УДК 934.5; 522

СИНТЕЗАТОР РЕЧИ КАК ЧАСТЬ ДИАЛОГОВОЙ СИСТЕМЫ НА РЕКУРРЕНТНОМ ПРОЦЕССОРЕ

Рождественский Ю.В., Дьяченко Ю.Г., Морозов Н.В.

Введение

Диалоговая система общения пользователя с электронной средой, реализованная на основе цифрового сигнального процессора (ЦСП) является базовым узлом многих систем контроля и ограничения доступа и управления технологическими процессами и приборами. Она включает в себя подсистемы распознавания речи (изолированных слов или последовательности слов-команд) и синтезатор речи. Разработка таких систем является актуальной задачей. В перспективе диалоговая система общения пользователя с электронной средой станет неотъемлемой составляющей среды обитания человека.

Данная работа посвящена проблемам реализации подсистемы синтеза речи на вычислительных средствах с ограниченными ресурсами и адаптации архитектуры параллельного динамического (рекуррентного) речевого процессора (РРП) для выполнения функций диалоговой системы общения пользователя с электронной средой. В этом отношении речевой синтез является хорошей тестовой задачей для апробации и оценки архитектуры рекуррентного процессора.

Основным носителем речевой информации, требующей озвучивания, является текстовая форма представления речи. Поэтому все проблемы речевого синтеза будут в дальнейшем рассматриваться на примере систем преобразования текста в речь.

По объему речевого словаря системы синтеза речи по тексту (ССРТ) разделяются на системы с ограниченным словарем и неограниченным языковым словарем. Системы с ограниченным словарем используются для синтеза сообщений, суммарный объем которых может в несколько раз превышать объем самого словаря. Они используют более простые алгоритмы для синтеза речи по сравнению с ССРТ с неограниченным словарем и при соблюдении определенных ограничений допускают реализацию на типовых современных ЦСП, в том числе и на РРП.

Модель формирования звуков человеческой речи

Непрерывная человеческая речь представляет собой сложный звуковой сигнал, который формируется голосовым аппаратом человека. Его акустическая модель представляет собой ряд последовательно-параллельных резонаторов, источник возбуждения акустических колебаний и источники шумовых звуков. Все акустические характеристики этой модели являются нестационарными:

- период основного тона (ПОТ) — период импульсных колебаний голосовых связок, определяющий частоту возбуждающего сигнала,
- формантные характеристики (частоту, ширину резонансного пика, интенсивность) — акустические характеристики резонаторов, моделирующих отдельные части голосового тракта,
- шумовую составляющую (частотный спектр, интенсивность) — акустические характеристики щелевых фрагментов речевого аппарата.

Вместе с тем, из всего многообразия звуков можно выделить от 20 до 60 характерных и устойчивых для конкретного языка звуковых структур, характеризующихся определенными особенностями звукообразования. Такие звуковые фрагменты получили название “фонем”. Фонемы могут использоваться в качестве

"кирпичиков", из которых строится синтезируемая звуковая фраза. Однако для синтеза "естественной" речи необходимо учитывать еще эмоциональные, просодические и синтаксические особенности озвучивания синтезируемой фразы [1, 2].

Корректное разрешение этих проблем чрезвычайно важно для построения качественной ССРТ. В настоящее время установлено, что основную роль в просодической и эмоциональной окраске текста играют ПОТ голосовых звуков, длительность и интенсивность звучания отдельных фонем и слов в целом. Эти факторы наиболее значимы для восприятия речи.

Таким образом, информационный поток (интерфейс) между лингвистическим анализатором текста и акустическим синтезатором речи должен содержать:

- признак текущей фонемы или паузы,
- длительность текущей фонемы,
- значение ПОТ в различных частях фонемы, если он меняется,
- энергетическую характеристику фонемы,
- дополнительные сведения о подтипе фонемы, если таковые имеются.

Методы и алгоритмы синтеза речи

Существующие методы и алгоритмы синтеза речи различаются по следующим критериям:

- качество и естественность синтезируемой речи,
- потребные вычислительные ресурсы для реализации алгоритмов,
- объем памяти данных для реализации ограниченного словаря.

В зависимости от особенностей реализации ССРТ и ограничений, накладываемых конкретной аппаратурой, на которой предполагается реализовать ССРТ, предпочтительным может оказаться тот или иной метод.

В настоящее время известны артикуляционные, формантные и конкатенационные методы синтеза речи.

Артикуляционные методы синтеза предполагают построение достаточно полной математической модели акустики человеческого речевого тракта и звукообразования. Такая модель отличается высокой вычислительной сложностью, но дает достаточно полную картину процессов, происходящих в голосовом аппарате человека, и считается весьма перспективной для построения высококачественных синтезирующих речевых систем в будущем.

Формантные методы синтеза речи [3] реализуют частотную модель голосового аппарата человека для формирования звуковых образов фонем, ставя в соответствие каждой языковой фонеме набор формантных характеристик, соответствующий определенному человеческому голосу. Они обладают небольшой вычислительной сложностью, но испытывают затруднение с описанием звукового перехода от одной фонемы к другой – дифона.

Дифоны отвечают за узнаваемость говорящего и естественность человеческой речи. Дифон выделяется из речевой фразы как отрезок речи между центрами двух соседних фонем. Для воспроизведения качественной человеческой речи обычно требуется 2000...3000 дифонов на один голос на неограниченном словаре. Качественный синтез дифонов из фонем требует большого количества правил построения дифонов. Таким образом, формантный синтезатор способен обеспечить синтез высококачественной речи, но при существенном увеличении вычислительных затрат и большом объеме обучения.

Конкатенационные методы синтеза речи опираются на лингвистический подход к построению человеческой речи. Устойчивой и эффективной частицей для

конкатенации является дифон и его производные — трифон и т.д. Использование дифона или трифона и в качестве "кирпичика" для синтеза позволяет обеспечить лучшее качество синтезируемой речи, поскольку в этом случае разрезание и "сшивание" слова (фразы) происходит по центру наиболее устойчивого звука, соответствующего фонеме, характеристики которого на протяжении некоторой длительности можно считать стационарными.

Проблемы дифонного конкатенационного синтеза в основном связаны с достаточно большим объемом памяти для хранения дифонов и привязки конкретного набора дифонов к конкретному голосовому аппарату. Поскольку виды существования одного и того же дифона сильно отличаются в зависимости от просодической информации, возникает необходимость в преобразовании библиотечных структур в дифоны другой длительности и с другим периодом основного тона. Для решения этой задачи в настоящее время наиболее широко используется метод PSOLA [4].

Сам по себе метод PSOLA не является методом синтеза речи. Он лишь обеспечивает повышение качества синтезированной речи путем привязки и настройки ПОТ и длительности конкатенируемых фрагментов речи (дифонов, аллофонов и т.д.) по контуру ПОТ и длительности звучания соответствующих речевых единиц, определенных подсистемой синтеза речи верхнего уровня на основе текстовой и просодической информации. Существует несколько версий метода PSOLA, каждый из которых ориентирован на применение в соответствующей области обработки речевого сигнала:

- *TD-PSOLA* (Time-Domain PSOLA) [4, 5] – преобразует речевой сигнал во временной области,
- *FD-PSOLA* (Frequency-Domain PSOLA) [6] – работает в частотной области,
- *LP-PSOLA* (Linear-Predictive PSOLA) [7] – работает с сигналом остатка, результатом предварительной фильтрации речевого сигнала,
- *Метод MBR-PSOLA* (Multi-Band Re-synthesis PSOLA) [8] – использует процедуру анализа/синтеза на основе многополосового возбуждения.

Наиболее эффективным методом является TD-PSOLA. Он характеризуется небольшими вычислительными затратами на его реализацию и оказывается весьма эффективным для преобразования ПОТ и длительности звучания отдельных фрагментов. Сглаживание границ конкатенируемых предзаписанных фрагментов речи обеспечивается использованием окна Хеннинга. Увеличение длительности звучания достигается повторением (мультипликацией) одного или нескольких коротких фрагментов речи. Работа метода TD-PSOLA по преобразованию ПОТ иллюстрируется на рис. 1 .

Конкатенационный синтез обеспечивает высокое качество и естественность звучания синтезируемой речи при относительно низких вычислительных затратах. Однако он требует сравнительно больших затрат памяти для хранения базы дифонов (около 6 Мбайт на один голос для высококачественной речи) и характеризуется привязанностью базы данных к голосу конкретного диктора и к скорости речевого потока.

Затраты памяти могут быть резко снижены за счет хранения в базе данных сжатых представлений дифонов, закодированных высококачественным вокодером, вместо хранения "живого" звука в традиционном формате импульсно-кодовой модуляции. В этом случае конкатенации дифонов предшествует этап декомпрессии (декодирования) дифонов. В зависимости от доступных вычислительных ресурсов (класса ЦСП) может использоваться тот или иной метод речевого кодирования (ADPCM, CELP, MELP и т.д.), обеспечивающий максимально достижимое качество декодированной речи при достаточной степени сжатия базовых дифонов. Это обеспечивает естественность и качество звучания взамен универсальности,

достигаемой в формантном синтезаторе.

Исходным материалом для синтеза речи в ССРТ методом конкатенации является предварительно лингвистически и просодически размеченная фраза, представленная в текстовой или закодированной компактной форме. Алгоритм такого синтеза показан на рис.2.

Этап выделения дифона определяет тип дифона для конкатенации и его характеристики: длительность, разметку ПОТ по длине дифона и т.д., – на основе информации о размеченной фразе, поступающей на вход синтезатора в текстовом или закодированном виде. Источником этой информации может служить либо результат преобразования текста в фонемное представление с интегрированной просодической информацией, полученный с помощью известных систем синтеза речи по тексту (например, MBROLA [10]), либо размеченное вручную на этапе обучения идентичное по содержанию высказывание, записанное диктором.

Дешифрация дифона заключается в поиске и считывании из дифонной базы кодировки дифона с нужными характеристиками.

Декодер преобразует сжатое представление дифона в обычный звуковой сигнал во временной области, сегментированный на 10–20-мсек фреймы. Получаемые при этом значения ПОТ, лежащие обычно в диапазоне от 2 до 20 мсек, интерполируются по всей длительности дифона.

Алгоритм TD-PSOLA выполняется поэтапно:

1) взвешивание фрагментов синтезированного речевого сигнала окном Хеннинга, W_H, (рис.1). При этом центр каждого окна совмещается с положением интерполированных импульсов основного тона. Ширина окна Хеннинга, L_H, выбирается в диапазоне от 2 до 4 локальных ПОТ. Окно Хеннинга описывается формулой:

$$W_H(n) = 0.5 \cdot (1 - cos\,(2\pi \cdot n / (L_H - 1))),$$

2) последовательное наложение взвешенных окном Хеннинга фрагментов с перекрытием, величина которого определяется соотношением ПОТ в исходной размеченной фразе и ПОТ в описании дифона в базе данных. Величина перекрытия соседних фрагментов определяется по формуле:

$$L_{OL} = (L_H/2 + \tau_D - \tau_S),$$

где L_H – размер окна Хеннинга, τ_D и τ_S – значения ПОТ в образце дифона в базе данных и в исходной размеченной фразе. Сами фрагменты при этом не изменяются. В результате амплитуда синтезированного сигнала автоматически интерполируется (сглаживается) на интервале L_{OL} в месте конкатенации двух дифонов, что приводит к уменьшению слышимых искажений звука.

Общее число накладываемых с перекрытием фрагментов, N_D, равно такому ближайшему целому их числу, чтобы получающаяся суммарная длительность дифона, L_D, удовлетворяла неравенству:

$$(L_S - L_H/2) \le L_D \le (L_S + L_H/2),$$

где L_S – длительность дифона, определенная по исходному описанию синтезируемой фразы. Разность ($L_S - L_D$) добавляется к длительности следующего дифона для соблюдения общей длительности синтезируемой фразы. Если длительность дифона в базе данных недостаточна для его воспроизведения в синтезируемой фразе, последние несколько взвешенных фрагментов повторяются с перекрытием требуемое количество раз до заполнения всего дифона. Если длительность дифона больше необходимой, лишние последние кодовые пакеты (фрагменты) отбрасываются.

Синтезированный речевой сигнал по фреймам передается на устройство

воспроизведения звука.

Отметим, что растягивание длительности дифона за счет повторения последних его фрагментов в общем случае ведет к появлению искажений синтезированной речи из-за нарушения контура изменения ПОТ по длине дифона, слышимых как гармонический шум. Поэтому желательно иметь в дифонной базе данных описания дифонов достаточной длительности для воспроизведения синтезированной речи в наиболее медленном темпе произнесения. Практика показывает, что темп произнесения 60 слов в минуту соответствует размеренной речи и достаточен для комфортного восприятия синтезированной речи.

Реализация синтезатора речи на ЦСП

Разработанный и описанный выше алгоритм синтеза речи методом конкатенации дифонов был реализован на типовых ЦСП, относящихся к различным классам: 8-разрядном микроконтроллере и 16-разрядном целочисленном ЦСП. Оценки требуемых ресурсов и объективного качества синтезируемой речи приведены в табл.1.

Таблица 1. Реализация синтезатора речи на ЦСП

ЦСП	**Метод сжатия**	**Вычисл. затраты, MIPS**	**Память программ, кбайт**	**Память данных, кбайт**	**Внешняя память, кбайт**	**MOS**
8-разрядный микрокон-троллер	CELP	9.9	14.46	1.4	105.0	3.74
16-разрядный ЦСП	MELP	5.6	26	8	63.4	3.86

Для реализации была выбрана ССРТ со словарем из 200 слов. Синтезируемая речь воспроизводится мужским голосом. Внешняя память используется для хранения дифонной базы данных. Оценка качества синтезированной речи показана в системе оценок MOS (Mean Opinion Score), в которой практически неискажающий вокодер ADPCM 32 кбит/с имеет оценку 4.1, а широко использующийся в GSM телефонии вокодер QCELP-8 имеет оценку 3.45.

Анализ табл.1 показывает, что синтезатор речи, обеспечивающий хорошее качество синтезируемой речи, может быть успешно реализован даже на 8-разрядном микроконтроллере с предельной производительностью 10 MIPS при использовании CELP-вокодера для декомпрессии дифонов. Недостатком такой реализации является необходимость использования значительной внешней памяти для хранения дифонной базы данных. Использование более мощного 16-разрядного ЦСП позволяет снизить требования к объему памяти для хранения дифонной базы и повысить качество синтезируемой речи за счет применения более сложного и качественного MELP-вокодера. При этом вычислительные ресурсы такого ЦСП задействуются не полностью, что предоставляет возможность одновременной реализации на нем нескольких разных приложений (например, распознаватель и синтезатор речи) или многоканального синтезатора речи.

Оба варианта синтезатора речи (8- и 16-разрядный) могут быть с успехом реализованы на РРП. В первом случае целесообразно использовать многопроцессорную реализацию РРП на основе 8-разрядных процессоров, имеющих общую память для хранения дифонной базы. Такая реализация обеспечивает увеличение числа каналов синтезатора речи при некотором ухудшении качества синтезируемой речи. Второй

вариант подразумевает использование многопроцессорную реализацию РРП на базе 16-разрядных процессоров. Такая реализация позволит сократить объем требуемой внешней памяти и одновременно повысить слышимое качество звучания синтезированной речи за счет небольшого сокращения числа каналов синтезатора.

Диалоговая система общения пользователя с электронной средой

Под электронной средой традиционно понимается вся совокупность электронных средств и приборов, окружающих человека в его повседневной жизни. К ней относятся как профессиональное оборудование, так и бытовая техника. Связанные в локальную сеть, эти устройства могут управляться одним центральным процессором, приемные микрофоны и датчики которого располагаются в различных помещениях жилого или офисного здания. Управляя работой "подчиненных ему" электронных приборов, центральный процессор находится в постоянном ожидании голосовой команды, произносимой пользователем. При ее получении процессор анализирует ситуацию и реализует требуемые для выполнения данной команды действия, оповещая пользователя о результатах ее исполнения. Кроме того, диалоговая система позволяет предупредить пользователя о возможном некорректном задании команды и запросить повторение или уточнение ее. Тем самым снижается уровень возможных ошибок и повышается эффективность голосового управления электронной средой.

Таким образом, диалоговая система общения пользователя с электронной средой предполагает реализацию как минимум двух подсистем: распознавателя и синтезатора речи. Распознаватель речи воспринимает команды пользователя и формирует "программу действий", требующихся для выполнения задания (приказа, запроса и т.д.), сформулированного пользователем в речевой фразе. Синтезатор речи служит средством голосового подтверждения принятой команды и, при необходимости, сообщения пользователю результатов ее выполнения.

Техническая реализация такой диалоговой системы зависит от вычислительных средств управления электронной средой. Роль центрального процессора может выполнять как РС (в дополнение к прочим своим функциям, обычным для РС), так и менее мощное вычислительное устройство, построенное на основе современного ЦСП (например, РРП) в качестве ядра. Выбор того или иного варианта определяется многими факторами, начиная от демографического состава коллектива пользователей (пол, возраст проживающих в доме) до финансового состояния пользователей.

Простейшая реализация, основанная на использовании 8-разрядных микроконтроллеров, позволит обеспечить весь спектр функций диалоговой системы общения пользователя с электронной средой при минимальных финансовых затратах и приемлемом качестве распознавателя и синтезатора речи. Использование более мощных ЦСП обеспечит более высокое качество работы диалоговой системы за счет некоторого удорожания реализации. Алгоритмы синтезатора речи могут быть настроены на любую из возможных реализаций и реализованы на РРП.

Заключение

Рассмотренные проблемы разработки алгоритмов синтеза речи и интеграции средств речевого анализа и синтеза показывают, что конкатенационный синтез на основе дифонов является наиболее эффективным средством реализации высококачественной диалоговой системы общения пользователя с электронной средой. Он позволяет реализовать диалоговую систему с хорошим качеством синтезируемой речи на РРП с различными функциональными возможностями и ограничениями, обусловленными областью применения синтезатора и степенью загруженности РРП

выполнением других задач в параллель с синтезатором. Разработанные алгоритмы составляют базу для выбора разрядности процессорных элементов параллельной архитектуры РРП, оценки эффективности использования различных вариантов архитектуры РРП для реализации речевых приложений и сравнения ее с типовыми ЦСП.

Предлагаемые решения основаны на учете особенностей слухового и речевого аппаратов человека. Они делают принципиально возможной реализацию диалоговой системы на простейших представителях ЦСП – 8-разрядных микроконтроллерах, обеспечивающих качество синтезированной речи на уровне 3.7 MOS при требуемой производительности 10 MIPS. Использование более мощных 16-разрядных ЦСП позволяет поднять качество синтезированной речи до уровня 3.8 MOS при меньшей требуемой производительности, что обеспечивает возможность реализации многоканального синтезатора речи. Оба технических решения являются основой для разработки вариантов параллельной архитектуры РРП, предназначенных для покрытия всего спектра речевых приложений, включая речевой синтез.

Предлагаемое техническое решение является реализацией синтезатора речи нижнего уровня, синтезирующего речь на основе заранее подготовленной информации о составе и особенностях произнесения синтезируемой фразы. Оно открыто для реализации проектов диалоговой системы общения пользователя с электронной средой различной сложности. В простейшем случае исходная кодировка фразы готовится заранее и хранится в памяти синтезатора. В более сложном случае диалоговая система может включать синтезатор верхнего уровня, который формирует кодировку синтезируемой фразы автоматически в зависимости от ее состава и привходящих обстоятельств, обеспечивая тем самым лучшее качество синтезируемой речи.

Литература

1. Klatt D., “*Review of Text-to-Speech Conversion for English*”, Journal of the Acoustical Society of America, JASA, 1987, Vol. 82 (3), pp.737-793.

2. “*DECTalk Software: Text-to-Speech Technology and Implementation*”, http://www.digital.com/info/DTJK01/

3. Klatt D., “*Software for a Cascade/Parallel Formant Synthesizer*”, Journal of the Acoustical Society of America, *JASA*, 1980, Vol. 67, pp. 971-995.

4. Charpentier F.J., Stella M.G., “*Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation*”, Proc. ICASSP’86, Tokyo, 1986, pp.2015 — 2018.

5. Lemmetti S., “*Methods, Techniques, and Algorithms*”, http://www.acoustics.hut.fi/~slemmett/dippa/chap5.html.

6. Moulines E., Emerard F., Larreur D., Le Saint Milon J., Le Faucheur L., Marty F., Charpentier F., Sorin C., “*A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech*”, Proceedings of ICASSP’90, 1990, Vol. 1, pp. 309-312.

7. Moulines E., Laroche J., “*Non-Parametric Techniques for Pitch-Scale Modification of Speech*”, Speech Communication, Vol. 16, 1995, pp. 175-205.

8. Dutoit T., Leich H., “*MBR-PSOLA: Text-to-Speech Synthesis Based on an MBE Re-Synthesis of the Segments Database*”, Speech Communication, Vol. 13, 1993, pp. 435-440.

9. Donovan R., “*Trainable Speech Synthesis”,* PhD. Thesis. Cambridge University Engineering Department, England, 1996.

10. *MBROLA Project Homepage*, 1998, http://tcts.fpms.ac.be/synthesis/mbrola.html.

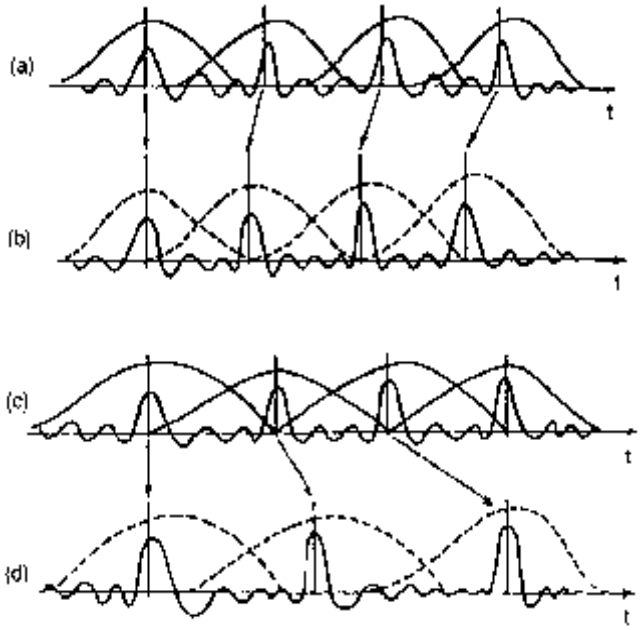

Рис.1. Модификация ПОТ методом PSOLA:
(a), (c) — исходный синтезируемый сигнал;
(b) — результат сокращения ПОТ;
(d) — результат увеличения ПОТ

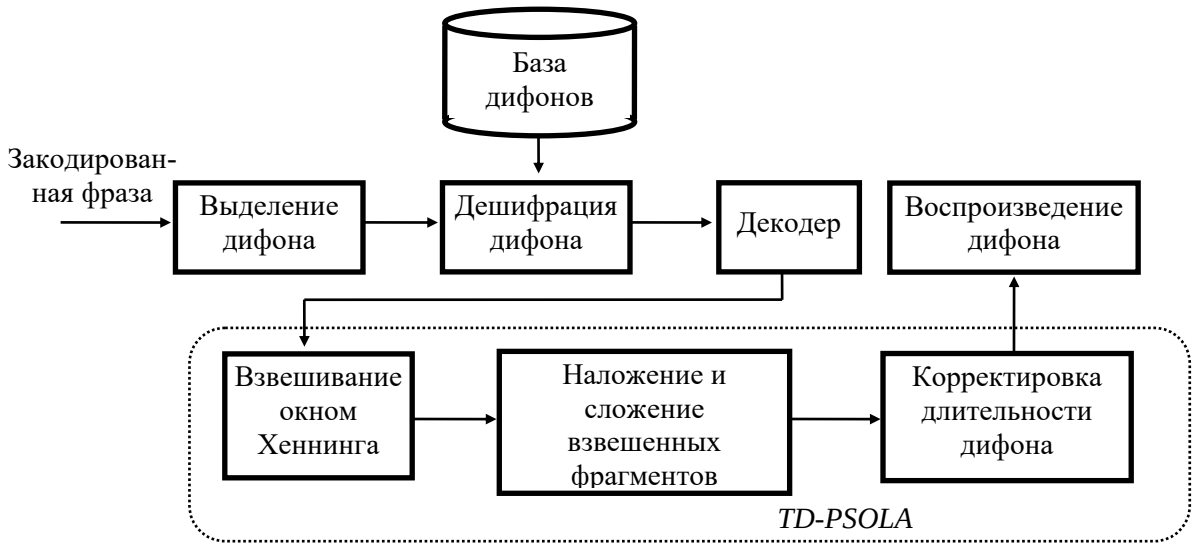

Рис.2. Алгоритм синтеза речи