003) // «Завтра», 2003.08.22]
9. Действительно, согласно ч. 2 ст. 89 **ЖК РСФСР** в случае выезда нанимателя и членов его семьи на другое постоянное место жительства договор найма считается расторгнутым со дня выезда. [К. Аксенова. Особенности жилищных споров (2002) // «Биржа плюс свой дом» (Н. Новгород), 2002.05.20]

10. Так, например, при аренде **BMW** 5-й серии на счёте замораживается около \$ 5 тыс. [Денис Литошик. Автомобиль напрокат\Тариф на час (2002) // «Автопилот», 2002.05.15]
11. В этом нет ничего странного, ведь "**ГУМ**" — это корпус гуманитарных факультетов Московского государственного университета им. Ломоносова. [Анна Семенова. Власть, дай порулить (2002) // «Вечерняя Москва», 2002.04.11]
12. Молодёжная конференция, посвящённая 10-летию **ИБТ СО РАН**, показала, что молодые учёные Сибири успешно осваивают современные информационные технологии и как разработчики, и как пользователи. [В. Барахнин. Будущее компьютерных технологий закладывается сегодня (2001) // «Наука в Сибири» (Новосибирск), 2001.03.07]
13. Несмотря на то, что страны-участницы **НАТО** такое решение не поддержали, Министерство обороны США выделило 8 млрд. долларов на разработку **ПРО**, что полностью противоречило всем соглашениям. [Ольга Божьева. Договор по ПРО: полгода до развязки // «Воздушно-космическая оборона», 2001]
14. Прежде всего указать соцпроисхождение командира, — то, что он из рабочей семьи, член **ВКП (б)**. [Василь Быков. Болото (2001)]
15. Место разреза должно находиться за **ЛЭП** по ходу движения поезда на расстоянии не менее 100 м. [Как защитить устройства АЛСН от помех (2001) // «Локомотив», 2001.06.28]
16. С 1998 г. — начальник Управления собственной безопасности **ФСБ РФ**, а через год — заместитель директора ФСБ — начальник Департамента экономической безопасности. [Новости ВКО // «Воздушно-космическая оборона», 2001]
17. Национальная ассоциация издателей провела консультации по проблеме фактического запрета рекламы алкогольных напитков в печатных **СМИ** с участием практически всех основных игроков рекламного рынка: представителей **РСР**, **КонфОПа**, **СЖР**, Ассоциации рекламодателей, **РАРА**, крупнейших издательских домов. [Вера Васильева, Людмила Сатушева. Наивное письмо гр. Путину (2000) // «Рекламный мир», 2000.02.15]
18. В те незапамятные времена, когда ещё не было ни **МАПа**, ни **ФЗ "О рекламе"**, трехпрограммник мог рекламировать на всю страну всё, что заблагорассудится. [Динозавры радиозэфира (2000) // «Рекламный мир», 2000.03.30]
19. В 1989 году окончила **МГПИ** им. Ленина, в 1998 году — Международную академию предпринимательства по специальности "Менеджмент в рекламе". [Кто куда (2000) // «Рекламный мир», 2000.03.30]

20. Конфликт начался с того, что член правления **РАО** Андрей Трапезников потребовал от **МПС** немедленной выплаты всех долгов (в **РАО ЕЭС** насчитали, что за этот год задолженность **МПС** перед энергетиками увеличилась на 2, 7 млрд руб.) [Неделя 22.08-2000.08.28 // «Коммерсантъ-Власть», 2000]
21. 33 тыс. заключённых, содержащихся в российских колониях и тюрьмах, заражены **ВИЧ**. [Павел Черников. Россия в цифрах // «Коммерсантъ-Власть», 2002]
22. А 12 июня, в День принятия декларации о государственном суверенитете РФ, в **ЦПКиО** им. Горького артисты Читинской области порадовали москвичей театрализованным праздником "Забайкалье— золотая моя колыбель". [Марина Ленская. Приглашаем в Забайкалье! (2002) // «Вечерняя Москва», 2002.06.13]
23. Появление электроники — читай систем **ABS**, **ESP** и пр. — вызвало к жизни ещё ряд мутаций. [Николай Качурин. Крутящий момент истины (2002) // «Автопилот», 2002.02.15]
24. И не один, а с **Ми-8**, **Ф-1**, **РПК**, **ПМ**, **АКМС**, **ПБС-1**, **Ми-24 П**, **ВВС**, **СП-5** и гранатомётом **РПГ-18 "Муха"** [Юрий Гладильщиков. Младший брат. О чем «Война» Алексея Балабанова (2002) // «Известия», 2002.03.15]
25. Уже есть предложения делать современные сиденья, ремни и подушки безопасности, выпускные системы, **АБС** и т. д. [Петр Меньших, Владимир Каданников. В Тольятти придут инвесторы... (2003) // «За рулем», 2003.05.15]
26. Вспомним, к примеру "Каштанку" екатеринбургского **ТЮЗа**, ставшую лауреатом прошлогодней "Золотой Маски". [Северная сказка (2004) // «Экран и сцена», 2004.05.06]
27. На пути нарушителя стоят **PIN**- и **PUK**-коды. [Дмитрий Лысак. Секретная миссия простого брелка (2003) // «Stuff», 2003.03.06]
28. В табл. 2 и на рис. 2 представлены полученные оценки интенсивностей переходов и начальных значений среднего **IQ** и среднеквадратического отклонения. [Л. С. Куравский, С. Б. Малых. Применение марковских моделей для анализа эволюции психологических характеристик в популяции // «Вопросы психологии», 2003]
29. Линолеум на вспененной основе (**ПВХ**) лучше всего соединять на стыках методом холодной сварки. [Елена Волкова. Спрашивают читатели -- отвечают специалисты (2002) // «Биржа плюс свой дом» (Н. Новгород), 2002.04.08]
30. На вопросы отвечали: Владимир Николаев (**НПП** Нижегородстройтехцентр), Игорь Печенов (**ЗАО** «Новтекс»), Алексей Мартынов (компания «Ордер»). [Елена

Волкова. Спрашивают читатели -- отвечают специалисты (2002) // «Биржа плюс свой дом» (Н. Новгород), 2002.04.08]

31.Теперь получили распространение **SIM**-карты удвоенной ёмкости (64 Кб) специально для размещения приложений на базе системы обмена **SMS**. [Сергей Потресов. SIM-SIM, откройся. Операторы мобильной связи нашли новый источник доходов (2002) // «Известия», 2002.02.19]

32.**ЦТК** внесёт в неё принадлежащие ей акции **МНПЗ**, «Моснефтепродукта» и **МТК** (сеть **АЗС** в Москве), а Sibir Energy — добывающие активы в Западной Сибири с запасами 170 млн тонн нефти. [На акции Sibir Energy нашелся новый покупатель (2002) // «Финансовая Россия», 2002.09.19]

33.Ford Focus (Всеволожск) 1.6, **МКПП**, 90 л. с., \$ 10,9 тыс. [Новости (2002) // «Автопилот», 2002.08.15]

34.Был ли электрифицирован участок **РЖД** Одноробовка — Готня во время вхождения части **ЮВЖД** в состав **ЮЖД** Украины? [коллективный. Однопутки, двухпутки, трёхпутки... (2008-2011)]

35.А вот о **СССР** вспоминают с ностальгией — бесплатная квартира, вкусная колбаса и т.п. [коллективный. Социализм vs Капитализм (2011)]

36.В этой цитате речь идёт только об осложнениях от прививок **БЦЖ**, т. е. лишь о меньшей части случаев зарегистрированных осложнений от всевозможных прививок за один год. [коллективный. Статистика инфекционной заболеваемости в РФ (2004-2011)]

37.Есть два кредита на **ИП** на общую сумму 1.8 млн руб и две кредитных карты на **физлицо** на сумму 300 т.р. [коллективный. Что делать или где взять кредит с плохой кредитной историей? (2011)]

38.Одновременно с этим в прошлом месяце **Центробанк КНР** впервые за последние три года повысил процентные ставки по годовым кредитам и годовым депозитам в национальной валюте на 0,25 п.п. [Евгений Басманов. Пекин ужесточил денежно-кредитную политику (2010.11.11) // <http://www.rbcdaily.ru/2010/11/11/world/526218.shtml>, 2010]

39.Рекордный объем зерновых — 175,6 тыс. т — **ОЗК** продала за рубеж в первой половине августа. [Наталья Копейченко. «Ростехнологии» ответят за томографы (2010.08.18) // <http://www.rbcdaily.ru/2010/08/18/market/503264.shtml>, 2010]

40.Искренне удивляюсь Вашей способности везде успевать и в «**РП**» Вы отметились, и здесь, и по ночной Москве на велике, и в кино, и т. д. и т. д... [коллективный. Горный двухподвесочный (2010)]

41. **АпДалСв АО** — ЭТО нанотехнологии! [коллективный. Золото Якутии :) (2010)]
42. Точно такую же картину можно нарисовать относительно **КПК** и т.д. и т.п. [коллективный. Об автоматическом (безоговорочном) признании ПВО (2010)]
43. Если у вас артрит в грудном отделе, то никакие **МРТ**, рентгены и т.п. не покажут ничего на начальной стадии, я подчёркиваю, на начальной стадии. [коллективный. Форум о медицинской диагностике (2010)]
44. Если квартиры в домах серии И-155 имеют среднюю площадь 66,6 кв. м в пределах **МКАД** и 78,1 кв. м в области до 5 км от **МКАД**, то П-44Т — 59,3 кв. м и 61,6 кв. м соответственно. [Яна Зайцева. Лидер панельного домостроения (2009.07.09) // <http://www.rbcdaily.ru/2009/07/09/market/422299.shtml>, 2009]
45. Он включает достройку Богучанской **ГЭС** установленной мощностью 3 ГВт и строительство алюминиевого завода производственной мощностью 600 тыс. т в год, который станет основным потребителем электроэнергии **ГЭС**. [Леонид Хомерики. Энергометаллургический залог (2009.03.17) // <http://www.rbcdaily.ru/2009/03/17/tek/406261.shtml>, 2009]
46. **СЗЛК** производит 150 тыс. т целлюлозы в год, производство бумаг и бумажно-беловых изделий составляет 170 тыс. т в год. [Маргарита Рахнова. Mercer International не нашла в России леса (2008.05.14) // <http://www.rbcdaily.ru/2008/05/14/industry/342741.shtml>, 2008]
47. Необычные выставки, лекции, **ДНК**-лаборатории, открытые для всех желающих, веб-сайты с полными текстами трудов основоположника теории эволюции и даже новый «Бигль»! [Т. Зимина. В центре внимания — Дарвин! // «Наука и жизнь», 2008]
48. Среди последних был ряд объективно выделенных «фирм» с именем (например, **МФТИ**, **МИФИ**, **МИЭМ**, **МИСиС**, **МИНХ**, **МИМО**, **Станкин**, **МАИ**, **МГПИ** и т. д.) [Ю. Неретин. ЕГЭ: перспективы и эволюция // «Наука и жизнь», 2008]
49. Т.е. чтобы принимали только по **ЕГЭ**. [коллективный. Ваше отношение к **ЕГЭ**? За или против (2005-2008)]
50. Желание уйти от рекомендаций общего характера приводит к рассмотрению классов задач или отдельных «представительных» задач и к демонстрации их возможностей параллельного решения по **SPMD**-технологии, привлекающей всё большее внимание как разработчиков **ВС**, так и математиков-программистов [А. Б. Барский. Применение **SPMD**-технологии при построении сетевых баз данных с циркулирующей информацией // «Информационные технологии», 2004]

Приложение Е. Ошибки озвучивания аббревиатур

[illegible]

36										
37					1	ип			1	физлИцо
38										
39	1	оск			1	оск	1	оск	1	оск
40										
41					1	Ао			1	Ао
42										
43										
44										
45					1	гээс				
46										
47	1	днк							1	днк
48	4	мИфи мИио мИсис стАнкин	1	мИмо	4	мифи мисмси мимо май	4	мИмо мИсис стАнкин мАи	4	мифи мимо станкин май
49										
50	1	спмд					1	воскресе нье	1	спмд

**Приложение Ж. Список предложений для тестирования расшифровки
цифровых обозначений и выражений**

1. Дата рождения— **08.04.1980** г. (**23** года), женат. Тел. **232-23-23**. [Резюме экономиста (2004)]
2. Из оставшихся **26521000** человек **13450000** составляют депутаты и госслужащие, **5 млн.**— военные, **3977368**— нелегальные эмигранты. [Галина Коренева. Голодный учитель-позор для России! (2003) // «Аргументы и факты», 2003.01.22]
3. В результате к **5 млн. 800 тысячам** рублей, запланированным на **2002** год из федерального бюджета на строительство хирургического корпуса Йошкар-Олинской горбольницы, к концу года Москва выделила ещё **6 млн.** рублей. [Здравоохранение: реалии дня. Отдадим долги и будем развиваться! (2003) // «Марийская правда» (Йошкар-Ола), 2003.01.17]
4. В минувшем году отметил своё **30-летие** московский Музей В. А. Тропинина и художников его времени. [Елена Чебышева. Дети рисуют всерьез (2002) // «Вечерняя Москва», 2002.01.10]
5. Глядя на сегодняшнюю Ирландию, трудно себе представить, что в середине **XIX** в. после разразившегося голода отчаявшиеся люди пересекали океаны и, терпя лишения и страдания, добирались до берегов Америки и Австралии, а ещё **10** лет назад не во всех деревнях было электричество. [Об Изумрудном острове, кельтах, Св. Патрике, погоде и современном образовании (2001) // «Туризм и образование», 2001.03.15]
6. Лампа коптила много столетий, потому что ламповое стекло изобрели лишь во второй половине **XVIII** в. [Приручение света // «Трамвай», 1990]
7. Взять $\frac{1}{2}$ стакана дрожжей, $\frac{1}{4}$ стакана воды, $\frac{1}{2}$ фунта сахара, $\frac{1}{3}$ фунта т. е. **1** стакан изюма, $\frac{1}{2}$ золотника шафрана, намоченного в $\frac{1}{4}$ стакана холодной воды или рома, **60** желтков, **2 $\frac{1}{2}$** стакана растопленного масла, муки **1** гарнец без **2** стаканов. [Елена Молоховец. Подарок молодым хозяйкам, или средство к уменьшению расходов в домашнем хозяйстве / Булки и куличи (1875-1900)]
8. Это лишний раз подтвердил **12-й** по счёту и первый в статусе республиканского фестиваль горномарийской песни, прошедший при

- аншлаге во Дворце культуры имени Я. Эшпая. [Песни при аншлаге (2003) // «Марийская правда» (Йошкар-Ола), 2003.01.15]
9. Эта этноязыковая общность существовала в Центральной Европе во II тысячелетии до н. э. и названа Г. Краэ древнеевропейской. [В. В. Седов. Этногенез ранних славян // «Вестник РАН», 2003]
 10. На требование остановиться **4,3**-тонный автомобиль разогнался почти до паспортных **90** км/ч, за что и был расстрелян из табельных милицейских автоматов и пистолетов на Нагатинской набережной. [Левый марш // «Автопилот», 2002]
 11. **15.09.2004, 17.21** СТОКГОЛЬМ, **15** сентября. [Борис Зайцев. ФИФА выиграла судебную тяжбу за свои права в Интернете (2004) // ИТАР-ТАСС, 2004.09.15]
 12. В Сиднее, завоевав **35** медалей, россияне были на **14** месте. [Сергей Латышев. Открытие Паралимпийских игр (2004) // ИТАР-ТАСС, 2004.09.18]
 13. Лобовые столкновения составляют около **50%** всех случаев аварий. [В. А. Овчинников. Применение генетических алгоритмов в задачах синтеза кузова автомобиля // «Информационные технологии», 2004]
 14. Мы вспоминаем **1950-1960-е** годы, когда сотрудничество между Советским Союзом и Египтом было на очень высоком уровне. [В. В. Путин. Вступительное слово в ходе российско-египетских переговоров в расширенном составе // «Дипломатический вестник», 2004]
 15. В дни празднования **60-летия** высадки союзных войск в Нормандии выражаю от имени всего российского народа огромную, искреннюю благодарность за ваш исторический подвиг. [В. В. Путин. Послание ветеранам государств -- союзников по антигитлеровской коалиции // «Дипломатический вестник», 2004]
 16. Среди респондентов **61%** девушек и **39%** юношей. [Ю. Ф. Флоринская, Т. Г. Рощина. Жизненные планы выпускников школ из малых городов (2004) // «Человек», 2005]
 17. Наш адрес: **119034**, Москва, Турчанинов пер., **6**. Тел.: **708-3102**; т/ф: **246-9323**.
 18. тел.: **(00222) 2-22-22, 40-01-06** [Договор на поставку оборудования (2003), телефон изменён]
 19. Ноутбуки оснащаются процессорами Pentium **4** с тактовой частотой до **3,2** ГГц и видеокартой ATI Radeon **9700**. [«Гигабитный» ноутбук // «Computerworld», 2004]

- 20.Надо: **1,5** кг картофеля **1** пучок зелёного лука [«Даша» делится секретами -- побалуй близких вкусным обедом! // «Даша», 2004]
- 21.В **2005** году общемировая численность абонентов вырастет до **2 млрд.**, а в **2009**— до **2,5 млрд.** По данным Gartner, за минувший год объём мирового рынка ИТ-услуг вырос на **6,2%**, составив **569 млрд.** долл. причём период наиболее быстрого роста пришёлся на вторую половину года.
[Аналитический дайджест // «Computerworld», 2004]
- 22.Вертолёт выполнял плановый полёт. Ми-**24** упал в районе населённого пункта Алхан-Кала примерно в **16.41** по московскому времени. [В Чечне разбился вертолет Ми-24, один летчик выжил (2004) // РБК, 2004.09.13]
- 23.В Республиканской детской больнице N **1** появился доставленный из Москвы спиральный компьютерный томограф фирмы стоимостью **38 млн.** руб. Один компьютерный томограф уже функционирует во **2-ой** детской больнице. [В детской больнице появился спиральный компьютерный томограф (2004) // «Московский комсомолец» в Саранске, 2004.12.23]
- 24.**9-10** июня В. В. Путин встретился с Премьер-министром Канады П. Мартином. [Встречи В.В. Путина на Си-Айленде // «Дипломатический вестник», 2004]
- 25.Говорят, память человека развивается только до **30-ти** лет. [Запись LiveJournal (2004)]

Приложение II. Результаты тестирования расшифровки цифровых обозначений и выражений

[illegible]

Приложение К. Список предложений для тестирования чтения особых символьных обозначений

1. Я вообще не понимаю, откуда в средствах массовой информации взялась цифра в \$ 5 млн. [Александр Садчиков. Партийный дестабилизатор. Александр Беспалов привел «Единую Россию» к краху (2003) // «Известия», 2003.02.24]
2. Высота рамы – 400 мм. Размер шин — 406×40 мм. Количество передач – 1 шт. [Руководство по эксплуатации велосипеда (1978)]
3. Когда металлический шар, площадь поверхности которого $S=100\text{ см}^2$, покрыли тонким слоем хрома, масса шара увеличилась на 36 мг. [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7-9 кл. (2003)]
4. Нагружающий механизм поддерживает нагрузки сжатия на образец величиной 1, 20, 100 и 200 кПа с точностью $\pm 5\%$. [Г. К. Мухамеджанов. Приборы для испытаний нетканых полотен // «Текстиль», 2003]
5. При этом температура Земли повысится на 6°C по сравнению с сегодняшним уровнем. [А. А. Макаров. Научно-технологические прогнозы и проблемы развития энергетики России до 2030 года. (2008) // «Вестник РАН», 2009]
6. Подливку посолить, поперчить, добавить лимонный сок, а после того как она остынет до $50\text{-}60^\circ$ - желтки. [Рецепты национальных кухонь: Франция (2000-2005)]
7. Банки, обслуживающие массовых клиентов, считали, что из них 20% приносят 80% доходов. [Валерий Коржов. Банк с роботизированным лицом // «Computerworld», 2004]
8. Температура ночью -15, -17 градусов, днем -12, -14 градусов. [Погода (2000) // «Сургутская трибуна», 2000.02.26]
9. Температура ночью — $+2\text{ }+7^\circ$, днем 6 — 11° . [Анатолий Гладилин. Прогноз на завтра (1972)]
10. Отправьте запрос *100#.
11. Вопрос №2.
12. Выберите: да/нет.
13. Red&White.
14. mail@mail.com.

- 15.Используется охранная сигнализация Бульдог®.
- 16.Был использован фотоаппарат Кодак™.
- 17.<http://www.test.ru/symbols>
- 18.Евро заменил в соотношении 1:1 европейскую валютную единицу (Википедия).
- 19.Выпускаются монеты достоинством 2 €. (Википедия).
20. $A > B$.
- 21.Представим сумму продуктов деления (кластеров) в виде суммы всех элементов таблицы Менделеева: **He, C, O, Mg, Si, Fe** и т. д. [Ядерная диссоциация химических элементов в геохимической истории развития Земли // «Геоинформатика», 2003]

Приложение Л. Результаты тестирования чтения символьных обозначений

Голос	Katya		Юлия		Tatyana		Alyona		iSpeech	
Номер предложения	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок
1					1				1	не там
2	1		1		1				1	
3									1	
4	1				1				1	
5	1									
6									1	
7	1									
8										
9	1									
10	2				1	*			2	
11									1	
12										
13	1								1	
14									1	
15	1									
16	1				1					
17			1						1	не читает слеши
18	1		1				1		1	
19					1				1	
20	1		1						1	

Приложение М. Список английских слов с их практической транскрипцией
[Черепанова 2015]

№	Английское словоупотребление	Практическая транскрипция
1	Twitter	тви+тэр.
2	WebMoney	вэб ма+ни
3	BBC Earth	би би си эрс
4	Holdings Incorporation	хо+улдингс инкорпорэ+йшн
5	The Astrophysical Journal	эстроуфи+зикал джо+рнал
6	trailblazer	трэ+йл блэ+йзэр.
7	facebook	фэйсбу+к
8	Samsung Electronics	самсу+нг илектро+никс
9	Manchester United	манче+стэр юна+йтэд
10	Special Edition	спешл иди+шн
11	Work and Travel	ворк энд трэвл
12	Microsoft Word	ма+йкрософт ворд
13	Black and White	блэк энд ва+йт
14	Social Security	со+ушл сикью+рити
15	Market Watch	ма+ркет вотч
16	Dream Works	дрим воркс
17	swift	свифт
18	Drag and Drop	дрэг энд дроп
19	International Business	интэрнэ+шнл би+знес
20	Nokia Partners	но+киа па+ртнэрс

21	Financial Journal	файнэ+ншл джо+рнал
22	World Congress	ворлд конгре+с
23	Milky Way	ми+лки вэй
24	Galaxy Note	гэ+лакси но+ут
25	email	имэ+йл
26	Adidas Global	адида+с гло+убал
27	Capital Partnership	кэ+питал па+ртнэршип

Приложение Н. Результаты тестирования по озвучиванию англоязычных вставок

0 – не точное соответствие или недопустимое прочтение

Голос	Katya		Юлия		Tatyana		Mary TTS		Alyona		iSpeech	
Номер предложения	Допустимость	Точность	Допустимость	Точность	Допустимость	Точность	Допустимость	Точность	Допустимость	Точность	Допустимость	Точность
1							0	0			0	0
2	0	0				0	0	0		0	0	0
3	0	0			0	0	0	0		0	0	0
4			0	0	0	0	0	0		0	0	0
5				0	0	0	0	0		0	0	0
6			0	0	0	0	0	0		0	0	0
7							0	0		0	0	0
8	0	0			0	0	0	0		0		
9		0			0	0	0	0		0	0	0
10					0	0	0	0		0	0	0
11				0	0	0	0	0		0	0	0
12				0		0	0	0		0	0	0
13		0			0	0	0	0			0	0
14					0	0	0	0			0	0
15	0	0				0	0	0		0	0	0
16					0	0	0	0		0	0	0
17						0	0	0				
18					0	0	0	0			0	0
19					0	0	0	0			0	0
20	0	0				0	0	0		0	0	0
21		0			0	0	0	0		0	0	0
22		0		0	0	0	0	0		0		0
23		0			0	0	0	0		0	0	0
24					0	0	0	0			0	0
25					0	0	0	0				
26	0	0			0	0	0	0			0	0
27	0	0			0	0	0	0		0	0	0

Приложение П. Список предложений для тестирования правильности снятия омонимии

Выделены проблемные слова.

1. А не много ли **берете** на себя? [Ю. О. Домбровский. Обезьяна приходит за своим черепом, часть 2 (1943-1958)]
2. И тут взвился какой-то юноша в зеленом **берете**. [Герман Садулаев. Шалинский рейд (2009) // «Знамя», 2010]
3. Единообразное медленное движение **весел** усугубляет печальную важность сего обряда. [Г. П. Каменев. Инна (1804)]
4. Когда человек доволен и **весел**, видит ли он, как летит время? [П. Ю. Львов. Даша, деревенская девушка (1803)]
5. Ну как с ними можно **вести дела**? [Сергей Доренко. Левые силы – перезагрузка (2003) // «Завтра», 2003.08.13]
6. Так вот, телеканал "Хороших новостей" будет передавать "**Вести** с полей" и сообщения о прочих трудовых подвигах. [Даешь «вести с полей» (2003) // «Совершенно секретно», 2003.04.08]
7. Была **вещая** правда в стихах Анны Ахматовой: [Дон Аминадо. Поезд на третьем пути (1954)]
8. Простуженно хрипели по скверам репродукторы, тревожно **вещая** о войне. [Михаил Бубеннов. Белая береза / части 1-2 (1942-1952)]
9. Едва потеряла она его из **виду**, то и упала в обморок. [А. Е. Измайлов. Бедная Маша (1801)]
10. При этом под «жизненными интересами» мы не имеем в **виду** некое право оказывать давление на наших соседей. [С. В. Лавров. Статья в газетах «Коммерсант» и «Уолл-стрит джорнал» // «Дипломатический вестник», 2004]
11. Среди депутатов он всегда оставался белой **вороной**. [Сати Спивакова. Не всё (2002)]
12. Рядом с Якобсом на **вороной** кобыле восседал сбежавший принц. [Андрей Белянин. Свирепый ландграф (1999)]
13. Меняется ли в течение дня длина тени от перекладины **ворот**? [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7-9 кл. (2003)]

- 14.Для облегчения подъёма грузов часто применяют **ворот**, состоящий из двух валов, неподвижно закреплённых на одной **оси**. [Владимир Лукашик, Елена Иванова. Сборник задач по физике. 7-9 кл. (2003)]
- 15.Впрочем, кто читал **все** книги Дюма? [Рецепты национальных кухонь: Франция (2000-2005)]
- 16.В нашем гнезде **все** размещено очень рационально. [Сочинение девочки по биологии (2003)]
- 17.Друг его, предуведомленный обо **всем**, искренно сожалеет о **нем**. [Н. Мамышев. Злосчастный (1807)]
- 18.Конечно, далеко не **всем** подойдёт самостоятельная сборка серверов на базе Itanium 2. [Михаил Кузьминский. Водопады Supermicro // «Computerworld», 2004]
- 19.Поисковые работы в 2004 **году** продлятся примерно 2 недели. [На полях былых сражений в Новгородской области в понедельник возобновятся поисковые работы (2004) // РИА «Новости», 2004.09.13]
- 20.К 2015 **году** следует ожидать увеличения этого слоя. [Янис Астафьев. Кто будет работать в России в 2015 году? // «Отечественные записки», 2003]
- 21.Потому что из-за неё всей семье горе. [Юрий Трифонов. Дом на набережной (1976)]
- 22.Людей тянуло туда, к круглой конусовидной **горе**, где им мерещились синезелёные призраки горных **духов**, плясавшие вокруг зеленоватого светящегося **облака**. [И. А. Ефремов. Озеро горных духов (1942-1943)]
- 23.Сегодня в 15 кинотеатрах **города** начинается демонстрация [Екатерина Барабаш. Это не «Брат-3». Это «Война» (2002) // «Вечерняя Москва», 2002.03.14]
- 24.Завоевать планету, зарабатывая деньги, завоевывая **города** и борясь за ресурсы в "Цивилизации" или WarCraft, — увлекательнейшее занятие. [Анна Майорова. Виртуальное отечество в опасности. Неполадки в экономике игрушечной страны ведут к реальному мошенничеству (2002) // «Известия», 2002.10.22]
- 25.Она не ожидала, что к её **горю** отнесутся с такой чёрствостью. [М. А. Булгаков. Театральный роман (1936-1937)]
- 26.**Я горю** желанием, не надо делать скоропостижных выводов. [коллективный. Как должна поступать нормальная альфа (2012)]
- 27.Ты когда-нибудь думал, куда **делись** последние пять лет? [Виктор Пелевин. Желтая стрела (1993)]
- 28.Нашёл деньги, ни с кем не **делись**. [Д. И. Фонвизин. Недоросль (1782)]
- 29.Ей это было по **душе**: держать в страхе и повиновении. [Токарева Виктория. Своя правда // «Новый Мир», 2002]

30. Я попытался расслышать, шипит ли еще в **душе** вода. [Александр Терехов. Каменный мост (1997-2008)]
31. Любая книга оказывалась **западней**, в которую он летел сломя голову. [Даниил Гранин. Искатели (1954)]
32. На этот раз не перевалили отрог, а обогнули его **западней**. [А. А. Фадеев. Разлив (1923)]
33. Их встретила огромная, размером с диван, тётка, Зинаида Никифоровна, **обитая** шёлковой цветастой тканью. [Людмила Улицкая. Казус Кукоцкого [Путешествие в седьмую сторону света] // «Новый Мир», 2000]
34. Он же со своей стороны доказывает, что, **обитая** в Средней Азии не менее 3500 лет, турки также являются «коренным населением». [Виктор Шнирельман. Символическое прошлое. Борьба за предков в Центральной Азии // «Неприкосновенный запас», 2009]
35. "Ирония судьбы, Или с лёгким **паром!**" в новогоднюю ночь — это уже даже не традиция. [Вадим Самодуров. Рождественские забавы: песни, порно, президент... (2002) // «Вечерняя Москва», 2002.01.10]
36. **Паром** пыхтел и свистел, отдаляясь от крутого **берега**, а противоположный так и оставался еле видимым. [Р. Б. Ахмедов. Промельки (2011) // «Бельские Просторы»,]
37. Ставит в ванную на **полку** лосьон, помазок для бритья. [И. Грекова. Перелом (1987)]
38. У нас в **полку** есть такие, что плачут без причины и поют петухами. [Василий Гроссман. Жизнь и судьба, часть 2 (1960)]
39. Я не выношу женских **слез**: при виде их я тотчас теряюсь. [И. С. Тургенев. Ася (1858)]
40. Он **слез** и ушел подальше, чтобы не слушать хруста голов под колесом и гогота. [Михаил Шишкин. Венерин волос (2004) // «Знамя», 2005]
41. Я перевела **стрелку** — и вечер со стремительностью экспресса изменил направление. [Анатолий Алексин. Раздел имущества (1979)]
42. Травкин сказал лучшему **стрелку** взвода — Марченко: — Пугни-ка их. [Э. Г. Казакевич. Звезда (1946)]
43. И правда, ведь имея среди нескольких ответов в **тесте** тот, который слегка помечен, мы именно на нем останавливаем свое внимание? [коллективный. Рецензия на фильм «Поймай меня, если сможешь» (2006-2011)]
44. Шашлык есть, сосиски в **тесте**, салат... [Олег Гладов. Любовь стратегического назначения (2000-2003)]

45. Дети доклады читают, а наши **целую** неделю гуляют, школе 20 лет исполнилось. [Наши дети: Подростки (2004)]
46. Люблю, **целую** и жду с нетерпением весточки Глафира! [Письмо студентки (2000)]
47. И конфетку к **чайку**. [Олег Павлов. Карагандинские девятины, или Повесть последних дней // «Октябрь», 2001]
48. Вместо ответа нам была оставлена тонкая книжка про **чайку**. [Мариам Петросян. Дом, в котором... (2009)]
49. На запах приятнее, **чем** на вкус, во рту чувствуется осадок. [Запись LiveJournal (2004)]
50. В **чем** парадокс современного российского социума? [Андрей Андреев. Будущее принадлежит нам! (2003) // «Завтра», 2003.08.22]

Приложение Р. Результаты тестирования правильности снятия омонимии

Голос	Katya		Юлия		Tatyana		Mary TTS		Alyona		iSpeech	
Номер предложения	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок	Кол-во ошибок	Уточнение мест ошибок
1							1					
2	1				1				1			
3	1						1					
4					1				1		1	
5					1	дела	2	дела				
6	1				1				1			
7	1								1		1	
8							1					
9	1											
10					1							
11	1				1							
12							1		1		1	
13							1					
14	2	оси			2	оси	1		1			
15	1											
16							1				1	
17	1				1	нем	2	нем			1	
18											1	
19												
20							1					
21							1				1	
22	2	облака			3	духов облака			2			
23					1							
24					1		1					
25					1		1					
26	1								1			
27					1		1	лёт			2	
28	1						1		1			
29					1		1				1	
30	1								1			
31					1		1		1			
32	1		1				1				1	
33	1				1		1		1		1	
34							2	сторон ы				
35	1				1				1			

36							1		1			
37					1		1					
38	1						1		1			
39	1						1		1			
40					1							
41												
42	1				1		1		1			
43					2	нем	2	нем	1		1	
44	1											
45	1											
46							1		1			
47	1				1		1		1			
48												
49												
50							1					

Приложение С. Список предложений для оценки правильности сегментной транскрипции

Выделены слова с вероятными ошибками.

1. В глубине квартиры затрещал **энергичный** женский голос. [Токарева Виктория. Своя правда // «Новый Мир», 2002]
2. До чего же хороша она, любимая моя **помощница**! [И. Грекова. Перелом (1987)]
3. Если же судить **нестрого**, Иисус был в законе. [Нодар Джин. Учитель (1980-1998)]
4. Прилетел в Москву и легко поступил в **Иняз**. [Сергей Шаргунов. Мой батюшка (2011)]
5. Плотный и **лёгкий** одновременно. [Наши дети: Подростки (2004)]
6. И пусть ни одного мелкого **чувства** не останется в эту прекрасную минуту в наших **сердцах**. [Виктор Пелевин. Чапаев и пустота (1996)]
7. А какао с сахаром я вообще никогда пить не могла. [коллективный. Диета Кима Протасова (2007-2010)]
8. Но имей в виду: если не выполнишь обещанное или еще хоть **раз ворвешься без** вызова, уволю без выходного пособия. [Семен Данилюк. Бизнес-класс (2003)]
9. На работе у Анны Фёдоровны тоже было неприятное **происшествие**. [Людмила Улицкая. Пиковая дама (1995-2000)]
10. Да и детки были маленькие: 1-5 класс. [Соломенные картины // «Народное творчество», 2004]
11. Наверное, это и был его идеал отношений **с** женой... [Ольга Зуева. Скажи что я тебе нужна... // «Даша», 2004]
12. Кто-то наверху решил, что типичный меньшевик — это Ширвиндт в пенсне... [Григорий Горин. Иронические мемуары (1990-1998)]
13. Ни просьбы, ни слезы, ни **отчаяние** — ничто не помогло, отец был неумолим. [Н. П. Милонов. История бедной Марьи (1805)]
14. Я **бухгалтер** в бюджетной сфере, а муж преподаёт в вузе. [Помогите, пожалуйста! (2002) // «Домовой», 2002.04.04]
15. Потащить и потянуться, затем подтащить и подтянуться.

**Приложение Т. Результаты оценки правильности сегментной транскрипции
в синтезаторах речи**

Голос	Katya		Юлия		Tatyana		Mary TTS		Alyona		iSpeech	
Номер предло- жения	Кол-во ошибок	Уточне- ние мест ошибок	Кол- во ошиб- бок	Уточ- нение мест ошибок	Кол-во ошибок	Уточ- нение мест ошибок	Кол-во ошибок	Уточ- нение мест ошибок	Кол-во ошибок	Уточ- нение мест ошибок	Кол-во ошибок	Уточ- нение мест ошибок
1											1	энергичный
2	1	помощница			1	помощница						
3	1	нестрово			1	нестро во	1	нестро во			1	нестро во
4					1	иняз без й	1	иназ	1	иняз без й	2	легко, иняз
5							1	логкий				
6												
7	1	какаА					1	какаА				
8	1	ворвЕшь ся									1	бес вызова
9							1	прое шестви е			1	прое шестви е
10	1	класС									1	класС
11	1	3 женой	1	3 женой	1	3 женой	1	3 женой				
12					1	пенснЕ	1	пенснЕ			1	пенснЕ

[illegible]

Приложение У. Набор фраз для тестирования интонации

Тип предложения	Количество	Примеры
1. Односинтагменное повествовательное предложение	3	Девочка упала. Она плачет. Завтра будет солнечно. Мы собираемся на природу. Обед готов. Я уже проголодался.
2. Специальный вопрос с вопросительным словом (+ ЛИ и НЕУЖЕЛИ)	4	Вы действительно знаете. Когда он приедет? Аня тебе рассказала? Что вчера произошло? Когда он вернётся? Вы встретитесь с ним? Я испугался. Любишь ли ты ещё меня?
3. Специальный вопрос + придаточное предложение	2	Откуда он знает, как это случилось? Когда нам сообщат, где получать посылку?
4. Общий вопрос с ударением на последнем слове	2	Маша не замужем? Она живёт одна. Его зовут Николай? Мы встречались на собрании.
5. Общий вопрос с переходом ударения	1	Мне подходят такие условия? Можно заказывать билеты.
6. Вопрос с ИЛИ	2	Вы уже в курсе? Он приедет завтра или задержится ещё на неделю? Вы будете жить в гостинице или в частной квартире? Когда вы поедете в командировку?
7. Уточняющий вопрос (типа «за неделю», «через 2 месяца»)	2	Когда вы вернётесь? Через два месяца? За неделю? Или за месяц?
8. Вопрос с А	2	А без паспорта? Можно будет купить билет?

		А если прийти завтра? Есть вероятность успеть?
9. Побудительные конструкции	2	Завтра нужно будет сдавать эту тему. Повторите её ещё раз! Не придавайте этому никакого значения! Всё в порядке.
10. Восклицания с негативной оценкой	2	Какой ужасный день! Ничего не получается. Я ничего тебе не дам. Напрасно стараешься!
11. Восклицания с положительной оценкой без вопросительного слова	2	Как вы провели выходные? Замечательно! Вот видите! Всё получилось.
12. Восклицания с вопросительным словом	2	Она вам рассказала? Какая приятная неожиданность! Он много читает. Сколько у него книг!
13. Незаконченное предложение	2	А может... Вы согласитесь? Сумеречная келья с крошечным – в альбомный лист – окном под самым потолком.
14. Первая часть сложного предложения (главное)	3	Он живёт в области, а работает в Москве. Понимаете, не хочу будить жену. Мне очень жаль, вы пришли слишком поздно.
15. Почти то же, только более длинное (или придаточное)	1	Когда ваши друзья приехали, было уже темно.
16. Пауза без знаков препинания	3	Главный корпус и примыкающие к нему служебные здания находятся в следующем квартале. Когда вы были в Москве последний раз? Два года тому назад я был здесь проездом домой. А если работающие в офисе люди в следующем месяце целую неделю не

		будут выполнять свои обязанности?
17. Перед двоеточием	1	Катя вернулась домой: она забыла выключить утюг.
18. Первое придаточное в сложноподчинённом вопросе	1	Когда он приедет, вы встретитесь с ним?
19. Перечисление с однородными членами	1	Студенты сдали последние экзамены, наконец-то получили дипломы и отметили выпускной.
20. Обращение в постпозиции, слова автора	2	Все собрались, а, мама? "Я знаю, – он сказал, – где искать клад".
ВСЕГО	40	

**Приложение Ф. Образец анкеты по оценке адекватности и естественности
интонации синтезированной речи**

	Добавьте в текст необходимые по смыслу (после прослушивания фразы) знаки препинания (! , : ; ... “ – ? или их комбинации). Для разных голосов знаки могут меняться. В каждой строке может быть несколько предложений	Оцените естественность интонации (1 – естественно, 0 – неестественно)
1.	а без паспорта можно будет купить билет	
2.	а если прийти завтра есть вероятность успеть	
3.	а если работающие в офисе люди в следующем месяце целую неделю не будут выполнять свои обязанности	
4.	а может вы согласитесь	
5.	Аня тебе рассказала что вчера произошло	

...

Приложение X. Образец бланка оценки качества речи

Дата _____

Аудитор _____

№№ фраз	номер варианта фразы				
	1	2	3	4	5
1					
2					
3					
4					
5					
6					
7					

Аудитор проставляет в бланке свои оценки качества речи **по пятибалльной шкале с точностью до 0,1**, руководствуясь характеристиками речевого сигнала, приведенными в таблице ниже. **Фразы желательно прослушивать по одному разу (не более двух!)**

Одинаковые оценки всех вариантов одной фразы не допускаются.

Характеристика речи	баллы
Естественное звучание речи, отдельные малозаметные искажения помехового типа. Искажения типа дребезжание, хрип отсутствуют. Высокая узнаваемость	> 4,5
Некоторое нарушение естественности и узнаваемости, слабое присутствие одного вида искажений (картавость, гнусавость, дребезжание, хрип и др.)	3,6 – 4,5
Заметное нарушение естественности и ухудшение узнаваемости, присутствие нескольких видов искажений (картавость, гнусавость, дребезжание, хрип и др.)	2,6 – 3,5
Постоянное присутствие искажений типа картавость, гнусавость, дребезжание, хрип и др. Существенное искажение естественности и ухудшение узнаваемости	1,7 – 2,5
Сильные механические искажения типа картавость, гнусавость, дребезжание, хрип и др.; механический голос. Наблюдается потеря естественности и узнаваемости	< 1,7

Пояснения

Узнаваемость голоса диктора – величина, характеризующая степень сохранения субъективно воспринимаемых индивидуальных признаков голоса.

Селективные признаки – специфические субъективно воспринимаемые характеристики звучания речи.

- **картавость** – искажение сонорных «л», «р»;
- **плаксивость** – дрожание основного тона;
- **гнусавость** – звучание речи с излишней назализацией («произношение в нос»);
- **механический голос** – монотонность речи, потеря естественности звучания речи;
- **дребезжание, хрип** – неровный, резкий и прерывистый звук;
- **помеха** – шум и призвуки в речевых паузах.