качестве при адаптации системы к диктору, а так же при изменении геометрии тракта одного и того же диктора во время произношения различных фонем, так как диапазон частот в мгновенном спектре сигнала и длина речеобразующего тракта L корреллируются. Поэтому в структуре анализатора в зависимости от L выбирается тот или иной ФНЧ, рассчитывается соответствующая частота дискретизации и количество отсчетов в кадре, подвергаемом спектральному анализу в данный момент.

ЛИТЕРАТУРА

1. Акинфиев Н.И. К вопросу построения теории речевых сообщений// Сб. научн. труд. ГОС НИИ МРТП СССР. - 1957. - Вып.4.- С.3-25.
2. А.с. № 143430 (СССР). Устройство для автоматического слежения за артикуляционными параметрами речи по речевому сигналу с возможностью выделения сигнала-остатка речи/ Акинфиев Н.И.. Опубл. в БИ, 1961. - № 24.
3. Saito S, Itakura F. The Theoretical consideration of Statistically Optimum Methods for Speech Spectral Density Report N3107. Electrical Communication Laboratory NT.T.-Tokyo,1966.
4. Atal B.S. Schroeder M.R. Predictive Coding of Speech Signals. Proc. Conf. Commun. and Process, 1967.- P. 360-361.
5. Маркел Д.Д., Грей А.Х. Линейное предсказание речи. - М.: Связь, 1980.
6. Фант Г. Анализ и синтез речи. - Новосибирск: Наука, 1970.
7. Методы автоматического распознавания речи/ Под ред. Ли У.- М.: Мир, 1983.
8. Paige A., Zue V. Calculation of Vocal Tract Lendth IEEE Transactions on Audio and Electroacoustics 1970.-Vol.AU-18.
9. Зааль Р. Справочник по расчету фильтров. - М: Радио и связь, 1983.

г. Минск

ЧАСТЬ II. СИНТЕЗ РЕЧИ

УДК 621.391

Б.М.Лобанов

МИКРОВОЛНОВОЙ СИНТЕЗ РЕЧИ ПО ТЕКСТУ

Введение

Синтез речи по тексту предполагает наличие определенных процедур (правил) модификации акустических характеристик каждой фонемы в зависимости от ее окружения, позиции в речевой единице, ударения, интонации и других факторов. Поэтому в системах синтеза речи по тексту чаще всего используют формантный синтез сигналов [1,2], позволяющий в широких пределах изменять акустические характеристики звуков и таким образом моделировать эффекты коартикуляции, ассимиляции, редукции фонем, управлять мелодическим, ритмическим и динамическим контурами речи. С использованием формантного синтезатора достигается высокое качество синтезированной речи, однако возможности дальнейшего совершенствования ограничиваются в настоящее время неполнотой моделей речеобразования как в целом, так и части моделирования индивидуальных свойств человеческого голоса. В современных формантных синтезаторах практически отсутствует учет взаимодействия источников возбуждения и речевого тракта, динамики изменения формы импульсов возбуждения, индивидуальных свойств речевого тракта и др. Поэтому с помощью формантного синтезатора с трудом удается синтезировать близкий к женскому голос, имитировать эстетически приятные голоса или просто копировать любой наперед заданный голос.

Выходом из этого положения могло бы стать использование в системах синтеза речи по тексту отрезков естественной речевой волны. При этом необходимо выбрать в качестве элемента речевой волны такой ее отрезок, который, с одной стороны, позволял бы путем комбинации элементов получать все необходимое многообра-

зие фонетически значимых звуков речи, а с другой - осуществлять их модификацию в соответствии с просодическими правилами. Очевидно, что в качестве этого элемента не могут быть выбраны традиционно-используемые при параметрическом синтезе речи такие ее элементы, как слоги, дифоны или даже сегменты фонем и аллофонов.

В настоящей работе в качестве элементов естественного речевого сигнала предлагается использовать короткие отрезки волн (микроволн), соизмеримые с периодом основного тона. Сущность микроволнового синтеза заключается в представлении речевого сигнала конечным числом заранее выбранных типов волновых форм (ВФ). Число различных ВФ должно быть выбрано таким, чтобы отразить все значимое разнообразие импульсных реакций вокального тракта в процессе речеобразования. Ориентировочное число ВФ, необходимое для синтеза речи, может быть подсчитано исходя из опыта формантного синтеза речевых сигналов. При синтезе звонких звуков используются 3 управляемых по частоте форманты: F_1, F_2, F_3. Опытным путем установлено, что хорошее качество речи сохраняется при квантовании этих параметров на следующее число градаций: F_1 - 8, F_2 - I6, F_3 - 4. Разборчивость синтезированной речи все еще остается хорошей, если число градаций будет снижено до 4 для F_1, 8 - для F_2 и 2 - для F_3. Если допустить, что при синтезе используется все возможное разнообразие значений этих параметров, то получим следующие оценки необходимого числа ВФ:

$$N_{max} = 8 \times 16 \times 4 = 512;$$

$$N_{min} = 4 \times 8 \times 2 = 64.$$

Реально при синтезе речи используется не более половины возможных комбинаций, так что требуемое число ВФ для синтеза звонких звуков лежит в пределах от 32 до 256. Для синтеза шумных звуков необходимо дополнить набор ВФ отрезками шумовых сигналов, число которых лежит в пределах от одного до нескольких десятков.

По-видимому, впервые идея использования в качестве элементов синтеза ВФ, соизмеримых с периодом, высказана в работе [4].

В работе [5] идея синтеза с использованием набора ВФ успешно апробирована в системе дифонного синтеза речи по тексту для мужского и женского голосов. Однако до сих пор не нашли удовлетворительного решения вопросы определения оптимального

набора ВФ, их плавного соединения в текущем речевом потоке, адекватной модификации параметров ВФ в соответствии с просодическими правилами изменения мелодики, ритмики и динамики речи. Решению этих задач, без которых невозможен качественный синтез речи по тексту в полном объеме, посвящена настоящая работа.

I. Микроволновое представление фонем

В основу микроволнового представления фонем положен принцип последовательного разложения фонем на аллофоны и аллофонов - на составляющие их сегменты. Возможна различная степень детальности разложения каждой фонемы на аллофоны и аллофонов на сегменты. Здесь мы ограничимся одним из возможных разложений, достаточным как для понимания сущности этой процедуры, так и для обеспечения необходимого многообразия реализаций каждой фонемы и аллофона при синтезе слитной речи.

Из всего множества аллофонов русских гласных фонем (U, O, A, E, Y) целесообразно в первую очередь взять их мягкие варианты /U̇, Ȯ, Ȧ, Ė, I /, а также соответствующие им назализованные варианты /Ū, Ō, Ā, Ē, Ȳ/ и /Ū̇, Ō̇, Ā̇, Ē̇, Ī/. Каждый аллофон целесообразно представить в виде, по крайней мере, 3 последовательных сегментов: начального, срединного и конечного. При этом тип срединного сегмента (стационарная часть гласной) зависит только от типа выбранного аллофона, а тип начального и конечного сегментов зависит, кроме того, от типа предшествующей и последующей фонемы. На языке формантного описания начальный, срединный и конечный сегменты задают соответственно начало, середину и конец формантных переходов на гласной. Начальный и конечный переходы определяются местом образования предшествующей и последующей фонем [2], так что число типов начального и конечного (переходных) сегментов гласной определяется числом типов фонем, отличающихся местом образования.

Для русских согласных необходимо различать губное, зубное, альвеолярное, велярное и латеральное место образования. Таким образом, для описания переходных сегментов каждого аллофона гласной необходимо иметь до 5 различных типов ВФ. Общее представление каждой гласной фонемы в виде набора ВФ, необходимых для описания всех сегментов и аллофонов, дано в табл. I.

Таблица I

Фонема	/U/				/O/			
Аллофон / Сегмент	U	$\dot{U}$	$\bar{U}$	$\bar{\bar{U}}$	O	$\dot{O}$	$\bar{O}$	$\bar{\bar{O}}$
срединный	W_{11}^{u}	W_{21}^{u}	W_{31}^{u}	W_{41}^{u}	W_{11}^{o}	W_{21}^{o}	W_{31}^{o}	W_{41}^{o}
переходн. губной	W_{12}^{u}	W_{22}^{u}	W_{32}^{u}	W_{42}^{u}	W_{12}^{o}	W_{22}^{o}	W_{32}^{o}	W_{42}^{o}
переходн. зубной	W_{13}^{u}	W_{23}^{u}	W_{33}^{u}	W_{43}^{u}	W_{13}^{o}	W_{23}^{o}	W_{33}^{o}	W_{43}^{o}
переходн. альвеоляр.	W_{14}^{u}	W_{24}^{u}	W_{34}^{u}	W_{44}^{u}	W_{14}^{o}	W_{24}^{o}	W_{34}^{o}	W_{44}^{o}
переходн. велярный	W_{15}^{u}	W_{25}^{u}	W_{35}^{u}	W_{45}^{u}	W_{15}^{u}	W_{25}^{o}	W_{35}^{o}	W_{45}^{o}
переходн. латеральн.	W_{16}^{u}	W_{26}^{u}	W_{36}^{u}	W_{46}^{u}	W_{16}^{u}	W_{26}^{o}	W_{36}^{o}	W_{46}^{o}

Из всего множества аллофонов русских согласных целесообразно в первую очередь взять их варианты, обусловленные эффектом коартикуляции с последующим гласным [2]. При этом для твердых согласных достаточно выделить три аллофона: согласный перед /O/, /U/ - C^{u}, перед /A/ - C^{a} и перед /E/, /Y/ - C^{e}. Кроме того, выделяется один вариант мягких согласных C'. Для каждого аллофона согласных определяются три временных сегмента - начальный, срединный и конечный. Таким образом, описание каждого согласного в виде набора ВФ может быть представлено с помощью табл. 2.

Описанное представление русских фонем в виде ВФ сегментов аллофонов допускает значительное изменение их количества как в сторону увеличения, так и в сторону уменьшения в зависимости от предъявляемых требований к качеству синтезированной речи либо к объему запоминаемой информации.

Таблица 2

Фонема	/Z/				/L/			
Аллофон / Сегмент	Z^{e}	$Z^{'}$	Z^{α}	Z^{u}	L^{e}	$L^{'}$	L^{α}	L^{u}
срединный	W_{11}^{z}	W_{21}^{z}	W_{31}^{z}	W_{41}^{z}	W_{11}^{L}	W_{21}^{L}	W_{31}^{L}	W_{41}^{L}
начальный	W_{12}^{z}	W_{22}^{z}	W_{32}^{z}	W_{42}^{z}	W_{12}^{L}	W_{22}^{L}	W_{32}^{L}	W_{42}^{L}
конечный	W_{13}^{z}	W_{23}^{z}	W_{33}	W_{43}^{z}	W_{13}^{L}	W_{23}^{L}	W_{33}^{L}	W_{43}^{L}

2. Соединение ВФ в речевом потоке

Для синтеза звонких звуков используются БФ, вырезанные из речевого сигнала на соответствующих стационарных и переходных участках звуков, с длительностью, равной периоду основного тона. Соединение ВФ на стационарных участках сводится просто к их последовательному считыванию. Для переходных участков такая процедура могла бы подойти только в случае предварительной записи нескольких ВФ на каждый тип переходного участка. Это практически трудноосуществимо как в плане трудоемкости подготовки и препарирования исходного речевого материала, так и в плане чрезмерного разрастания требуемого объема памяти и количества правил синтеза переходных участков.

Существует одна интересная возможность обойти указанные трудности, основанная на использовании инерционных свойств слухового восприятия человека, аналогичных зрительному. Хорошо известно, что впечатление плавного замещения одной слайд-картинки другой может быть достигнуто путем плавного уменьшения яркости (от исходной до нуля) одного изображения и одновременного увеличения яркости (от нуля до необходимой) другого изображения, спроектированных на один и тот же экран. Проведенные нами исследования показали, что аналогичный эффект замещения присущ и слуховому восприятию звуков. Слуховой эффект плавного замещения достигается путем создания интервала перекрытия двух звуков с постепенным уменьшением амплитуды первого звука и одновременным увеличением амплитуды второго на интервале перекрытия (рис.I,а).

В результате суммирования (рис.I,б) в звуковом поле на участке перекрытия образуется сложный звук, воспринимаемый как плавный переход от первого ко второму звуку. В какой-то мере плавность этого перехода фиксируется и на сонаграмме (рис.I,в).

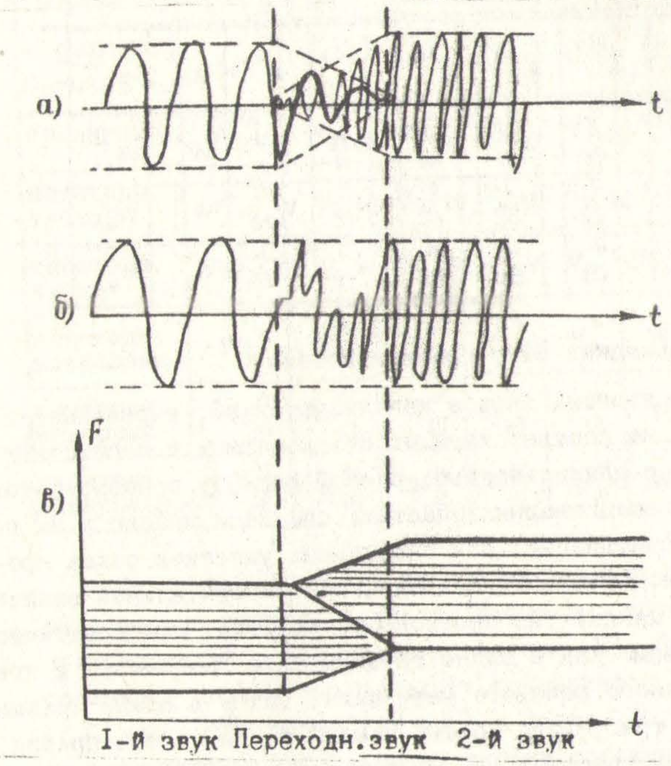

Рис. I

Математическое описание алгоритма, реализующего механизм плавного замещения последовательности ВФ, может быть дано следующим образом. Пусть для определенности требуется осуществить плавный переход на интервале гласной фонемы от начальной волны W_{12} к срединной W_{11} и затем к конечной W_{13}. Это соответствует (см.табл.I) переходному процессу на гласной в сочетании: "губная согласная - гласная - зубная согласная" (например, в слове "РАТ").

Обозначим длительности начального перехода T_{12}, срединного стационара T_{11} и конечного перехода T_{13}. Требуемые значения длительностей берутся из таблиц, подобных табл. I,2. В соответствии с изложенными выше правилами плавного замещения ВФ на участке гласной необходимо сформировать два сигнала:

$$S_1(t)=\begin{cases} W_{11}(t)\frac{1}{T_{12}}t, & 0 \le t \le T_{12} \\ W_{11}(t), & T_{12} < t < T_{12}+T_{11} \\ W_{11}(t)\frac{1}{T_{12}}(T_{12}-t), & T_{12}+T_{11} \le t \le T_{12}+T_{11}+T_{13}, \end{cases} \qquad (I)$$

$$S_2(t)=\begin{cases} W_{12}(t)\frac{1}{T_{12}}(T_{12}-t), & 0 \le t < T_{12} \\ 0, & T_{12} < t < T_{12}+T_{11} \\ W_{13}(t)\frac{1}{T_{13}}(t-T_{12}-T_{11}), & T_{12}+T_{11} \le t \le T_{12}+T_{11}+T_{13}. \end{cases} \qquad (2)$$

Огибающие этих сигналов изображены на рис.2.

Сигнал $S(t)$, в котором реализуется эффект плавного замещения последовательности ВФ W_{12}, W_{11}, W_{13} на интервале звучания гласной фонемы, образуется из сигналов (I), (2) путем простого суммирования: $S(t)=S_1(t)+S_2(t)$.

Исследования, проведенные с использованием различных типов ВФ, показали, что слуховой эффект их плавного замещения возникает лишь в том случае, когда соответствующий им переходный процесс в естественной речи характеризуется одновременным линейным движением формантных частот (рис.3,а). Если же в естественной речи при переходе от ВФ W_1 к W_2 (рис.3,б) движение формантных частот не удовлетворяет этому условию, то для достижения слухового эффекта плавного замещения ВФ - W_1 и W_2 - необходимо ввести промежуточную ВФ - W_{12}. В частности, рис.3,а иллюстрирует переход от гласной /А/ к /Е/, а рис 3,б - от /А/ к /I/.

3. Управление просодическими параметрами

Просодика речи (мелодика, динамика, ритмика) задается путем текущего управления частотой основного тона, амплитудой и длительностью звуков. Рассмотрим особенности и способы управления этими параметрами при микроволновом способе синтеза.

Простейшим способом управления частотой основного тона является следующий. Пусть исходная ВФ имеет длительность T_0', причем T_0' выбрана внутри определяемого просодическими правилами диапазона изменения длительности периодов основного тона:

$$T_{0min} < T_0' < T_{0max}.$$

В качестве конкретного значения T_0' может быть взято среднестатистическое значение периода основного тона речи диктора, используемого при формировании набора ВФ. Если текущее значение $T_0 = T_0'$,

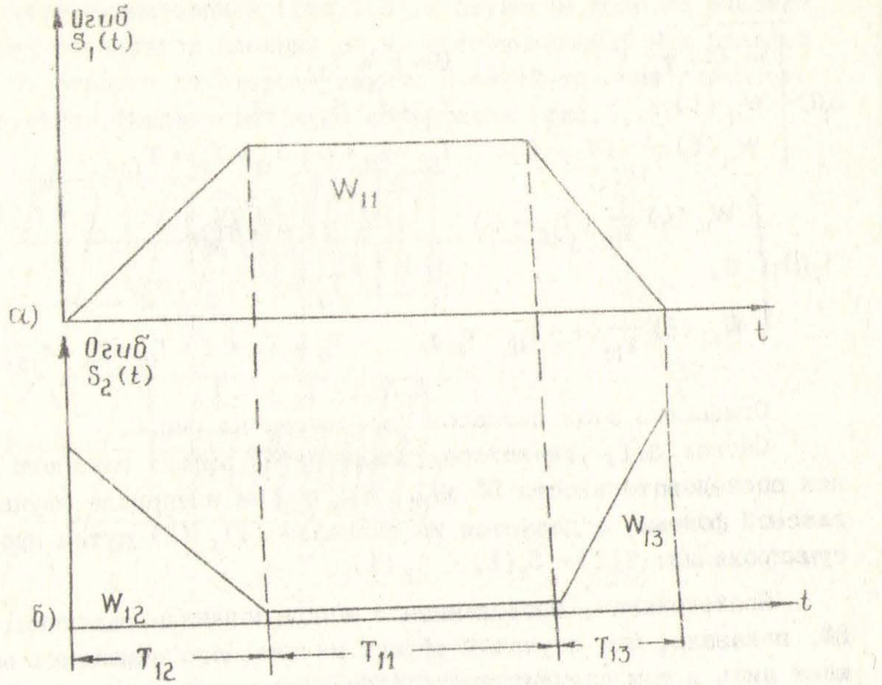

Рис. 2

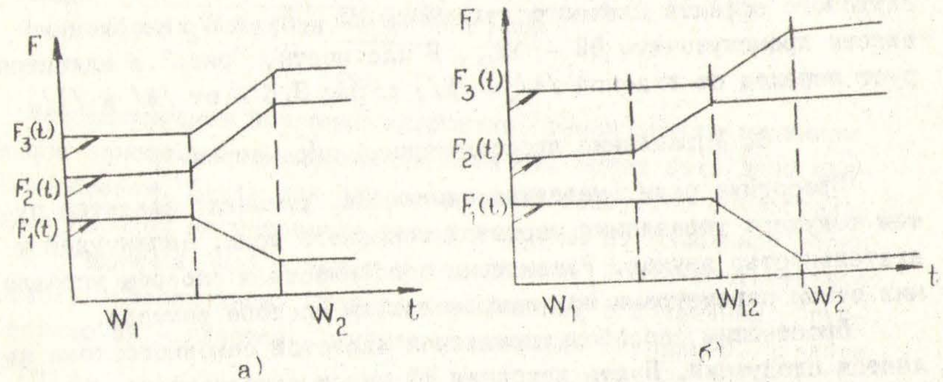

Рис. 3

то речевой сигнал образуется путем простого повторения заданий ВФ (рис.4,а). При $T_0 \geq T_0'$ повторное считывание ВФ начинается спустя временной интервал $T_0 - T_0'$, а сам интервал заполняется нулями. При $T_0 < T_0'$ в момент $t = T_0$ мгновенно прекращается считывание ВФ и начинается процесс ее повторного считывания.

Экспериментальное исследование данного метода управления частотой основного тона показало, что он обеспечивает достаточно высокое качество синтезированного звука. Это определенно можно утверждать для случая $T_0 > T_0'$, если интервал $T_0 - T_0'$ не превышает 30% от длительности периода T_0. Для случая $T_0 < T_0'$ искажения не заметны на слух, если момент прекращения считывания приходится на значение ВФ вблизи нуля (10% - 20% от амплитуды ВФ). В противном случае (см.рис.4,в) наблюдается отчетливое искажение звучания, напоминающее назализацию (гнусавость). Этот дефект может быть устранен путем соответствующего сглаживания процесса резкого прекращения считывания. Это может быть достигнуто, например, следующими двумя способами. В первом способе в моменты начала повторного считывания включается фильтр 2-го порядка с постоянной времени $\tau = 1/4\ T_0$ (рис.5,б). Во втором способе перед началом повторного считывания ($\tau = 1/4\ T_0$) сигнал ВФ умножается на гладкую единичную функцию (рис.5,в). Это может быть, например, функция вида

$$y = e^{-\alpha t^2}.$$

Использование одного из этих методов для случая $T_0 < T_0'$ дает вполне удовлетворительные результаты. Для случая $T_0 > T_0'$ можно использовать способ дополнения периода нулями (см.рис.4,б) при условии, что выбирается $T_0' \approx 0{,}7\ T_{0\,max}$.

Рассмотрим далее особенности управления двумя другими просодическими характеристиками: длительностью и амплитудой звуков. Определяемая ритмическим контуром требуемая длительность звонких звуков может быть получена путем задания необходимого числа периодов на каждом фонемном сегменте. Если требуемая длительность i-го фонемного сегмента равна T_i, то в среднем для ее реализации необходимое число $\bar{n}_i$ периодов определяется по формуле

$$\bar{n}_i = \frac{T_i}{T_{0i}},$$

где T_{oi} - среднее значение длительности периода основного тона на i-м фонемном сегменте, задаваемого мелодическим контуром. Точное значение необходимого числа периодов определяется путем сравнения требуемой длительности сегмента T_i и суммарной текущей длительности периодов основного тона, реализуемых на этом сегменте. При этом длительность сегмента задается с погрешностью, равной длительности последнего периода основного тона. Длительность шумных звуков задается путем простого подсчета необходимого количества отсчетов соответствующего шумового сигнала.

Последний просодический параметр - текущая амплитуда звука - задается путем умножения последовательности отсчетов звука на текущие значения динамического контура.

4. Синтез последовательности ВФ по фонемному тексту

Для синтеза речи по фонемному тексту необходимо реализовать процедуру генерации последовательности ВФ, описывающих каждую фонему с учетом ее текстового окружения. Процедуру генерации ВФ удобно описать, базируясь на понятии портрета фонемы [2]. Под портретом фонемы понимается некоторое универсальное описание, заданное в виде набора констант или некоторых функций, достаточное для генерации множества ее временных сегментов с учетом позиционно-комбинаторной изменчивости. В отличие от формантных портретов фонем [2] волновые портреты могут быть описаны существенно более компактно. Это связано с тем, что в ВФ уже заложена информация о формантных характеристиках звука. В волновом портрете фонемы нужно лишь указать, какой конкретной ВФ W^j описывается тот или иной сегмент фонемы, на какой длительности T и какой амплитуды A^j необходимо задать выбранную ВФ для описания данного сегмента, а также за какое время τ^j должно установиться стационарное значение ВФ для данного сегмента. Обобщенный волновой портрет фонемы, задаваемый на трех последовательных временных сегментах, представлен в табл.3.

Из табл.3 видно, что конкретный выбор ВФ на каждом сегменте определяется не только типом текущей синтезируемой фонемы, но и ее непосредственным окружением в тексте. В общем случае выбор требуемой ВФ осуществляется с помощью многозначной функции

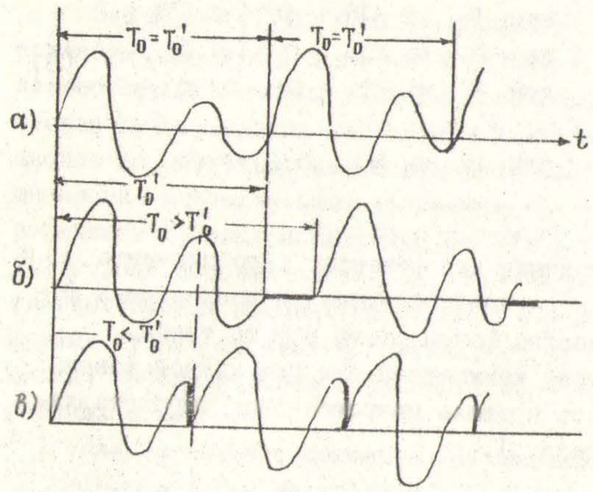

Рис. 4

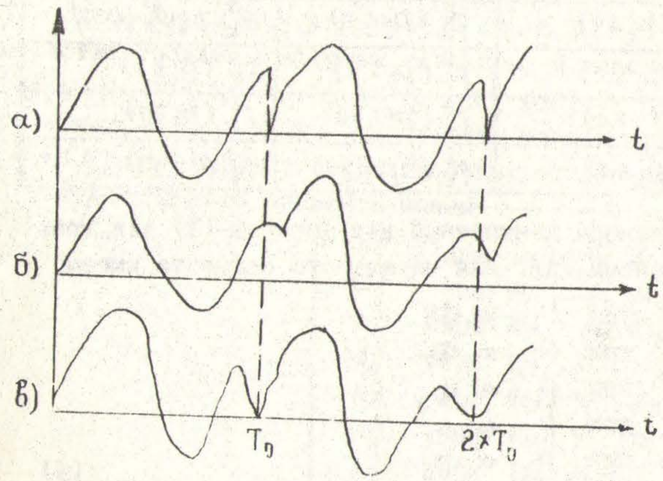

Рис. 5

$$W_i^j = \begin{cases} W_{11}^j & \text{при } p_{i-1} \in \Phi_1,\ p_{i+1} \in \Phi_1; \\ W_{12}^j & \text{при } p_{i-1} \in \Phi_1,\ p_{i+1} \in \Phi_2; \\ W_{21}^j & \text{при } p_{i-1} \in \Phi_2,\ p_{i+1} \in \Phi_1; \\ \dots\dots\dots\dots\dots\dots & \\ W_{m,n}^j & \text{при } p_{i-1} \in \Phi_m,\ p_{i+1} \in \Phi_n; \\ \dots\dots\dots\dots\dots\dots & \end{cases} \qquad (3)$$

Здесь W_i^j - ВФ, необходимая для синтеза j-го сегмента i-й фонемы текста; p_{i-1} , p_{i+1} - предшествующая и последующая фонемы текста; Φ_m, Φ_n - множества фонем m-го и n-го типов.

Формула (3) приобретает конкретный вид для каждой фонемы и ее сегментов, если учесть правила микроволнового представления фонем, изложенные в разд. I.

Таблица 3

Параметр \ Сегмент	начальный	срединный	конечный
W_i^j	$f_w^1(p_i \pm 1)$	$f_w^2(p_i \pm 1)$	$f_w^3(p_i \pm 1)$
T_i^j	$f_T^1(p_i \pm 1)$	$f_T^2(p_i \pm 1)$	$f_T^3(p_i \pm 1)$
A_i^j	$f_A^1(p_i \pm 1)$	$f_A^2(p_i \pm 1)$	$f_A^3(p_i \pm 1)$
τ_i^j	$f_\tau^1(p_i \pm 1)$	$f_\tau^2(p_i \pm 1)$	$f_\tau^3(p_i \pm 1)$

Приведем для примера конкретный вид формулы (3) для трех сегментов гласной фонемы /A/. Для начального сегмента имеем

$$W_i^i = \begin{cases} W_1^1 & \text{при } p_{i-1} \in \Phi_1 \\ W_2^1 & \text{при } p_{i-1} \in \Phi_2 \\ W_3^1 & \text{при } p_{i-1} \in \Phi_3 \\ W_4^1 & \text{при } p_{i-1} \in \Phi_4 \\ W_5^1 & \text{при } p_{i-1} \in \Phi_5 \\ W_6^1 & \text{при } p_{i-1} \in \Phi_6 \\ W_7^1 & \text{при } p_{i-1} \in \Phi_7 \\ W_8^1 & \text{при } p_{i-1} \in \Phi_8 \\ W_9^1 & \text{при } p_{i-1} \in \Phi_9 \\ W_{10}^1 & \text{при } p_{i-1} \in \Phi_{10} \\ W_{11}^1 & \text{при } p_{i-1} \in \Phi_{11}. \end{cases} \qquad (4)$$

Здесь $\Phi_1 = \{P, B, F, V, L\}$ - множество твердых губных и боковых согласных, $\Phi_2 = \{T, D, S, Z, R, C, CH, ZH, K, G, X\}$ - множество зубных, альвеолярных и небных твердых согласных, $\Phi_3 = \{P', B', F', V'\}$ - множество губных мягких согласных, $\Phi_4 = \{T', D', S', Z', R', SH', CH'\}$ - множество мягких зубных и альвеолярных согласных, $\Phi_5 = \{K', G', X'\}$ - множество мягких небных согласных, $\Phi_6 = \{L'\}$ - единичное множество мягких боковых согласных, $\Phi_7 = \{M\}$ - единичное множество твердых губных носовых согласных, $\Phi_8 = \{N\}$ - единичное множество твердых зубных носовых согласных, $\Phi_9 = \{M'\}$ - единичное множество мягких губных носовых согласных, $\Phi_{10} = \{N'\}$ - единичное множество мягких зубных носовых согласных, $\Phi_{11} = \{U, O, A, E, Y, \#\}$ - множество гласных и паузы.

Для срединного сегмента /A/ имеем

$$W_i^2 = \begin{cases} W_1^2 & \text{при } p_{i-1} \in \Phi_{12} \\ W_2^2 & \text{при } p_{i-1} \in \Phi_{13} \\ W_3^2 & \text{при } p_{i-1} \in \Phi_{14} \\ W_4^2 & \text{при } p_{i-1} \in \Phi_{15}. \end{cases} \qquad (5)$$

Здесь $\Phi_{12} = \{U, O, A, E, Y, L, R, V, Z, ZH, B, D, P, T, G, F, K, S, SH, X, C\}$ - множество твердых неносовых согласных и гласных, $\Phi_{13} = \{J, L', R', V', Z', B', D', G', P', T', K', F', S', SH', X', CH\}$ - множество мягких согласных, $\Phi_{15} = \{M', N'\}$ - множество мягких носовых согласных, $\Phi = \{M, N\}$ - множество твердых носовых согласных.

Для конечного сегмента имеем

$$W_1^3 = \begin{cases} W_1^3 & \text{при } p_{i+1} \in \Phi_1 \\ W_2^3 & \text{при } p_{i+1} \in \Phi_2 \\ W_3^3 & \text{при } p_{i+1} \in \Phi_3 \\ W_4^3 & \text{при } p_{i+1} \in \Phi_4 \\ W_5^3 & \text{при } p_{i+1} \in \Phi_5 \\ W_6^3 & \text{при } p_{i+1} \in \Phi_6 \\ W_7^3 & \text{при } p_{i+1} \in \Phi_7 \\ W_8^3 & \text{при } p_{i+1} \in \Phi_8 \\ W_9^3 & \text{при } p_{i+1} \in \Phi_9 \\ W_{10}^3 & \text{при } p_{i+1} \in \Phi_{10} \\ W_{11}^3 & \text{при } p_{i+1} \in \Phi_{16} \\ W_{12}^3 & \text{при } p_{i+1} \in \Phi_{17} \\ W_{13}^3 & \text{при } p_{i+1} \in \Phi_{18} \\ W_{14}^3 & \text{при } p_{i+1} \in \Phi_{19} \\ W_{15}^3 & \text{при } p_{i+1} \in \Phi_{20}. \end{cases} \qquad (6)$$

Здесь множества Φ_1 -Φ_{10} те же, что в формуле (4), а множества Φ_{16}, Φ_{17}, Φ_{18}, Φ_{19}, Φ_{20} являются единичными и содержат соответственно гласные /U, O, A, E, Y/.

С учетом конкретных правил аллофонической изменчивости аналогичные формулы выбора ВФ сегментов могут быть записаны для каждой фонемы.

Выбор длительности и амплитуды сегментов фонем осуществляется по тем же формулам, что и выбор ВФ, так что каждой ВФ может быть сопоставлена длительность ее повторения и амплитуда.

Как уже указывалось, процесс установления ВФ на каждом сегменте может быть реализован путем моделирования слухового эффекта плавного замещения ВФ. Выбор времени замещения τ^j осуществляется по формулам, аналогичным (но не идентичным) формулам для выбора ВФ. Для гласных фонем $\tau_г^1 = T^1$, $\tau_г^2 = 0$, $\tau_г^3 = T^3$, т.е. первый сегмент замещается вторым за время, равное длительности первого сегмента, на втором сегменте процесс замещения отсутствует, а на интервале длительности третьего сегмента осуществляется процесс замещения второго сегмента третьим (рис.6). Для согласных фонем устанавливаются фиксированные значения τ_c^1, τ_c^2, τ_c^3, зависящие только от типа согласной фонемы. При этом τ_c^1 устанавливает время замещения ВФ последнего сегмента (i-I)-й фонемы первым сегментом i-й согласной фонемы, τ_c^2 устанавливает время замещения первого сегмента вторым и третьим. Особая роль отведена τ_c^3, которое определяет опережающее замещение третьего сегмента i-й согласной фонемы первым сегментом (i + I)-й гласной фонемы и реализуется только в случае, если (i + I)-я фонема является гласной (см.рис.6).

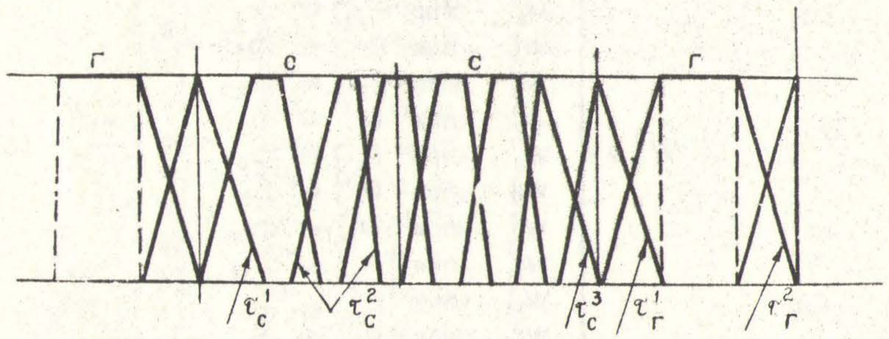

Рис. 6

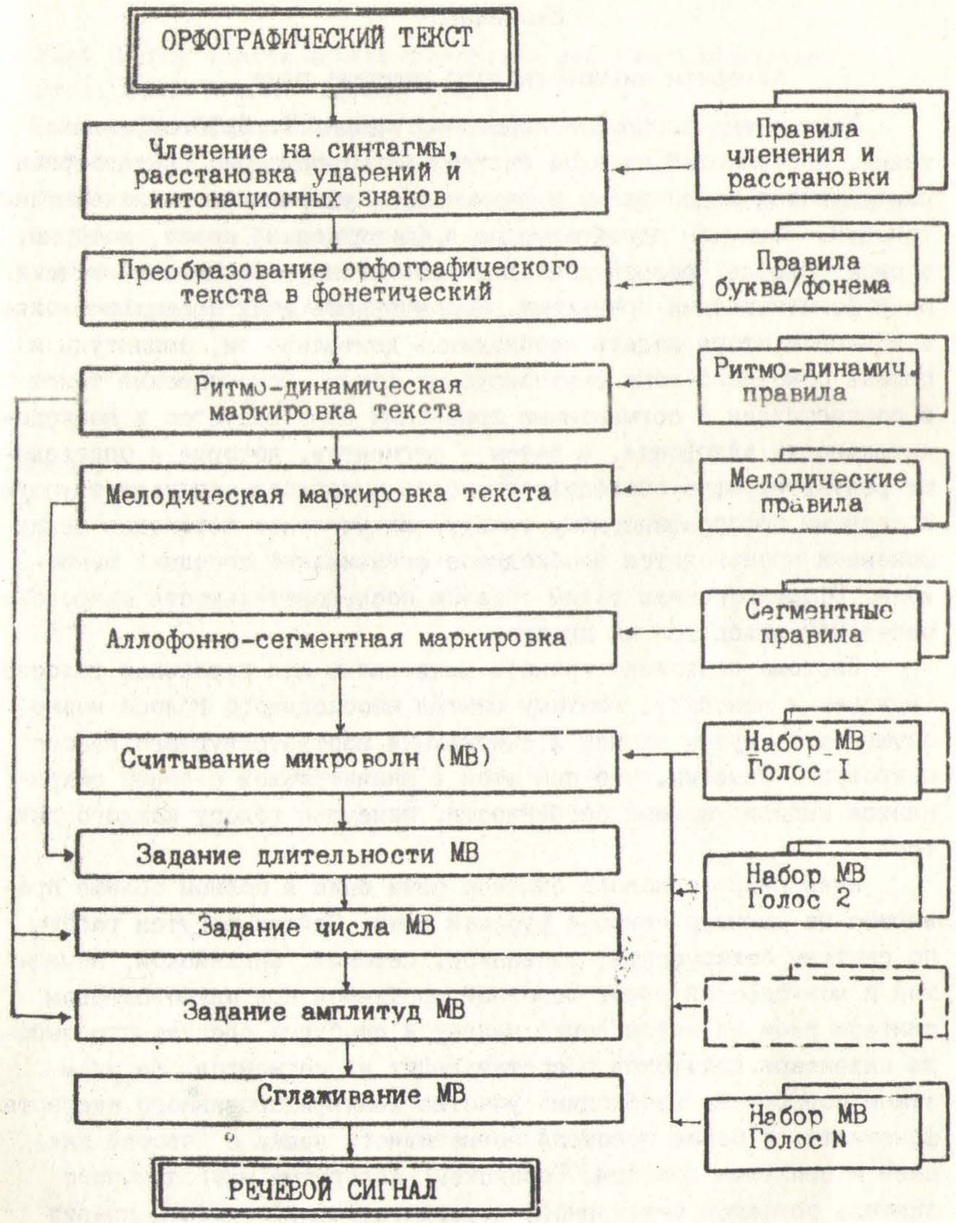

Рис. 7. Блок-схема метода микроволнового синтеза речи

Заключение

Алгоритм микроволнового синтеза речи

Блок-схема алгоритма приведена на рис.7. Орфографический текст, поступающий на вход системы микроволнового синтеза речи, расчленяется на синтагмы и размечается ударениями и знаками интонации. Затем он преобразуется в фонетический текст, который, в свою очередь, размечается в соответствии с ритмо-динамическими и фонетическими правилами. Совокупность этих правил позволяет в конечном итоге задать необходимые длительности, амплитуды и период основного тона синтезируемых фонем. Фонетический текст в соответствии с сегментными правилами преобразуется в последовательность аллофонов, а затем - сегментов, которые и определяют результирующую последовательность микроволн, соответствующую исходному орфографическому тексту. На участках переходов между фонемами производится необходимое сглаживание соседних микроволн. Сформированная таким образом последовательность микроволн через ЦАП выводится на динамик.

Система позволяет хранить микроволны для различных голосов (мужских и женских), поэтому синтез необходимого голоса можно осуществить путем записи и считывания соответствующего набора микроволн. Заметим, что при этом в значительной степени сохраняются индивидуальные особенности, присущие голосу каждого диктора.

Идея микроволнового синтеза речи была в полном объеме проверена на примере синтеза русской речи. Сейчас ведутся работы по синтезу белорусской, словацкой, чешской, английской, немецкой и монгольской речи. Основной проблемой при микроволновом синтезе речи на новом языке является проблема адекватного выбора инвентаря аллофонов и составляющих их сегментов. На этом этапе совершенно необходимо участие квалифицированного эксперта-фонетиста, а также носителя нормативного языка с хорошей дикцией и приятным голосом. Блок-схема алгоритма многоязычного синтеза остается неизменной, изменяются только наборы правил и наборы необходимых микроволн.

ЛИТЕРАТУРА

1. Klat D. The klattalk text-to-speech convesation system. Proc. IEEE ICASSP, Paris, 1982.
2. Lobanov B.M. The Phonemafone text-to-speech system. Proc. ICPhS, Tallinn,1987.
3. Morel M. Synthese vokale par recordment de segment d'oooscillogrames Revue d'Acoustique, vol.14, no 56, 1981.
4. Hamon C. Synthese par concatenation de formes d'ondes. Note thechnique, NT/LAA/TSS/355 CNET, 1988.

г. Минск