Технология синтеза речи в историко-методологическом аспекте

Соломенник А.И., *аспирант филологического факультета МГУ им. М.В. Ломоносова*

В статье рассматривается зарождение и развитие технологии синтеза речи, начиная от первых механических устройств для порождения отдельных звуков речи и заканчивая современными методами синтеза слитной речи на основе селективного выбора единиц и скрытых Марковских моделей. Даётся характеристика различных методов и подходов к решению задачи порождения естественно звучащего речевого сигнала, кратко обсуждаются достоинства и недостатки этих методов, их историческая преемственность.

• синтез речи • история синтеза речи • методы синтеза речи • формантный синтез • формантный синтез • артикуляционный синтез • конкатенативный синтез • селективный синтез • статистический параметрический синтез.

The paper deals with speech synthesis technology origin and development from the first mechanical synthesizers to modern unit selection and hidden Markov models (HMM)-based text-to-speech synthesis. Different methods of speech synthesis are described; their advantages and imperfections are discussed.

• speech synthesis • history of speech synthesis • speech synthesis methods • formant synthesis • articulatory synthesis • concatenative synthesis • unit selection • HMM-based synthesis.

Введение

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 (ок. 946 – 1003), Альбертом Великим (1198 — 1280) и Роджером Бэконом (1214 — 1294) [Mattingly 1974]. 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.

Первые механические синтезаторы

Первые синтезаторы, появившиеся во второй половине XVIII века, были механическими, они могли порождать 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.

В 1779 году Петербургская академия наук объявила ежегодную премию за объяснение разницы между пятью гласными звуками и за конструирова-

ние 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), при помощи пульсирующего воздушного потока порождавших русские гласные. Воздушный поток порождался вибрирующими язычками, подобными голосовым связкам человека [2].

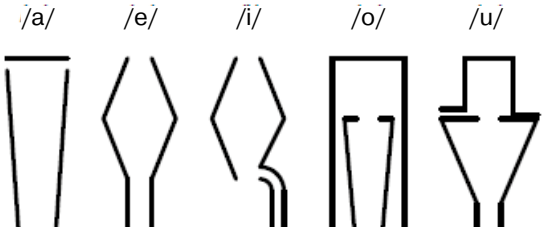

Рис. 1. Система резонаторов Кратценштейна [3]

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 [4] и короткие фразы на нескольких языках [1]. Для управления «говорящей машиной» требовался хорошо обученный оператор, порождение речи можно было сравнить с игрой на органе. Усовершенствованный вариант машины Кемпелена (рис. 2) был создан в 1837 году английским физиком Чарльзом Уитстоном (1802 — 1875). Также под впечатлением от машины Уитстона американский учёный и изобретатель Александр Грэм Бэлл (1847 — 1922) собрал собственную аналогичную модель [4].

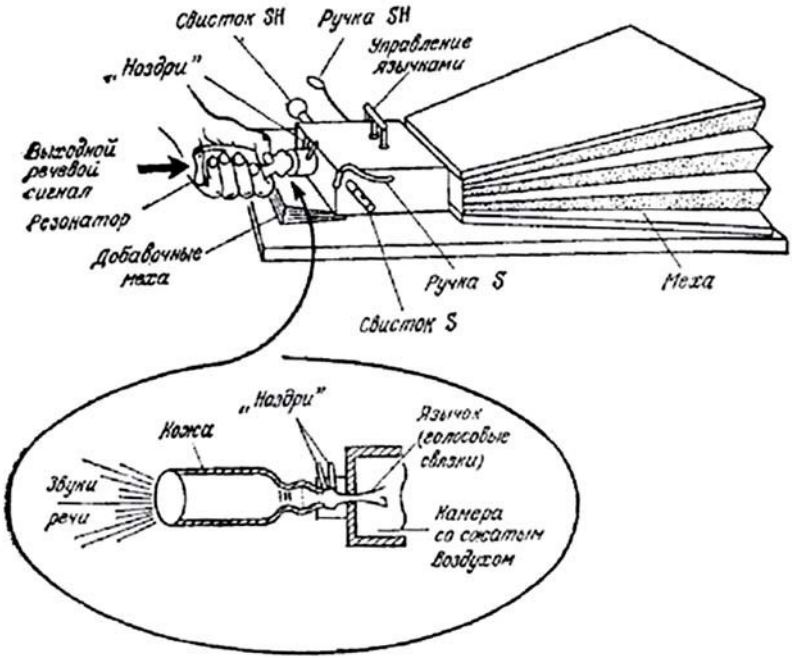

Рис. 2. Говорящая машина Кемпелена, построенная Уитстоном [4]

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 (ок. 1800 — ок. 1850) представил свою говорящую машину под названием «Эйфония», которая по сообщениям современников могла производить обычную и шёпотную речь, а также исполнять песни [1].

В XX веке, несмотря на развитие электрических методов синтеза речи, разработка механических синтезаторов речи проводилась до 60-х годов [3]. Это было связано, с одной стороны, с малой доступностью сложных электрических компонентов [4], 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. 3). По форме он практически повторял голосовой тракт человека, был выполнен из резины и металла и управлялся клавишами, подобными клавишам трубы [4].

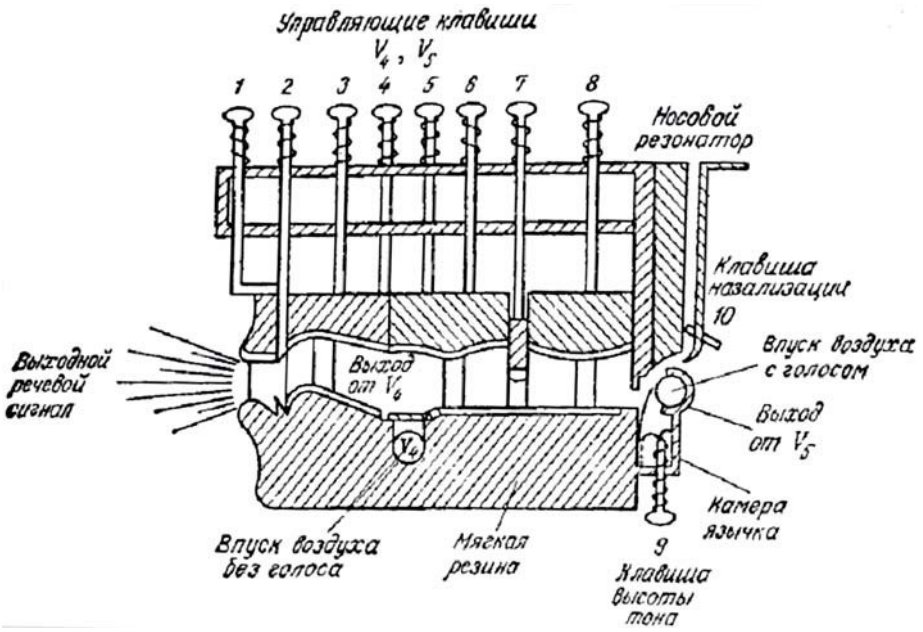

Рис. 3. Механический синтезатор Риша [4]

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

Первые электрические синтезаторы

В XX веке с освоением электрических устройств и зарождением электроники начались попытки построить синтезаторы речи — электрические аналоги речепроизводящей системы. Самый первый электрический синтезатор был создан Дж. Стюартом в 1922 году [5]. Его схема (рис. 4) включала в себя электрический зуммер для моделирования голосовых связок и пару индуктивно-ёмкостных резонаторов для моделирования резонансов горла и ротовой полости [4]. Таким образом генерировались первые две форманты (резонансные частоты голосового тракта), то есть устройство могло синтезировать только гласные звуки.

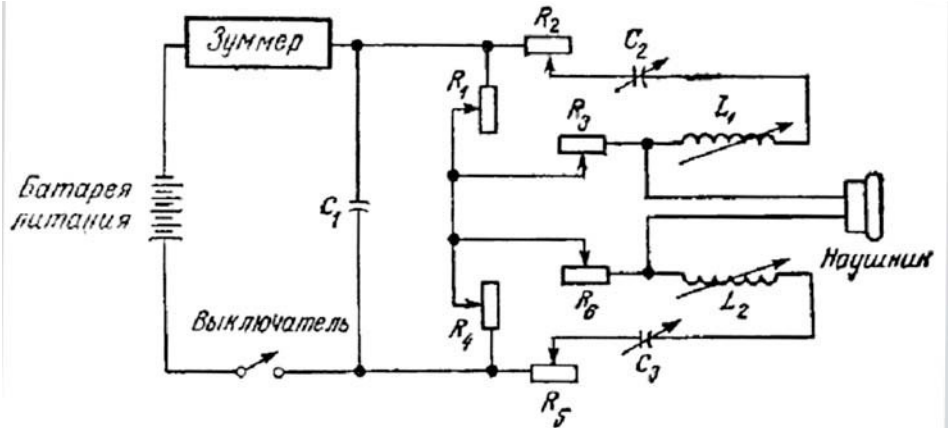

Рис. 4. Электрическая модель голосового тракта Стюарта [4]

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 [6].

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, 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. 5, с. 46) [2].

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

В литературе [7] упоминаются попытки синтеза русской речи при помощи первых музыкальных синтезаторов. «Вариофон» Е. А. Шолпо (1891 — 1951), сконструированный в 1931 году, представлял собой оптический синтезатор. Звук записывался на движущуюся плёнку с помощью вырезанных зубчатых дисков разной формы, изменявших очертания звуковой дорожки и трансмиссии, позволявшей синхронизировать контур и подачу плёнки. Первый электронный музыкальный синтезатор АНС был спроектирован Е. А. Мурзиным

(1914 — 1970) в 1938 году и построен в 1958. АНС содержал 720 звуковых дорожек чистых тонов, которые можно было накладывать друг на друга. Клавиатуры не было, на стекле, покрытом специальной непрозрачной мастикой, прочерчивалась линия, через которую пускался световой луч на фотоэлементы.

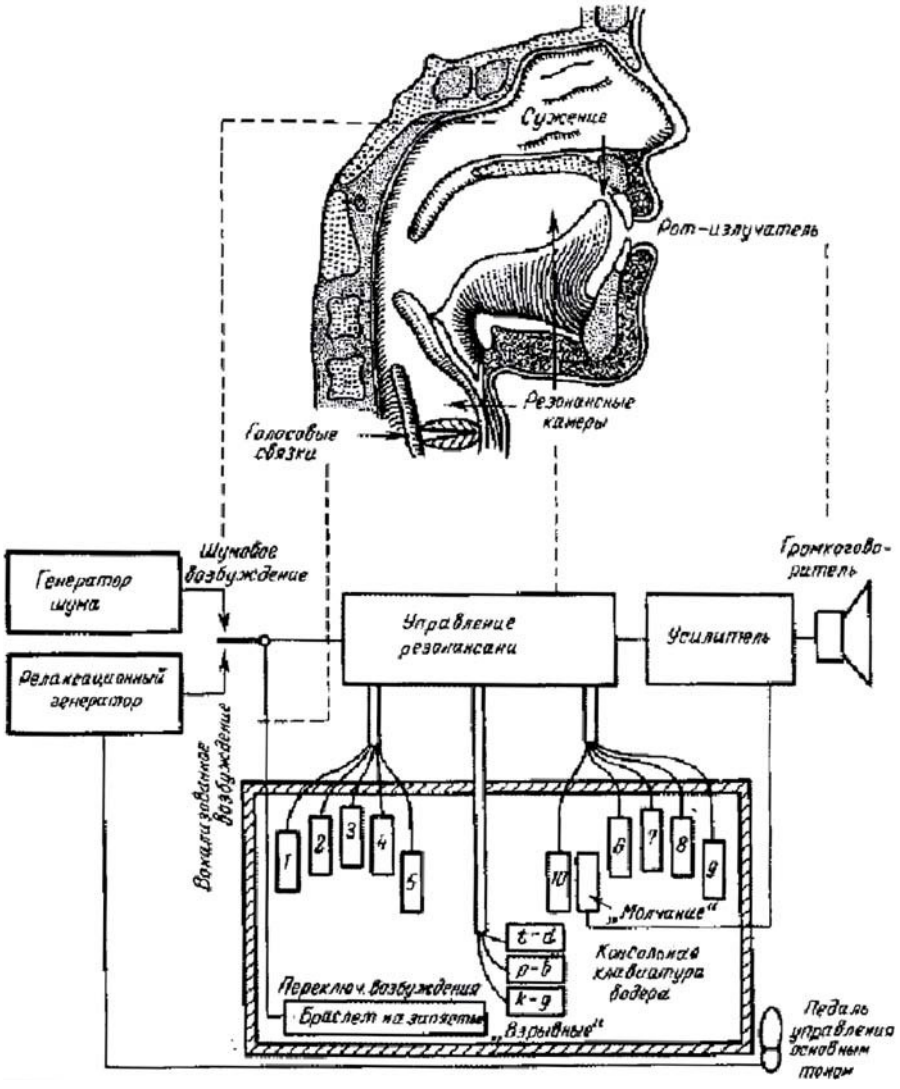

Рис. 5. Схема синтезатора «водер» [Фланаган 1968]

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 дру-

гую сторону спектрограммы, являлись источником управляющих сигналов для набора тех же полосовых фильтров, что и в водере. Дополнительные дорожки на спектрограмме управляли переключением тона и шума и несли информацию о частоте основного тона. Подобный метод использовался Дж. Борстом и Ф. Купером в устройстве «водек» (1957 год) [2].

Наиболее известный «проигрыватель» спектрограмм, синтезатор Pattern Playback (рис. 6), был представлен американскими исследователями Ф. Купером, А. Либерманом и Дж. Борстом в 1951 году. Он состоял из оптической системы для динамической модуляции амплитуд гармоник основного тона в 120 Гц в зависимости от изображений на движущейся прозрачной ленте [5].

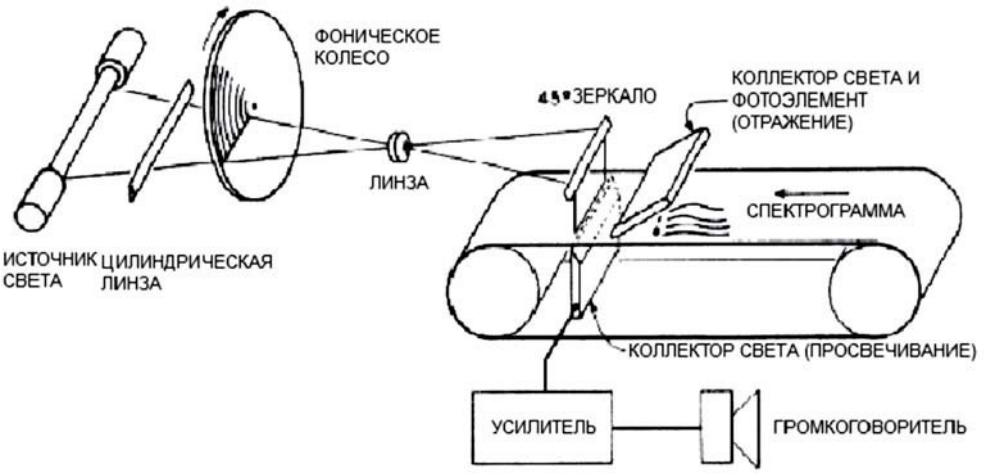

Рис. 6. Синтезатор Pattern Playback [5]

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 не моделируется напрямую 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 определённые заранее 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 разработки основанных на ней формантных и артикуляционных синтезаторов, а также синтезаторов, использующих линейное предсказание. Эти три метода называют также технологиями синтеза первого поколения [8].

XX век: синтезаторы первого поколения

Синтезаторы первого поколения можно на основании используемых ими методов разделить на две большие группы: акустические и артикуляционные. К направлению акустического синтеза относится формантный синтез и синтез с использованием линей-

ного 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 [9].

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 [99]. 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 белого шума к соответствующей точке линии [2].

Первый артикуляционный синтезатор с динамическим контролем (рис. 7) 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 [3].

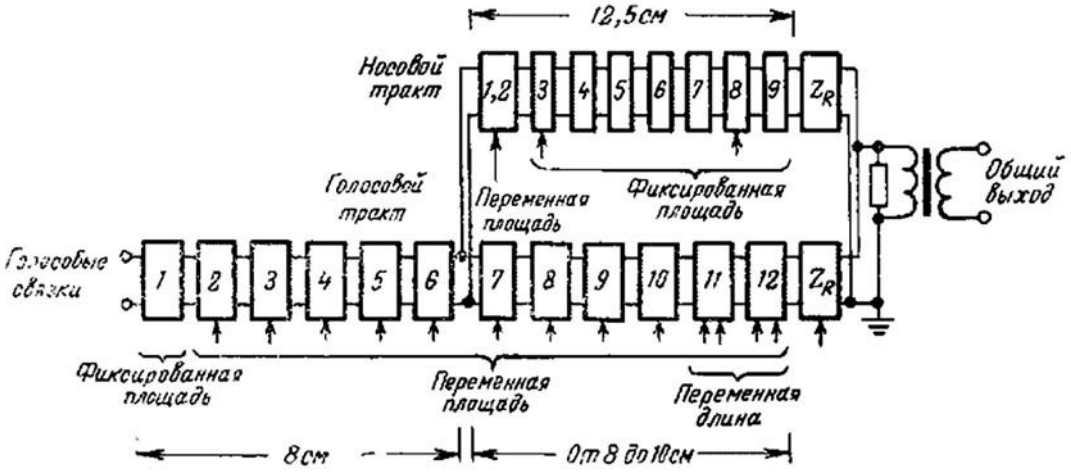

Рис. 7. Аналог голосового тракта с линией передачи, управляемый непрерывно [2]

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 [8].

Формантный синтез

Первым формантным синтезатором стал PAT (Parametric Artificial Talker) английского исследователя У. Лоуренса, представленный в 1953 году. Этот синтезатор состоял из трёх электронных формантных резонаторов, соединённых параллельно, на вход которым подавался шум или гармонический сигнал. 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 [5]. Синтезатор Лоуренса был первым из параллельных формантных синтезаторов. Их главное преимущество состояло в относительной простоте 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 [10].

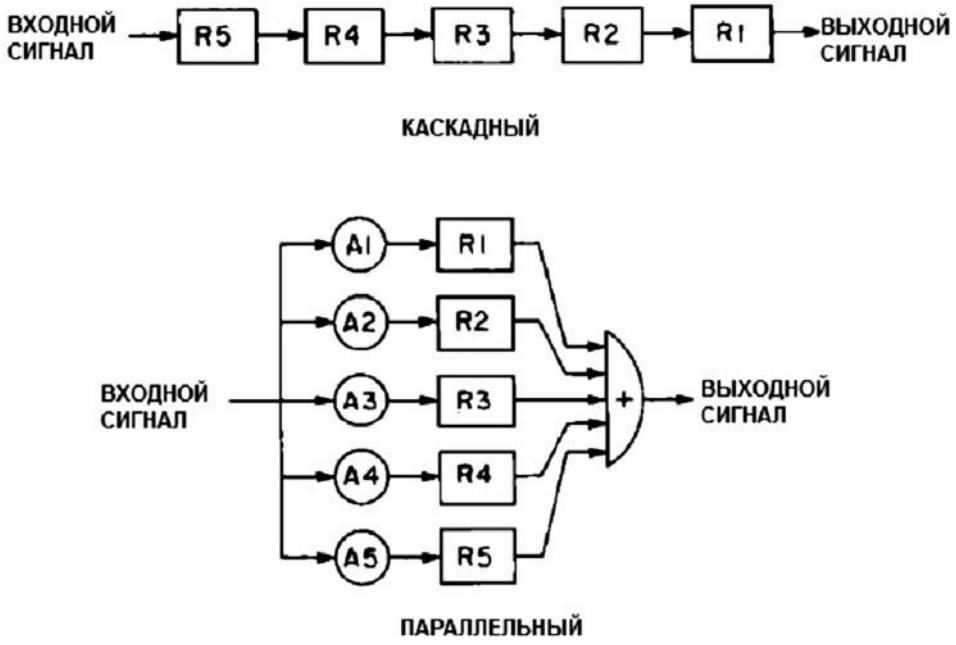

Рис. 8. Каскадный и параллельный синтезаторы. В параллельном синтезаторе амплитуда каждого формантного резонатора должна контролироваться отдельно. В каскадном выходной сигнал каждого резонатора является входным сигналом следующего [10]

В том же 1953 году известный шведский исследователь речи, автор классической акустической модели речеобразования «источник-фильтр» Гуннар Фант продемонстрировал свой каскадный формантный синтезатор OVE I (Orator Verbis Electris). В нём частота двух нижних резонаторов контролировалась механической рукой, а амплитуда и частота основного тона определялись ручными потенциометрами [5].

В дальнейшем оба типа синтезаторов усложнялись и совершенствовались, позволяя каждой новой версии звучать всё ближе к естественной человеческой речи. В 1973 году английскому исследователю Дж. Холмсу удалось вручную настроить на своём синтезаторе (рис. 9) произнесение предложения «I enjoy the simple life» так хорошо, что обычный слушатель не мог отличить его от произнесения того же текста живым человеком [3]. Однако оставалась проблема с автоматическим контролем работы синтезатора, который не мог пока приблизиться к ручной настройке произнесения.

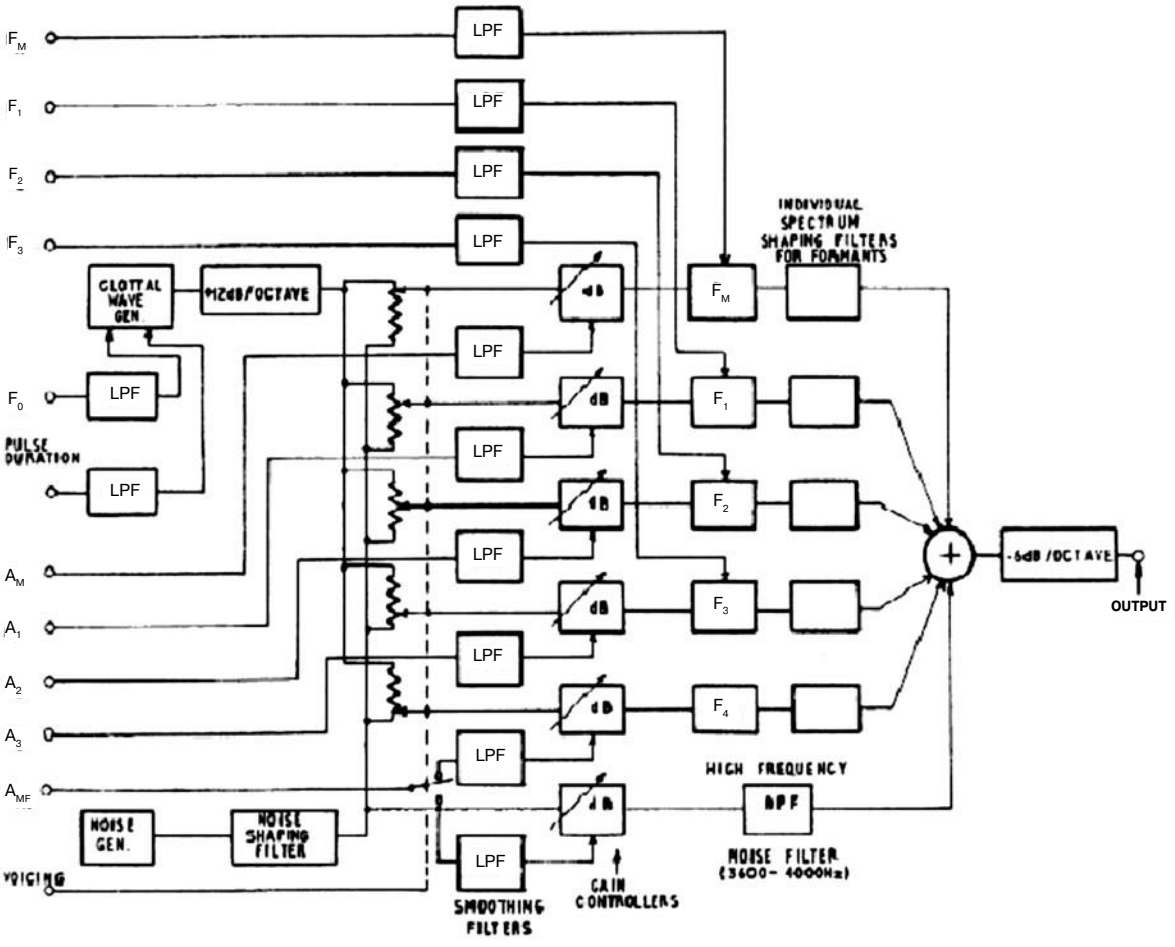

Рис. 9. Формантный синтезатор Холмса, состоящий из резонаторов для четырёх формант и носовой форманты, каждый из которых возбуждался вариативной смесью сигналов шумового и голосового источников [5]

С развитием компьютерной техники и появлением вычислительных машин в середине 50-х годов электрические аналоговые синтезаторы стали посте-

пенно 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 [5].

Первая модель формантного синтезатора русской речи «Фонемофон-1» (рис. 10) была разработана в Минске в начале 70-х годов, а в его последующих версиях удалось добиться синтеза русской речи по произвольному тексту весьма высокого качества [7].

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

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) [3].

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 [6]. В общем виде простейший КЛП-синтезатор имеет достаточно сложную структурную схему, представленную на рис. 11.

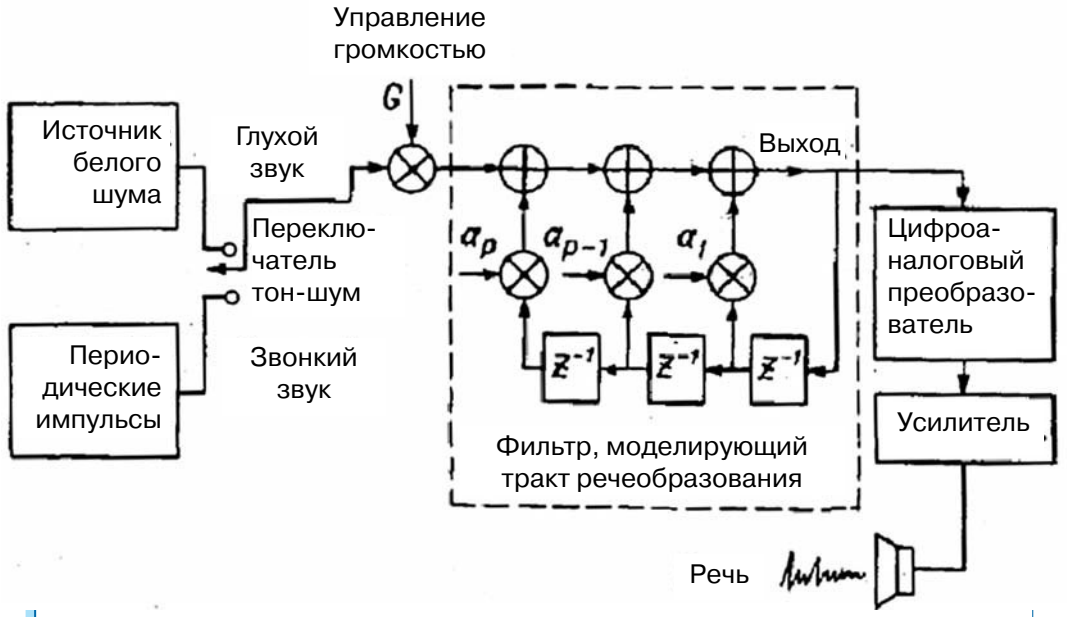

Рис. 11. Структурная схема КЛП-синтезатора [6]

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 речи.

Синтезаторы первого поколения обычно требовали детального фонетико-акустического описания того, что должно быть произнесено, и не включали какого-либо автоматического способа получения подобного описания для произвольного сообщения или текста.

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 «Текст–Речь» (Text-to-Speech или сокращённо TTS). В идеале такие устройства должны имитировать деятельность человека, который читает письменное сообщение или текст любой степени сложности [9]. 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 (рис. 12), 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.

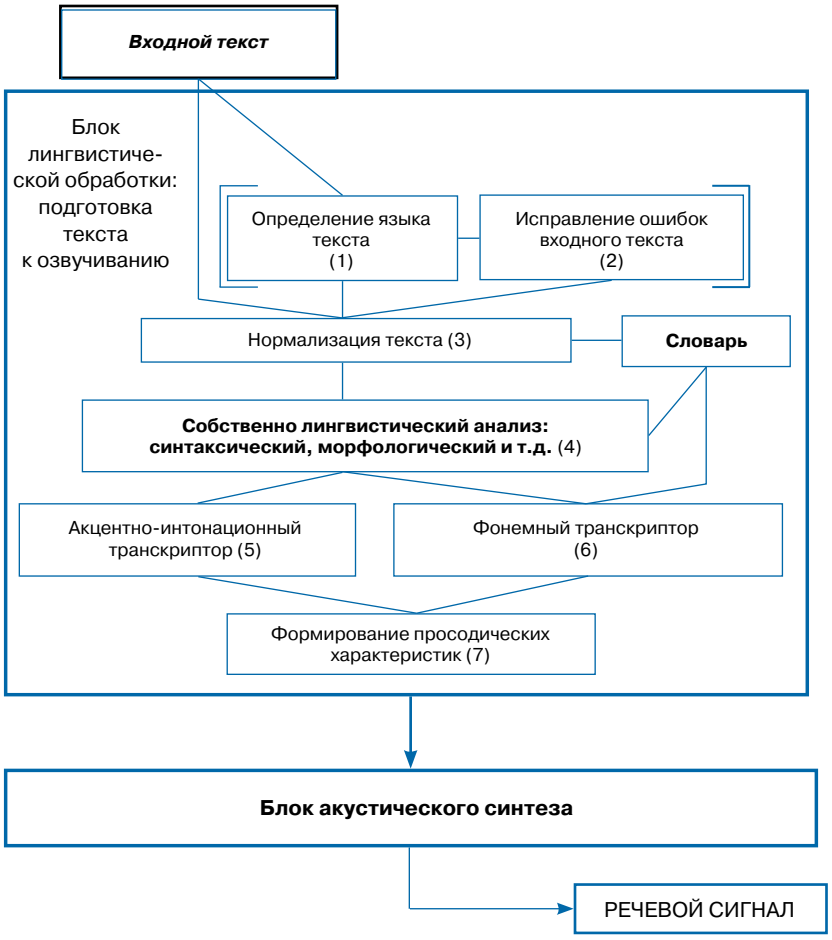

Рис. 12. Лингвистический этап автоматического синтеза речи [9]

Первая полноценная система «Текст-Речь» 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 [5].

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 интонационно-просодических 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 для блока 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) синте-

зированной речи. Так называемый ресинтез, то есть подача на вход синтезатора акустических характеристик естественной речи, является также надёжным способом понять, насколько хорошо работает его акустический компонент.

Конкатенативный синтез

Конкатенативный (или компилятивный) синтез, называемый также техникой второго поколения [8], смог появиться благодаря тому, что перед создателями систем синтеза уже не стояли такие жёсткие ограничения по доступной компьютерной памяти (как в 70-е и 80-е годы) и появилась возможность хранить большие объёмы речевых данных. В отличие от систем первого поколения в них не используется упрощённая классическая модель «источник-фильтр». Вместо этого в памяти компьютера хранятся фрагменты реальных акустических сигналов (либо в виде оцифрованных фрагментов звуковой волны, либо в параметризованной форме, полученной в результате акустического анализа исходных «живых» образцов) из речи определённого «диктора-донора», из которых путём склейки (или конкатенации) и создавалась первичная основа синтезируемого акустического сигнала. В дальнейшем эта основа подвергается модификации по правилам, функция которых состоит в том, чтобы придать склеенным фрагментам акустического сигнала нужные просодические характеристики [9].

Различные системы конкатенативного синтеза используют в качестве базовых элементов для склейки звуковые единицы различного размера: фрагменты фонемной размерности (акустические аллофоны), полуслоги, слоги и образцы смешанных типов. Наиболее часто в таких системах используются дифоны — отрезки, начинающиеся в середине одного звука и заканчивающиеся в середине следующего. Дифоны как оптимальная единица для учёта эффектов коартикуляции в речевом сигнале были впервые предложены американским исследователем Дж. Петерсоном с коллегами в 1958 году [5].

На качество речи, производимой конкатенативным синтезатором, влияет как качество и количество самих единиц для конкатенации (степень покрытия всех необходимых сегментных единиц), так и используемые алгоритмы просодической модификации речевого сигнала. Наиболее широко используемым методом модификации речи во временной и частотной области является алгоритм PSOLA (Pitch Synchronous Overlap and Add), разработанный в 1985 году, и его последующие варианты [3].

По современным меркам объём звуковой базы для обычного конкатенативного синтеза речи является относительно небольшим, что позволяет построить синтезатор высокого качества довольно быстро. Однако главным недостатком систем такого типа является то, что они, в отличие от, например, формантного синтеза по правилам, не обладают достаточной гибкостью в изменении тембра голоса, так как для этого необходимо создавать новую базу акустических образцов для другого диктора-донора [9].

XX век: синтезаторы третьего поколения

К третьему поколению технологий автоматического синтеза речи обычно относят синтез на основе скрытых Марковских моделей и селективный синтез речи [8]. Их общей чертой является использование больших объёмов речевых данных, а также высокая естественность синтезированной речи.

Селективный синтез речи

В настоящее время доминирующей технологией автоматического синтеза речи является так называемый селективный синтез, так как он позволяет получать синтезированную речь, которая по своим характеристикам наиболее приближена к естественной [8].

Селективный синтез речи (в англоязычных источниках называемый unit selection) является разновидностью конкатенативного синтеза, то есть при генерации речевого сигнала используются заранее сделанные звукозаписи естественной речи. В отличие от более ранних аллофонных или дифонных синтезаторов речи, порождающих итоговый речевой сигнал из отдельных и специально подготовленных звуковых единиц, выделенных из небольшого и тщательно подобранного набора 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 оптимальную последовательность звуковых единиц (рис. 13), учитывая одновременно и то, насколько кандидат подходит под описание необходимых характеристик целевого звука (стоимость замены), и то, насколько хорошо выбранные элементы будут конкатенироваться с соседними (стоимость связи). При этом с учетом указанных стоимостей из базы в качестве оптимальных могут быть выбраны не отдельные звуки, а их цепочки или даже целые предложения. Такой подход позволяет минимизировать модификации речевого сигнала, что повышает естественность синтезируемой речи.

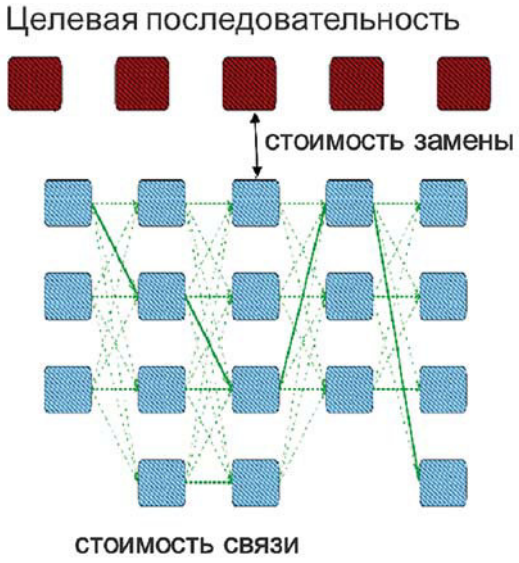

Рис. 13. Выбор целевой последовательности при селективном синтезе речи

Первыми системами селективного синтеза стали n-Talk (1992) [Sagisaka et al. 1992] и CHATR (1994) [12], 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 [13].

Статистический параметрический синтез

Статистический параметрический синтез, так же как и описанный выше конкатенативный, является методом, основанным не на правилах, а на имеющихся акустических данных. Однако в отличие от конкатенативного метода, при котором необходимые для синтеза

параметры речевого сигнала уже присутствуют в самих хранимых в памяти компьютера единицах конкатенации, в статистическом параметрическом синтезе делается попытка машинного обучения системы на имеющихся речевых данных с целью получения модели соответствия характеристик речи, поступающих на вход акустического блока, физическим параметрам целевых звуковых единиц. Полученная акустическая модель даёт два преимущества: уменьшение памяти для хранения модели вместо самой речевой базы и возможность её параметрической модификации, например, быстрого изменения тембра голоса [8].

Наиболее распространённой техникой в данном направлении синтеза является метод, основанный на использовании скрытых Марковских моделей (HMM — hidden Markov models). В HMM представлена не только последовательность фонем, но и различная лингвистическая информация (та же, что и для селективного синтеза), а акустические параметры, сгенерированные HMM, используются для управления вокодером, т.е. для порождения речевого сигнала используются параметры речевого тракта и параметры возбуждения [14] 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 [15]. 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.

Заключение

Как видно из сказанного выше, в уже довольно длительной истории технологий синтеза речи значительно менялись приоритеты и направления исследований. Это связано и с целями, которые ставились перед синтезаторами: от демонстрации возможности получения звуков, подобных человеческой речи, и моделирования процессов речеобразования до получения разборчивого, а затем и естественного выразительного чтения компьютером произвольного текста. Нельзя не отметить также, что история и успехи разработок в области синтеза речи тесно связаны с развитием других научных дисциплин: физики (механики, электродинамики, акустики), математики (статистики), информатики, физиологии, психологии и, конечно же, лингвистики (фонетики, автоматической обработки естественного языка).

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 [8]. Предметом широких исследований в последние годы является также и оценка качества работы синтезаторов речи: за рубежом активно ведутся работы по стандартизации оценок. Для русскоязычных синтезаторов существуют отдельные перспективные разработки, но есть потребность в выработке единого стандарта для оценки качества синтеза [16].

Литература

1. *Кейтер Дж*. Компьютеры — синтезаторы речи. М.: Мир, 1985.
2. *Кодзасов С. В., Кривнова О. Ф*. Общая фонетика. Москва, 2001. 592 с.
3. *Лобанов Б. М., Цирульник Л. И*. Компьютерный синтез и клонирование речи. Минск, «Белорусская Наука», 2008. 316 с.
4. *Обжелян Н. К., Трунин-Донской В. И*. Машины, которые говорят и слушают. Кишинев, 1987.
5. *Фланаган Дж*. Анализ, синтез и восприятие речи. М.: Связь, 1968.
6. *Black A., Taylor P*. CHATR: A Generic Speech Synthesis System // COLING94, Japan, 1994.
7. *Hunt A., Black A*. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database // Proceedings of ICASSP 96, 1996. P. 373–376.
8. *Klatt D*. Review of Text-to-Speech Conversion for English // JASA vol. 82 (3), 1987. P. 737–793.
9. *Klatt D. H*. Software for a cascade/parallel formant synthesizer // JASA. 1980. V. 67. P. 971–995.
10. *Lemmetty, S*. Review of Speech Synthesis Technology. Master's Thesis, Helsinki University of Technology, 1999. 104 p.
11. *Mattingly, I. G*. Speech Synthesis for Phonetic and Phonological Models // Current Trends in Linguistics, edited by T. S. Sebeok, Vol. 12, 1974. Mouton, The Netherland. P. 2451–2487.
12. *Sagisaka, Y. et al*. ATR — n-Talk speech synthesis system // Proceedings of ICSLP92, Banff, Canada, 1992. P. 483–486.
13. *Taylor P*. Text-to-Speech Synthesis. Cambridge University Press, 2009. 474 p.
14. *Tokuda K., Masuko T., Yamada T*. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features // Proceedings of Eurospeech-1995, 1995.
15. *Соломенник А.И., Таланов А.О., Соломенник М.В., Хомицевич О.Г., Чистиков П.Г*. Оценка качества синтезированной речи: проблемы и решения. Изв. вузов. Приборостроение. Тематический выпуск «Речевые информационные системы». № 2, 2013. С. 38–42.

Сведения об авторе

Соломенник Анна Ивановна —
аспирант кафедры теоретической и прикладной лингвистики филологического факультета МГУ им. М.В. Ломоносова, научный сотрудник ООО "Речевые технологии" (Минск, Беларусь). Научные интересы: автоматический синтез речи (в частности, оценка качества синтезированной речи), идентификация диктора по голосу. Электронная почта: solomennik-a@speechpro.com