МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

С.В. Рыбин

СИНТЕЗ РЕЧИ

Учебное пособие

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург

2014

Рыбин С. В. СИНТЕЗ РЕЧИ Учебное пособие по дисциплине "Синтез речи".
– СПб: Университет ИТМО, 2014. – 92 с.

В учебном пособии рассматриваются технологии синтеза интонационной речи. Синтез речи является одной из важнейших задач речевой обработки и имеет широкое применение в современных информационных технологиях. Материал пособия разбит на 6 разделов. Изложены история вопроса и основные этапы разработки систем автоматического синтеза. Пособие может быть использовано при подготовке магистров по направлению 230400.68 “ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ”, а также магистров по направлению 230100.68 “ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА” и аспирантов.

Рекомендовано Советом факультета Информационных технологий и программирования 25.02.2014 г., протокол № 2

УНИВЕРСИТЕТ ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Оглавление

ВВЕДЕНИЕ 5

1. СИСТЕМЫ СИНТЕЗА РЕЧИ: ИСТОРИЯ РАЗВИТИЯ, СОВРЕМЕННОЕ СОСТОЯНИЕ 6

1.1. Первые механические синтезаторы 6

1.2. Первые электрические синтезаторы 9

1.3. XX век: синтезаторы первого поколения 13

1.3.1. Артикуляционный синтез 13

1.3.2. Формантный синтез 14

1.3.3. Синтезаторы, использующие линейное предсказание 16

1.4. XX век: синтезаторы второго поколения 17

1.5. XX век: синтезаторы третьего поколения 18

1.5.1. Селективный синтез речи 18

1.5.2. Статистический параметрический синтез 19

1.6. Перспективные направления синтеза 19

2. ОБЗОР ТЕХНОЛОГИЙ TTS 20

2.1. Типы синтезаторов 20

2.1.1. Параметрический синтез 20

2.1.2. Компилятивный синтез 20

2.1.3. Синтез речи по фонетическим правилам 21

2.2. Оценка качества синтеза речи 21

2.3. Структура TTS 23

3. ЛИНГВИСТИЧЕСКИЙ ТЕКСТОВЫЙ ПРОЦЕССОР 26

3.1. Задачи лингвистического процессора 26

3.2. Нормализация текста (графематический анализ) 28

3.2.1. Выделение предложений, слов, символов, знаков препинания 29

3.2.2. Обработка пользовательской разметки 30

3.2.3. Расшифровка нестандартных записей 30

3.3. Использование словарей в синтезе речи 32

3.4. Обработка незнакомых слов 33

3.5. Снятие омонимии (омографии) 34

3.6. Методы разрешения неоднозначности при анализе текста 35

3.6.1. Синтаксический и морфологический анализ предложения 35

3.6.2. Статистические методы 35

4. ПРОСОДИЧЕСКИЙ ПРОЦЕССОР 36

4.1. Определение границ синтагм 36

4.1.1. Установка пауз по правилам 37
4.1.2. Установка пауз на основе статистических моделей 38

4.2. Определение интонационного контура 39
4.2.1. Генерация контура F0 методом ресинтеза 40
4.2.2. Формирование контура F0 для произвольного предложения 42
4.2.3. Генерация тонального контура в системах инженерного типа 43
4.2.4. Генерация тонального контура на основе лингвистических моделей интонации 45

4.3. Примеры интонационных контуров 47

5. ФОНЕТИЧЕСКИЙ ПРОЦЕССОР 48

5.1. Построение транскрипции 48

5.2. Вычисление физических параметров 50

6. АКУСТИЧЕСКИЙ ПРОЦЕССОР 53

6.1. Оптимальный выбор звуковых элементов методом Unit Selection 53
6.1.1. Стоимость замены 55
6.1.2. Стоимость связи 57
6.1.3. Поиск по алгоритму Витерби 58
6.1.4. Речевая база и качество синтеза для метода Unit Selection 58
6.1.5. Основные сложности и ограничения применения метода Unit Selection 60

6.2. Сглаживание энергетической огибающей 60

6.3. Модификация звуковых элементов 61
6.3.1.Алгоритм TD-PSOLA 61
6.3.2. Алгоритм SPECINT (Spectrum Interpolation) 63
6.3.3. Алгоритм LP-PSOLA 68
6.3.4. Экспериментальные сравнения 71

6.4. Объединение элементов в единый звуковой поток 74

6.5. Звуковые эффекты, используемые при синтезе речи 75
6.5.1. Параметрический эквалайзер 76
6.5.2. Ревербератор 78

7. СИНТЕЗ, ОСНОВАННЫЙ НА МОДЕЛЯХ 79

ЛИТЕРАТУРА 85

ВВЕДЕНИЕ

Автоматический синтез речи - это технология, позволяющая преобразовать входную текстовую информацию в звучащую речь. При этом одним из важнейших аспектов является качество синтезируемой речи. Именно от качества зависит пригодность использования технологии синтеза речи на современном коммерческом уровне. Под **системами автоматического синтеза речи** (иначе их еще называют **синтезаторами речи**) в настоящее время понимают системы, преобразующие орфографический текст и другую информацию в звучащую речь. Общепринятое в английской литературе обозначение – TTS (**Text To Speech**) System – системы преобразования текста в речь.

Технология автоматического синтеза речи может быть полезна в самых различных отраслях и направлениях, таких как:

- телекоммуникации,
- мобильные устройства,
- промышленные и бытовые электронные устройства,
- автомобильная индустрия,
- образовательные системы,
- компьютеризированные системы,
- Internet-сервисы,
- системы ограничения доступа,
- аэрокосмическая промышленность,
- военно-промышленный комплекс.

Синтезаторы речи обладают широкими возможностями применения. Например, в call-центрах и автоинформационных системах. Технология синтеза речи многого достигла в своем развитии. Синтезированную речь сегодня часто сложно отличить от естественной речи. Позвонив в информационную службу, мы уже слышим не роботизированную речь, а приятный естественный голос. Технология синтеза речи, интегрированная в автоинформационную систему, «охотно» вступит в беседу с каждым дозвонившимся и поможет в получении информации. На 90% запросов к любым информационно-справочным системам способен отвечать компьютер. Автоинформационная система с синтезом речи освобождает операторов от ответов на часто повторяющиеся вопросы такого плана как курс доллара, точное время, прогноз погоды и многое другое. Технология синтеза речи открывает широкие возможности для людей с физическими недостатками. Разработаны говорящие машины для слепых и слабовидящих. Для немых предусмотрены портативные устройства синтеза речи, в которых сообщение набирается на клавиатуре, что позволяет общаться с другими людьми.

На сегодняшний день благодаря электронным словарям и переводчикам на основе технологии синтеза речи возможно изучение иностранных языков с постановкой правильного произношения. Электронный словарь помещается в кармане и может быть использован в любом месте, а не только за рабочим

столом, как это обычно бывает с традиционным книжным словарем. Еще одним примером синтеза речи могут служить различные системы звукового оповещения: телефонная справочная информация, объявление станций в метро, информация об отправлении автобуса или поезда, реклама в универмаге.

На основе технологии синтеза речи созданы «говорящие» книги (аудиокниги). Такие книги позволяют по-новому воспринять литературное произведение – в его звуковом оформлении. Многие люди полагают, что напечатанный текст не передает всей полноты ощущений. В то время как элементарная разница в произношении или, например, интонации героев делает произведение более живым. Можно с уверенностью сказать, что системы синтеза речи в различных формах своего практического применения прочно вошли в нашу повседневную жизнь.

1. СИСТЕМЫ СИНТЕЗА РЕЧИ: ИСТОРИЯ РАЗВИТИЯ, СОВРЕМЕННОЕ СОСТОЯНИЕ

Синтез речи, то есть в широком смысле искусственное создание звучащей речи, подобной человеческому голосу, – задача, которая издавна интересовала людей (возможно, как часть идеи создания искусственного человека). Существуют легенды о «говорящих головах», умевших отвечать на вопросы, которые были созданы Гербертом Орильякским (946 – 1003), Альбертом Великим (1198 – 1280) и Роджером Бэконом (1214 – 1294). Но и 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, занимающихся созданием и развитием синтеза речи.

1.1. Первые механические синтезаторы

Первые синтезаторы, появившиеся во второй половине XVIII века, были механическими, они могли порож7дать отдельные звуки или небольшие фрагменты слитной человекоподобной речи подобно музыкальным инструментам, то есть требовали участия оператора-исполнителя. Очень важным является то, что уже в них посредством различных механических приспособлений воспроизводились основные процессы, происходящие при производстве речи человеком. Первые механические машины являлись скорее музыкальными инструментами, чем сложными техническими системами.

В 1779 году Петербургская Академия наук объявила ежегодную премию за объяснение разницы между пятью гласными звуками и за конструирование устройства, их порождающего. Немецкий учёный Христиан Готлиб Кратценштейн (1723 – 1795), работавший в то время в Петербурге, предложил лучшее решение. Он создал систему резонаторов (рис. 1.1), при помощи пульсирующего воздушного потока порождавших русские гласные.

Воздушный поток порождался вибрирующими язычками, подобными голосовым связкам человека [2].

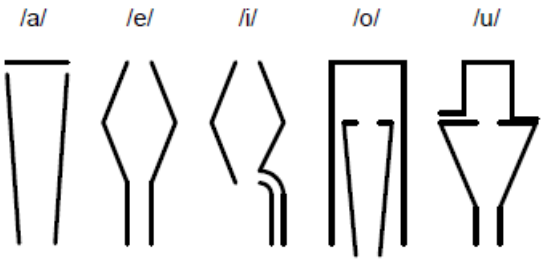

Рисунок 1.1. Система резонаторов Кратценштейна

Ещё ранее и независимо от Кратценштейна над механической системой синтеза речи стал работать и представил результат своих трудов в 1791 году австрийский изобретатель Вольфганг фон Кемпелен (1734 – 1804). Его машина могла произносить различные звуки и их комбинации. В ней моделировалось продвижение струи воздуха через голосовой тракт человека: имелись меха для подачи воздуха на язычок, который возбуждал резонатор, управляемый рукой. Согласные, в том числе и носовые, получались с помощью четырёх каналов, зажимаемых пальцами [2]. По утверждению самого Кемпелена, его машина производила 19 хорошо различимых согласных звуков и короткие фразы на нескольких языках. Для управления «говорящей машиной» требовался хорошо обученный оператор, порождение речи можно было сравнить с игрой на органе. Усовершенствованный вариант машины Кемпелена (рис. 1.2) был создан в 1837 году английским физиком Чарльзом Уитстоном (1802 – 1875). Также под впечатлением от машины Уитстона американский учёный и изобретатель Александр Грэм Бэлл (1847 – 1922) собрал собственную аналогичную модель.

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 голосового тракта. В своих работах 1828 года «О гласных звуках» и «О механизме гортани» Уиллис описал механизм извлечения гласных звуков по аналогии со звукоизвлечением органа.

В 1840 году Джозеф Фабер (1800 – 1850) представил свою говорящую машину под названием «Эйфония», которая по сообщениям современников могла производить обычную и шепотную речь, а также исполнять песни. Машина Фабера состояла из воздушного меха, который приводился в движение ножной педалью. Это были «как бы» легкие человека. Процесс производства звука был следующим: в вытесняемый из меха воздух направлялся в различные по объёму трубки, при помощи ряда клавиш. Каждая

из этих трубок отвечала за разные положения голосовой щели и полости рта. То есть Фабер хотел просто механически воспроизвести весь голосовой аппарат человека. Правда, ему это не удалось до конца осуществить. Голосовой тон в машине Фабера производился при помощи вибрации тонкой пластинки из слоновой кости на каучуковой подкладке. Звук получался очень резким и крикливым. Проще говоря, очень неприятным для человеческого уха. В источниках, описывающих работу машины Фабера указано, что звук «р» производится вибрацией не «языка», а твердой пластинки, помещенной за гортанью. А вот трубка, изображающая нос, находилась не сверху надставной трубы, изображающей рот, а ниже её и. Механизм, в целом, был очень тяжелым, и требовал серьезных усилий со стороны исполнителя, который пытался извлечь звуки речи.

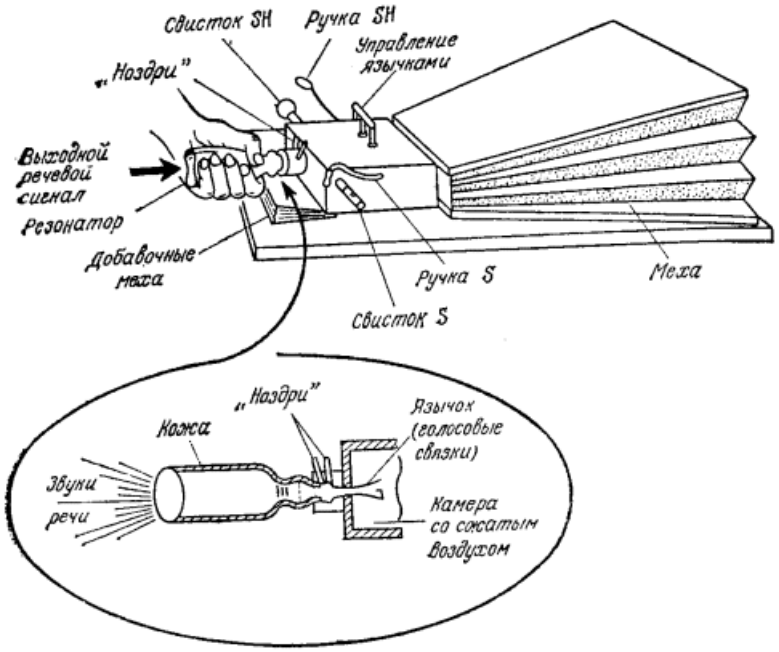

Рис. 1.2. Говорящая машина Кемпелена, построенная Уитстоном

В XX веке, несмотря на развитие электрических методов синтеза речи, разработка механических синтезаторов речи проводилась до 60-х годов. Это было связано, с одной стороны, с малой доступностью сложных электрических компонентов, 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 [2]. Среди наиболее известных устройств следует упомянуть механический синтезатор Р. Риша, продемонстрированный им в 1937 году (рис. 1.3). По форме он практически повторял голосовой тракт человека, был выполнен из резины и металла и управлялся клавишами, подобными клавишам трубы.

Таким образом, общим методом создания механических синтезаторов стала имитация или прямое моделирование голосового тракта человека.

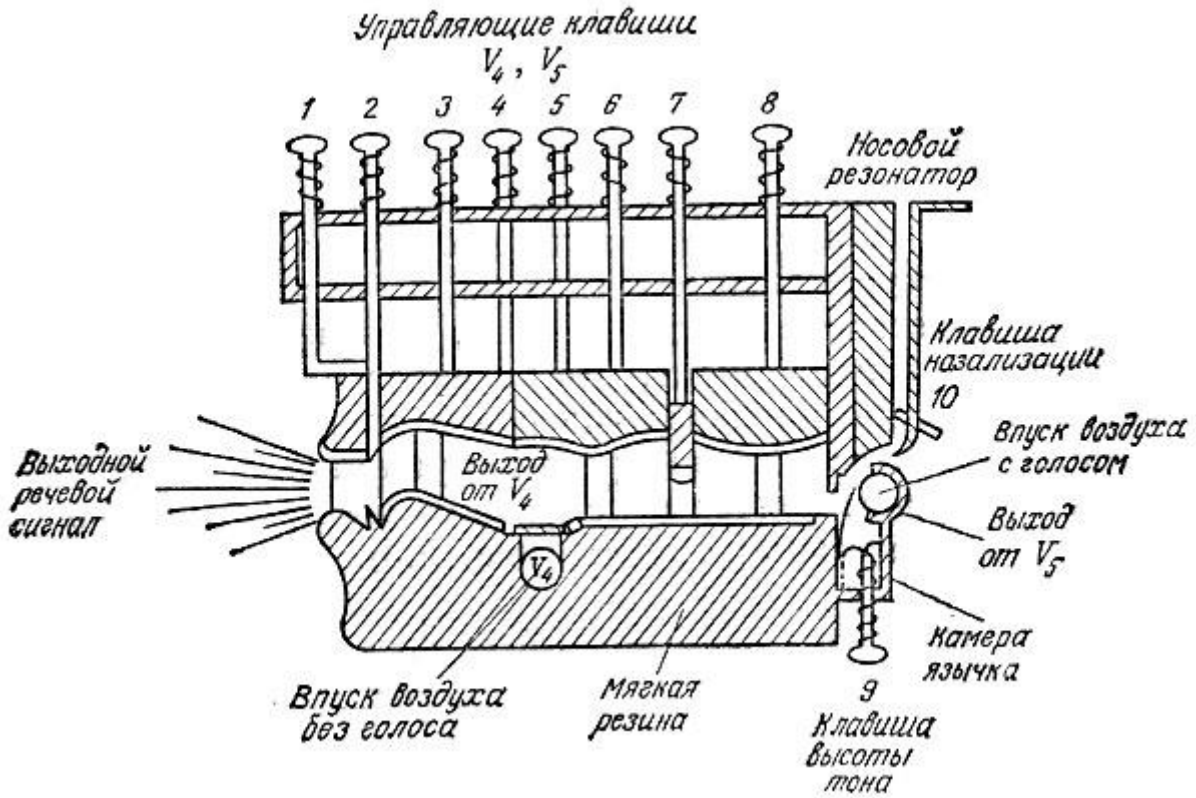

Рис. 1.3. Механическое говорящее устройство Риша

Основными рабочими компонентами таких моделей были: устройство для подачи воздуха (аналог лёгких), вибрирующая часть (аналог гортани) и система резонаторов, в большей или меньшей степени точно воссоздававших форму голосового тракта человека. Механические синтезаторы стали прототипом современного артикуляционного синтеза. Всё более новые и всё более сложные механические синтезаторы регулярно появлялись примерно до середины XX века.

1.2. Первые электрические синтезаторы

В XIX веке появление резонаторной теории Гельмгольца дало новый толчок в развитии речевых исследований. Речевой тракт человека рассматривался как последовательность резонаторов. Было установлено, что гласные звуки различаются резонансными частотами, названными впоследствии **формантами.** Вокальный тракт может быть рассмотрен как простая акустическая труба с открытым концом или резонатор. Форманты образуются при прохождении звуковой волны от звукового источника (голосовых складок) к губам. Звук частично отражается от губ говорящего и идет к слушателю, а частично отражается от губ и идет в обратном направлении к голосовым складкам.

Начались попытки построить синтезаторы речи – электрические аналоги речепроизводящей системы. Самый первый электрический синтезатор был создан Дж. Стюартом в 1922 году. Его схема (рис. 1.4) включала в себя электрический зуммер для моделирования голосовых связок и пару индуктивно-ёмкостных резонаторов для моделирования резонансов горла и

ротовой полости. Таким образом, генерировались первые две форманты, то есть устройство могло синтезировать только гласные звуки.

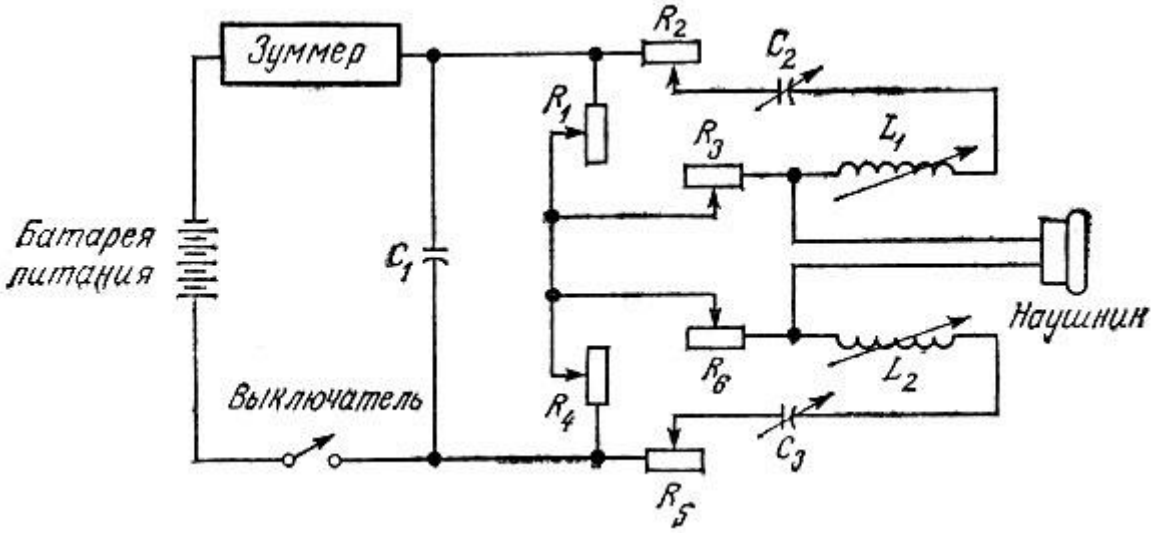

Рис. 1.4. Электрическая модель голосового тракта Стюарта

Аналогичный синтезатор, состоящий из четырёх подключенных параллельно резонаторов, возбуждаемых прерывателем тока, был создан немецким инженером Карлом Вилли Вагнером (1883 – 1953) в 1936 году [2].

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. Первым электрическим синтезатором, способным генерировать фрагменты связной речи, стал «водер» (Voder – Voice Operating Demonstrator), созданный американским инженером Гомером Дадли (1896 – 1987), Р. Ришем и С. Уоткинсом. Водер был основан на вокодере (название произошло от двух слов voice — голос, и coder — кодировщик), созданном в Bell Laboratories 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. 1. 5) [2].

Водер управлялся от ручной клавиатуры и синтезировал сигналы с заданным спектром. Десять параллельно соединенных полосовых фильтров составляли блок управления резонансами. Переключение источника возбуждения - шумового или импульсного генератора - осуществлялось браслетом на запястье оператора, а управление частотой импульсов - ножной педалью. На выходе фильтров стояли потенциометры, управлявшиеся десятью пальцами и изменявшие напряжение сигнала каждого фильтра. Для имитации взрывных согласных использовались еще три дополнительные клавиши. Обучение операторов "игре" на водере требовало значительного времени, но зато в итоге получалась довольно качественная речь с высоким уровнем разборчивости. Усовершенствованный вариант вокодера Дадли, VODER, был представлен на Нью-Йоркской Всемирной выставке 1939 года.

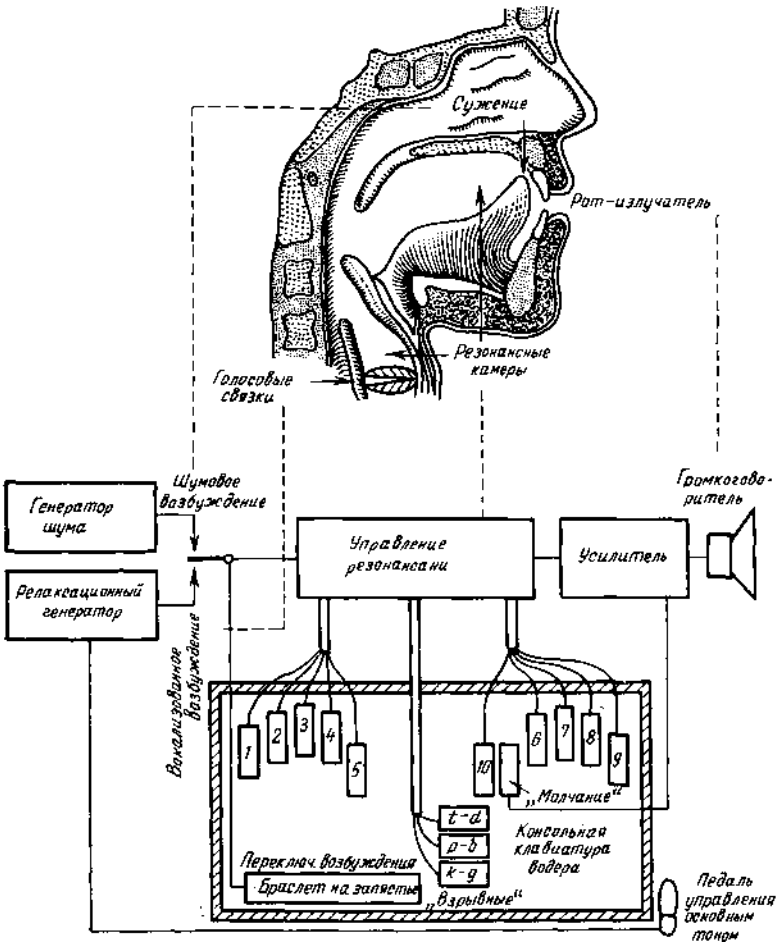

Рис. 1.5. Схема вокодера Дадли

В 1938 году советским ученым Е.А.Мурзиным был создан синтезатор звуков АНС. Евгений Мурзин, назвал свое изобретение в честь Александра Николаевича Скрябина.

В этом устройстве речь генерировалась с помощью рядов Фурье как сумма гармоник — элементарных спектральных составляющих, то есть чистых тонов. Банк тонов был записан на покрытый фотоэмульсией стеклянный диск, очень похожий на современный компакт-диск. Он был покрыт фотоэмульсией, и с помощью специального станка на него концентрическими кольцами были записаны 144 фотооптические звуковые дорожки "чистых тонов". Как происходил синтез звука показано на рис. 1.6.

Важным этапом в развитии методов экспериментальных фонетических исследований и синтеза речи стала разработка звукового спектрографа в 1946 году. 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 Дадли. Дополнительные дорожки на спектрограмме управляли переключением тона и шума и несли информацию о частоте основного тона (ЧОТ). Подобный метод использовался Дж. Борстом и Ф. Купером в устройстве, названным ими «водек» (1957 год) [2].

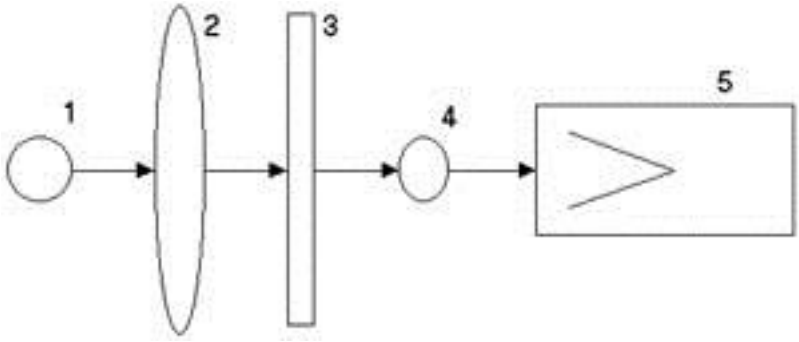

Рис. 1.6. Синтезатор Мурзина

Наиболее известный «проигрыватель» спектрограмм, синтезатор Pattern Playback (рис. 1.7), 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.

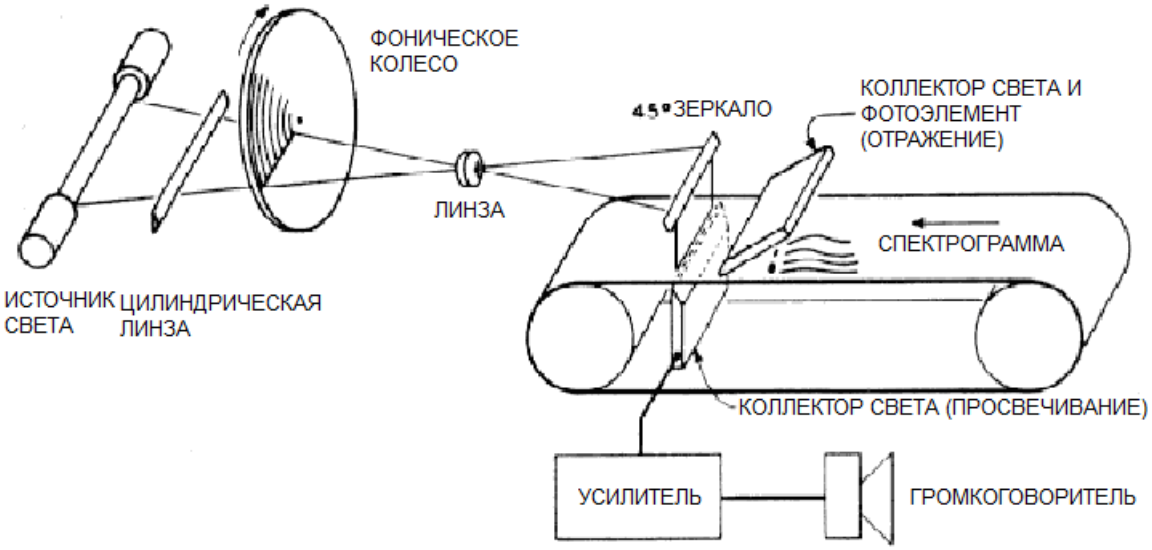

Рис. 1.7. Синтезатор Pattern Playback

При помощи этого синтезатора, позволявшего производить монотонную 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 напрямую не моделируется голосовой 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 заранее определённые частотные составляющие. Электрические синтезаторы стали прототипом современного компьютерного параметрического синтеза.

Следующей важной вехой в истории синтеза речи стало развитие акустической теории речеобразования (1960), создавшей необходимую теоретическую базу для создания основанных на ней формантных и артикуляционных синтезаторов, а также синтезаторов, использующих линейное предсказание. Эти три метода называют также технологиями синтеза первого поколения [3].

1.3. XX век: синтезаторы первого поколения

Синтезаторы первого поколения можно на основании используемых методов 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 задачи непосредственного отражения в синтезе процессов, связывающих артикуляцию с акустикой речевого сигнала, а вместо этого они выявляют и 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.

1.3.1. 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. В артикуляционной модели трубка, соответствующая голосовому тракту, обычно разделяется на множество небольших секций, и таким образом может быть представлена в качестве неоднородной электрической линии передачи [2].

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 к соответствующей точке линии белого шума [2].

Первый артикуляционный синтезатор с динамическим контролем (рис. 1.8) DAVO (Dynamic Analog of the VOcal tract) был разработан в 1958 году в Массачусетском технологическом институте Д. Розеном. Он управлялся записанными на ленту контролирующими сигналами, созданными вручную.

С течением времени артикуляционные синтезаторы развивались, в них вводилось дополнительное моделирование ослабления сигнала в голосовом 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 рода 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 [3].

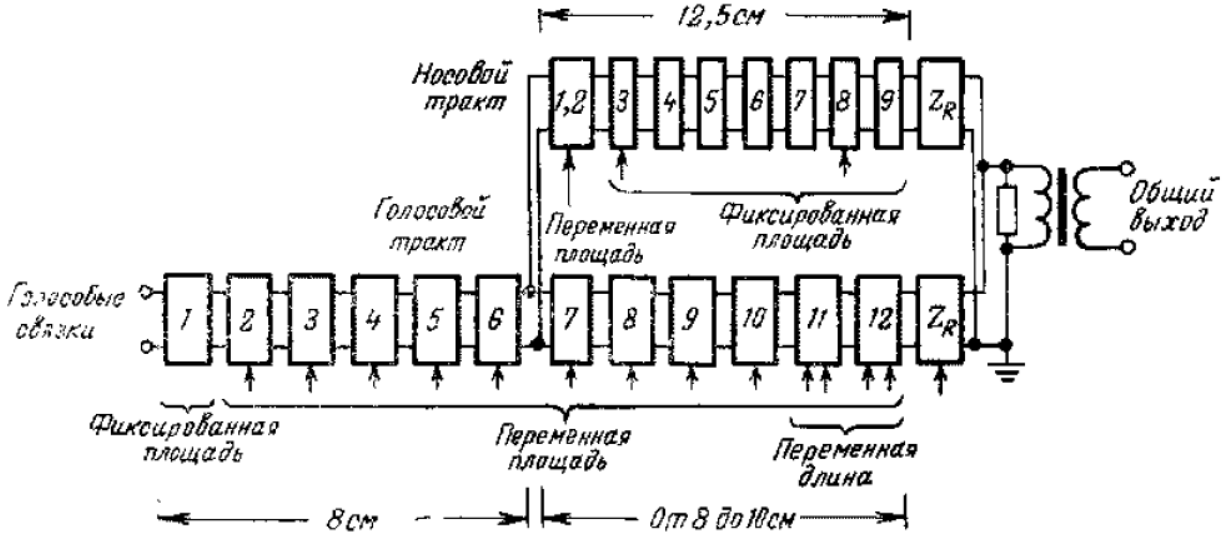

Рис.1.8 . Аналог голосового тракта с линией передачи, управляемый непрерывно

1.3.2. Формантный синтез

Первым формантным синтезатором стал PAT (Parametric Artificial Talker) английского исследователя Уолтера Лоуренса, представленный в 1953 году. Этот синтезатор состоял из трёх электронных формантных резонаторов, соединённых параллельно, на вход которым подавался шум или гармонический сигнал. Он управлялся шестью временными функциями (три форманты, частота основного тона, амплитуда шума и амплитуда голосового

источника), которые считывались с нарисованных на движущейся стеклянной дорожке шаблонов. Этот синтезатор был первым из параллельных формантных синтезаторов. Их главным преимуществом была относительная простота управления. Вторым типом формантных синтезаторов, позволяющим более точно моделировать передаточную функцию голосового тракта, но имеющих зачастую более сложную структуру, стали каскадные синтезаторы, в которых формантные резонаторы были соединены последовательно (рис. 1.9).

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 контролировались механической рукой, а амплитуда и частота основного тона определялись ручными потенциометрами.

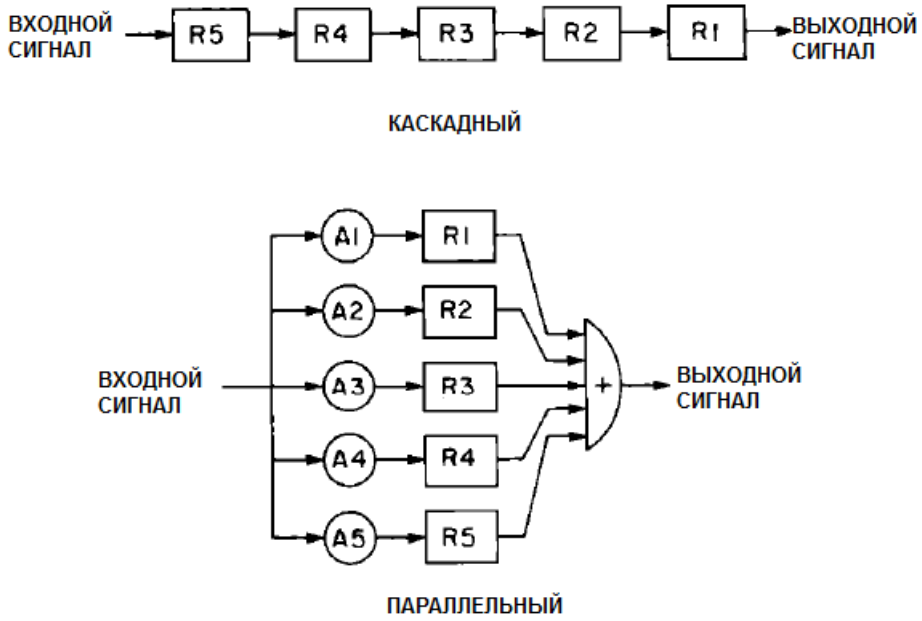

Рис.1.9. Каскадный и параллельный синтезаторы

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

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.

Первая модель формантного синтезатора русской речи «Фонемофон-1» (рис. 1.10) была разработана в Минске в начале 70-х годов, и успех в её создании был связан, прежде всего, с разработкой принципов формантного синтеза речевых сигналов. В последующих версиях удалось добиться синтеза русской речи по произвольному тексту весьма высокого качества [4]. В это же самое время начались исследования в области синтеза речи на кафедре фонетики филологического факультета ЛГУ. В конце 80-х годов на основе накопленного опыта стало возможным развитие проекта системы автоматического синтеза речи по произвольному русскому тексту. В 1989 году был изготовлен первый образец компилятивного синтеза.

Рис. 1.10. Внешний вид синтезатора «Фонемофон-1»

1.3.3. Синтезаторы, использующие линейное предсказание

Метод линейного предсказания позволяет напрямую использовать при синтезе искусственной речи параметры передаточной функции голосового тракта и является своеобразной альтернативой 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). Для синтеза речевого сигнала в КЛП-синтезаторе используются следующие изменяющиеся во времени параметры: период основного тона, средняя громкость звука, признак тон-шум и определённое заранее количество коэффициентов линейного предсказания. При этом качество синтезированной речи зависит от числа коэффициентов, точности их вычисления, а также от того, насколько хорошо моделируются источники возбуждения.

Обычно для работы КЛП-синтезатора из оцифрованной речи человека вычисляются необходимые параметры, а далее все необходимые единицы синтеза (слова или более короткие единицы) записываются в параметризованном виде в память и затем при синтезе извлекаются и соединяются, или конкатенируются, в определённом порядке. Таким образом, модель линейного предсказания косвенно поспособствовала развитию технологии конкатенативного синтеза. Синтезаторы первого поколения обычно требовали детального описания того, что должно быть произнесено, и не включали какого-либо автоматического способа получения подобного описания для произвольного сообщения или текста.

1.4. 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 широкая задача озвучивания любого сообщения, вводимого в компьютер в текстовом виде и неизвестного заранее системе синтеза. Это привело к развитию синтезаторов типа TTS. В идеале такие устройства должны имитировать деятельность человека, который читает письменное сообщение или текст любой степени сложности. Поэтому в синтезаторах такого типа (то есть синтезаторах речи в современном понимании этого термина) появился блок лингвистической обработки, независимый от акустического блока и метода генерации речевого сигнала, тогда как самые ранние синтезаторы и синтезаторы первого поколения были ориентированы в основном или полностью на модельную разработку акустического блока, то есть на задачу генерации речевого сигнала.

Первая полноценная система TTS для английского языка была создана в 1968 году в Японии Норико Умеда и его коллегами. Она была основана на артикуляционной модели акустического блока. Анализ текста и расстановка пауз производились при помощи сложных правил. По свидетельству специалистов, речь, производимая этой системой, была разборчивой, но довольно монотонной. 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 интонационных моделей (длительность, частоту основного тона и громкость).

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

1.5. XX век: синтезаторы третьего поколения

К третьему поколению технологий автоматического синтеза речи обычно относят синтез на основе скрытых Марковских моделей (HMM – hidden Markov models) и селективный синтез речи [2]. В англоязычных источниках метод называют **unit selection.** Их общей чертой является использование для автоматического синтеза речи больших объёмов речевых данных, а также высокая естественность синтезированной речи.

1.5.1. Селективный синтез речи

Метод Unit Selection является в настоящее время основной технологией автоматического синтеза речи, так как он позволяет получать синтезированную речь, которая по своим характеристикам наиболее приближена к естественной [2].

Метод является разновидностью конкатенативного синтеза речи, то есть в процессе синтеза речевого сигнала используются заранее сделанные звукозаписи естественной речи. В отличие от более ранних аллофонных[1] или дифонных[2] синтезаторов речи, порождающих итоговый речевой сигнал из отдельных и специально подготовленных звуковых единиц, выделенных из небольшого и тщательно подобранного набора слов, при селективном синтезе для каждой базовой единицы синтеза производится выбор наиболее подходящего кандидата из множества вариантов, взятых из озвученных предложений естественного языка. Для этого записываются специальные звуковые базы, размер которых может составлять до нескольких десятков часов звучащей речи. В процессе акустического синтеза алгоритм строит

[1] Аллофон — реализация фонемы, один из ее вариантов, обусловленный конкретным фонетическим окружением. В отличие от фонемы, является не абстрактным понятием, а конкретным речевым звуком.

[2] Дифон — сегмент речи между серединами соседних аллофонов.

оптимальную последовательность звуковых единиц, учитывая одновременно и то, насколько кандидат подходит под описание необходимых характеристик 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 модификации речевого сигнала, что повышает естественность синтезируемой речи.

Первыми системами селективного синтеза стали n-Talk (1992) и CHATR (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 [5].

1.5.2. Статистический параметрический синтез

Статистический параметрический синтез является методом, основанным не на правилах, а на имеющихся акустических данных. При этом подходе делается попытка машинного обучения системы на имеющихся речевых данных с целью получения модели соответствия характеристик речи, поступающих на вход акустического блока, физическим параметрам звуковых единиц. Получаемая модель даёт два преимущества: уменьшение памяти для хранения модели вместо самой речевой базы и возможность её параметрической модификации, например, быстрого изменения тембра голоса [2].

Наиболее распространённой техникой в данном направлении синтеза является метод, основанный на использовании скрытых Марковских моделей. Скрытые Марковские модели звуковых единиц применялись в системах распознавания речи с конца 70-х годов. Работу над автоматическими системами синтеза речи, основанными на HMM, начали в 1995 году японские учёные К. Токуда с коллегами [6]. Возможность использования статистического подхода в применении к синтезу речи обусловлена возросшим быстродействием вычислительных машин и объёмов носителей информации для хранения больших речевых баз, необходимых для обучения акустических моделей.

1.6. Перспективные направления синтеза

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 [2].

2. ОБЗОР ТЕХНОЛОГИЙ TTS

2.1. Типы синтезаторов

Речевые синтезаторы принято делить на два типа: с ограниченной и неограниченной словарной базой. В синтезаторах с ограниченным словарем речь хранится в виде отдельных слов или предложений, которые выводятся в определенной последовательности в процессе синтеза речевого сообщения. Речевая база в системах такого типа произносится диктором заранее, а затем преобразуется в цифровую форму с использованием различных методов кодирования, для уменьшения необходимого объема для ее хранения на носителе. В синтезаторах с неограниченным словарем элементами речи являются фонемы или слоги, поэтому в них применяется метод синтеза по фонетическим правилам, а не простая компоновка. Данный метод весьма перспективен, т.к. обеспечивает работу с любым необходимым словарем, однако качество речи значительно ниже, чем при использовании метода компоновки.

На сегодняшний день в сфере синтеза речи можно выделить три основные группы методов:

- параметрический синтез;
- компилятивный синтез;
- синтез речи по фонетическим правилам.

Каждый из подходов характеризуется наличием ряда достоинств и недостатков.

2.1.1. Параметрический синтез

Параметрический синтез речи является итоговой операцией в вокодерных системах, где речевой сигнал представлен набором непрерывно изменяющихся во времени параметров. Данный метод речевого синтеза целесообразно использовать в случаях, когда набор текстовых сообщений ограничен и редко подвержен изменению. К достоинствам данного метода относится возможность записать речь для любого языка и любого диктора. В зависимости от степени сжатия информации в параметрическом представлении качество синтезируемой речи может достигать очень высокого уровня. Недостатком такого подхода является невозможность применять параметрический синтез для заранее не заданных сообщений.

2.1.2. Компилятивный синтез

Компилятивный синтез сводится к составлению сообщения из предварительно записанного словаря исходных элементов синтеза. Очевидно, что содержание синтезируемых сообщений фиксируется объёмом словаря. Как правило, число единиц словаря не превышает нескольких сотен слов.

Основная проблема в компилятивном синтезе — объёмы памяти для хранения словарной базы. Для решения этой проблемы используются разнообразные методы сжатия/кодирования речевого сигнала. Компилятивный синтез имеет широкое практическое применение. За рубежом разнообразные устройства (от военных самолётов до бытовых устройств) оснащаются системами речевого ответа. Примером компилятивного синтеза речи являются объявления в транспорте: фразы «Осторожно, двери закрываются», «Следующая остановка:», «Остановка…» и названия остановок записаны диктором заранее и лишь соединяются вместе для оповещения по команде водителя или кондуктора.

2.1.3. Синтез речи по фонетическим правилам

В зависимости от размера исходных элементов здесь различают следующие виды синтеза:

- микросегментный;
- аллофонный;
- дифонный;
- полуслоговый;
- слоговый;
- синтез из различных единиц произвольного размера.

Часто в качестве таких элементов используются полуслоги — сегменты, содержащие половину согласного и половину примыкающего к нему гласного. При этом появляется возможность синтезировать речь по заранее не заданному тексту, но возникают проблемы управления интонационными характеристиками. Качество такого синтеза не совсем соответствует качеству естественной речи, поскольку на границах «сшивки» дифонов часто возникают искажения.

Однако и компиляция речи из заранее записанных словоформ также не решает проблемы высококачественного синтеза произвольных сообщений, поскольку акустические и просодические (длительность и интонация) характеристики слов изменяются в зависимости от типа фразы и места слова во фразе. Это положение не меняется даже при использовании больших объемов памяти для хранения словоформ.

2.2. Оценка качества синтеза речи

При разработке систем автоматического синтеза речи очень важным является вопрос оценки качества синтеза речи. В процессе оценки качества учитываются следующие основные характеристики:

- разборчивость речи;
- естественность (натуральность) речи;
- мультимодальность речи;

- многоязычие.

Рассмотрим указанные характеристики более подробно.

Основным критерием оценки качества синтеза речи является **разборчивость** синтезированной речи. Очевидно, что чем выше разборчивость речи, тем более высокого класса синтезатор. Существуют различные способы (типы) оценки разборчивости, основными из которых являются следующие:

- звуковая (не менее 75 %);
- слоговая (не менее 85 %);
- словесная (не менее 99 %);
- фразовая (98 – 99 %);
- смысловая.

Обычно разборчивость измеряется (оценивается) следующим образом. Речевая система синтезирует различные неожиданные для слушателя фразы. Указанные фразы фиксирует группа слушателей, каждый из которых пытается разобрать (распознать), что сказала машина. Например, предоставляется возможность прослушать 100 фраз, и слушатели поняли 98 – 99 % – это очень хорошая фразовая разборчивость.

При оценке **словесной разборчивости** слушателям предлагаются отдельные слова, самые разные, но осмысленные, никак между собой не связанные, из разных областей. Слушатели записывают слова, которые они поняли (расслышали). Затем подсчитывается количество понятых слов и рассчитывается процент относительно общего количества предложенных слов. При словесной разборчивости хорошим считается результат не менее 99 %.

В случае оценки **слоговой разборчивости** произносятся бессмысленные слоги (например, «псу», «ваз», «дус», «гры» и т.д.). При этом используются специальные таблицы частотной встречаемости слогов, с учетом которых формируются тестовые последовательности, которые подаются на вход речевой системы. Слушателям опять-таки следует записать все понятые ими слоги. Затем подсчитывается, сколько слогов услышано правильно: если примерно 85 % и выше, то разборчивость считается достаточной.

Звуковая разборчивость оценивается по тем же бессмысленным слогам, но считается не число неправильных восприятий (если хотя бы один звук в слоге был воспринят неправильно, то слог уже неправильный), а считается число звуков, т.е. фонем, воспринятых неправильно. Поскольку слоги состоят из различных звуков, то считается, что 75 % правильно воспринятых звуков – это уже неплохо.

При оценке разборчивости используются также градации, т.е. более градуированные оценки – какая разборчивость считается отличной, хорошей, удовлетворительной, неудовлетворительной.

Следующей характеристикой, используемой для оценки качества синтезатора речи, является **естественность** (натуральность) **речи**. Это субъективная характеристика, которая оценивается слушателями на основании их личного восприятия речи. Натуральность синтезированной речи зависит от многих факторов, например, могут быть натуральные звуки, но ритмическая сторона речи может быть сильно испорчена. Иначе говоря, слушатель может слышать понятную и разборчивую речь, произносимую вполне естественным голосом, но интонация при этом какая-то неестественная, «роботная». Это может также проявляться в том, что машина путает или “съедает” ударения.

Натуральность речи можно оценить, но нет объективных критериев – это только субъективное впечатление слушателя. Ведь даже разные люди имеют разное произношение, которое иногда может даже показаться неестественным. При реализации речевых синтезаторов интонация и ритмика речи определяются исходя из анализа входного предложения. Расстановка ударений в словах осуществляется в соответствии со словарём и с учётом синтаксических правил.

Под **мультимодальностью речи** понимают отражение эмоционального состояния говорящего, индивидуальность его голоса, стиль речи, акцент и т.п. В системах автоматического синтеза речи эта характеристика выражается в возможности синтеза различных типов голосов и их индивидуальных особенностей. Эта возможность относится к экстралингвистическим способностям системы, так как не связана с языковыми и собственно речевыми особенностями реализации. К мультимодальности речи, в частности, относят поддержку мужских и женских голосов, различные голосовые модуляции (бас, баритон). Сюда же относится возможность синтеза некоторых эмоциональных компонент, содержащихся в речи, таких, как волнение, гнев, ласка и т.п. Не всегда хорошо, когда синтезатор «говорит» безразлично и монотонно; во многих случаях полезно, когда он «говорит», моделируя эмоции человека.

В отличие от мультимодальности, **многоязычие** относится к лингвистическим способностям и подразумевает возможность синтеза речи на нескольких естественных языках. Например, на русском, английском и т.д.

2.3. Структура TTS

Система синтеза речи обычно состоит из четырех основных частей, будем называть их процессорами.

Лингвистический текстовый процессор. Он предназначен для решения следующих задач.

- **Выделение предложений** в тексте и разбиение их на **отдельные слова**; разметка текста на буквы, специальные символы, цифры и знаки пунктуации. Данный шаг необходим для успешности дальнейшей обработки текста.
- **Учёт разметки текста**, проставленной пользователем (пользовательские теги, пользовательский знак ударения). Пользовательская разметка имеет приоритет над обработкой текста по умолчанию.
- **Нормализация текста**. Текст, подаваемый на синтез, часто содержит большое количество обозначений, которые не могут быть прочитаны (т. е. транскрибированы) в исходном виде. Требуется их расшифровка. Эта процедура называется нормализацией текста.
- **Определение места ударения** и морфо-грамматических характеристик слов в предложении. Для определения места ударения в слове система синтеза речи по тексту использует морфограмматический словарь.
- **Снятие омонимии (омографии).** Снятие омонимии (омографии) представляет из себя выбор одной из нескольких словоформ, соответствующих тому или иному слову текста. Эти словоформы могут отличаться ударением (*з<u>а</u>мок* или *зам<u>о</u>к*), наличием буквы ё (*все* или *всё*), грамматическими характеристиками (*стали* - глагол или существительное). Выбор словоформы производится с помощью анализа контекста: лексического окружения слова, а также его грамматической позиции в предложении.

Просодический процессор. Просодическая обработка текста заключается в придании тексту интонационного оформления. Сюда относится деление текста на просодические единицы – синтагмы, определение длины пауз между синтагмами и выбор интонационного контура для каждой из синтагм.

Синтагма – основная единица реализации интонации. Характеризуется интонационной и смысловой целостностью, единым мелодическим и динамическим контуром, акцентно-ритмической структурой. Границы синтагмы могут маркироваться паузами; внутри синтагмы паузы недопустимы. В составе синтагмы выделяется главное слово, получающее т.н. **синтагматическое ударение**, в то время как остальные словесные ударения могут существенно ослабляться – определяет степень централизации синтагмы (очень высокая для русского языка).

Деление предложения на синтагмы осуществляется в первую очередь с опорой на знаки препинания. В большинстве случаев наличие знаков препинания является надежным сигналом о наличии паузы. В то же время некоторые отдельные случаи, такие как вводные слова (возможно, к сожалению, и т.п.), обрабатываются по особым правилам, поскольку выделение их запятыми не обязательно обозначает возможность паузы при чтении. Длинный отрезок предложения, не разделенный знаками препинания,

делится на синтагмы по особому алгоритму, включающему в себя анализ синтаксических связей между словами. Деление на синтагмы сопровождается также выбором места фразового ударения, то есть основного ударения в синтагме. В большинстве случаев в русском языке фразовое ударение падает на последний ударный слог синтагмы, например, *студент читает книгу*. Однако в некоторых случаях фразовое ударение может переноситься на другой слог, например, когда последним словом в синтагме является местоимение: «Вы можете прочесть ее».

Для каждой синтагмы, выделенной в процессе анализа текста, выбирается наиболее подходящий интонационный контур (ИК). Набор интонационных контуров, используемый в системе синтеза речи, основан на стандартной классификации Е.А.Брызгуновой [1] и включает в себя такие интонационные типы, как повествовательное предложение, общий вопрос, частный вопрос, восклицание и т.п. Выбор ИК осуществляется на основе знаков препинания (вопросительный знак, восклицательный знак, запятая, тире и т.п.), а также лексического содержания предложения (например, наличия вопросительных слов).

Фонетический процессор. Задачей фонетического процессора является:

- построение транскрипции по правилам и учет исключений транскрипции;
- вычисление физических параметров интонации для синтагм синтезируемого текста.

Акустический процессор. Основная часть акустической обработки – оптимальный выбор звуковых элементов из базы диктора, осуществляющийся по методу Unit Selection. Далее может быть произведена модификация звуковых элементов по частоте основного тона, длительности и тембру, а также добавлены звуковые эффекты (например, с помощью инструментов реверберatора и эквалайзера).

Структурно, схему взаимодействия процессоров можно представить следующим образом (рис.2.1).

В качестве формата передачи данных между процессорами в данной схеме предполагается XML-формат данных. XML — текстовый формат, предназначенный для хранения структурированных данных (взамен существующих файлов баз данных), для обмена информацией между программами, а также для создания на его основе более специализированных языков разметки (например, XHTML), иногда называемых словарями. XML — это иерархическая структура, предназначенная для хранения любых данных, визуально структура может быть представлена как дерево. Важнейшее обязательное синтаксическое требование — то, что документ имеет только один корневой элемент.

3. ЛИНГВИСТИЧЕСКИЙ ТЕКСТОВЫЙ ПРОЦЕССОР

3.1. Задачи лингвистического процессора

Лингвистический текстовый процессор производит предварительную обработку текста, необходимую для построения его транскрипции и дальнейшей генерации звучащей речи. Так, в потоке текста должны быть выделены элементы, с которым система синтеза речи должна будет работать в дальнейшем: абзацы, предложения, слова, буквы и другие символы. При этом должна быть учтена пользовательская разметка текста (при ее наличии), выполненная, например, с помощью специализированных тегов **SSML** (Speech Synthesis Markup Language). Кроме того, орфографический текст, который должна обработать система синтеза речи, сам по себе содержит недостаточно информации для создания правильной **транскрипции** (условной записи, по которой непосредственно будет формироваться звучание текста).

Не все слова орфографического текста могут быть «прочитаны» синтезатором в том виде, в котором они представлены изначально. В естественных текстах часто встречаются следующие виды нестандартных записей: цифры; включения других алфавитов (для русского языка – в первую очередь, латиница); специальные знаки, не являющиеся ни буквами, ни цифрами; сокращения и аббревиатуры (акронимы); и др. Все эти записи должны быть превращены в «обычные», стандартные слова языка, на котором производится синтез, или в подобные им записи (например, транслитерацию иного алфавита).

Далее производится подготовка слов текста к транскрибированию. Для русского языка транскрипция в общем случае достаточно регулярна и осуществляется по заданным правилам на отдельном этапе анализа текста, однако на этапе лингвистического анализа могут быть определены места ударных гласных, а также наличие буквы *ё*, которая обычно передается на письме как *е*. Для языков с нерегулярной орфографией, таких как английский, список транскрипций слов может быть задан в списке (словаре). Однако и в том и в другом случае основной проблемой являются слова, не найденные в словаре, а также случаи, когда одному написанию соответствуют два или более разных слов (**омонимы**, например, «замОк-зАмок»).

Слова, различающиеся звучанием, но пишущиеся одинаково (замОк-зАмок), называются омографами; слова, пишущиеся и звучащие одинаково, но имеющие разное значение и/или разные грамматические характеристики, называются омонимами (например, печь – существительное и печь – глагол). Для синтеза речи наиболее важно снятие омографии, однако и различение омонимов может быть важным для грамматического анализа текста. Часто омонимы и омографы объединяют под общим наименованием омонимов.

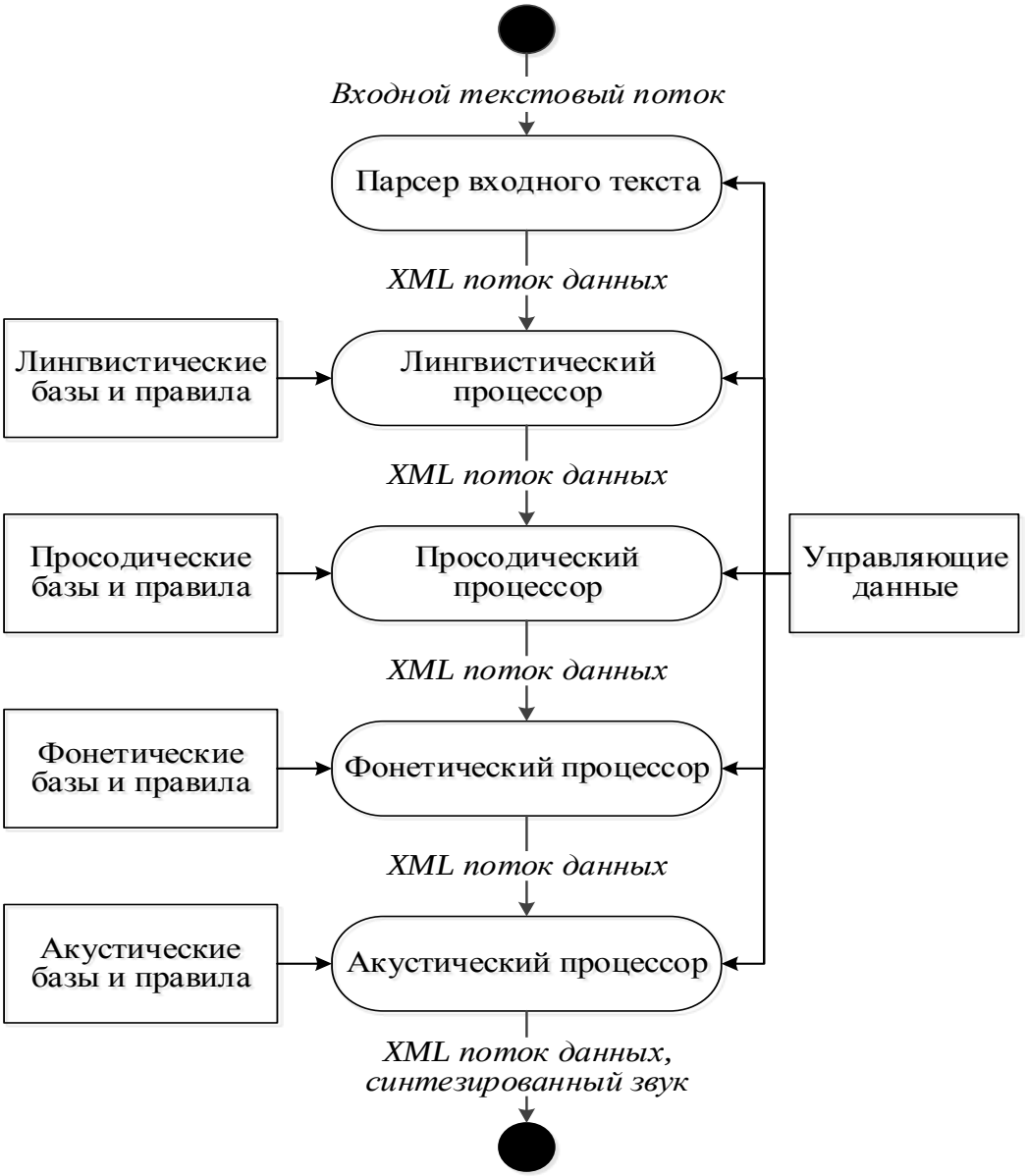

Рис. 2.1. Схема синтезатора речи

Для выбора между омонимичными словоформами (**снятие омонимии**) применяется как анализ лексического контекста, так и анализ грамматической структуры предложения. Для определения грамматических характеристик слов может использоваться морфограмматический словарь либо статистические методы определения части речи и других грамматических категорий слова (**Part-of-speech tagging**). Общая схема работы процессора представлена на рис. 3.1.

Итак, в задачи лингвистического процессора входит:

- первичная обработка текстовых данных и представление их в едином формате для последующей обработки;
- выделение предложений в тексте и разбиение их на отдельные слова;
- разметка текста на буквы, специальные символы, цифры и знаки пунктуации;

- учет разметки текста, проставленной пользователем;
- расшифровка сокращений, аббревиатур, числительных и других нестандартных записей, поиск и исправление орфографических ошибок и опечаток;
- определение места ударения и морфо-грамматических характеристик слов в предложении;
- снятие омонимии (омографии).

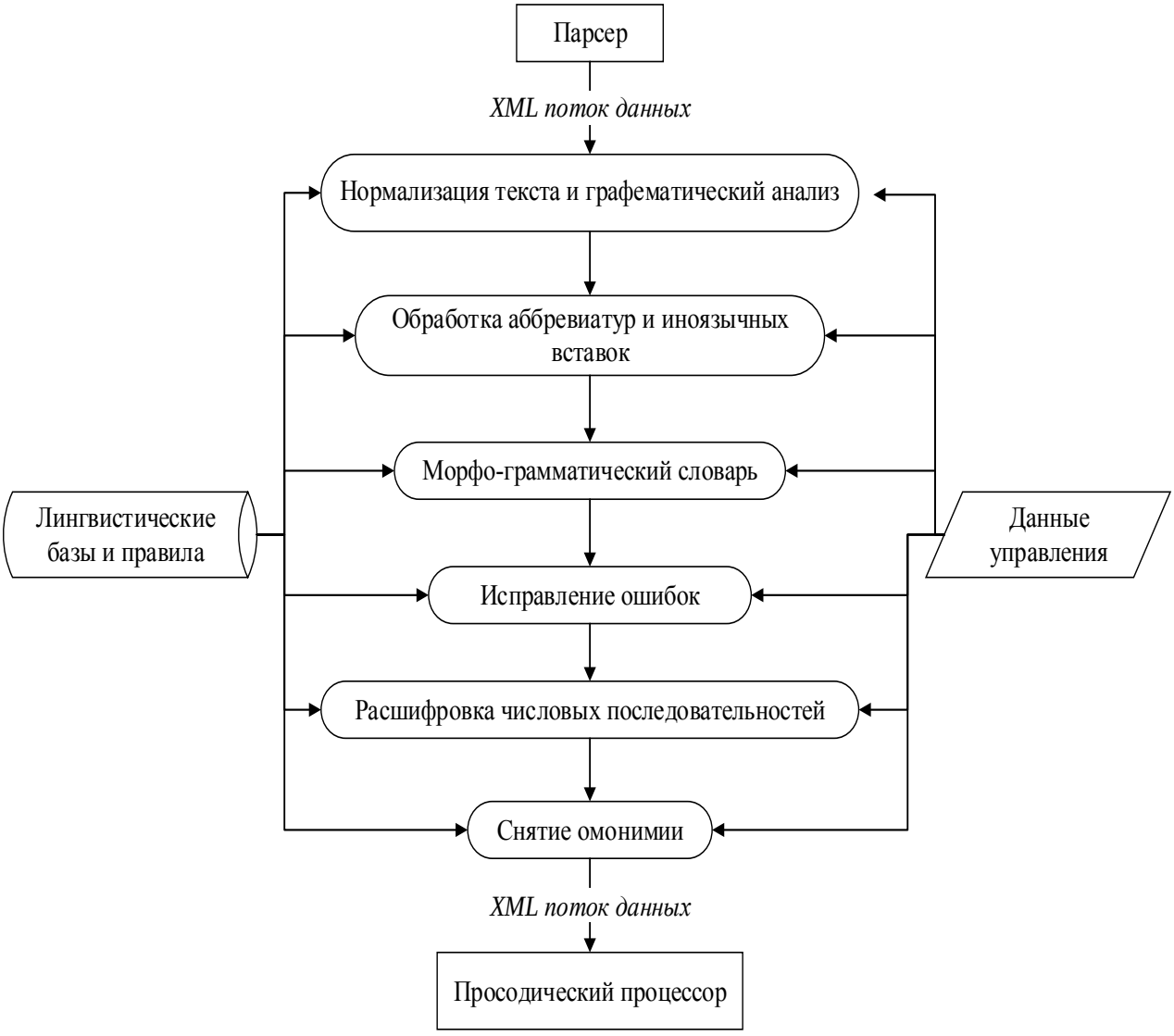

Рис. 3.1. Схема работы лингвистического процессора

3.2. Нормализация текста (графематический анализ)

Задачи, входящие в нормализацию текста:

- выделять предложения в тексте;
- разбивать предложения на слова;
- разбивать текст на буквы, цифры, знаки пунктуации, специальные символы;

- учитывать информацию, проставленную пользователем (место ударения, наличие паузы и др.);
- расшифровывать специальные символы (%, *, №...), сокращения («т.е.», «км/ч» и т.п.) и др.;
- выделять специальные форматы записей (дата, время, интернет-адрес...).

Рассмотрим эти задачи более подробно.

3.2.1. Выделение предложений, слов, символов, знаков препинания

Для правильного синтеза речи необходимым подготовительным этапом является деление текста на отрезки, которые будут впоследствии основными единицами синтеза. Это, прежде всего, слова и предложения. В самом первом приближении словами можно считать отрезки текста между пробелами, а предложениями – отрезки текста между точками и другими конечными знаками препинания. Однако для анализа реальных текстов такого подхода недостаточно. Рассмотрим примеры текста, поступающего на вход программы синтеза речи:

Порядка 22% от объема полученных в 2011 году средств – 172 тыс. долларов – были переданы 39 благотворительным организациям.

Write a cheque from acc 3949293 (code 84-15-56), for $114.34, sign it and take it down to 1134 St Andrews Dr, or else!!!! (пример взят из [21]).

Анализируя эти примеры, можно сделать следующие наблюдения:

- Знаки препинания, как правило, не отделены пробелами от слов и должны быть проанализированы специальным образом. Далее знаки препинания могут быть сохранены как отдельные символы либо сохраняться в виде флагов соответствующего слова. Следует учесть, что после слова может следовать несколько знаков препинания. Корректная информация о знаках препинания крайне важна для синтеза речи, т.к. свидетельствует о смысловом и интонационном делении текста.

- Некоторые другие выражения, объединенные в одно графическое слово, также должны быть отделены друг от друга для индивидуальной обработки (например, сочетания цифра+процент – «22%»).

- Точки не всегда сигнализируют конец предложения; в русском языке точка часто является элементом сокращения – например, «тыс.». Необходим анализ элемента, который оканчивается точкой (сокращение это или нет) и анализ дальнейшего контекста (следует ли далее слово с большой буквы). Даже при учете этих факторов могут возникнуть сложные случаи, которые

сложно решить без семантического анализа предложения (ср.: *В 1868 г. Лев Толстой закончил «Войну и мир»*). Точка также может быть элементом интернет-ссылки или адреса электронной почты (mail.ru), инициалов (А. С. Пушкин), а также дроби (1.5), даты (19.11.2012) и т.п.

- Изредка встречаются случаи, когда порядок следования элементов в тексте должен быть изменен для правильного синтеза (ср.: *$114.34*).

3.2.2. Обработка пользовательской разметки

В большинстве коммерческих приложений для синтеза речи пользователь может влиять на результат синтеза с помощью специальной разметки. Общепринятым форматом пользовательской разметки для синтеза речи является, как уже отмечалось ранее, SSML, основанный на XML.

Язык SSML был создан рабочей группой консорциума по голосовым браузерам и позволяет программно контролировать самые разные параметры процесса синтеза речи.

SSML основан на языке разметки Java Synthesis Markup Language (JSML), разработанном Sun Microsystems. Он охватывает практически все аспекты синтеза речи, хотя в некоторых областях остались неопределенные аспекты, поэтому каждый производитель принимает иной вариант языка. Кроме того, в отсутствие разметки, синтезатор, как ожидается, должен выполнить свое собственное толкование этого текста. Так SSML не является таким жестким в плане синтаксиса, как язык C, или даже HTML.

С помощью тегов SSML могут быть определены такие параметры синтеза речи, как выбор языка, голоса, темпа, тембра, мест пауз, произношения слова и т.п. Анализ данных тегов производится на этапе разбора текста.

3.2.3. Расшифровка нестандартных записей

Можно выделить следующие типы нестандартных записей в тексте:

1) Сокращения.

- Сокращения – элементы, которые должны быть заменены на полное слово или словосочетание.

- Сокращение может быть из одного слова (км, стр.) или из нескольких (т.е., к. ф. н., до Р.Х.).

- Сокращение может использоваться без точки (Гб, мск) либо содержать точку (одну или после нескольких частей, см. примеры выше), слеш (км/ч, в/ч, ж/д), дефис (кто-л.). При расшифровке сокращения эти знаки должны быть исключены из дальнейшей обработки предложения.

При расшифровке сокращений перед разработчиком системы синтеза речи встают следующие трудности. Во-первых, сокращения зачастую неоднозначны: например, английское «m» может обозначать “million” (миллион) или “metre” (метр), русское «г.» - «город» или «год». Во-вторых, в языках с богатой морфологией, таких как русский, приходится производить выбор правильной падежной формы слова (например, 1 км=километр, 2 км=километра, 5 км=километров, к 5 км=километрам и т.п.).

2) Специальные знаки – знаки, не являющиеся ни буквами, ни цифрами, но нуждающиеся в словесной расшифровке ($, *, #, % и т.п.). Во многом такие знаки ведут себя подобно сокращениям и тоже нуждаются в выборе правильной формы в зависимости от контекста (например, 1%, 2%, 5%...).

3) Цифровые записи – это цифры, встречающиеся в естественных текстах, которые должны быть преобразованы в соответствующие слова (числительные).

- Цифры бывают арабские и римские. Римские цифры могут быть преобразованы в арабские в соответствии с заданным алгоритмом.

- В текстах могут встретиться стандартные последовательности цифр или слов и цифр, требующие чтения по определенному формату. Самые распространенные из таких форматов: дата, время, номер телефона, индекс. Они должны быть выявлены в процессе нормализации текста и «прочитаны» по заданным правилам (например, в дате типа 19.11.2012 цифра 11 заменяется названием месяца – «ноября», добавляется слово «года» и т.п.).

- Цифры, не подпадающие под заданные форматы, могут обозначать количество обозначаемых предметов (количественные числительные, например, «десять») или номер следования предмета (порядковые числительные, например, «десятый»). При расшифровке числительных производится выбор между этими разрядами (например, «10 человек» или «10 этаж»). При расшифровке числительного, записанного в виде цифры с аффиксом (например, «10-ый», «20-ти», в английском: 2nd, 10th), аффикс должен быть отброшен, а числительному придана соответствующая форма.

- В русском языке расшифровка цифровых записей также сопряжена с выбором правильной формы (падежа и, для некоторых числительных, рода, например, «2 кота», «2 кошки», «2 кошкам» и т.п.).

4) Для русского языка: латиница. Хотя система синтеза русской речи прежде всего ориентирована на озвучивание текста, написанного кириллицей, однако в реальных текстах часто встречаются элементы, написанные латинским алфавитом. Это могут быть иностранные имена или названия зарубежных фирм и товаров, адреса интернет-сайтов и электронной почты и

т.п. Такие элементы должны быть транслитерированы (**транслитерация** – передача слова, написанного иноязычным алфавитом, с помощью русских букв). Правила транслитерации задаются с учетом того, слова какого языка чаще всего могут встретиться в синтезируемых текстах (чаще всего – английского).

5) Аббревиатуры и инициалы. Аббревиатуры (акронимы) – слова, образованные путем сложения начальных букв или слогов исходного словосочетания, которые читаются по буквам либо как единое слово (например, *АТС*, *ООО*, *ВАЗ*) (в отличие от сокращений, которые при чтении расшифровываются до исходного слова или словосочетания, например, *км* – километр, *и т.п.* – и тому подобное). Выбор способа чтения аббревиатуры задается по правилам, следует также определить, куда будет падать ударение в полученном слове. Инициалы тоже должны быть прочитаны как отдельные буквы (при этом точки, на которые заканчиваются инициалы, удаляются из дальнейшей обработки предложения).

3.3. Использование словарей в синтезе речи

Машиночитаемые электронные словари могут выполнять различную роль в системе синтеза речи. Так, в английском языке, имеющем бедную морфологию, но крайне нерегулярную орфографию, словарь, как правило, содержит список словоформ с соответствующими им транскрипциями (в словаре может также содержаться информация о части речи слова, однако, поскольку большинство английских словоформ может относиться к различным частям речи, то определение частей речи обычно выделяется в отдельный модуль программы). Ниже приводится фрагмент популярного словаря английского языка CMUDICT; каждая строка содержит в левой части словоформу, в правой части – ее транскрипцию в принятой в словаре нотации:

```
SPEECH  S P IY1 CH
SPEECHES  S P IY1 CH AH0 Z
SPEECHIFY  S P IY1 CH AH0 F AY2
SPEECHIFYING  S P IY1 CH AH0 F AY2 IH0 NG
SPEECHLESS  S P IY1 CH L AH0 S
```

Для русского языка не является целесообразным хранить в словаре всю информацию о транскрипции каждой формы, поскольку, во-первых, транскрипция с достаточной степенью достоверности выводится из орфографического облика слова, во-вторых, количество отдельных словоформ на то же количество лексем в русском языке значительно больше благодаря развитой морфологии (так, у английского глагола в общем случае четыре формы, а у русского их несколько десятков, если учитывать причастия). В морфограмматическом словаре русского языка (например, словарь группы АОТ www.aot.ru) слова хранятся в форме основ и соответствующих им словоизменительных парадигм (то есть целиком каждая форма не хранится). При этом для каждой словоформы доступна следующая информация:

- место ударения в данной словоформе;
- наличие буквы *ё* (*ё* должно определяться, даже если слово написано через *е*);
- грамматические характеристики словоформы (например, род, число и падеж для существительного, лицо, число и время для глагола и т.п.);
- некоторые сведения о семантике (например, является ли слово фамилией, географическим названием и т.п.).

Ниже приводится фрагмент машиночитаемого словаря русского языка. Каждая строка словаря содержит основу и цифровой или буквенный код, являющийся ссылкой на определенную словоизменительную и акцентную парадигму, то есть на набор окончаний, прибавляемых к данной основе, и место ударения для каждой полученной словоформы (сами парадигмы хранятся в отдельной части словаря).

```
синтез 67 29 Фа
синтезатор 67 29 Фа
синтезаторн 106 121
синтезир 100 108 Уп
```

При обработке текста в системе синтеза речи на вход модуля словаря поступает нормализованный текст. Каждое слово этого текста ищется в словаре, при нахождении ей присваиваются соответствующие характеристики (номер ударного гласного, замена *е* на *ё*, грамматические категории и т.п.) Если в словаре данная словоформа встречается более одного раза (например, *дом* – им.пад. ед.ч. и вин.пад. ед.ч.; *берег* – сущ., произносится с *е* и *берёг* – глагол, произносится с *ё*), учитываются все словарные вхождения.

3.4. Обработка незнакомых слов

Каким бы объемным ни был словарь, использующийся в системе синтеза речи, тем не менее в текстах, как правило, будут встречаться слова, которые не будут находиться в словаре. Однако для таких слов также необходимо определить место ударения, построить транскрипцию и – в идеале – определить грамматические характеристики. Для обработки незнакомых слов в системе синтеза речи могут использоваться следующие методы:

1) Морфологический анализ и правила. Зачастую слова, не найденные в словаре, могут быть правильно интерпретированы с помощью **морфологического** анализа (анализ слова на составляющие его элементы – основу и аффиксы), который позволяет связать незнакомое слово с уже известным словом, содержащимся в словаре. Так, например, если в словаре нет слова *супервыставка*, мы можем выделить в нем префикс *супер-* и основу *выставка*, которая в словаре есть. Если морфологический анализ не позволяет разложить слово на знакомые элементы, то место ударения или транскрипция

может быть выведено с помощью правил (например, поиска внутри слова элементов, сигнализирующих о месте ударения, и т.п.).

2) Статистические методы. При наличии достаточно большого словаря транскрипций можно обучить статистическую модель, выводящую транскрипцию слова или место ударения автоматически. Этот способ удобен тем, что не требует участия эксперта для формирования правил, однако может давать сбой на словах, имеющих редкий для обрабатываемого языка орфографический облик (например, иностранные имена и названия, часто встречающиеся среди незнакомых слов).

Незнакомые слова, попадающиеся в текстах, могут быть не только обозначениями каких-то новых предметов и понятий, но и словарными словами, написанными с ошибкой или опечаткой (например, *интелект вместо «интеллект», *карова вместо «корова»). Ошибки в подающихся на синтез словах могут быть исправлены как с помощью списков частотных ошибок и опечаток (либо просто добавления распространенных ошибочных вариантов, типа *исскуство вместо «искусство», в общий словарь), так и с помощью правил, проверяющих наличие распространенных ошибок (например, для русского языка – написание двойных согласных, безударных гласных и т.п.).

3.5. Снятие омонимии (омографии)

Снятие омонимии – одна из наиболее важных и сложных задач для автоматического синтеза речи. Словоформы, имеющие одинаковое написание, но разное прочтение, встречаются во многих языках. Однако для русского языка эта проблема особенно важна, поскольку количество омонимов очень велико. Омонимы в русском языке могут различаться ударением (например, *ст<u>о</u>ит-сто<u>и</u>т*), а также наличием буквы *ё*, которая в современной орфографии чаще всего передается как *е* (*все-всё*), либо и тем и другим (*б<u>е</u>рег-бер<u>ё</u>г*). Омонимичные словоформы могут иметь одинаковые грамматические признаки (например, *з<u>а</u>мок-зам<u>о</u>к, з<u>а</u>мка-замк<u>а</u>*…), либо различаться грамматическими характеристиками. В последнем случае омонимичными могут быть как различные словоформы внутри одной парадигмы (например, род.п. ед.ч. – им.п. мн.ч.: *облак<u>а</u>-<u>о</u>блака, стран<u>ы</u>-стр<u>а</u>ны*…), так и формы разных парадигм (например, существительное-инфинитив: *вест<u>и</u>-в<u>е</u>сти, проп<u>а</u>сть-пр<u>о</u>пасть*…).

С формальной точки зрения, мы считаем, что слово в тексте представляет собой омоним, если одному слову соответствует несколько словарных вхождений (словоформ), в том числе, когда варианты произношения этого слова различаются ударением или/и наличием буквы *ё*. В задачи модуля снятия омонимии входит выбор правильного вхождения для данного контекста.

Разрешение омонимии, как и расшифровка специальных обозначений, производится при помощи анализа контекста [10,11]. На уровне индивидуальных слов-омонимов производится поиск в предложении

ключевых слов или выражений. Этот этап включает анализ слов непосредственно рядом с текущим, как, например, в случае устойчивых выражений: *скрыто за семью замк<u>а</u>ми, в четырех стен<u>а</u>х*. Также анализируется состав предложения целиком, например, *дверь была заперта на необычный зам<u>о</u>к* (ключевое слово *заперта*).

На уровне классов словоформ производится анализ грамматического окружения, то есть поиск согласованных слов в предложении. Для формализации этого принципа в [10,11] предлагается ввести грамматические правила, увеличивающие условный «вес» словоформы в зависимости от ее окружения. Правила хранятся в формализованном виде, позволяющем быстро оценивать и корректировать работу системы.

3.6. Методы разрешения неоднозначности при анализе текста

Выбор правильной формы слова при расшифровке сокращений, числительных и других нестандартных элементов, а также при снятии омонимии сводится к задаче разрешения неоднозначности текста (определенный элемент может быть интерпретирован тем или иным образом, и программа должна выбрать один из возможных вариантов). Эта задача может быть решена двумя различными способами.

3.6.1. Синтаксический и морфологический анализ предложения

Сюда может относиться как полный анализ структуры предложения (**<u>парсинг</u>**), так и анализ окружения конкретного слова, задаваемый в виде контекстных правил. Например, выбор формы числительного может зависеть от наличия предлога слева (*до 10 раз, в 10 раз*), формы согласованного слова справа (*1 полосатая кошка*, *1 полосатый кот*) и т.п. Выбор формы омонима может осуществляться разными способами. В случае с омонимами, одинаковыми по грамматическим характеристикам, разрешение омонимии может осуществляться только с помощью анализа лексического содержания предложения (ключевые слова, устойчивые выражения и т.п.). Если же грамматические характеристики различаются, то можно использовать и анализ грамматического окружения слова для выбора омонима, подходящего к синтаксическому контексту. Усложняет проблему то, что омонимичные словоформы могут существенно различаться по частотности (например, *<u>у</u>х<u>а</u>-<u>у</u>ха*, *сорок<u>а</u>-сор<u>о</u>ка*, *кред<u>и</u>т-кр<u>е</u>дит*, *мо<u>ю</u>-м<u>о</u>ю*…). В таком случае зачастую становится продуктивным подход, когда задаются специальные условия для нахождения низкочастотного омонимичного варианта, а в остальных случаях по умолчанию берется вариант с высокой частотностью.

3.6.2. Статистические методы

Статистические методы, основаны на обучении вероятностной модели на основе речевых корпусов. Такие методы, по сути, также основываются на анализе контекста рассматриваемых слов (например, HMM-модели, основанные на n-граммах, или деревья решений, использующие в качестве

признаков характеристики соседних слов), однако контекст здесь учитывается автоматически, без участия эксперта. Могут быть получены хорошие результаты при наличии достаточно большой обучающей выборки, в которой в достаточном количестве встречаются все нужные элементы (при этом обучающая база данных должна содержать их расшифровку или правильное произношение); однако проблемы появляются при недостаточности данных.

4. ПРОСОДИЧЕСКИЙ ПРОЦЕССОР

4.1. Определение границ синтагм

Под синтагмой понимается самостоятельная в интонационном смысле часть предложения или всё предложение. Установка границ синтагм влияет на передачу интонационных характеристик при синтезе речи, а также на передачу смыслового содержания. При разбиении текста на синтагмы важно не поставить границу синтагмы там, где она может нарушить смысловое восприятие речи (или передачу смыслового содержания текста). Синтагмы в речи разделяются паузами. Такие паузы делают речь более понятной и естественной, разрешая неоднозначные трактовки смысла предложений.

Отметим, что процесс определения синтагм должен удовлетворять решению двух основных задач: установить границы синтагм в тех местах, где они обязательно должны присутствовать, и не устанавливать границу синтагмы там, где она может нарушить смысловое восприятие речи.

Многие системы синтеза речи при определении мест пауз опираются только на знаки препинания. Однако большие участки текста, расположенные между этими знаками, могут звучать монотонно и осложнять восприятие речи, что делает актуальной задачу определения мест пауз на подобных участках. При синтезе русской речи дополнительно возникает другая проблема – пунктуация традиционно используется для обособления различных вводных конструкций, таких как, например, «может быть», «конечно» и т.д., которые не выделяются паузами в устной речи.

Кроме того, системы синтеза речи должны не только определять места пауз, но и их продолжительность как внутри предложений, так и между ними. Самым простым решением данной задачи является задание различных констант, регламентирующих длительность пауз. Но, так как длительность естественных (производимых человеком) пауз является очень вариативной величиной, необходим специальный метод, позволяющий вычислять длительность пауз в зависимости от типа контекста и структуры предложения.

Принципы, описывающие расстановку пауз в естественной речи, зависят от ряда факторов. Наиболее значимым из них является синтаксическая структура предложения: паузы зачастую располагаются между синтаксически связными компонентами. Однако длина предложения, семантика определенных слов и другие особенности также имеют значение. В системах синтеза речи эти факторы могут быть учтены путем задания правил,

определяющих, после какого слова в предложении должна стоять пауза, или путем обучения статистических моделей на большом речевом корпусе, на основе которых будут вычисляться вероятности наличия пауз после того или иного слова.

4.1.1. Установка пауз по правилам

Процесс определения синтагм в этом случае можно условно разбить на три основные части.

1. Расстановка пауз.
2. Расстановка фразовых ударений.
3. Особые случаи расстановки ударений

В свою очередь, этап **расстановки пауз** делится на следующие, последовательно выполняемые, этапы:

1) определение связей в каждой паре слов;
2) грамматический анализ;
3) установка ударений для всех слов в предложении согласно информации, поступившей от лингвистического процессора.
4) установка пауз вокруг больших групп слов.
5) удаление пауз для однородных членов и деепричастных оборотов после служебных слов.
6) установка ударений для слов, которые находятся между пауз.
7) установка пауз на основании синтаксических связей
8) установка пауз на длинном отрезке без пауз.

Просодический процессор выделяет в каждом предложении последовательности слов, связанные синтаксической связью, которые, скорее всего, будут представлять из себя цельные просодические единицы (синтагмы). Между парами слов устанавливаются связи того или иного типа, что позволяет определить, может ли внутри данной пары слов быть установлена пауза. Этот этап является подготовительным перед определением местоположения и длины пауз в предложениях.

Далее проводится неполный (поверхностный) синтаксический анализ предложения. Для правильного деления предложения на синтагмы не нужно производить полный анализ синтаксической структуры: достаточно выделить самостоятельные группы слов, между которыми в принципе возможна постановка паузы, а внутри которых пауза маловероятна. Поиск таких групп слов осуществляется при помощи сопоставления словам синтаксических шаблонов – заранее заданных последовательностей частей речи и/или грамматических форм, соответствующих различным часто встречающимся в текстах словосочетаниям. При построении системы шаблонов учитываются следующие частеречные категории, грамматические характеристики слов: род, число, падеж и др., а также согласование между различными частями речи. Дополнительные характеристики включают отдельные семантические признаки слов, а также возможность задания правила для конкретного слова.

Кроме того, отдельно анализируются особые синтаксические структуры, такие как однородные члены предложения, вводные слова, сложные предлоги и т.д.

4.1.2. Установка пауз на основе статистических моделей

Установка пауз по правилам работает достаточно хорошо, однако невозможно учесть все, в особенности, сложные случаи, встречающиеся в различных текстах. Также разработка подобных правил для новых языков требует большого количества времени. Преимуществом методов машинного обучения является простота применения, при условии наличия размеченного речевого корпуса достаточного объема. Ожидается, что статистические модели будут более детально имитировать поведение человека, нежели правила, основанные на знаниях экспертов.

Для определения мест пауз и их длительностей в [9] предлагается использовать следующие классификаторы: CART[7] и RF[8]. Классификатор CART применяется как для определения мест пауз, так и для определения их длины: для каждой границы слов определяется длительность паузы между ними (там, где она равна нулю или меньше заданного порога, пауза отсутствует). Также данный тип классификатора применялся только для определения длин пауз. В этом случае предсказывается длительность только между теми словами, куда была поставлена пауза на предыдущих этапах обработки текста. Классификатор RF применялся только для определения мест пауз в виду его специфики.

Классификатор CART – рекурсивный метод разбиения набора данных на основе минимизации критерия (4.1):

$$G(C_1, C_2) = \frac{D(C_1)T(C_1) + D(C_2)T(C_2)}{T(C_1) + T(C_2)}, \tag{4.1}$$

где

$$D(C) = \frac{2\left(\sum_{i=1}^{|C|}\sum_{j=i}^{|C|} d(U_i, U_j)\right)}{|C|^2 - |C|}, T(C) = \frac{|C|^2 - |C|}{2},$$

$|C|$ – размер кластера C, $d(U,V)$ – расстояние между векторами признаков U и V, критерием завершения служит минимальное количество элементов в кластере (в [9] рекомендуется значение три).

Классификатор RF выполняет классификацию данных на основе множества признаков путем создания иерархии («деревьев») запросов на основе предсказанных значений признаков в каждой точке. Лист каждого из деревьев содержит информацию обо всех наблюдениях характеризуемой величины, признаки которой лежат в одной области значений. В [9] применялся «лес решений», содержащий 100 деревьев, где каждое дерево построено на 60% случайно выбранных данных, что снижает чувствительность алгоритма к шуму в обучающих данных. Данные параметры были выбраны на основе максимизации качества результата.

Для решения задачи классификации в [9] использовались следующие признаки:

- пунктуация: знак препинания после текущего слова, после двух предыдущих и после двух следующих слов;
- количество слов и слогов: количество слов и слогов в предложении, количество слов и слогов от предыдущей паузы до текущего слова и от текущего слова до конца предложения;
- грамматические признаки: часть речи, падеж, признак является ли слово собственным существительным (имена, названия и т.д.);
- признаки согласования: согласуется ли грамматическая форма текущего слова с двумя последующими словами;
- регистр первой буквы в слове: является ли первая буква в двух предыдущих, в текущем или двух следующих словах заглавной или нет.

Для минимизации ошибок вычисления грамматических признаков необходима процедура разрешения неоднозначности для слов-омонимов и омографов (замОк - зАмок). Предполагается использовать подход, предложенный в работе [10], точность работы которого составляет 96%.

Сравнивая подходы на основе классификаторов CART и RF, можно отметить следующее. Очевидным преимуществом использования CART является маленький размер модели, что является важным показателем при реализации системы синтеза речи. Однако RF дает лучшие результаты при определении мест пауз. Более того, не все ошибки одинаково критичны: в некоторых случаях пауза недопустима, в то время как в других имеет право быть. CART допускает более критические ошибки по сравнению с RF, хотя это может быть выявлено только на основе экспертных оценок. В основном ошибки CART выражаются в виде пауз внутри синтаксически связанных цепочек: после предлогов, союзов и других служебных слов, использующихся для связи последовательности слов; между модификатором (прилагательное, наречие и т.д.) и существительным или глаголом, к которому он относится. Такого рода ошибки практически отсутствуют при использовании классификатора RF. Кроме того, модель RF является более гибкой, т.к. она может быть настроена с целью увеличения или уменьшения количества пауз в синтезируемой речи, что может быть полезно для практических приложений системы синтеза речи. Например, увеличение количества пауз снижает темп речи.

4.2. Определение интонационного контура

Базовые интонационные модели, из ограниченного набора которых исходят создатели синтезаторов, реализуются на практически бесконечном множестве предложений. Даже в языках, где тональный параметр не используется для создания лексических противопоставлений, реализация базовой модели в конкретном предложении может зависеть от таких фонетических свойств, как длина предложения, количество, место и степень

выраженности словесных ударений, число слогов в использованных словах, структура слогов и даже их звуковой состав. В результате у разных предложений наблюдаемый контур F0 (контур основной частоты голоса) может иметь весьма разнообразную и сложную форму: интонационно мотивированные изменения тона (подъемы и падения) могут чередоваться в ровными (платообразными) участками; в контуре могут присутствовать "дырки" и локальные падения, обусловленные глухостью/звонкостью согласных; контур в целом может располагаться в разных областях голосового диапазона говорящего; параметры тонального пространства, занимаемого контуром (его рабочая зона), могут меняться от начала к концу предложения, например, контур может одновременно понижаться и сужаться и т. д. Воспроизведение подобных поверхностных эффектов при синтезе речи, с одной стороны, необходимо, так как от этого сильно зависит естественность конечного результата, а с другой – представляет значительные трудности. Это заставляет разработчиков либо создавать самим, либо искать в лингвистической фонетике какие-то интонационно-просодические модели, которые могли бы послужить основой для автоматического порождения тональных контуров. Элементы модельных представлений содержатся даже в простейших системах, которые обеспечивают только просодический ресинтез.

4.2.1. Генерация контура F0 методом ресинтеза

В системах, основанных на просодическом ресинтезе, в памяти системы хранятся детальные количественные данные о контурах основной частоты, интенсивности и длительности для некоторого фиксированного набора фраз, полученные в результате измерения их естественных произнесений. Например, контур основной частоты может быть запомнен в виде последовательности чисел, представляющих результат попериодного измерения звуковой волны на вокальных отрезках фразы, или же как последовательность значений, измеренных через небольшие временные интервалы (например, каждые 10мс) по контуру F0, полученному с помощью каких-либо автоматических методов акустического анализа речи. Затем эти данные воспроизводятся без изменения при генерации синтетических отрезков, не выходящих, как правило, за пределы того набора фраз, для которых в системе имеются готовые просодические образцы. Несмотря на очевидные ограничения, описанные системы (на Западе их называют copy-синтезаторами) находят свое применение. В частности, они оказываются полезными при тестировании качества синтезаторов в озвучивании сегментного состава речевых отрезков, т. е. помогают оценить степень естественности синтезированной речи, состоящей из искусственных звуков и естественной просодии. В этом случае синтезироваться может любой речевой отрезок, однако для получения просодических данных для ресинтеза он должен быть сначала произнесен человеком, т.е. стать известным синтезирующей системе. Для понимания закономерностей просодического

оформления речевых отрезков подобный ресинтез не представляет особого интереса.

Ресинтез известных просодических образцов используется также в системах, основанных на так называемых методах стилизации тонального контура – акустических или перцептивных. Цель акустической стилизации состоит в том, чтобы сократить детальную информацию, которая содержится в контурах F0 естественных фраз путем автоматического выделения некоторого набора опорных (целевых) точек, аппроксимирующих контур в целом. Стилизация может быть широкой или узкой, в зависимости от разрешенной максимальной плотности опорных точек. Узкая разновидность стилизации часто реализуется в виде выбора трех точек контура на отрезке каждого отдельного гласного фразы – начальной, экстремальной (или серединной) и конечной. Опорные точки при аппроксимации контура соединяются прямыми линиями.

При широкой стилизации в качестве опорных точек часто выбираются локальные экстремумы контура (пики и впадины). Переходы между ними интерполируются либо прямыми линиями, либо более сложными функциями. При таком подходе в качестве особых характеристик контура могут использоваться также прямые, отражающие динамику изменения общего тонального пространства контура во времени. Линия, соединяющая локальные максимумы кривой F0, образует верхнюю границу этого пространства (**topline**). Нижняя граница (**baseline**) задается локальными минимумами. Нисходящий характер обеих линий отражает общее смещение контура F0 вниз, которое часто наблюдается при произнесении повествовательных предложений во многих языках и называется деклинацией.

Перцептивная стилизация отличается от чисто акустической тем, что при выборе способа аппроксимации наблюдаемой кривой F0 учитываются данные восприятия. Наиболее известным примером применения метода перцептивной стилизации является модель, разрабатываемая с 1960 г. в Институте перцептивных исследований (IPO) в Голландии. Исходный контур F0 сначала аппроксимируется вручную последовательностью прямых отрезков (тональных сегментов), которые не соотносятся каким-то специальным образом с сегментной основой анализируемой фразы. Затем фраза с аппроксимированным контуром ресинтезируется, далее с помощью повторных ресинтезаций находится такая аппроксимация контура, которая содержит минимальное количество тональных сегментов и на слух не отличается от исходного контура. Примечательно, что в экспертных экспериментах с перцептивной стилизацией было обнаружено, что модификации кривой F0 на участках глухих и звонких согласных и смежных с ними гласных (так называемые микропросодии) практически не влияют на восприятие тонального контура фразы. "Голландский" метод аппроксимации контура F0 можно рассматривать как широкую разновидность перцептивной стилизации.

В некоторых публикациях описаны методы автоматической перцептивной стилизации, основанные на подходах, отличных от голландского метода. Иногда принимается, что минимальным носителем тональных различий является слог. Восприятие тона в рамках слога зависит не только от F0, но и от других фонетических характеристик (длительности, интенсивности, звуковой структуры и т. п.). По мнению указанных авторов, перцептивная стилизация тонального контура фразы должна представлять собой последовательность тонированных слогов, а тональный контур слога следует интерпретировать с учетом воздействия всех акустических факторов на восприятие высоты тона, а также с учетом известных психоакустических данных (абсолютных и относительных слуховых порогов оценки тональных изменений). Описанный метод является примером узкой разновидности перцептивной стилизации, он был реализован в автоматическом режиме и интенсивно тестировался на материале французского языка.

Судя по имеющимся в литературе оценкам, все методы стилизации контуров F0 позволяют генерировать ресинтезированную речь высокого качества. При создании систем TTS получение качественного тонального ресинтеза с помощью тех или иных автоматических методов не является целью разработок, однако выполняет важную подготовительную функцию. Во-первых, любой метод стилизации (при условии высокого качества ресинтеза) позволяет получить такое представление наблюдаемого контура F0, которое освобождено от ненужных акустических деталей и параметризованно, т. е. содержит количественную спецификацию конечного числа опорных тональных элементов (точек или отрезков), с помощью которых аппроксимируется контур. Во-вторых, выбор опорных элементов стилизации зачастую отражает теоретические представления (или допущения) исследователей о том, что представляет собой глубинная интонационная характеристика предложения, которая получается (или может быть получена) на выходе лингвистического блока подготовки текста к озвучиванию. В этом случае ресинтез на основе выбранного метода стилизации позволяет дать предварительную оценку сложности параметрического просодического интерфейса и активно используется для текущей отладки правил генерации тонального контура. В то же время ясно, что ресинтез сам по себе не может обеспечить порождение тонального оформления произвольного предложения.

4.2.2. Формирование контура F0 для произвольного предложения

В конкретных системах автоматического синтеза речи содержание и сложность просодических правил, порождающих тональный контур предложения по его интонационному описанию, зависит как от практических возможностей лингвистического блока системы, так и от того, что понимается под интонационной структурой предложения. Минимальная интонационно значимая информация включает: указание на коммуникативный тип предложения (**sentence mode**), интонационное членение и расположение акцентированных (или просто лексически ударных) слогов в пределах каждой

интонационной группы. В рамках этого общего минимального требования имеющиеся приложения делятся на две большие группы в зависимости от того, используется ли в них собственно интонационная транскрипция, базирующаяся на некотором фиксированном наборе интонационных единиц – общих моделей или более элементарных просодических элементов, входящих в интонационную систему синтезируемого языка. Условно системы синтеза, в которых интонационная транскрипция на входе просодических правил в явном виде не используется, могут рассматриваться как реализации инженерного подхода, в отличие от систем, опирающихся на транскрипцию. Последние системы называются лингвистически (фонологически) ориентированными. Рассмотрим основные особенности этих подходов.

4.2.3. Генерация тонального контура в системах инженерного типа

В эту группу, прежде всего, попадают приложения, которые опираются на узкую акустическую стилизацию тональных контуров. Алгоритмы автоматического получения контуров F0 (**pitch extraction**) и автоматической сегментации речевого сигнала создают возможность построения больших, просодически ориентированных баз данных, в которых фиксируются частотные значения опорных точек контура F0 для каждого гласного или отдельного слога в составе предложения. Соответствие между минимальной интонационно значимой информацией, которая дается для каждого предложения в базе данных, и тональными параметрами гласных или слогов (с учетом большого набора поверхностных фонетических переменных – типа слога, его положения в слове и интонационной группе и т.п.) устанавливается с помощью статистических классификационных методов или методов, применяемых в системах распознавания речи, в частности нейроподобных сетей. После такого предварительного анализа или обучения реальный синтез произвольного предложения получается путем конкатенации тональных слоговых контуров, выбранных из базы с учетом как интонационных признаков, так и поверхностных фонетических факторов, влияющих на акустическую реализацию слогового контура F0. Нетрудно видеть, что просодические тональные правила заменяются в системах описанного типа хранением обширного инвентаря тональных слоговых контуров, которые конкатенируются "склеиваются", образуя сложный тональный контур предложения.

По имеющимся в печати отзывам, синтез на основе узкой акустической стилизации и тональной конкатенации обеспечивает очень высокую естественность синтезированной речи. Разработки в этом направлении начались сравнительно недавно, их технологичность, значительная доля автоматизации подготовительной работы привлекают исследователей, занимающихся речевыми технологиями, и специалисты прогнозируют бурный рост соответствующих приложений. В то же время с лингвистической точки зрения подобные системы мало интересны: фактически в них можно усматривать представление об интонации как о некотором акустическом гештальте, который развертывается в виде сложной тональной схемы на

слоговой цепочке предложения. Однако возможно, что некоторые речевые единицы, ритуальные или несущие сильную эмоциональную окраску, действительно запоминаются и используются в речи, снабженные подобными "гештальтными схемами-мелодиями", находящимися за пределами собственно интонационной системы языка. Безусловный интерес для лингвистически ориентированных исследований интонации представляет компьютерный инструментарий, который используется при создании послоговых конкатенативных систем тонального синтеза.

Кроме приложений, основанных на конкатенации слоговых тональных контуров, к системам инженерного типа относится и ряд разработок, которые на самом деле занимают промежуточное положение между чистой тональной конкатенацией и лингвистически ориентированными моделями тонального синтеза.

В приложениях такого типа наиболее часто используется артикуляционно-акустическая модель тонального контура (**production-oriented model**), предложенная известным японским специалистом в области речевых технологий Х. Фуджисаки [12]. Основное допущение этой модели состоит в том, что тональный контур, непрерывный по своей природе, является на самом деле реализацией локальных физиологических событий, которые осуществляются разными ларингальными механизмами. Различаются два типа событий – фразовые и акцентные тональные команды, которые моделируются соответственно импульсной и ступенчатой функциями. Кроме этого, вводится один глобальный параметр, который фиксирует нижнюю границу рабочей области голосового диапазона, на нее накладываются фразовые и акцентные команды. Локальные компоненты модели описываются несколькими параметрами, которые задают относительную амплитуду тонального изменения и временные моменты реализации команд (таймирование) относительно границ фразы (для фразовых импульсов) и границ акцентированного слова для акцентных. Результирующий тональный контур получается путем сложения всех компонентов, имеющих, как следует из сказанного выше, разные области реализации во времени. В связи с этим модель Фуджисаки часто относят к суперпозиционным фонетическим моделям интонации.

При создании системы синтеза для конкретного языка используются просодические базы данных, где каждое предложение содержит, по крайней мере, минимальную интонационную информацию. При анализе корпуса предложений фразовые команды соотносятся с границами интонационного членения, а акцентные – с акцентированными слогами. Амплитудные и временные параметры аппроксимирующих функций подбираются по базе данных с помощью статистических методов. Модель тестировалась в системах синтеза для весьма разных языков: японского, английского, китайского, немецкого и ряда других.

4.2.4. Генерация тонального контура на основе лингвистических моделей интонации

В лингвистически ориентированных системах тонального синтеза контур F0 рассматривается как акустическая манифестация интонационной структуры предложения, которая может быть представлена в виде определенной конфигурации абстрактных интонационных элементов, которые должны фиксироваться в выходной транскрипции лингвистического блока синтезатора. В разработке таких систем активное участие принимают лингвисты. В соответствии с теоретическими направлениями, существующими в западной интонологии, можно выделить два типа моделей, которые не только находят применение в системах синтеза речи по тексту, но и благодаря этому активно развиваются. Это так называемые суперпозиционные (**layred components**) и линейные или последовательные (**tone sequences**) модели. Оба типа моделей исходят из представления о комбинаторной природе интонации: интонационная структура предложения рассматривается как конструкция, состоящая из нескольких функционально самостоятельных тональных элементов. Оба типа моделей признают существование и лингвистическую значимость локальных тональных объектов, имеющих фиксированную временную привязку в предложении, и глобальные тональные признаки, которые характеризуют тональное пространство, в рамках которого реализуется контур в целом. Однако функциональная интерпретация локальных и глобальных тональных элементов и их взаимодействие в предложении трактуются в этих моделях по-разному.

В суперпозиционных моделях интонационная структура предложения рассматривается как иерархическая просодическая структура, определяемая в каждой точке предложения одновременно тремя тональными объектами, каждый из которых имеет свою сферу реализации. Тональные составляющие описываются следующим образом: выделяются глобальные тональные признаки, характеризующие тональное пространство, в котором реализуется предложение в целом, глобальные тональные признаки пространства, занимаемого последовательными интонационными группами в предложении, и тональные фигуры, которые реализуются на составляющих, называемых акцентными группами. Интонационные контуры основных коммуникативных типов предложений отличаются только глобальным тональным признаком, отражающим частотное смещение тонального контура во времени (его наблюдаемым коррелятом служит линия деклинации, соединяющая акцентированные слоги в предложении). Так, повествовательные предложения имеют наиболее резкий наклон деклинационной линии, а общий вопрос характеризуется отсутствием наклона (плоской линией деклинации). Реализационной базой лингвистических моделей суперпозиционного типа является описанная выше модель Фуджисаки.

Линейные модели восходят к работам Ж. Пьерхумберт [13], посвященным первоначально интонации американского варианта английского

языка. В лингвистическом плане интонационная модель Пьерхумберт опирается на идеи метрической и автосегментной фонологии, развиваемые в США. В качестве минимальных элементов в модели выделяются два одинарных тона, отличающиеся тональным уровнем – высокий (H) и низкий (L). Интонационные тоны рассматриваются как абстрактные тональные цели (мишени), ближайшим отражением которых в наблюдаемом тональном контуре являются точки переломов (изменений) F0.

На основе этих тональных примитивов формируются тональные единицы следующих функциональных типов:

1) тональные акценты – одинарные (H*, L*) и битональные (аналоги контурных тонов) (H*+L, H+L*, L*+H, L+H*). Знак * обозначает привязку тона к акцентированному лексически ударному слогу;
2) фразовые тоны – два типа тональных движений (H-, L-), которые реализуются между последним тональным акцентом интонационной группы и граничным тоном;
3) граничные тоны – тоны, соотнесенные с начальным (%H, %L) и конечным (H%, %L) слогами интонационной группы.

Возможные комбинации перечисленных тональных единиц образуют грамматику интонационной структуры фразы, которая состоит из четырех следующих компонентов: начальный тон, тональные акценты, фразовый тон, конечный тон.

Абстрактные тональные репрезентации, которые условно можно рассматривать как маршрут или схему движения в целевом тональном пространстве, преобразуются в наблюдаемые контуры F0 с помощью просодических правил двух типов: тонального шкалирования и таймирования. Правила тонального шкалирования определяют для абстрактных целевых тонов конкретные значения F0, которые считаются зависимыми от двух факторов: степени выделенности слога, несущего тон, и тональной спецификации предшествующего тона. Таким образом, частотная спецификация последовательности тонов осуществляется строго слева направо (отсюда название "линейная" модель). Правила таймирования задают с учетом разных поверхностных фонетических факторов координаты временных точек, в которых должна достигаться тональная цель. Кроме просодических правил, используются адаптирующие функции, с помощью которых в контуре F0 целевые тональные точки соединяются тональными переходами и контур в целом сглаживается.

Глобальные тенденции, наблюдаемые в контурах F0, в крайних вариантах линейной модели описываются также исключительно локально. Например, деклинация считается поверхностным результатом локального взаимодействия определенных смежных тонов (аналогично **downstep** в африканских языках), а не глобальным тональным признаком, распространяющимся на всю интонационную группу. Локальная

интерпретация глобальных тенденций является наиболее дискуссионной стороной строго линейных моделей и причиной построения различных гибридных моделей, авторы которых вводят в линейную модель и глобальные тональные признаки. В целом, надо сказать, что на Западе, особенно в США, линейная модель Пьерхумберт получила очень большой резонанс как в фонологических исследованиях, так и прикладных разработках. Эта модель в адаптированном виде использовалась в системах синтеза для английского, немецкого, китайского, японского и шведского языков. При создании приложений все просодические правила и адаптирующие функции настраиваются автоматически с помощью обширных аннотированных баз данных. Для интонационной аннотации речевых корпусов была создана широко известная просодическая транскрипционная система ToBI (сокр. англ. Tones and Break Indices). В то же время нельзя не отметить, что лингвистический (функциональный) потенциал линейной модели даже для английского языка в полной мере не проверен и не используется в системах синтеза, так как до сих пор не сформулированы правила выбора тонов, образующих тональный компонент интонационной структуры предложения.

4.3. Примеры интонационных контуров

Рассмотрим несколько примеров интонационных контуров (ИК) для русского языка, по классификации Брызгуновой [1].

Второй тип ИК (рис. 4.1): синтагматическое ударение на вопросительном слове. На гласном центра ровный или нисходящий высокий тон, затем дальнейшее падение.

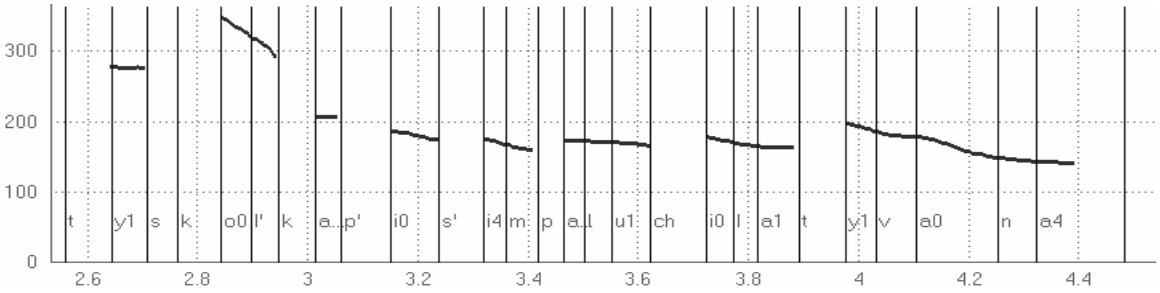

Рис. 4.1. ИК для фразы «Ты сколько писем получил от Ивана?»

Третий тип ИК (рис 4.2): восходящий тон с последующим падением.

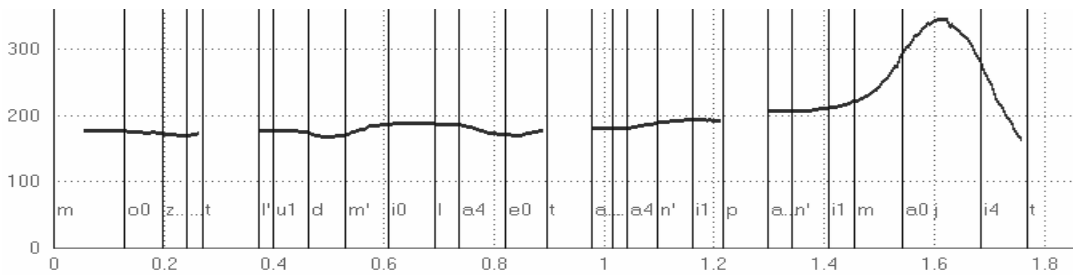

Рис. 4.2. ИК для фразы «¬Может Людмила этого не понимает?»

Синтагматическое ударение ставится на последнем слове предложения.
Четвертый тип ИК (рис 4.3): нисходяще-восходящий тон.

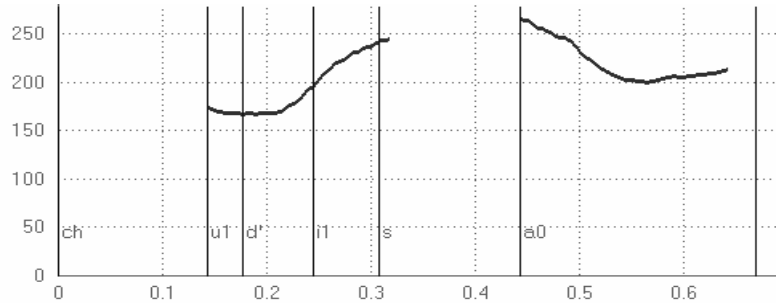

Рис. 4.3. ИК для фразы «Чудеса!»

5. ФОНЕТИЧЕСКИЙ ПРОЦЕССОР

5.1. Построение транскрипции

Рассмотрим алгоритм построения транскрипции на примере русского языка. Русский язык в этом отношении является достаточно регулярным, что позволяет описать практически весь алгоритм набором правил.

На вход транскриптора подается текст, в котором указаны места, в которых при произношении будут сделаны паузы, и для каждого слова указано, какая из его гласных находится под ударением. На выходе транскриптор выдает последовательность фонем и аллофонов, соответствующих входному тексту и определяющих его произношение.

Место ударения в каждом из слов существенно влияет на то, как будет произноситься данное слово — очевидно, что речь, в которой все гласные буквы имеют одинаковую продолжительность и интенсивность, будет звучать весьма неестественно. В русском языке безударные гласные имеют редукцию. Чем больше степень редукции, тем меньше длительность произношения данной буквы, и тем меньше возможность различить произносимые буквы между собой, например, «(водное) поло» — «(из-под) пола».

Редукция гласных в русском языке вычисляется следующим образом. Ударные гласные не редуцируются — степень редукции у них 0. Гласная в первом предударном слоге (предшествующем ударному слогу) имеет степень редукции 1. Гласные во втором и следующих предударных слогах имеют степень редукции 2, а гласные в заударных слогах (после ударного слога) — степень редукции 4. Например, в слове «бородатый» редукция гласных в слогах будет, соответственно, 2-1-0-4. При этом безударная гласная, являющаяся первой буквой в слове, как, например, «а» в слове «аллофон», редуцируется только до степени 1.

Что касается транскрибирования согласных букв, то здесь также есть свои нюансы. Так, в положении перед гласной «а» все согласные различимы между собой в произношении, тогда как в других положениях, например, на конце слова или перед согласными, произношение согласных может меняться. Например, слова «рог» и «рок» не различаются по произношению — в слове «рог» происходит оглушение звонкой согласной «г» на конце слова. Таким же образом может происходить и озвончение согласных — например, в словосочетании «этот звон» последняя «т» в слове «этот» озвончается и произносится как «д». Правила оглушения и озвончения согласных реализованы в транскрипторе.

Наличие или отсутствие пауз между словами во многом определяет особенности транскрибирования на стыках слов. В случае, если между словами нет паузы, имеет место взаимовлияние соседних звуков, принадлежащих разным словам. Кроме того, предлоги, предшествующие словам, или частицы, следующие за словами, при произношении объединяются с тем словом, с которым соседствуют, и становятся составляющими единого фонетического слова, как, например, в сочетаниях слов «по воде» или «могли бы». Редукция гласных в таких случаях рассчитывается для всего фонетического слова как целого. Если между словами присутствует пауза, то она делает невозможным влияние друг на друга звуков в словах, находящихся по разные стороны от нее.

Некоторые слова русского языка произносятся не так, как должны были бы произноситься согласно обычным правилам произношения — например, слово «принтер» произносится как [принтыр], а не [принтер]. Эти слова-исключения вместе со своими транскрипциями хранятся в отдельном словаре.

При обработке текста транскриптором производится следующая последовательность действий.

1. Устанавливается степень редукции гласных влево и вправо от ударной гласной каждого слова. При этом:
   1.1. Каждой ударной гласной присваивается степень редукции 0.
   1.2. Гласным слева от ударной гласной в слове присваивается степень редукции, увеличивающаяся от 1 до 2.
   1.3. Гласным справа от ударной гласной присваивается степень редукции 4.
   1.4. Если первая буква в слове — безударная гласная, то ей присваивается степень редукции 1.
   1.5. Те слова, в которых нет своего ударения (предлоги, частицы и т. п.), рассматриваются как единое фонетическое слово вместе со словом, к которому они относятся. При этом степени редукции гласных в фонетическом слове (то есть имеющим основное ударение) расставляются аналогично тому, как это делается для обычных слов.
2. Производится транскрибирование текста в фонемы в соответствии с правилами преобразования буква-фонема, которые подгружаются из внешнего файла.

3. Производится транскрибирование фонем в аллофоны в соответствии с правилами преобразования фонема-аллофон, которые также задаются во внешнем файле. Определение того, какой именно аллофон должен соответствовать данной фонеме, производится в зависимости от контекста — от того, какие фонемы или паузы стоят перед и после данной фонемы.
4. Исключения из обычных правил произношения, существующие для некоторых слов русского языка, обрабатываются отдельно от основной последовательности действий. Примеры таких слов: принтер, Габриель, Фред и т.п.

5.2. Вычисление физических параметров

На вход алгоритму построения физических параметров подается текст, в котором указаны места, в которых при произношении будут сделаны паузы, для каждого слова указано, какая из его гласных находится под ударением и какой силы — это ударение, а также какой тип интонационного контура лежит на этом слове, для каждой буквы - ей соответствующая фонема и аллофон. На выходе построитель физических параметров выдает последовательность аллофонов, соответствующих входному тексту, определяющих его произношение с указанием для каждого из них значения частоты основного тона, отклонения энергии и длительности звучания.

Сила ударения и тип интонационного контура в каждом из слов существенно влияет на то, как будет произноситься данное слово — очевидно, что речь, в которой все гласные буквы имеют одинаковую продолжительность и интенсивность, будет звучать весьма неестественно. В русском языке безударные гласные имеют редукцию. Чем больше степень редукции, тем меньше длительность произношения данной буквы, и тем меньше возможность различить произносимые буквы между собой, например, «(водное) поло» — «(из-под) пола».

Наличие или отсутствие пауз между словами во многом определяет особенности формирования физических параметров на стыках слов.

В случае, если между словами нет паузы, имеет место взаимовлияние соседних звуков, принадлежащих разным словам. Кроме того, предлоги, предшествующие словам, или частицы, следующие за словами, при произношении объединяются с тем словом, с которым соседствуют, и становятся составляющими единого фонетического слова, как, например, в сочетаниях слов «по воде» или «могли бы». Физические параметры в таком случае вычисляются для всего фонетического слова как целого. Если между словами присутствует пауза, то она делает невозможным влияние друг на друга звуков в словах, находящихся по разные стороны от нее.

Функциональность описываемого алгоритма состоит в том, чтобы определить временные (в мс) и мелодические (в Гц) характеристики базовых элементов компиляции, которые при обработке синтагмы выбираются в нужной последовательности специальным процессором (блоком кодировки).

Необходимые для этого предварительные операции над синтезируемым текстом: выделение синтагм, выбор типа мелодического контура, определение степени выделенности (ударности-безударности) гласных — осуществляются предшествующими модулями.

Правила временного оформления синтагмы сформулированы отдельно для гласных и согласных. Правила, задающие временные характеристики гласных в обрабатываемой синтагме, учитывают степени выделенности (редукции) гласного (4 градации). Кроме того, для ударного гласного последнего полнозначного слова учитывается число слогов в слове и количество ударных гласных, предшествующих данному в синтагме. Предусмотрено также продление гласных (независимо от степени их редукции и фонетического качества) в позиции абсолютного конца синтагмы. Что касается влияния согласных на длительность гласных, то оно учитывается лишь в наиболее ярких случаях, прежде всего, для гласных в позиции перед интервокальными вибрантами.

Для последовательностей гласных, образующих единый элемент компиляции (заударные флексии), действует правило аддитивного сложения длительностей, задаваемых правилами формирования длительностей.

Правила, определяющие временные характеристики согласных учитывают следующие факторы: позиция согласного относительно границ синтагмы и фонетического слова; интервокальная-неинтервокальная позиция; позиция в кластере (стечения согласных); простой-сложный состав базовых элементов компиляции, необходимых для звукового синтеза согласных.

В алгоритм формирования физических параметров входят также правила, задающие длительность паузы после окончания синтагмы (конечной-неконечной), которые необходимы для синтеза связного текста.

Правила мелодического оформления синтагмы задают два значения частоты основного тона (F0) для каждого выбранного элемента компиляции, которые образуют его начальную и конечную мелодические характеристики. Вычисление этих “физических” значений происходит на основе предварительного определения по правилам мелодических характеристик транскрипционных аллофонов в полутоновой шкале (Т-значения). Полутоновые характеристики (начальная и конечная) каждого аллофона формируются текущим образом (слева направо) слоговыми циклами, т.е. в рамках последовательности (Cn)Г, где Cn — любое число согласных, в том числе 0, предшествующих гласному.

Алгоритм формирования физических параметров содержит правила для формирования следующих типов мелодических контуров:

- повествовательное предложение,
- повествовательное предложение, в котором есть слово с особым выделением,
- вопрос с вопросительными словами,
- восклицательное предложение,

- восклицательное предложение с вопросительными словами (Какая погода!), "Какая погода...",
- простой вопрос,
- вопрос со значением противопоставления,
- пунктуация - запятая тире,
- пунктуация двоеточие,
- пунктуация тире,
- пунктуация запятая,
- пауза есть,
- пунктуации нет.

Для всех контуров, кроме вопроса с вопросительными словами, учитывается возможность разного положения главноударного слога (мелодического центра) синтагмы. Специальный вопрос формируется для случая совпадения мелодического центра с вопросительным местоимением.

При определении мелодических характеристик элементов компиляции, входящих в обрабатываемый слог, учитываются следующие факторы:

- тип мелодического контура синтагмы;
- положение слога относительно мелодического центра контура (совпадение, слева, справа);
- положение слога относительно начальной и конечной границы синтагмы;
- степень выделенности (редукции) гласного в обрабатываемом слоге;
- степень выделенности (редукции) гласного, непосредственно предшествующего обрабатываемому слогу;
- число символьных элементов в слоге;
- тип символьного элемента слога (согласный, гласный) и положение этого элемента относительно начала слога (первый - не первый);
- фонетическое качество согласных в слоге (глухость-звонкость);
- простой - сложный состав базовых элементов компиляции, необходимых для звукового синтеза согласных в слоге.

Результат применения правил к любой затранскрибированной нужным образом синтагме может быть представлен в виде таблицы стандартного формата, пример которой приводится ниже (таблица 5.1) для фразы “Мама мыла малину?” (в мужском произнесении).

Таблица 5.1

Звук	**Длительность**	**Значение ЧОТ в Гц**	
	(в мс)	**нач.**	**кон.**
м	60	120	120
а+	100	125	125

м	60	120	120
а	75	125	125
м	60	125	180
ы?	80	180	240
л	40	240	225
ъ	50	225	225
м	60	225	120
а	75	120	110
л’	50	110	110
и+	95	110	110
н	60	110	85
у2	110	85	80

В таблице 5.1 представлены звуки, определяющие произносительный вариант фразы, для которых вычислены параметры длительностей и частоты основного тона с которыми должны быть синтезированы звуки во фразе целиком. Знаком " + " обозначаются ударные гласные, " ' " - мягкие согласные, а " ? " - фразовое ударение.

6. АКУСТИЧЕСКИЙ ПРОЦЕССОР

Схема работы акустического процессора представлена на рис. 6.1.

6.1. Оптимальный выбор звуковых элементов методом Unit Selection

После того, когда требуемые параметры звуковых элементов, необходимых для синтеза определенного предложения, получены, наступает очередь применения метода Unit Selection (US) для выбора оптимальной последовательности их реализаций из звуковой базы данных [14].

Для того чтобы определить, насколько тот или иной элемент базы подходит для синтеза данной единицы, вводятся понятия **стоимости замены** (англ. **target cost**) и **стоимости связи** (англ. **concatenation cost**). Стоимость замены для элемента из базы u_i по отношению к искомому элементу t_i вычисляется по формуле

$$C^t(u_i, t_i) = \sum_{k=1}^{p} w_k^t C_k^t(u_i, t_i), \qquad (6.1)$$

где

- C_k^t — расстояние между *k*-ими характеристиками элементов,
- w_k^t — вес для *k*-ой характеристики.

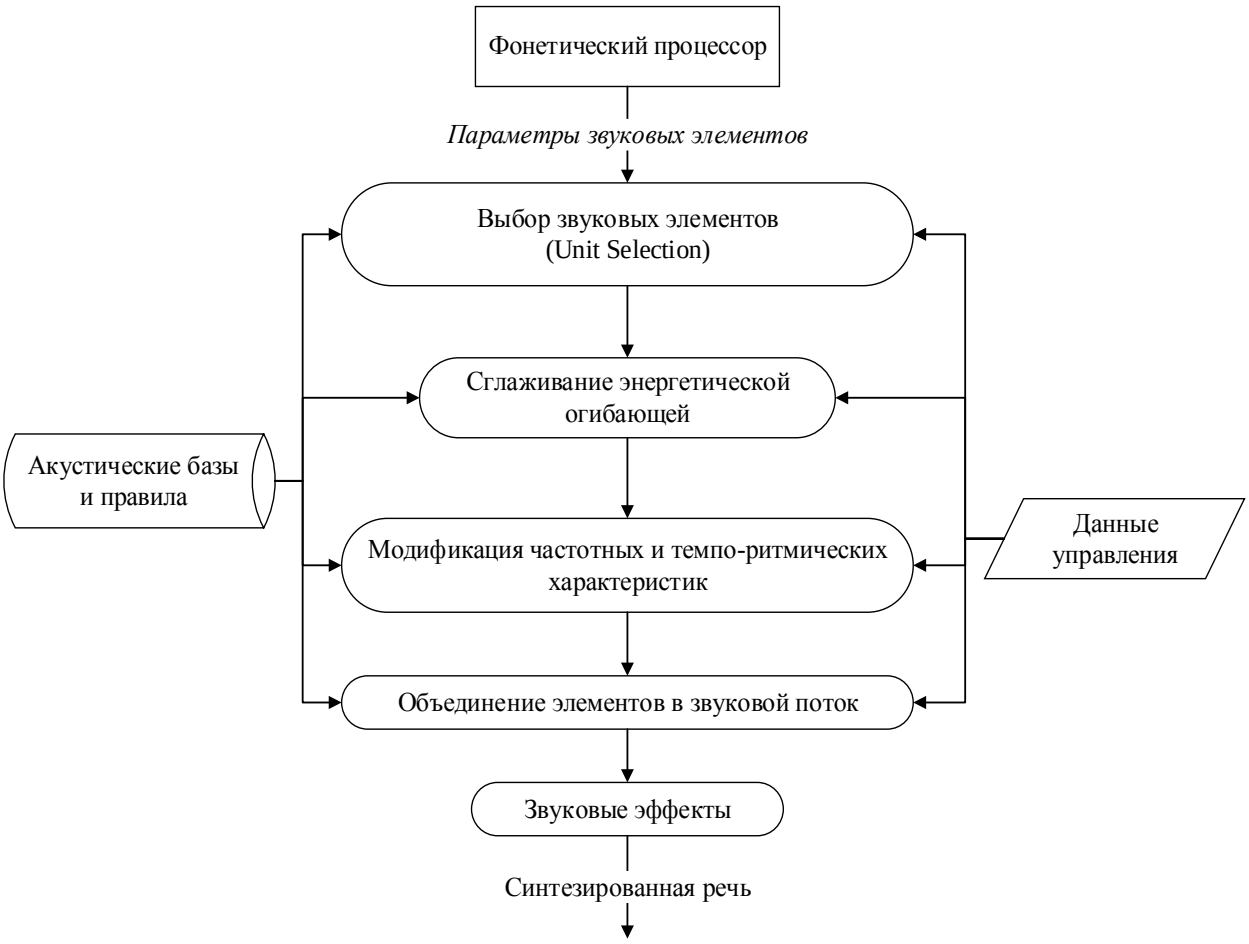

Рис.6.1.Схема работы акустического процессора

Другими словами, это есть взвешенная сумма различий в признаках между требуемым элементом и конкретным элементом речевой базы. В качестве признаков могут выступать любые уместные, с точки зрения разработчика, просодические и лингвистические характеристики элементов. Как правило, используется следующая информация: частота основного тона (ЧОТ), длительность, контекст, позиция элемента в слоге, слове, количество ударных слогов во фразе и другие [15].

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

$$C^c = (u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c (u_{i-1}, u_i), \qquad (6.2)$$

где

- C_k^c — расстояние между *k*-ими характеристиками элементов,

- w_k^c — вес для *k*-ой характеристики.

Общая стоимость для целой последовательности из n элементов есть сумма введенных выше стоимостей

$$C(u,t)=\sum_{i=1}^{n}C^t(u_i,t_i)+\sum_{i=2}^{n}C^c(u_{i-1},u_i). \qquad (6.3)$$

Задача метода US — выбрать такое множество элементов базы $u_1,u_2,\ldots,u_n$, которое бы минимизировало общую стоимость согласно формуле (6.3).

6.1.1. Стоимость замены

Основное назначение функции стоимости замены — оценивать, в какой мере подходит данная единица речевой базы к требуемому элементу. В связи с этим, стоимость замены должна отражать, как сильно различия в характеристиках влияют на восприятие замены одного элемента другим. При построении этой функции, как правило, руководствуются одним из следующих принципов:

- независимых признаков,
- акустического пространства.

Принцип независимых признаков

В этом случае расстояние для каждого признака считается независимо от других, взвешивается и затем общая стоимость считается как некоторая функция полученных расстояний. В качестве такой функции можно использовать простую сумму (6.1). Функции C_k^t определяют расстояния для каждой отдельно взятой характеристики. Для категориальных это может быть простое бинарное решение, совпадают они или нет. Для непрерывных (например, ЧОТ) это может абсолютное расстояние или его логарифм. Различия в одних характеристиках оказывают больше влияния на восприятие замены, чем в других. Эта разница отражается в выборе весов w_k^t для конкретного расстояния. Для установки весов существует несколько подходов:

1) автоматический подбор на основе объективной меры,
2) перцепционный,
3) ручная настройка.

Автоматический подбор на основе объективной меры. Суть этого подхода заключается в попытке найти такой набор весов, который минимизировал бы акустическое расстояние между синтезированным и эталонным выражениями. Для оценки близости требуется метрика, поставляющая расстояния между синтезированными и эталонными высказываниями. Высказывания, воспринимаемые на слух как сходные, должны иметь маленькое расстояние между собой. Для нахождения оптимальных весов достаточно воспользоваться методом линейной регрессии.

Задача определения такой метрики является отдельной проблемой [16]. При таком подходе веса могут подбираться индивидуально для каждой единицы базового типа.

Перцепционный. Слабое место предыдущего подхода заключается в том, что разработчик во многом полагается на акустическую меру, которая лишь частично соответствует человеческому восприятию. В рамках данного подхода ставится эксперимент, в котором людей просят оценить синтезированные предложения, а затем тренируют модель согласно полученным оценкам. Очевидный недостаток — большие временные затраты и сложность в организации эксперимента.

Ручная настройка. Проектировщик системы полностью полагается на свой опыт. В ходе тестирования системы веса постепенно уточняются. Главное преимущество - полный контроль над процессом.

Очевидным плюсом принципа независимых признаков при построении функции стоимости замены является небольшое число подлежащих настройке весов (равное количеству используемых признаков). Однако предположение независимого влияния весов на общую стоимость является слишком сильным. Яркой демонстрацией слабости этого принципа является тот факт, что два различных набора характеристик будут неминуемо иметь ненулевое расстояние. Это противоречит нашим знаниям о речи, которые как раз говорят о том, что различные комбинации характеристик зачастую проецируются в одну акустическую реализацию.

Принцип акустического пространства

Главная идея этого подхода заключается в кластеризации единиц базового типа по просодическому и фонетическому контекстам. Блэк и Тэйлор предложили следующую схему кластеризации.

Вводится объективная мера для измерения расстояний между единицами одного базового типа. Опять же, выбор подходящей акустической меры — отдельное поле для исследований. В своей работе авторы используют взвешенное расстояние Махаланобиса на коэффициентах MFCC (**Mel Frequency Cepstral Coefficients**), ЧОТ, мощности и их дельтах (производных первого порядка). Акустическое расстояние между двумя единицами $Adist(V,U)$ — это среднее по всем фреймам внутри единиц плюс среднее по X% фреймов единиц, предшествующих рассматриваемым (близкие единицы будут иметь сходный левый контекст):

$$Adist(V,U) = \frac{WD \cdot U}{|V|} \cdot \sum_{i=1}^{|U|} \sum_{j=1}^{n} \frac{w_j \left| F_{ij}(U) - F_{(i \cdot |V|/|V|)j}(V) \right|}{SD_j \cdot n \cdot |U|}, \qquad (6.4)$$

где

- U,V — элементы одного базового класса,
- $|U| < |V|$ — количество фреймов в U и V,
- $F_{xy}(U)$ — признак y фрейма x элемента U,

- SD_j — стандартное отклонение признака j,
- w_j — вес для признака j,
- WD — взвешивает разницу в продолжительности элементов.

Введенная мера используется для вычисления «загрязненности» $Impurity(C)$ кластера C как среднего акустического расстояния между элементами кластера:

$$Impurity(C) = \frac{2 \cdot \left(\sum_{i=1}^{|C|} \sum_{j=i}^{|C|} Adist(u_i, u_j) \right)}{|C|^2 - |C|}.$$

Затем с помощью стандартной техники деревьев решений кластер разбивается на две части наилучшим образом. Качество разбиения $Goodness(C_1, C_2)$ кластера C на две части $C_1 \, и \, C_2$ задается формулой

$$Goodness(C_1, C_2) = \frac{Adist(C_1)T(C_1) + Adist(C_2)T(C_2)}{T(C_1) + T(C_2)}, \quad T(C) = \frac{|C|^2 - |C|}{2}.$$

В качестве критерия разбиения используются бинарные вопросы, которые касаются характеристик, применяемых для вычисления стоимости замены (фонетический контекст, просодический контекст (ЧОТ и длительность для элемента и его соседей), ударение, позиция в слоге, позиция в слове, позиция в предложении). На каждом этапе выбирается вопрос, дающий лучшее разбиение. Разбиение обычно продолжается до тех пор, пока не будет достигнут какой-либо порог (например, минимальное количество элементов в листе).

6.1.2. Стоимость связи

Основное назначение функции стоимости связи — оценивать, насколько хорошо два элемента соединяются друг с другом. Идеальной была бы функция, имеющая высокую корреляцию с восприятием речи слуховой системой человека. Обычно общая стоимость складывается из нескольких слагаемых, основанных на спектральных и просодических характеристиках фреймов речи с обеих сторон соединяемых элементов. Как правило, учитываются:

1. Разница в ЧОТ.
2. Разница в энергии.
3. Нестыковка различных спектральных параметров:
   (a) MFCC (Mel Frequency Cepstral Coefficients);
   (b) LPC (Linear Predictive Coding Coefficients);
   (c) LSF (Line Spectral Frequencies);
   (d) MCA (Multiple Centroid Analisys).
   (e)

Так же, как и при кластеризации речевой базы, вводится акустическая мера на спектральных параметрах. За последние годы было проведено большое количество исследований с целью выяснить, какая комбинация спектральное представление/метрика дает лучшую корреляцию с человеческим восприятием. К единому мнению по этой проблеме ученые так и не пришли. Можно лишь отметить, что расстояние Махаланобиса на коэффициентах MFCC в большинстве тестов показывает неплохие результаты.

6.1.3. Поиск по алгоритму Витерби

Согласно классическому алгоритму Ханта и Блэка [14] общая стоимость последовательности элементов из базы $u = (u_1, ..., u_n)$ для данной спецификации $t = (t_1, ..., t_n)$ задается формулой (6.3). Эта формула дает стоимость для любой фиксированной последовательности элементов базы $u = (u_1, ..., u_n)$. Цель состоит в том, чтобы найти такую последовательность, стоимость которой будет минимальна. Задача поиска оптимальной последовательности сводится к поиску пути наименьшей стоимости на графе.

Хотя алгоритм Витерби и превосходит в значительной степени поиск полным перебором (квадратичная оценка против экспоненциальной), в своей чистой реализации, и он может не дать необходимой скорости вычислений. В этом случае следует воспользоваться одной из техник отсечения (англ. **pruning**), целью которых является уменьшение количества рассматриваемых последовательностей. При этом отсечение некоторого подмножества последовательностей приводит к риску исключить оптимальный путь, в то время как полный поиск по алгоритму Витерби гарантированно найдет траекторию с наименьшей стоимостью. Последствия зависят от того, много ли найдется в базе путей, имеющих стоимость близкую к оптимальной.

Выделяются две основные техники отсечения: предварительный отбор (англ. **pre-selection**) и отсечение лучей (**beam pruning**). В первом случае для каждого элемента спецификации отбирается фиксированное количество лучших кандидатов. Во втором случае рассматривается только фиксированное количество локально оптимальных путей.

Схематично, процесс работы метода Unit Selection представлен на рис.6.2.

6.1.4. Речевая база и качество синтеза для метода Unit Selection

Метод Unit Selection критически зависит от речевой базы. Качественный синтез возможен только на основе полной, сбалансированной и корректно размеченной базы данных. С ростом объема базы возрастает темповая и интонационная вариативность речи диктора. Иными словами, чем больше база, тем больше вероятность того, что в ней найдется элемент в необходимом контексте с необходимой длительностью и контуром ЧОТ. Как следствие,

меньше искажения от цифровой модификации сигнала и выше естественность синтезируемой речи.

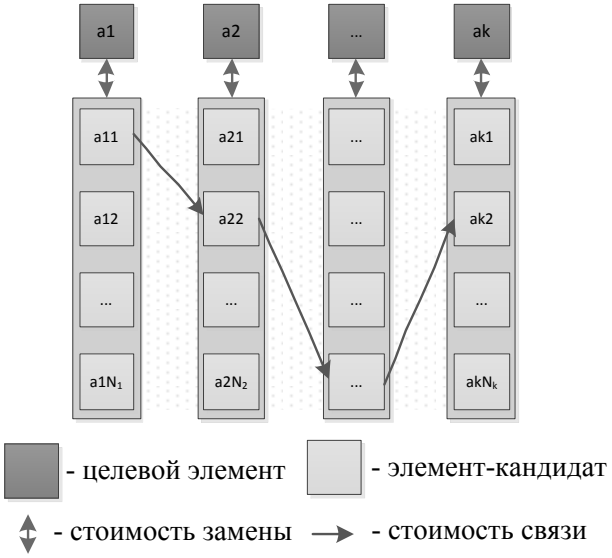

Рис. 6.2. Схема работы метода Unit Selection

В процессе подготовки речевой базы на предварительных этапах желательно проводить запись большого числа дикторов. Запись каждого диктора представляет собой чтение фонетически представительного текста. Запись желательно осуществлять в заглушенной камере с использованием высококачественных средств записи и оцифровки речевого сигнала.

Полученные предварительные записи большого числа дикторов необходимы для получения максимально качественного итогового набора дикторов, голоса которых будут использоваться в системе синтеза речи. Наличие относительно широкого круга дикторов на начальном этапе позволяет осуществить осознанный выбор и минимизировать риск того, что голос того или иного диктора окажется малопригодным для использования в системе синтеза речи.

Отобранные на предварительном этапе дикторы используются для записи больших звуковых баз данных, которые в дальнейшем сегментируются на различных уровнях анализа. В такой ситуации ошибка в выборе диктора на поздних этапах может вылиться в существенные материальные и временные затраты.

Для повышения качества синтеза база сегментируется на разных уровнях. В качестве меток используются реальная и каноническая транскрипции, орфографические слова с отметками логического и синтагматического ударения, типы интонационных контуров. Также размечаются речевые явления: смех, кашель, причмокивания и др.

В целом, при использовании корректно размеченной, сбалансированной базы, качество синтезируемой речи можно субъективно охарактеризовать как очень хорошее. Однако оно не является постоянной величиной. В какой-то степени такое поведение заложено в самой технологии: когда на выходе образуются немодифицированные фрагменты непрерывной речи, качество

будет соответствовать записям базы. С другой стороны, в базе просто может не быть хороших соответствий спецификации. И в этом случае синтез будет звучать менее естественно, с заметными искажениями.

6.1.5. Основные сложности и ограничения применения метода Unit Selection

Как уже отмечалось выше, качество синтеза методом Unit Selection в большой степени зависит от качества используемой речевой базы. Одним из ключевых факторов является размер базы. Чем больше размер базы, тем больше имеется вариантов для синтеза, тем выше вероятность гладкой стыковки фрагментов. С другой стороны, с увеличением базы возрастают затраты на вычисление стоимостей связи и замены, поэтому для устройств с ограниченными вычислительными ресурсами приходится идти на компромисс между производительностью и качеством.

6.2. Сглаживание энергетической огибающей

На данном этапе происходит выравнивание энергетической огибающей полученной звуковой последовательности. В силу ограниченности звуковой базы данных довольно часто возникают ситуации, когда один звук гораздо громче или гораздо тише соседнего. Данные разногласия в амплитуде будут восприниматься слушателем как неестественные артефакты. Пример такой ситуации представлен на рис. 6.3.

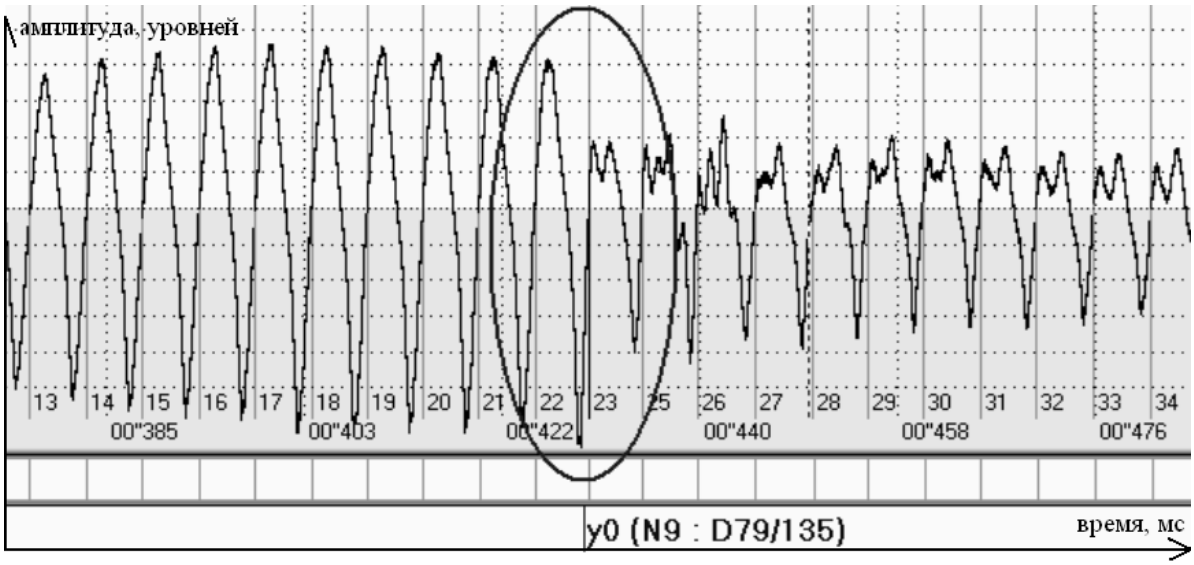

Рис. 6.3. Нарушение энергетической гладкости сигнала

Исправление подобных ситуаций происходит путём плавного приведения амплитуды более громкого звука к более тихому. Результатом работы данного этапа для примера, представленного на рис.6.3, будет следующий звуковой сигнал (рис.6.4).

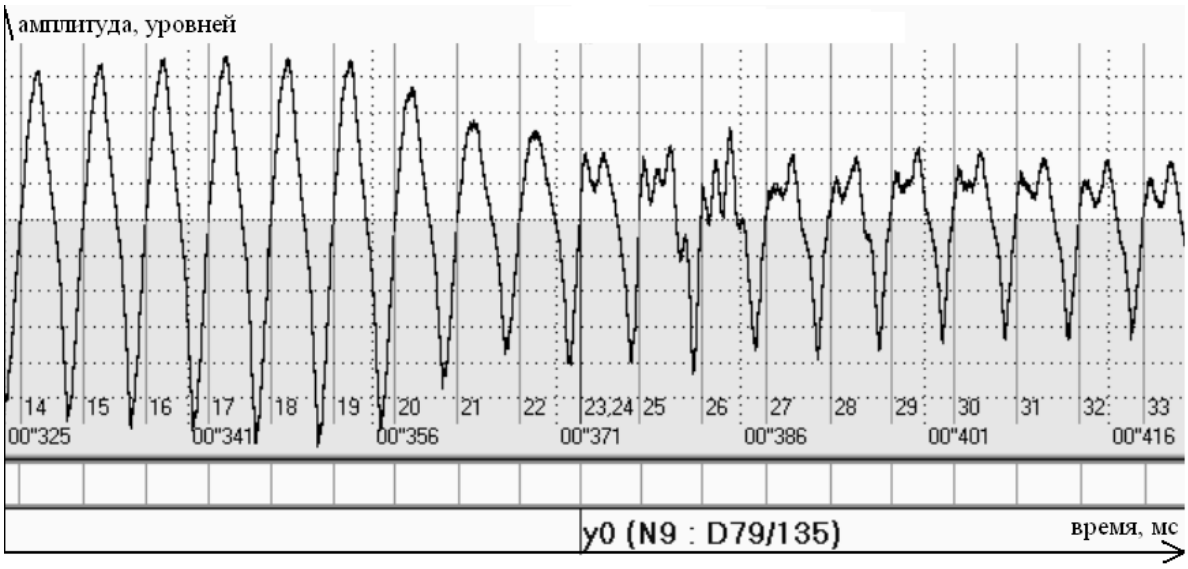

Рис. 6.4. Обеспечение энергетической гладкости

6.3. Модификация звуковых элементов

На данном этапе происходит исправление темпо-ритмических и частотных артефактов [17], проявляющихся в нарушениях плавности интонационной огибающей сигнала и ритмических соотношений между элементами в звуковой последовательности, которые также воспринимаются слушателем, как неестественные образования в потоке речи.

Корректировка происходит путём модификации длительности и частоты основного тона отдельных звуковых единиц, длительность или частота основного тона которых, выходит за границы предсказанного допустимого коридора для данной конкретной фразы, интонационной модели и диктора.

6.3.1.Алгоритм TD-PSOLA

Широко распространены алгоритмы, работающие во временной области, наиболее популярным из которых является технология TD-PSOLA (Time-Domain Pitch-Synchronous-Overlap-Add) [18]. Данный алгоритм работает периодосинхронно, т.е. каждый обрабатываемый фрагмент представляет собой один период. Обязательным условием для этого является возможность определить частоту основного тона сигнала с высокой точностью, т.к. от этого напрямую зависит качество работы этого алгоритма. Границами периодов основного тона служат места закрытия гортани. Далее сигнал разбивается на фрагменты, взвешенные окном Хеннинга, которое захватывает два соседних периода с перекрытием в один период, как показано на рис. 6.5.

Эти взвешенные фрагменты затем могут быть перекомбинированы путём перемещения их центров и наложением с добавлением перекрывающихся частей (отсюда и название, **overlap and add** – перекрытие и добавление). Несмотря на то, что после выполнения данных операций, форма результирующего сигнала становится не в точности такой, какая была

прежде, процедура перекрытия с добавлением позволяет получить достаточно близкий результат, что бы различия не были заметны.

Непосредственная модификация частоты основного выполняется путём распределения полученных взвешенных фреймов на новые значения частоты, предоставляющей собой множество расстояний между окнами им соответствующее. Для примера рассмотрим участок речи с частотой основного тона 100Гц, границы периодов буду лежать с интервалом в 10мс. Взяв эти периоды за основу, проанализируем их и разделим на описанные выше периодосинхронные фрагменты, взвешенные окнами Хеннинга. Далее создадим новое множество периодов, границы которых буду располагаться ближе друг к другу, скажем через каждые 9мс. Далее, если перераспределить подготовленные фреймы путём перекрытия с наложением, мы получим сигнал, который будет иметь частоту основного тона, равную 1.0/0.009 = 111Гц. Если производить обратную операцию – создать множество периодов, границы которых будут располагаться дальше друг от друга, и перераспределить фреймы с перекрытием, мы получим синтезированный сигнал с более низкой частотой основного тона. Процедура уменьшения частоты основного тона частично объясняет причину использования двух периодов во взвешенных фреймах; это делается для того, чтобы не оставалось пустых мест в результирующем сигнале при увеличении расстояния между центрами фреймов.

При сохранении длительности фонограммы, в целом слушатели не замечают неестественностей в сигнале при небольших модификациях частоты основного тона.

Когда алгоритм применяется для модификации хорошо размеченной на периоды основного тона речи, качество его работы чрезвычайно высоко, и пока степень изменения частоты основного тона не слишком значительна (скажем +/- 10% от оригинала), качество речи может быть «идеальным», в том смысле, что слушатель не может заметить в речи какой-то неестественности. С точки зрения вычислительной нагрузки на аппаратные ресурсы, сложно представить какой-либо алгоритм, работающий быстрее. Поэтому зачастую TD-PSOLA рассматривается как приемлемое решение для проблемы модификации частоты основного тона. Однако, конечно алгоритм не идеален во многих ситуациях, не потому, что он не выполняет поставленную задачу, а потому, что на практике, как минимум, нам приходится модифицировать частоту основного тона более чем на 10%, например, чтобы гарантировать гладкость интонационного контура в синтезированной речи в случаях отсутствия звуковых элементов с требуемой частотой в базе данных. Так же, работая во временной области, он вносит неконтролируемые искажения в сигнал и, при уменьшении частоты основного тона, существенно редуцируется энергия на границах "склеек" фреймов.

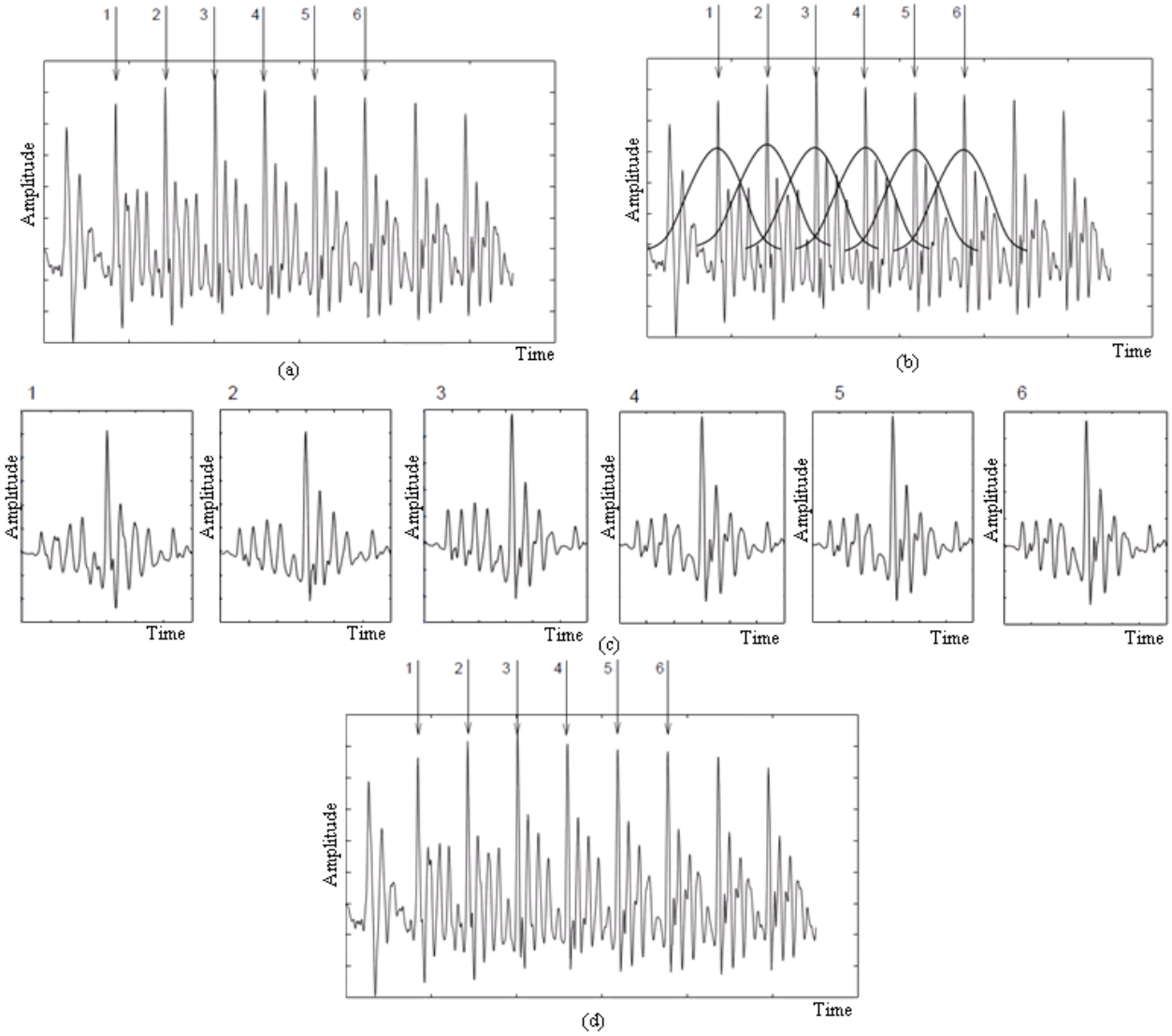

Рис.6.5. Основные операции алгоритма PSOLA:

(a) участок вокализованного сигнала, размеченный на периоды основного тона, (b) взвешивающие окна Хеннинга, центрированные на каждом периоде. (c) полученная последовательность пар периодов после процедуры взвешивания окном (d) ресинтезированный путём перекрытия с добавлением сигнал

6.3.2. Алгоритм SPECINT (Spectrum Interpolation)

В связи с психоакустическими эффектами малейшие искажения в относительном положении формант, изменения огибающей основного тона, ведут к побочным эффектам, из-за которых речь становится неестественной, непривычной для нашего восприятия, как следствие человек при её прослушивании быстро утомляется и не может длительное время внимательно её воспринимать. Вследствие этого одним из основополагающих действий является получение огибающей основного тона исходного сигнала и её воспроизведение на сигнале новой длины.

Немаловажно сохранение энергетической огибающей, поскольку при увеличении или уменьшении частоты основного тона появляются неизбежные её искажения, что также приводит к снижению естественности речи.

Перед тем как понизить, или повысить основной тон, увеличить, или уменьшить длительность, необходимо получить значения основного тона на всём модифицируемом участке. При модификации изменить требуемые характеристики аллофонов так, чтобы траектория основного тона осталась прежней, т.е. измениться должен только масштаб (частоты и времени), иначе при малейшем изменении спектральной картины мы услышим режущие слух, новые интонации в речи даже при незначительных модификациях.

Для этого анализируется сигнал с целью получения вектора значений частоты основного тона на всём его протяжении. В системе синтеза русской речи это аллофон. То есть на каждом периоде аллофона вычисляется значение его основного тона, заполняется некоторый массив данных (вектор значений). Далее полученная огибающая изменяется по тону (поднимается или опускается), затем путём сплайн-интерполяции она растягивается или сжимается на требуемую длину. В итоге получаем модель аллофона после модификации, под которую мы должны модифицировать исходный аллофон.

Модификация сигнала посредством периодосинхронного дискретного преобразования Фурье

Модификация сигнала под требуемую модель происходит следующим образом [19]. Каждый период модифицируется под параметры, смоделированные выше. Рассмотрим этот процесс на примере некоторого периода. Путём дискретного преобразования Фурье получаем спектр сигнала, рассматриваем отдельно вещественные и мнимые его составляющие (рис. 6.6 и рис. 6.7 соответственно).

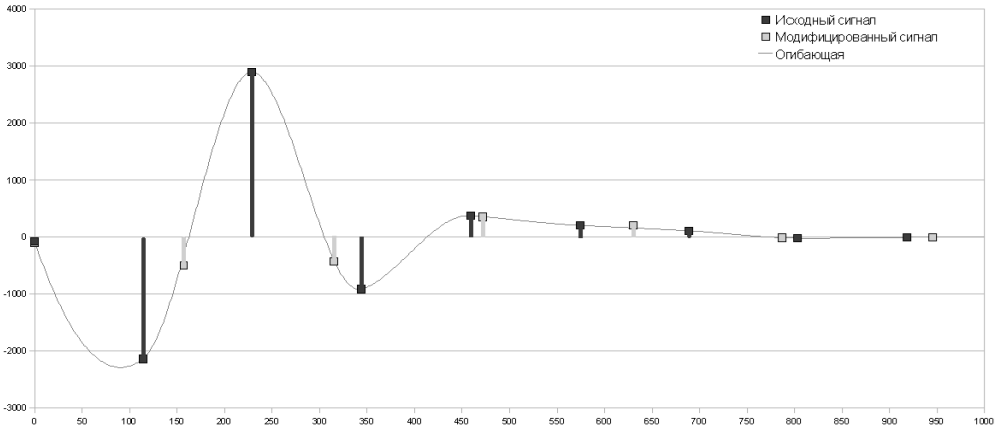

Рис.6.6. Вещественная часть сигнала после ДФП (до и после интерполяции)

Очевидно, что в спектральной области мы получим пики на частотах, кратных частоте периода. Далее мы интерполируем пики на весь диапазон частот, равный половине частоты дискретизации, и вычисляем значения сплайнов в точках, соответствующих пикам нового периода. Далее, выполнив

обратное дискретное преобразование Фурье, мы получим период с требуемой частотой.

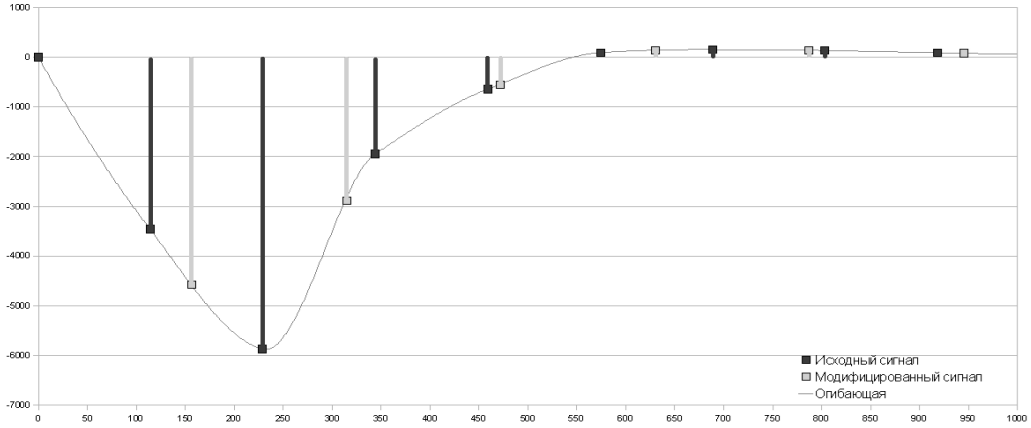

Рис.6.7. Мнимая часть сигнала после ДФП (до и после интерполяции)

Однако при таком подходе без дополнений мы не можем контролировать амплитуду результирующего сигнала. Точнее огибающая амплитуды у нас сохранится, но абсолютное её значение будет отличным от исходного, что сделает сигнал громче или тише, т.к. этот параметр напрямую зависит от того, повышается или понижается основной тон. С увеличением частоты основного тона амплитуда уменьшается, с уменьшением — увеличивается.

Для сохранения исходных величин амплитуды вычисляется нормирующий коэффициент, на который домножаются значения коэффициентов вещественной и мнимой части. В результате получаются пики, находящиеся на огибающей, которая нормирована таким образом, чтобы после обратного ДФП получились те же значения амплитуд, как и в исходном сигнале (рис. 6.8 и рис. 6.9).

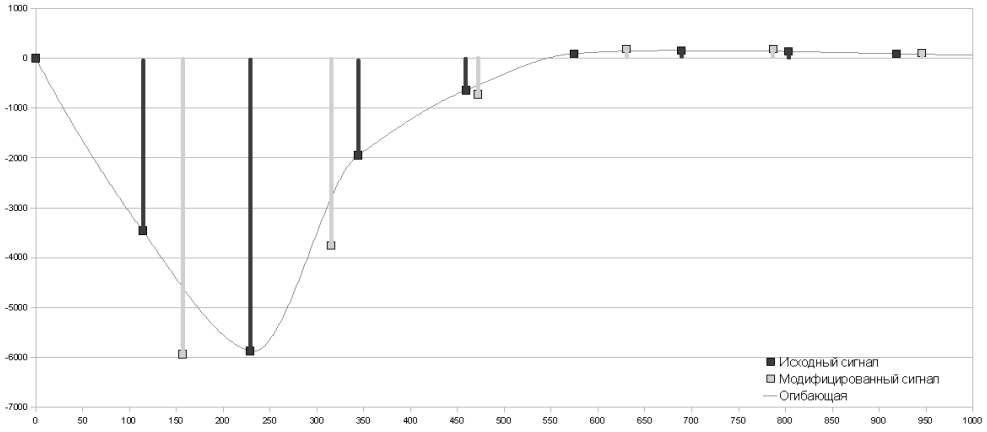

Рис.6.8. Вещественная часть спектра сигнала после ДФП (до и после интерполяции с нормировкой)

Спектры мощности сигнала до и после модификации отображены на рис.6.10. Из рисунка легко заметить, что период был модифицирован примерно со 115Гц на 155Гц. Его поведение во временной области показано на рис. 6.11.

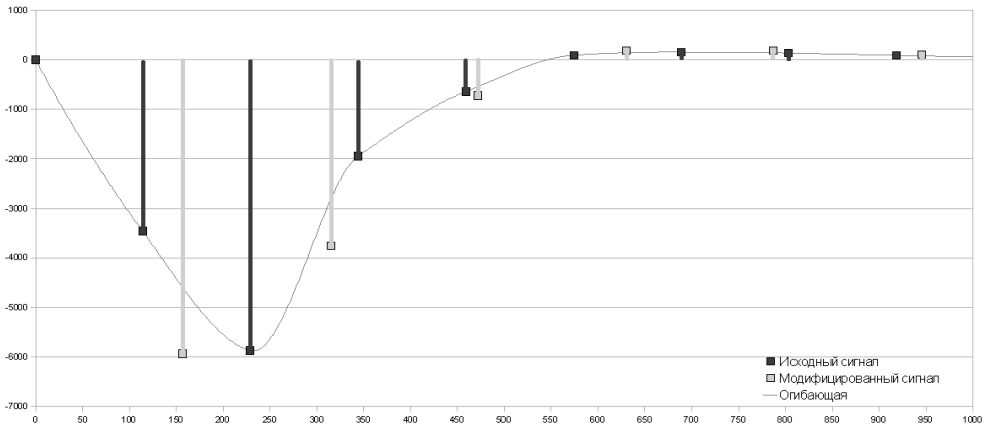

Рис.6.9. Мнимая часть спектра сигнала после ДФП (до и после интерполяции с нормировкой)

Со всеми остальными периодами сигнала производится аналогичные действия.

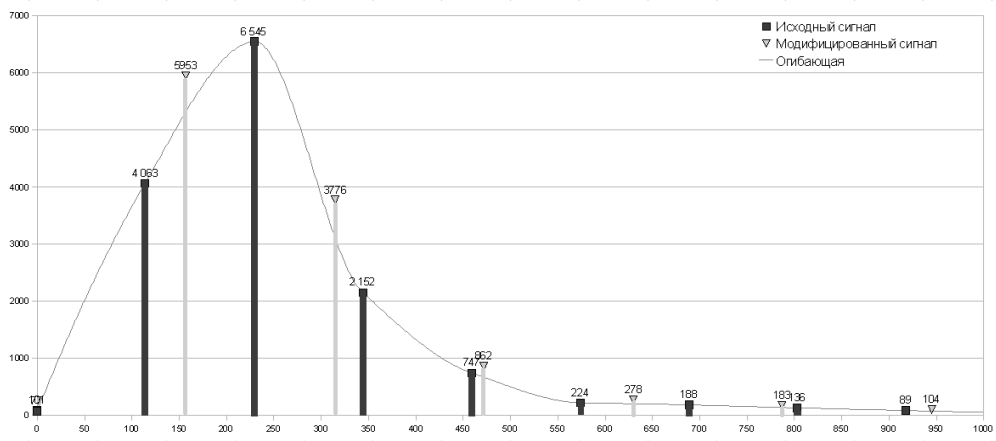

Рис.6.10. Спектры мощности исходного и модифицированного сигнала

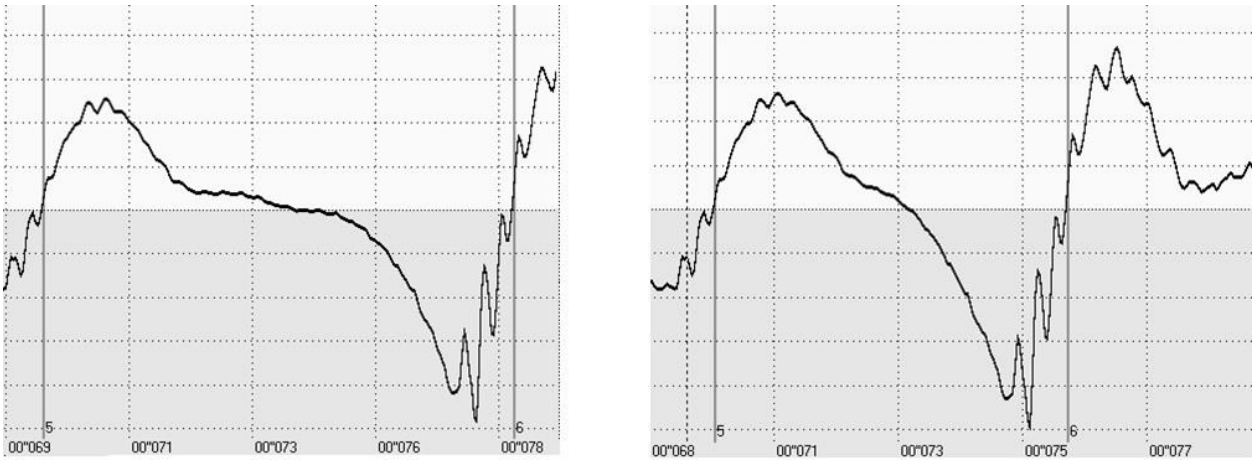

Рис.6.11. Один период во временной области (слева — исходный, справа — после модификации)

Модификация длительности

Изменения основного тона приводят к изменению длины аллофона, звук которого подвергается модификации. Это обуславливает потерю

естественности, диктор начинает говорить то быстрее, то медленнее. Подобное явление нередко возникает и при компилятивном синтезе. В таких случаях также необходимо исправлять длительность аллофонов.

Повышение основного тона периодического сигнала по вышеописанному алгоритму уменьшает его длительность. Для её восстановления обычно используется повтор периодов сигнала. При этом необходимо избежать возникновения двух основных дефектов снижающих качество синтезированного сигнала. Первый связан с тем, что при каждом повторе сбивается фаза сигнала (рис.6.12), что выражается в характерном потрескивании при воспроизведении сигнала. Второй связан с тем, что многократное повторение одного периода человеческое ухо воспринимает как гудение или звон.

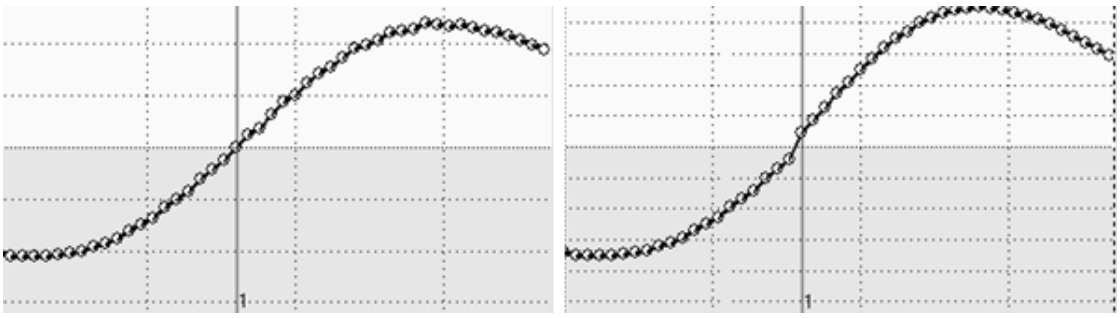

Рис.6.12. Слева - пример корректной стыковки периодов, справа - пример некорректной стыковки периодов

В [19] была предложена и реализована следующая схема повтора периодов. Пусть $L_n, n = 0,\ldots N$ — массив значений левого периода (рис.6.13), а $R_p, p = 0,\ldots P$ — массив значений правого периода (рис.6.14).

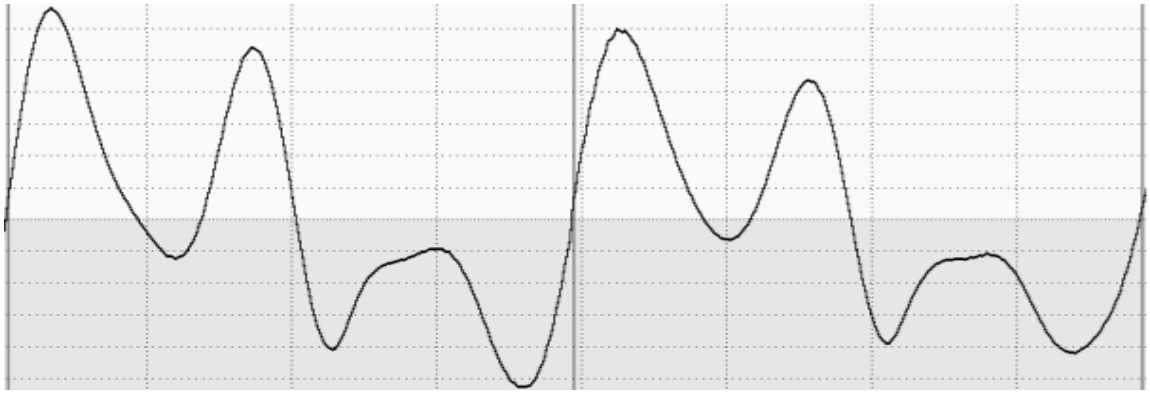

Рис.6.13. Периоды до удлинения сигнала.

Тогда массив значений нового периода, который необходимо вставить между правым и левым определяется суперпозицией:

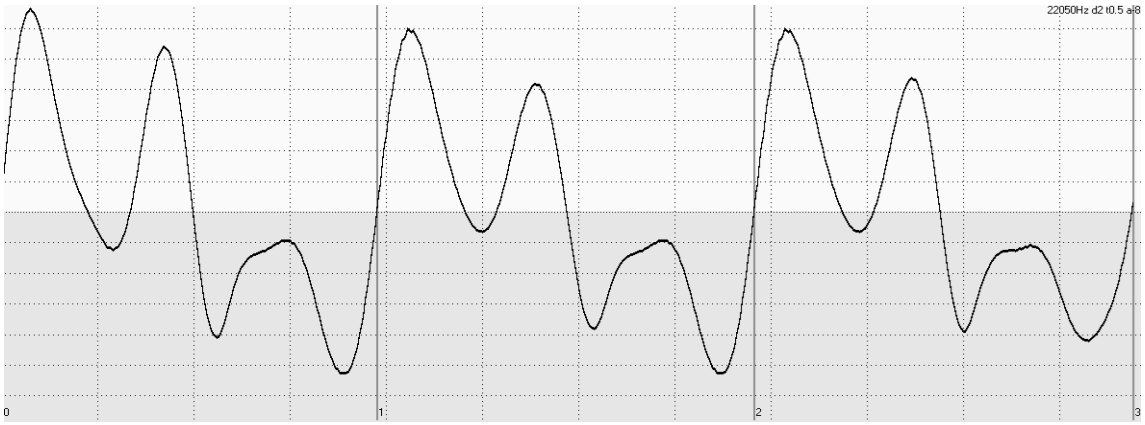

Рис.6.14. Периоды после удлинения аллофона. Добавленный период в центре

$$M_k = TR_k + TL_k, k = 0 \ldots K, \quad \begin{cases} TL_k = 0, k = 0..G \\ TL_k = L_{k-G} W^1_{k-G}, k = G..K \end{cases}, \quad \begin{cases} TR_k = 0, k = F..K \\ TR_k = R_k W^2_k, k = 0..F \end{cases},$$

где

$$W^1_n = \frac{1}{2}\left(1 - \cos\left(\frac{\pi n}{S}\right)\right), n = 0 \ldots S, \quad W^2_n = \frac{1}{2}\left(1 + \cos\left(\frac{\pi n}{F}\right)\right), n = 0..F,$$

$$G = \max(0, K - N), \quad S = \min(N, K), \quad F = \min(P, K).$$

Полученный в результате период (рис. 6.11) идеально стыкуется по фазе как с правым, так и с левым своим соседом, при этом не является повтором ни того, ни другого, что и является необходимым нам решением вышеописанных дефектов. Алгоритм пригоден и для многократного повторения, если его применить последовательно для вставки каждого нового периода.

Соответственно, понижение высоты основного тона периодического сигнала вышеизложенным алгоритмом увеличивает его длительность. В этом случае, для компенсации (уменьшении длительности) также важно избежать сбоя фазы, поэтому используется аналогичный подход, массив значений нового периода, который необходимо вставить вместо имеющихся двух определяется суперпозицией:

$$M_k = TR_k + TL_k, k = 0 \ldots K, \quad \begin{cases} TR_k = 0, k = 0 \ldots G \\ TR_k = R_{k+P-K} W^1_{k-G}, k = G \ldots K \end{cases}, \quad \begin{cases} TL_k = 0, k = F..K \\ TL_k = L_k W^2_k, k = 0..F \end{cases},$$

где

$$W^1_n = \frac{1}{2}\left(1 - \cos\left(\frac{\pi n}{S}\right)\right), n = 0 \ldots S, \quad W^2_n = \frac{1}{2}\left(1 + \cos\left(\frac{\pi n}{F}\right)\right), n = 0 \ldots F,$$

$$G = \max(0, K - P), \quad S = \min(P, K), \quad F = \min(N, K).$$

6.3.3. Алгоритм LP-PSOLA

Данный подход [20] комбинирует в себе основные идеи методов TD-PSOLA и SPECINT. Применяется LP модель, изображённую на рис. 6.15, для получения сигнала ошибки $e[n]$.

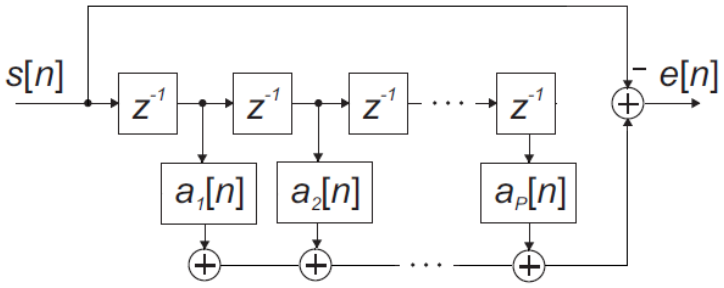

Рис.6.15. Структурная схема блока LP фильтра

Далее он модифицируется методом, представленным в разделе 6.3.1. И в заключение, полученная модифицированная функция ошибки предсказания $e'[n]$ используется для восстановления исходного сигнала с новой частотой основного тона.

Формула для вычисления $e[n]$ дана в (6.5):

$$e[n] = s[n] - \bar{a}^T \cdot \bar{s}[n-1], \tag{6.5}$$

где

$$\bar{s}[n-1] = \left[s[n-1], s[n-2], \ldots, s[n-P] \right]^T, \bar{a} = [a_1, a_2, \ldots, a_P]^T.$$

Результирующий вектор коэффициентов линейного предсказания вычисляется по формуле

$$\bar{a}_n = \bar{R}^{-1}[n-1]\bar{p}[n], \tag{6.6}$$

где

$$\bar{R}^{-1}[n-1] = \sum_{i=0}^{n} \bar{s}[i-1]\bar{s}[i-1], \quad \bar{p}[n] = \sum_{i=0}^{n} \bar{s}[i-1]s[i].$$

Значения $\bar{p}[n]$ в выражении (6.6) можно вычислить рекурсивно, для того чтобы избежать дополнительных накладных вычислительных расходов, как показано в формуле (6.7):

$$\bar{p}[n] = \bar{s}[n-1]s[n] + \bar{p}[n-1] \tag{6.7}$$

Далее, полученные LP коэффициенты на этапе анализа сигнала, применяются в обратном LP фильтре к модифицированной функции ошибки $e'[n]$ для того, чтобы получить модифицированный сигнал $s'[n]$ с желаемой частотой основного тона, как показано в (6.8):

$$s'[n] = e'[n] + \bar{a}^T \cdot \bar{s}'[n-1] \tag{6.8}$$

Общая схема алгоритма представлена на рис. 6.16. Модифицированную функцию ошибки $e'[n]$ можно получить из $e[n]$, используя алгоритм TD-PSOLA, как показано далее. LP модель определяется для каждого отсчёта

сигнала n, что позволяет добиться плавных переходов с соседними моделями. Качество результирующей модели зависит от выбора порядка её порядка P.

После правильного определения меток частоты основного тона $p_m[n]$ и периодов частоты основного тона $p[n]$ в исходном сигнале, контур частоты основного тона может быть модифицирован желаемым образом. С этой целью определяются новые метки $p'_m[n]$, соответствующие значениям новым периодом основного тона $p'[n]$ так, что

$$p'[n] = \beta[n]\, p[n],$$

где $\beta[n]$ это степень модификации периода основного тона, который может быть различным для естественной просодической модификации, автоматической коррекции частоты основного тона и так далее.

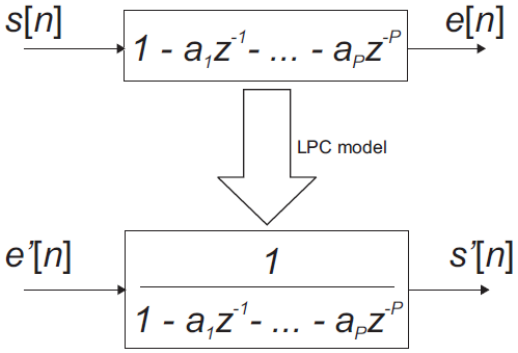

Рис.6.16. Схема анализа и синтеза нового сигнала с использованием LP-модели

Новые метки частоты основного тона $p'_m[n]$ определяются путём добавления в интервал $p'[n]$ отсчётов между двумя соседними метками так, что метка частоты основного тона будет перемещена в позицию $n + p'[n]$, если n содержит метку (т.е. $p'_m[n + p'[n]] = 1$, если $p'_m[n] = 1$, где позиция метки частоты основного тона равна 1). На следующем шаге необходимо соединить каждую новую метку частоты основного тона $p'_m[n]$ с соответствующим ей ближайшим пиком в оригинальном сигнале $p_m[n]$. Это делается путём непосредственного сравнения временных индексов $p_m[n]$ и $p'_m[n]$, как показано на рис. 6.17.

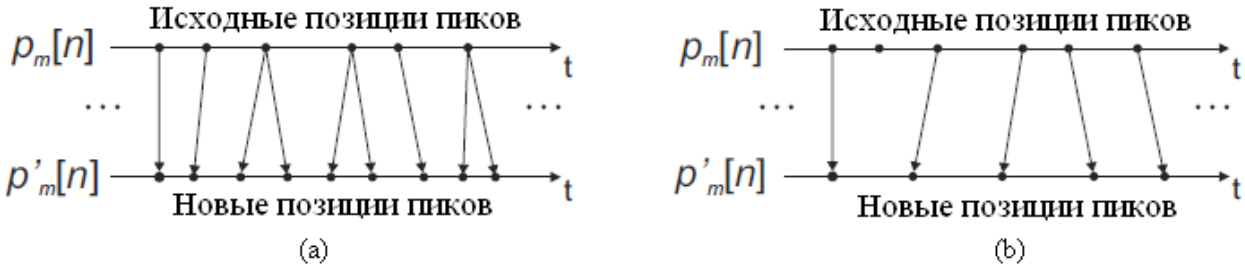

Рис.6.17. Распределение меток частоты основного тона при модификации сигнала: (a) увеличение частоты, (b) уменьшение частоты

В заключение, для генерации итоговой функции ошибки, сигнал разбивается на взвешенные окном Хеннинга фрагменты по парам периодов с перекрытием в один период, т.е. для каждой метки, начиная с предыдущей и заканчивая следующей. Данные сегменты соединяются согласно процедуре перекрытия с наложением так, чтобы соответствовать новым периодам частоты основного тона $p'[n]$, полученным ранее, как показано на рис. 6.18.

Данный подход так же обладает рядом проблем: аппаратная сложность вычисления LP коэффициентов и обратного LP-фильтра, в сигнале возбуждения остаётся информация о голосовом тракте, ведущая к появлению непонятных тембральных артефактов при восстановлении сигнала по модифицированной функции ошибки.

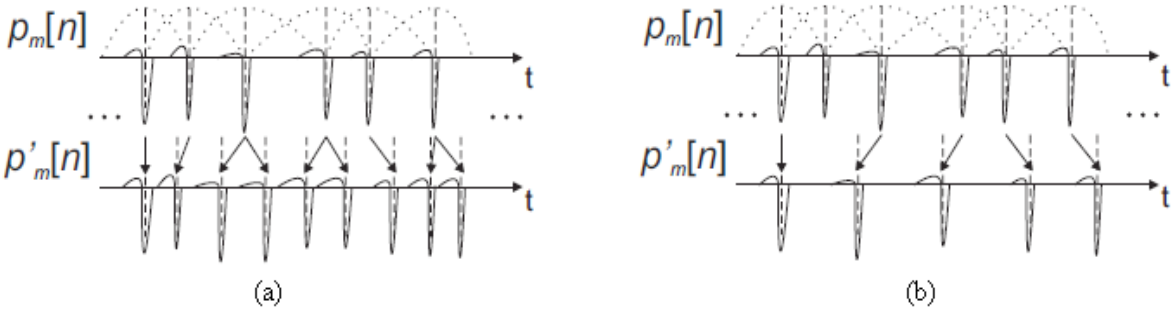

Рис.6.18. Получение модифицированного сигнала: (a) увеличение частоты, (b) уменьшение частоты

6.3.4. Экспериментальные сравнения

Рассмотрим ряд экспериментов модификации частоты основного тона представленными алгоритмами и сравнение результатов их работы.

Пример 1. Участок синтезированной женским голосом речи был модифицирован с $\beta[n]=2$ и $\beta[n]=0.5$. Рис.6.19 демонстрирует небольшие участки исходного и модифицированного сигналов методами TD-PSOLA, SPECINT и LP-PSOLA. На рис.6.20 приведены им соответствующие спектрограммы:

a) исходный сигнал,

b) модифицированный сигнал методом LP-PSOLA с $\beta[n]=2$,

c) модифицированный сигнал методом SPECINT с $\beta[n]=2$,

d) модифицированный сигнал методом TD-PSOLA с $\beta[n]=2$,

e) модифицированный сигнал методом LP-PSOLA с $\beta[n]=0.5$,

f) модифицированный сигнал методом SPECINT с $\beta[n]=0.5$,

g) модифицированный сигнал методом TD-PSOLA с $\beta[n]=0.5$.

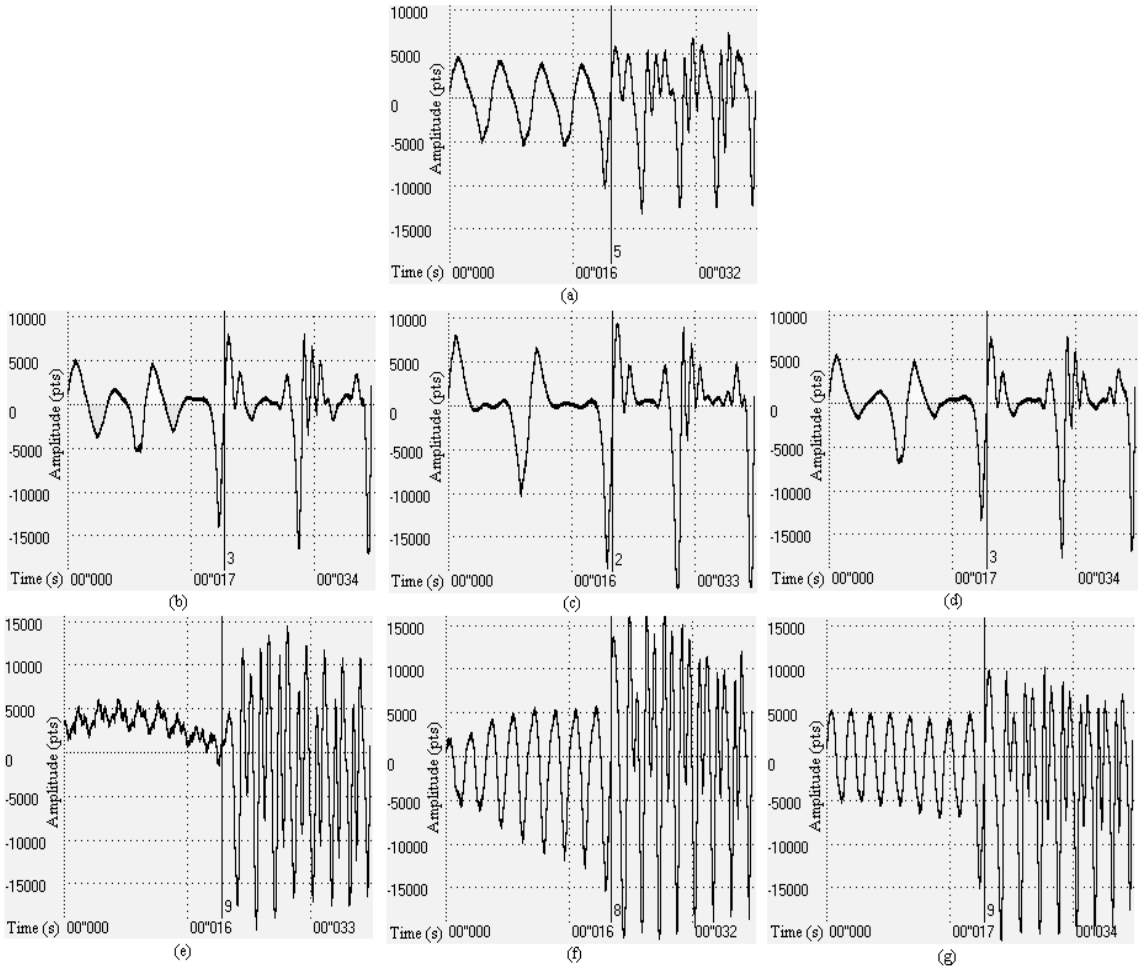

Рис.6.19.Фрагменты сигналов в примере 1

Исходя из данных на рис.6.19, можно сделать вывод, что все алгоритмы работают корректно - пики периодов основного тона становятся дальше или ближе друг от друга при модификации с $\beta[n]=2$ и $\beta[n]=0.5$ соответственно.

Аналогичный вывод можно сделать и проанализировав частотные полосы, представленные спектрограммами на рис.6.20. Здесь:

a) исходный сигнал,
b) модифицированный сигнал методом LP-PSOLA с $\beta[n]=2$,
c) модифицированный сигнал методом SPECINT с $\beta[n]=2$,
d) модифицированный сигнал методом TD-PSOLA с $\beta[n]=2$,
e) модифицированный сигнал методом LP-PSOLA с $\beta[n]=0.5$,
f) модифицированный сигнал методом SPECINT с $\beta[n]=0.5$,
g) модифицированный сигнал методом TD-PSOLA с $\beta[n]=0.5$.

Пример 2. Вновь характеристики частоты основного тона были модифицированы с $\beta[n]=2$ и $\beta[n]=0.5$. Однако в данном примере использовалась речь, синтезированная мужским голосом. Вид результатов

во временной и частотной области практически аналогичен тем, что изображены на рис.6.19 и рис.6.20 соответственно.

Так же, как и в примере 1, по результатам не сложно определить, что частота основного тона модифицирована должным образом. В то время как спектральная огибающая осталась неизменной.

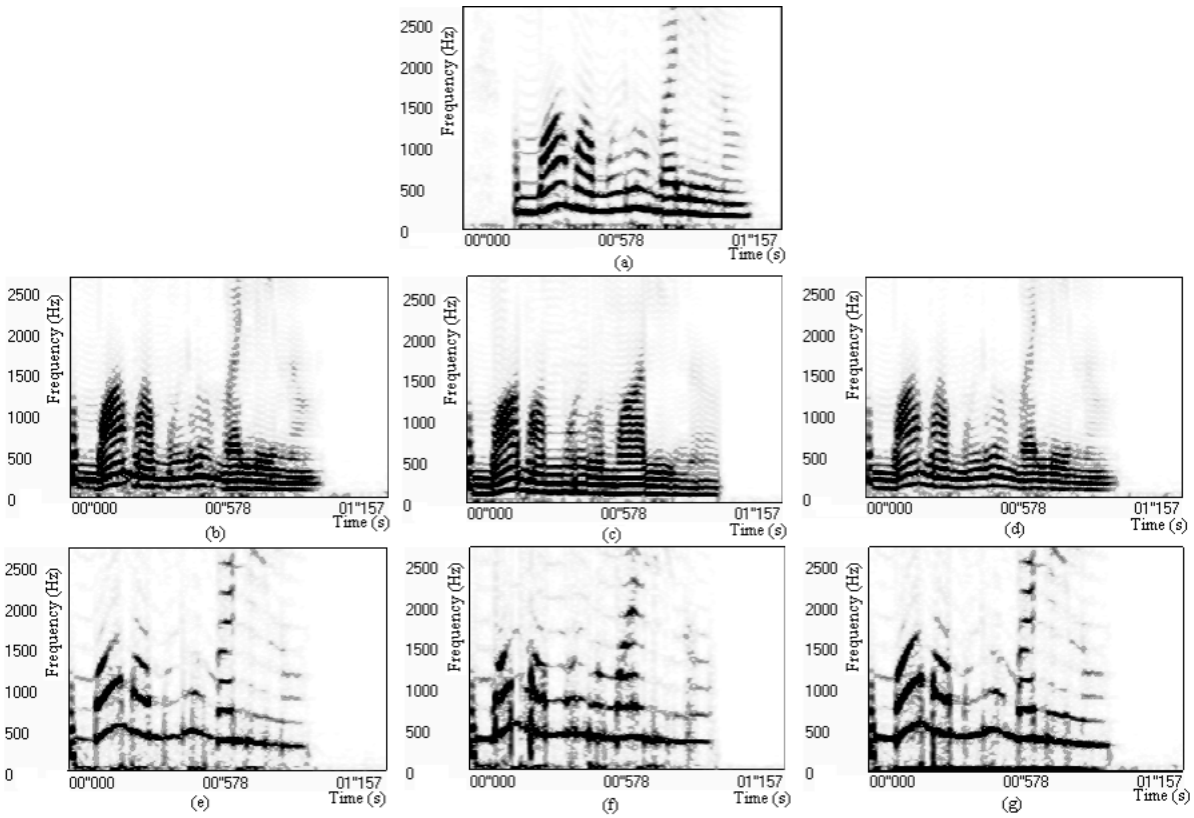

Рис.6.20. Спектрограммы сигналов в примере 1

Пример 3. На рис.6.21 приведено сравнение результатов представленных методов во временной области с $\beta[n]=2$ на сигнале из предыдущего примера 1. Здесь:

a) исходный сигнал,
b) модифицированный сигнал методом TD-PSOLA,
c) модифицированный сигнал методом SPECINT,
d) модифицированный сигнал методом LP-PSOLA.

Данный пример иллюстрирует значительные недостатки алгоритма TD-PSOLA, которые заключаются в существенной редукции энергии сигнала между соединяемыми пиками периодов частоты основного тона при $\beta[n]>1$. Хотя большие окна анализа могли бы исправить эту проблему, они могли бы стать причиной появления ложных пиков в модифицированном сигнале, т.к. они могут не полностью редуцироваться окнами анализа. А такие ложные пики ведут к грубостям и неестественностям в сигнале.

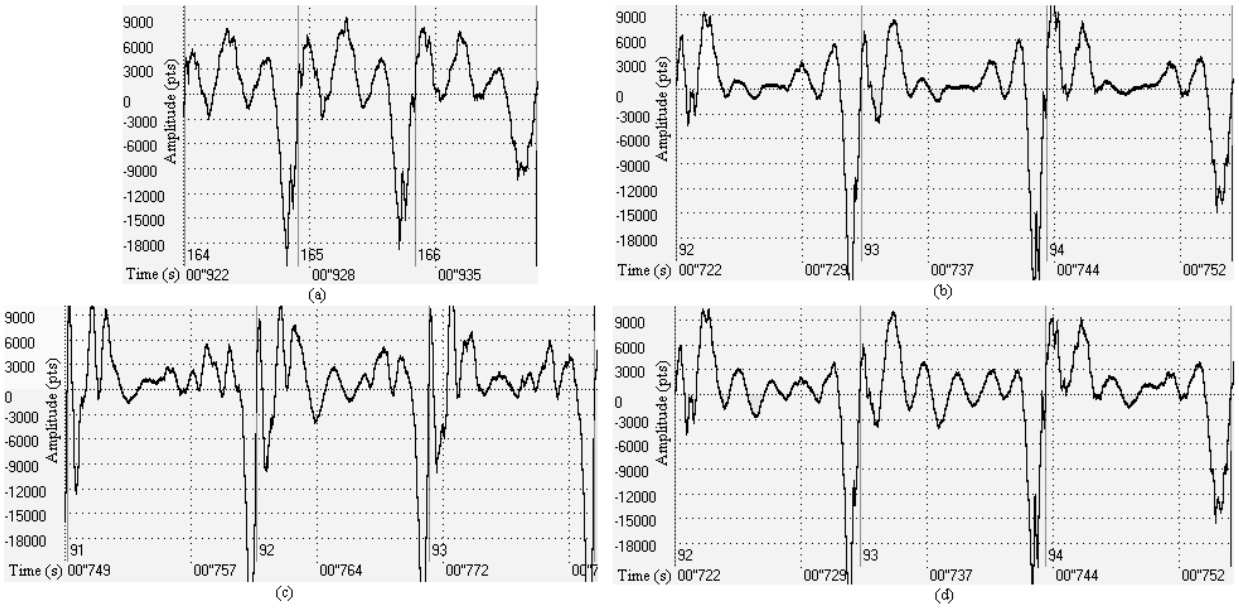

Рис.6.21. Фрагменты сигналов в примере 3

Проведенные эксперименты показали, что наилучшие результаты для модификации частоты основного тона в диапазоне $0.5 \le \beta[n] \le 2$ достигаются при использовании алгоритма LP-PSOLA. Преимущество этого метода заключается в сохранении индивидуальности каждого гортанного импульса. Однако, поскольку в сигнале ошибки частично остаётся формантная структура, метод LP-PSOLA не является совершенным и зависит от порядка LP фильтра.

6.4. Объединение элементов в единый звуковой поток

Рассмотрим обработку данных на данном этапе на примере пары звуковых единиц, не умаляя общности задачи, т.к. на данном этапе происходит стыковка каждой пары звуковых элементов при формировании единого звукового потока.

При конкатенативном синтезе можно выделить следующие типы таких пар, требующих различной логики стыковки:

- оба аллофона взяты из одного файла и идут подряд — в этом случае никакой стыковки вообще не происходит;
- оба аллофона взяты в точных контекстах;
- первый аллофон взят в точном контексте, второй — в близком или дальнем;
- первый аллофон взят в близком или дальнем контексте, а второй — в точном;
- оба аллофона взяты в близких/ом (дальних/ем) контекстах.

Каждый из представленных вариантов можно разделить на следующие типы, в зависимости от вида аллофона:

1) невокализованный — невокализованный (С – С);
2) невокализованный — вокализованный (С – V);
3) вокализованный — невокализованный (V – С);
4) вокализованный — вокализованный (V – V).

Основная задача выполнения данного шага - обеспечить плавность переходов (границ) между аллофонами. Достигается это путём сохранения перехода, который является естественным для данной звуковой единицы и плавным наложением одной звуковой единицы на другую. Схематично данные операции представлены на рис.6.22 и 6.23 для вариантов стыковки типа "точный контекст — точный контекст», «близкий (дальний) контекст — точный контекст", "точный контекст — близкий (дальний) контекст" соответственно.

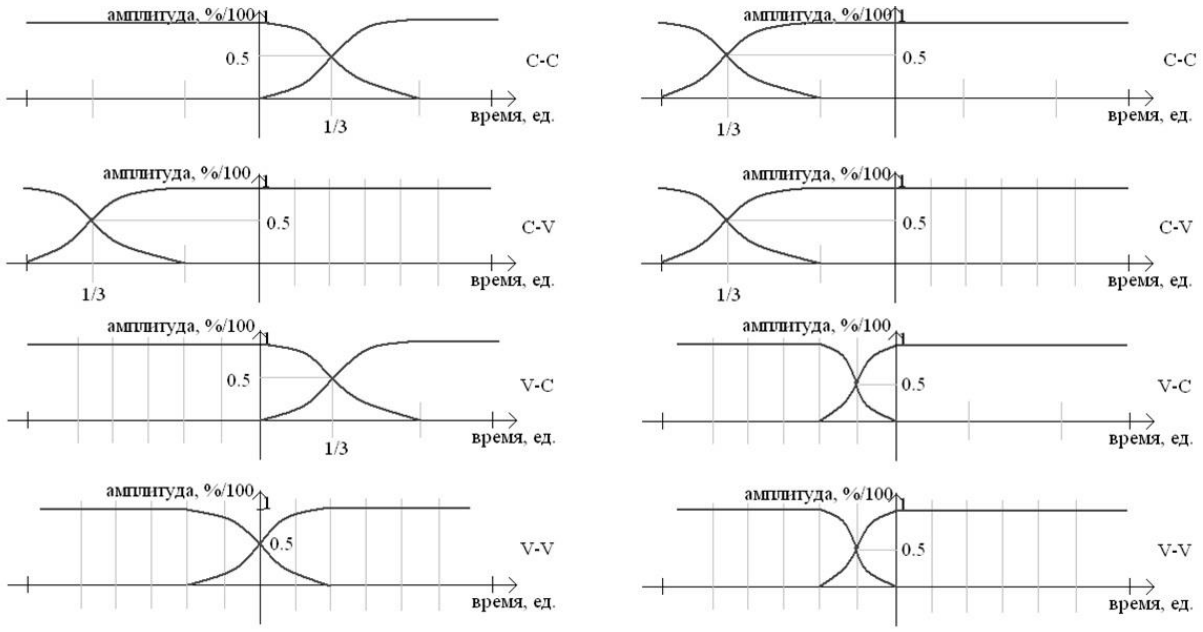

Рис.6.22. Слева - стыковка типа "точный контекст — точный контекст", справа - типа "близкий (дальний) контекст — точный контекст"

В вариантах стыковки типа "точный — точный" и "точный — близкий" для контекстов типа С – С, V – С в случае если длина невокализованного аллофона, содержащего "реальный" переход меньше чем 2/3 длины второго (правого) невокализованного аллофона, то перекрытие будет происходить по типу "близкий (дальний) — близкий (дальний)".

При реализации стыковки типа "близкий (дальний) - близкий (дальний) контексты" первый аллофон накладывается на второй на один период в вокализованном случае и на 20мс - в невокализованном.

6.5. Звуковые эффекты, используемые при синтезе речи

Алгоритмы постобработки звука в системах синтеза речи должны удовлетворять следующим требованиям:

- не искажать речевой сигнал;

- иметь минимальный набор регулируемых параметров;
- небольшой временной ресурс реализации.

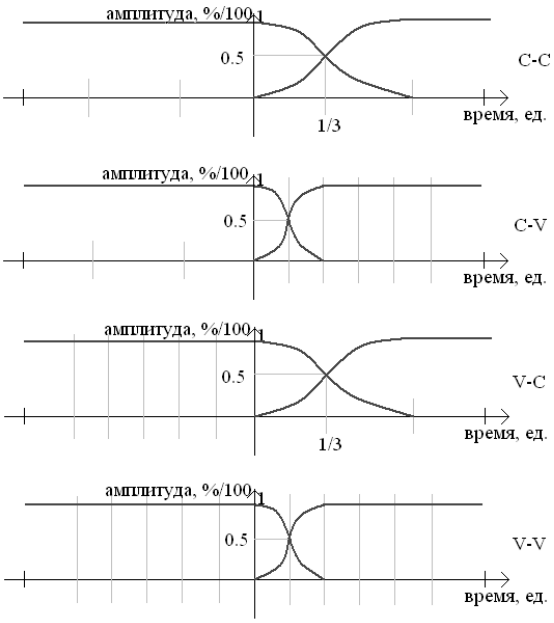

Рис.6.23. Стыковка типа " точный контекст — близкий (дальний) контекст"

6.5.1. Параметрический эквалайзер

Большое распространение в области обработки речевых сигналов имеют устройства, позволяющие изменять тембр сигнала. Это объясняется тем, что частотная характеристика сигнала, особенно после модификации, требует коррекции в той или иной степени. Для выравнивания спектральной картины необходимы частотные корректоры.

Наиболее распространенным частотным корректором является эквалайзер. Он представляет собой многополосный регулятор тембра, позволяющий осуществлять одновременную и взаимонезависимую регулировку усиления или ослабления сигнала сразу в нескольких частотных полосах.

Первые эквалайзеры были довольно простыми устройствами. Как правило, это была одна ручка, при помощи которой можно было вырезать определенную часть высокочастотной составляющей. Первый "серьезный" эквалайзер был изобретен Питером Бэксендалом (Peter J Baxandall). Его эквалайзер предусматривал отдельную регуляцию низких и высоких частот, причем можно было делать как усиление, так и ослабление. Регулятор оставался в своем среднем положении, если эквализации не требовалось.

Эквалайзеры делятся на два больших класса - графические эквалайзеры и параметрические эквалайзеры. Первые из них являются многополосными регуляторами тембра с фиксированными полосами частот коррекции. В таких эквалайзерах имеется возможность регулировать только величину подъема и спада амплитудно-частотной характеристики (АЧХ). Параметрические эквалайзеры позволяют регулировать, как ширину, так и усиление/ослабление

сигнала в каждой полосе. Это позволяет получить достаточно вариативную геометрию АЧХ.

При работе с эквалайзером очень важно понимать, что усиление какой-либо частотной полосы приводит к усилению общего уровня аудио сигнала, и чрезмерное усиление полос может зачастую привести к искажениям звукового сигнала. В связи с этим ослабление «ненужных» частот зачастую дает более качественный результат, нежели усиление «нужных». Поэтому эквалайзером следует пользоваться очень аккуратно и не использовать его, если в этом нет очевидной надобности. Важный - диапазон регулировки. Он обычно составляет плюс/минус 12 или 15 децибел. Как правило, больше и не требуется

- **Золотое правило применения эквалайзера** - "лучше меньше, да лучше". Если вам нравится, как звучит речь в необработанном виде, то применение эквалайзера едва ли сделает ее еще лучше, а вот испортить может вполне.
- **Второе правило - серебряное** - звучит так: "Лучше вычитать, а не добавлять".

Простейшим для реализации в системе синтеза речи представляется трехполосных параметрический эквалайзер. Он представляет собой три параллельно включенных фильтра: фильтр нижних частот (ФНЧ), полосовой фильтр (ПФ) и фильтр верхних частот (ФВЧ).

Разбиение спектра на три полосы производится заданием пользователем левой и правой частот: F_L и F_R. В каждой спектральной полосе предусмотрена независимая пользовательская регулировка усиления/ослабления: $\alpha_L, \alpha_C, \alpha_R$.

Возникает вопрос выбора фильтров для реализации эквалайзера.

- БИХ фильтры (фильтры с бесконечной импульсной характеристикой) требуют существенно меньших вычислительных затрат при фильтрации по сравнению с фильтрами с конечной импульсной характеристикой (КИХ фильтрами).
- Расчет КИХ фильтров более сложен, чем расчет БИХ фильтров с аналогичными частотными характеристиками.
- У БИХ фильтров неизбежны фазовые искажения, существенно влияющие на восприятие речевого сигнала. Их легко исключить при использовании КИХ фильтров
- Возможна неустойчивость работы БИХ фильтра.
- КИХ фильтры, реализуемые не рекурсивно, т. е. с помощью прямой свертки, всегда устойчивы.

Данные соображения позволяют сделать выбор в пользу реализации эквалайзера КИХ фильтрами. На рис.6.24 и 6.25 представлены типовые АЧХ (амплитудно-частотная характеристика фильтра) для ФНЧ и полосового фильтра (64 коэффициента):

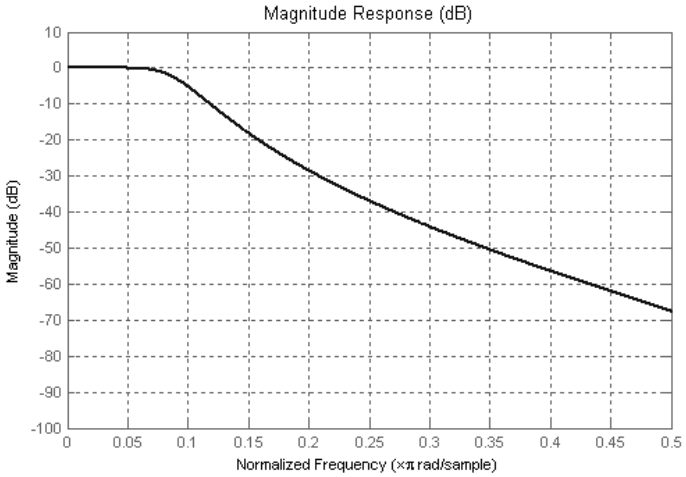

Рис.6.24. АЧХ для фильтра нижних частот эквалайзера

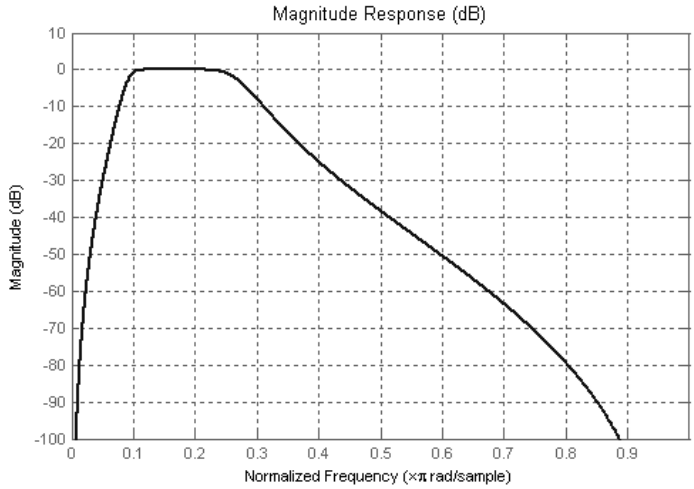

Рис.6.25. АЧХ для полосового (центрального) фильтра эквалайзера

На рис.6.26 представлена схема работы вышеописанного трехполосного эквалайзера для системы синтеза речи.

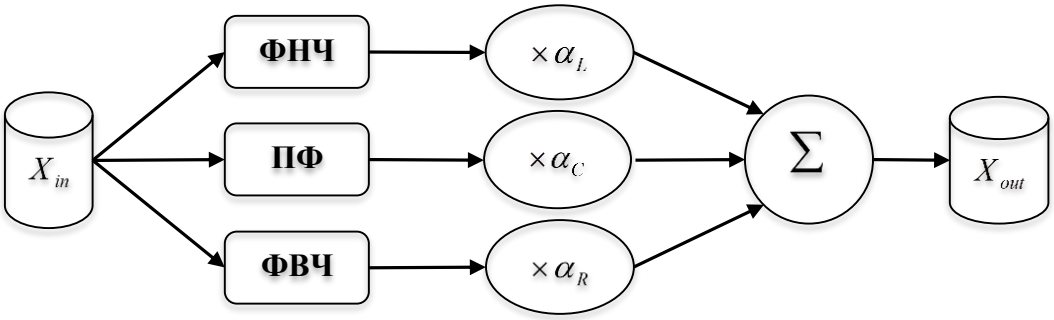

Рис.6.26. Схема работы трехполосного эквалайзера

6.5.2. Ревербератор

Реверберация (от латинского reverberatus, "повторный удар") - это процесс продолжения звучания после окончания звукового импульса или колебания благодаря отражениям звуковых волн от поверхностей. Поэтому

реверберация имеет место только в закрытых помещениях, хотя в особых условиях некоторые ее виды могут иметь место и на открытом пространстве (например, узкое горное ущелье, стадион, городская площадь и т.п.). Различные реализации реверберататоров очень востребованы у специалистов, занимающихся обработкой музыкальных и речевых сигналов. Эффект реверберации проявляется в более сочном гулком объемном звучании, обычно более приятном для восприятия, чем исходный «сухой» звук.

Простейший реверберататор для системы синтеза речи можно реализовать, как набор линий задержки $\Delta_1, \Delta_2, \ldots, \Delta_N$, каждая со своим весовым коэффициентом α_i.

В реальных условиях эффект реверберации наблюдается не сразу - ведь звуковой волне первоначально требуется какое-то время для того, чтобы достигнуть отражающей поверхности и возвратиться обратно. Линии задержки как раз имитируют этот процесс. Типичное значение первой задержки: 20-50 мс.

Перед добавлением к исходному сигналу, выход реверберататора фильтруется ФНЧ фильтром первого порядка. Желаемое соотношение между "сухим" и обработанным сигналом задается коэффициентом применимости реверберататора: $\gamma \in (0,1)$.

Реверберататор обычно может работать в двух режимах: ручной и с использованием готовой конфигурации. Для ручного режима регулируется первая (минимальная) задержка Δ_1, а величины остальных задержек определяются автоматически:

$$\Delta_i = \Delta_1 2^{\frac{1-i}{N}}, 2 \le i \le N$$

Практически, ручной режим реализует так называемый эффект задержки («эффект эхо»).

Режим готовой конфигурации предусматривает использование заранее рассчитанных параметров, имитирующих акустическую модель определенного помещения, например, «закрытая небольшая комната», «зал», «открытое пространство» и т.п. В этом случае все параметры задержек Δ_i и весовых коэффициентов α_i имеют свои уникальные значения. На рис.6.27 приведена схема работы предложенного реверберататора.

7. СИНТЕЗ, ОСНОВАННЫЙ НА МОДЕЛЯХ

Данный тип синтеза является гибридом подходов, основанных на правилах и речевом корпусе. В этом случае происходит описание звуковой базы данных параметрической моделью. Параметры (например, спектральные характеристики, частота основного тона, длительность и т.д.) обобщаются множеством статистических моделей, представляющими собой скрытые марковские модели, которые содержат в себе шаблоны речевых элементов.

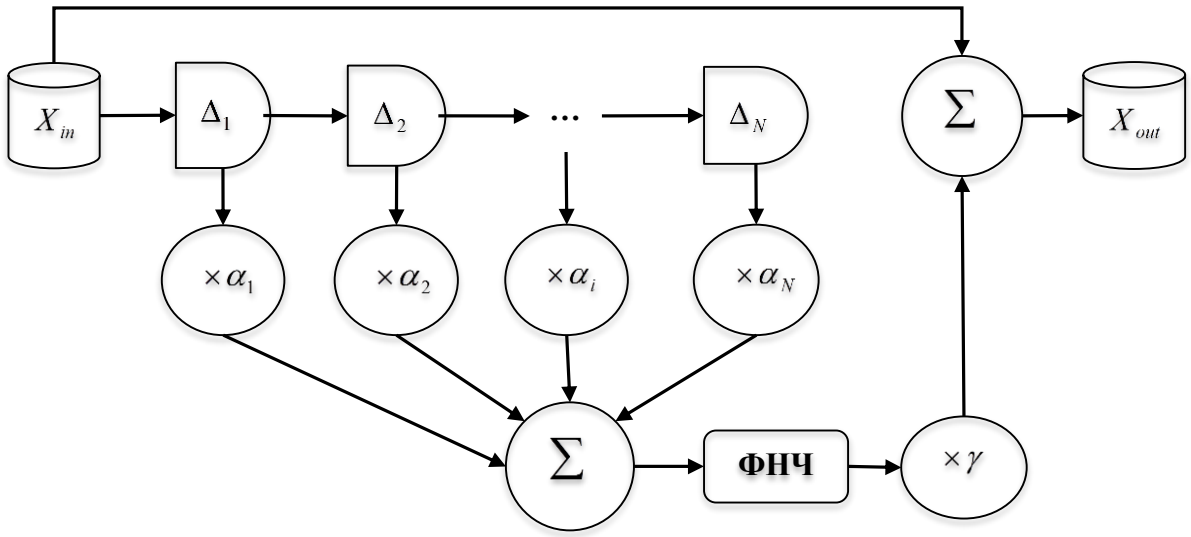

Рис.6.27. Схема работы ревербератора

Определение параметров речевого сигнала происходит на основе критерия максимального правдоподобия применительно к этим моделям. Так для модели, имеющей N состояний с M компонентами на состояние и вектор наблюдений $\mathbf{O} = [o'_1, o'_2, \dots, o'_T]'$, функция правдоподобия наблюдения o_t состояния Sj определяется следующим выражением:

$$p(o_t|q_j = S_j) = \sum_{m=1}^{M} c_{jm} b_{jm}(o_t) = \sum_{m=1}^{M} c_{jm} N(o_t; \mu_{im} \Sigma_{jm}) \quad (7.1)$$

Для модели, вектор наблюдения которой не содержит динамических признаков, т.е. $o_t=c_t$, наблюдения, которые максимизируют $p(O|\lambda)$, являются наиболее вероятной последовательностью:

$$p(O|\lambda) = p(O|Q,\lambda)P(Q|\lambda), \quad (7.2)$$

где Q – последовательность состояний и компонент: Q=(q,i), $q=\{q_1,q_2,\dots,q_T\}$, $i=\{i_1,i_2,\dots,i_T\}$.

После получения Q, максимизация $p(O|\lambda)$ и $p(O|Q,\lambda)$ выполняется следующим образом:

$$\log p(O|Q,\lambda) = \log \prod_{t=1}^{T} b_{q_t,i_t}(o_t) = -\frac{1}{2}(O-\mu)'\Sigma^{-1}(O-\mu) -$$

$$\frac{1}{2}\sum_{t=1}^{T} \log|\Sigma_{q_t}| - \frac{1}{2} TD \log(2\pi), \quad (7.3)$$

где

$$\mu = \left[\mu'_{q_1,i_1}, \mu'_{q_2,i_2}, \dots, \mu'_{q_T,i_T}\right]'; \Sigma = \mathrm{diag}\left[\Sigma_{q_1,i_1}, \Sigma_{q_2,i_2}, \dots, \Sigma_{q_T,i_T}\right],$$

T – длина последовательности векторов наблюдений в количестве фреймов; D – размерность статических векторов параметров.

В выражении 7.3 $\log p(O|Q,\lambda)$ максимален, если его производная равна 0:

$$\frac{\partial(\log p(O|Q,\lambda))}{\partial c} = -\Sigma^{-1}c + \Sigma^{-1}\mu = 0$$

Таким образом, максимум достигается при c = μ, т.е., другими словами, наиболее вероятная получаемая последовательность наблюдений является последовательностью векторов средних, независимо от ковариации Σ.

При введении динамических признаков, вектор наблюдений можно определить, как $o_t = [c_t', \Delta c_t', \Delta^2 c_t']'$, где:

$$\Delta^{(n)} c_t = \sum_{\tau=-L(n)}^{L(n)} \omega^{(n)}(\tau) c_{t+\tau} \quad (n = 0,1,2)$$

При введении нового определения вектора признаков, выражение 7.3 примет следующий вид:

$$\log p(O|Q,\lambda) = (O-\mu)'\Sigma^{-1}(O-\mu) - \frac{1}{2}\log|\Sigma| - \frac{3TD}{2}\log(2\pi)$$

$$= (Wc-\mu)'\,\Sigma^{-1}(Wc-\mu) - \frac{1}{2}\log|\Sigma| - \frac{3TD}{2}\log(2\pi)$$

$$= e(c) - \frac{1}{2}\log|\Sigma| - \frac{3TD}{2}\log(2\pi)$$

где:

- μ и Σ определяются так же, как и в 7.3;
- $e(c) = (Wc-\mu)'\Sigma^{-1}(Wc-\mu)$;
- $W = [\omega_1, \omega_2, \dots, \omega_T]'$;
- $\omega_t = \left[\omega_t^{(0)}, \omega_t^{(1)}, \omega_t^{(2)}\right]$;
- $\omega_t^{(n)} = [0_{M\times M}, \dots, 0_{M\times M}, \omega(n)\left(-L^{(N)}\right)I_{M\times M}, \dots, \omega(0)I_{M\times M},$
$\dots, \omega(n)\left(L^{(N)}\right)I_{M\times M}, 0_{M\times M}, \dots, 0_{M\times M}]$

Минимизация выполняется следующим образом:

$$\frac{\partial \log(O|Q,\lambda)}{\partial c} = 0,$$

$$(W'\Sigma^{-1}W)c - W'\Sigma^{-1}\mu = 0, \quad (7.4)$$

Выражение 7.4 можно представить в следующем виде:

$$Rc = r,$$

где:

$$R = W'\Sigma^{-1}W,\ r = W'\Sigma^{-1}\mu$$

При решении данной задачи напрямую требуется $O(T^3M^3)$ операций, однако существует более быстрый алгоритм, позволяющий существенно снизить вычислительную сложность вычислений.

Синтез речи, основанный на моделях, представлен в системе HTS, являющейся единственной открытой системой. Также подобные системы были реализованы в компании Microsoft и Whistler.

У истоков данного подхода к синтезу речи лежит концепция линейного предсказания значения текущего отсчета на основе линейной комбинации набора предшествующих ему отсчетов. Данная концепция предполагает расчет коэффициентов авторегрессионного фильтра:

$$\frac{\sigma}{A_p(z)}, \qquad A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$$

путем минимизации сигнала остатка, который вычисляется путем вычитания из исходного сигнала сигнала, предсказанного AR фильтром. Минимизация выполняется путем решения уравнения:

$$\sum_{j=1}^{p} a_j \varphi_x(i-j) = -\varphi_x(i), (i = 1, \dots, p),$$

где p – порядок предсказания, соответствующий количеству резонансных частот, порождаемых фильтром, от 0 до половины частоты дискретизации; a_i – коэффициенты линейного предсказания, первый коэффициент $a_0=1$; $\varphi_x(i)$ – значение автокорреляционной функции сигнала x.

Исходный речевой сигнал получается путем подачи сигнала остатка на AR фильтр, параметры которого содержатся в моделях той или иной звуковой единицы.

В основе такого подхода лежит упрощенная модель человеческого вокального тракта, где сигнал возбуждение и форма вокального тракта моделируются независимо. Предполагается, что функция возбуждения бывает двух типов: вокализованная, порождаемая периодическими единичными импульсами, моделирующими открытие гортани; невокализованная, порождаемая белым шумом, моделирующим шумовые завихрения воздуха в гортани. Стоит отметить, что такая замена оригинальной функции возбуждения ведет к существенному снижению естественности синтезированной речи. Далее сигнал возбуждения (вокальная и шумовая составляющие) обрабатывается дополнительными фильтрами, моделирующими вокальный тракт, а именно: гортань (G(z)), носовые и ротовые особенности (V(z)), губы (R(z)), как представлено на рисунке 7.1.

Данный набор фильтров можно представить следующим выражением

$$G(z)V(z)R(z) = \frac{1}{(1-\alpha z^{-1})(1-\beta z^{-1})} \frac{B}{\prod_{k=1}^{K}(1+b_{1,k}z^{-1}+b_{2,k}z^{-2})} c(1-z^{-1})$$

$$\approx \frac{\sigma}{A_p(z)} \quad (1.5)$$

При моделировании вокализованной речи, количество коэффициентов AR фильтра p следует выбирать, основываясь на следующих соотношениях: 2 коэффициента для формирования гортанных искажений, 2 – на каждую

форманту, которые условно распределены через каждые 1000 Гц спектра. Таким образом, например, при частоте дискретизации 22050Гц следует выбрать 24-25 коэффициентов.

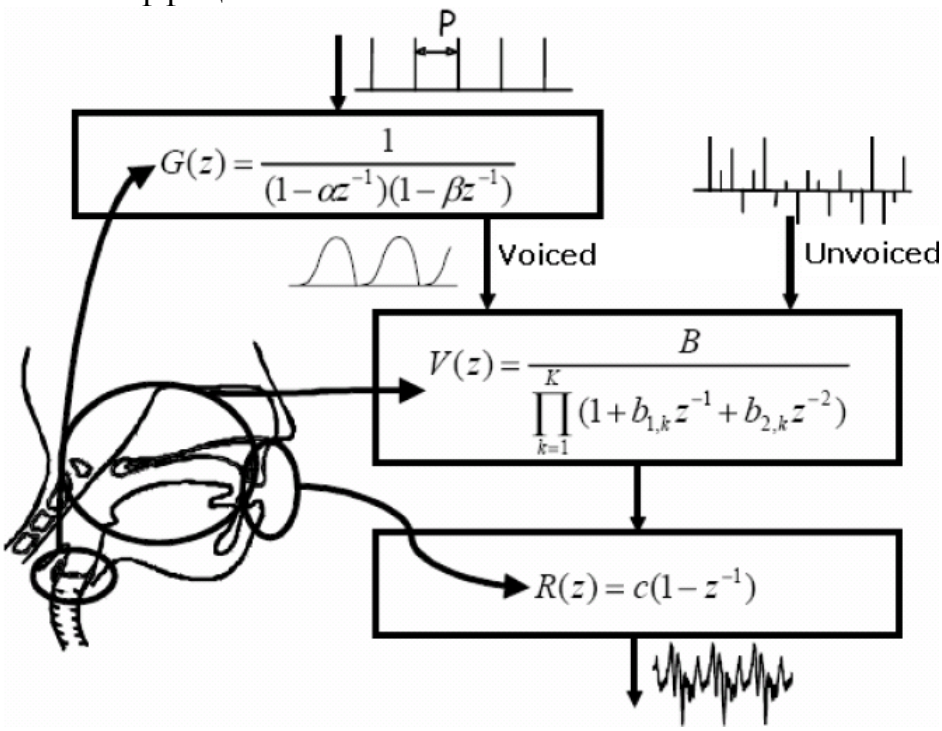

Рис.7.1. Модель речевого тракта

По своей природе речевой сигнал является нестационарным. Однако, для применения техники на основе линейного предсказания необходимо иметь стационарный сигнал. Для этого речь разбивают на перекрывающиеся между собой окна, внутри которых, в силу их небольшой длины, предполагается, что сигнал стационарен. Выбор размера такого окна (фрейма) и величины смещения напрямую влияет на вычислительную сложность алгоритма (меньшее смещение фреймов ведет к большему их количеству для обработки) и точность извлечения параметров (в зависимости от степени стационарности речи на некотором участке, использование большего или меньшего окна ведет к более точному вычислению параметров).

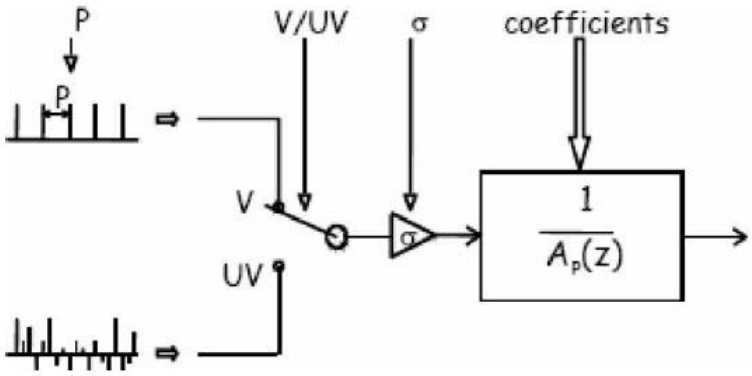

Рис.7.2. Синтез речи на основе линейного предсказания

На рисунке 7.2 представлена общая схема генерации речи не основе линейного предсказания: сигнал возбуждения, представленный либо единичными импульсами в случае вокализованного фрейма, либо белым шумом в случае невокализованного фрейма, является входом для AR фильтра, моделирующего вокальный тракт. Применение такой параметрической модели позволяет далее легко производить какие-либо просодические модификации. Так, частота основного тона определяется расстоянием между последовательными единичными импульсами, которое может быть задано требуемым значением T. Для сохранения исходной амплитуды сигнала используется масштабный коэффициент σ_0, который должен быть заменен на

$$\sigma = \sigma_0 \sqrt{\frac{T_0}{T}},$$

где T_0 – длина периода исходного сигнала. Длина сигнала модифицируется путем дублирования или удаления фреймов. Так же благодаря применению параметрической модели, гладкость спектральных характеристик может быть достигнута путем линейной интерполяции сегментов. Готовые фреймы складываются с перекрытием, формируя итоговый сигнал.

Главным недостатком применения подхода на основе линейного предсказания является сильная роботизированность голоса. Эксперименты показывают, что данный недостаток не может быть устранен путем увеличения порядка модели. В первую очередь, отдельное моделирование параметров фильтра и сигнала возбуждения может ухудшить результат, если их характеристики изменяются независимо. Также, функция возбуждения, которая имеет различную природу для вокализованных и невокализованных звуков, не может должным образом моделировать смешанные звуки, такие как, например, «в». Однако стоит отметить, что последнее можно исправить путем применения смешанной функции возбуждения [ссылки на статьи про смешанный сигнал возбуждения].

По сравнению с методом Unit Selection, подход, в основе которого лежат модели речи, имеет следующие преимущества и недостатки.

1. Автоматическое обучение параметров моделей, которое возможно выполнять на относительно небольшом речевом материале, позволяет существенно сократить объем требуемой памяти, а также позволяет разрабатывать новый голос за гораздо меньшее время.
2. Речь, полученная на основе моделей, более искусственна, однако в ней не наблюдаются разрывы, присутствующие при конкатенативном синтезе. Кроме того, при применении технологии Unit Selection качество синтеза существенно ухудшается в случае отсутствия подходящего звукового элемента в базе данных. При применении моделей отсутствующие в обучающей выборке звуковые элементы синтезируются на основе средних значений, максимально приближенных к требуемым, благодаря применению технологии кластеризации контекстов, основанной на деревьях. Это позволяет добиться разборчивости при ограниченном количестве контекстов.

3. Синтез, основанный на моделях, позволяет легко модифицировать характеристики голоса путем применения адаптации/интерполяции диктора, в то время как метод Unit Selection порождает речь, стиль которой не может быть отличен от стиля, представленного в речевом корпусе.

В заключение, достоинства и недостатки подходов, основанного на моделях и Unit Selection, можно представить в виде таблицы (таблица 7.1).

Таблица 7.1. Достоинства и недостатки современных подходов к синтезу речи

Критерий сравнения	Подход Unit Selection	Основанный на моделях подход
Сложность создания нового голоса	Высокая	Невысокая
Объем требуемой памяти	Высокий	Невысокий
Качество синтезируемой речи	Высокое	Роботизированная речь
Ощущение разрывов речи	Возможно	Менее вероятно
Качество синтеза элементов, отсутствующих в звуковом корпусе	Низкое	Высокое
Стиль речи	Фиксирован	Может быть модифицирован

В качестве схем, объединяющих данные подходы, могут применяться следующие: генерация физических параметров звуковых элементов на основе скрытых марковских моделей для последующего вычисления стоимости замены для метода Unit Selection; использование значений вероятностей переходов как «стоимости»; использование звуковых элементов, размер которых сопоставим с размером состояния модели; использование статистических моделей для вычисления стоимости связи между элементами.

Сами по себе системы статистического моделирования развиваются независимо от систем синтез речи. Они включают в себя такие приложения, как кодирование речи, преобразование параметров диктора и «подделку» речи диктора в задачах идентификации/верификации.

ЛИТЕРАТУРА

1. Брызгунова Е.А. (1982) Интонация // Русская грамматика. Том 1. М. Наука. С. 96-122.
2. Фланаган Дж. Анализ, синтез и восприятие речи. М.: Связь, 1968.

3. Taylor P. Text-to-Speech Synthesis. Cambridge University Press, 2009. 474 p.
4. Лобанов Б.М., Цирульник Л. И. Компьютерный синтез и клонирование речи. Минск, «Белорусская Наука», 2008. 316 с.
5. Hunt A., Black A. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database// Proceedings of ICASSP 96, 1996, pp. 373-376.
6. Tokuda K., Masuko T., Yamada T. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features // Proceedings of Eurospeech-1995, 1995.
7. Loh W.-Y. Classification and Regression Tree Methods // Encyclopedia of Statistics in Quality and Reliability, Wiley. 2008. P. 315-323.
8. Breiman L. Cutler *A.* Random Forests [Электронный ресурс], Режим доступа: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, свобо -дный. – Загл. с экрана. – Яз. англ.
9. Чистиков П.Г., Хомицевич О.Г., Рыбин С.В. Статистические методы определения мест и длительностей пауз в системах синтеза речи. Изв. вузов. Приборостроение. Тематический выпуск "Речевые информационные системы" (в печати).
10. 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.
11. Аничкин И.М., Чистиков П.Г. Формализация правил автоматического снятия омонимии в системе синтеза речи по тексту // Труды XXXVIII международной филологической конференции, 2008.
12. Fujisaki H. Dynamic characteristics of voice fundamental frequency in speech and singing // Production of Speech. N.Y. 1983.
13. Pierrehumbert, J. (1980) The phonology and phonetics of English intonation. PhD thesis, MIT. Distributed 1988, Indiana University Linguistics Club.
14. Black A.W., Hunt A.J. Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // In Proceedings of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1, pp. 373-376.
15. Conkie A. A robust unit selection system for speech synthesis // In Proceedings of Joint Meeting of ASA, EAA and DAGA. Berlin, Germany, 1999. Paper 1PSCB-10.
16. Vepa J. Join Cost for Unit Selection Speech Synthesis. University of Edinburgh, 2004.
17. Чистиков П.Г., Рыбин С.В. "Проблемы естественности речевого сигнала в системах синтеза", журнал "Информационные технологии в образовании", Санкт-Петербург, 2011.

18. Moulines E., Verhelst W. "Time-domain and frequency-domain techniques for prosodic modification of speech in Speech Coding and Synthesis", pp. 519–555, Netherland, 1995.
19. Главатских И.А., Чистиков П.Г. "Метод модификации физических параметров речевого сигнала на основе периодосинхронного Фурье-анализа", труды XXXVII международной филологической конференции, формальные методы анализа русской речи, Россия, 2009.
20. Rafael C. D. de Paiva, Luiz W. P. Biscainho and Sergio L. Netto "On the application of RLS adaptive filtering for voice pitch modification", in proceedings of the 10th International Conference on Digital Audio Effects, France, 2007.
21. Taylor P. Text-to-Speech Synthesis. Cambridge University Press; 1 edition, 2009.

УНИВЕРСИТЕТ ИТМО

Миссия университета – генерация передовых знаний, внедрение инновационных разработок и подготовка элитных кадров, способных действовать в условиях быстро меняющегося мира и обеспечивать опережающее развитие науки, технологий и других областей для содействия решению актуальных задач.

КАФЕДРА РЕЧЕВЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

О кафедре

Кафедра речевых информационных систем (РИС) создана в 2011 году на факультете Информационных технологий и программирования (ФИТиП).

Организатором создания кафедры выступает «Центр речевых технологий» (www.speechpro.ru). Заведующий – генеральный директор ООО «Центр речевых технологий», кандидат технических наук Хитров Михаил Васильевич, вице-президент консорциума «Российские речевые технологии», член ISCA, IEEE.

Кафедра РИС обеспечивает подготовку докторантов, аспирантов и магистров. Для тех, кто имеет высшее образование, но хотел бы связать свое будущее с речевыми технологиями, имеются курсы дополнительного профессионального образования.

Обучение на кафедре

Кафедра «Речевые информационные системы» (базовая кафедра «Центра речевых технологий») Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (ИТМО) в рамках направления 230400.68 «Информационные системы и технологии» открывает прием в магистратуру по новой образовательной программе 230400.68.04 «Речевые информационные системы».

Срок обучения 2 года. Обучение завершается защитой магистерской диссертации.

Целевая установка магистратуры – подготовка специалистов, способных участвовать в исследовательской и проектной работе в области речевых информационных технологий со специализацией в направлениях распознавания и синтеза речи, распознавания личностей по голосу, мультимодальной биометрии, в области проектирования и разработки информационных систем и программного обеспечения.

Область профессиональной деятельности выпускников кафедры РИС включает:

– исследование, разработка, внедрение речевых информационных технологий и систем;

– методы и алгоритмы цифровой обработки речевых сигналов;

– автоматизированные системы обработки речевых сигналов;

– программное обеспечение автоматизированных речевых информационных систем;

– системы автоматизированного проектирования программных и аппаратных средств для речевых информационных систем и информационной поддержки таких средств.

Объектами профессиональной деятельности выпускников кафедры РИС являются:

– информационные процессы, технологии, системы и сети, предназначенные для обработки, распознавания, синтеза речевых сигналов;

– инструментальное (математическое, информационное, техническое, лингвистическое, программное, эргономическое, организационное и правовое) обеспечение речевых информационных систем;

– способы и методы проектирования, отладки, производства и эксплуатации информационных технологий и систем в областях обработки, распознавания, синтеза речевых сигналов, телекоммуникации, связи, инфокоммуникации, медицины.

Широкий профиль подготовки, знание универсальных методов исследования и проектирования информационных систем, практические навыки работы с современным программным обеспечением – все это позволяет выпускникам кафедры найти работу в научных институтах и университетах, в фирмах, на производственных предприятиях, а также в коммерческих структурах. Студентам, которые хорошо проявляют себя в учебе, предлагается работа в ООО «Центр речевых технологий».

Учебный план предусматривает, в частности, следующие курсы:

– Информационные технологии:

– Системный анализ и моделирование информационных процессов и систем;

– Проектирование информационных систем;

– Организация проектирования и разработки программного обеспечения распределенных систем;

– Организация проектирования и разработки программного обеспечения встроенных систем;

– Тестирования программного обеспечения;

– Управление качеством разработки программного обеспечения.

Речевые технологии:

– Цифровая обработка сигналов;

– Цифровая обработка речевых сигналов;

– Математическое моделирование и теория принятия решений;

– Распознавание образов;

– Распознавание и синтез речи;

– Распознавание диктора (говорящего по голосу);

– Мультимодальные биометрические системы.

К преподаванию привлекаются ведущие специалисты «Центра речевых технологий», преподаватели НИУ ИТМО, а также специалисты, работающие в известных научных, производственных и коммерческих организациях.

Рыбин Сергей Витальевич

СИНТЕЗ РЕЧИ

Учебное пособие

В авторской редакции
Редакционно-издательский отдел НИУ ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ № 3227
Тираж 50
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49