УДК 004.934.5, 004.522

Б.М. Лобанов[1], Л.И. Цирульник[1], М. Железны[2], З. Крноул[2], А. Ронжин[3], А. Карпов[3]

СИСТЕМА АУДИОВИЗУАЛЬНОГО СИНТЕЗА РУССКОЙ РЕЧИ

Описываются имитационная и компиляционная модели аудиовизуального синтеза русской речи и созданный на их основе аудиовизуальный синтезатор речи. Рассматриваются преимущества и недостатки каждой из моделей, а также особенности их построения для русской речи.

Введение

Мировая тенденция развития речевых технологий указывает на актуальность органичного включения визуальной информации в качестве дополнительного канала восприятия и распознавания речи [1]. Визуальная информация очень важна при распознавании и восприятии речи в шумах и незаменима для людей с ограниченным слухом или дефектами произношения. Число исследований в области аудиовизуального распознавания и синтеза речи постоянно увеличивается. Разработкой бимодальных систем синтеза речи занимаются, в частности, научные коллективы Кембриджского университета, Великобритания; Политехнического университета, Монз, Бельгия; Университета Крита, Греция; Университета г. Загреб, Хорватия; Университета Западной Богемии, Чехия.

Для решения задачи создания системы аудиовизуального синтеза речи (часто называемой «говорящая голова») существует два подхода: *имитационный*, при котором создается 2D- или 3D-модель лица и настраиваются управляющие параметры для передачи мимики, выражения лица и движения губ при речевой деятельности [2, 3], и *компиляционный*, при котором «говорящая голова» формируется путем выбора соответствующих видеофрагментов или изображений из визуальной базы данных (БД) конкретного диктора [4, 5].

Преимуществом первого подхода является меньший физический объем данных, необходимых для синтеза визуальной речи. К недостаткам имитационного подхода можно отнести большую вычислительную сложность его реализации, а также недостаточно реалистичные результаты при персонализации «говорящей головы» (рис. 1), связанные с неизбежной схематичностью отображения речедвижений.

При компиляционном подходе требуемый объем БД возрастает; при этом, однако, вычислительная сложность реализации существенно уменьшается. Кроме того, компиляционный подход представляется более предпочтительным при создании системы персонализированного аудиовизуального синтеза речи по тексту, которое стало возможным благодаря успешному развитию теории и технологии компьютерного клонирования персональных характеристик голоса и речи [6–9].

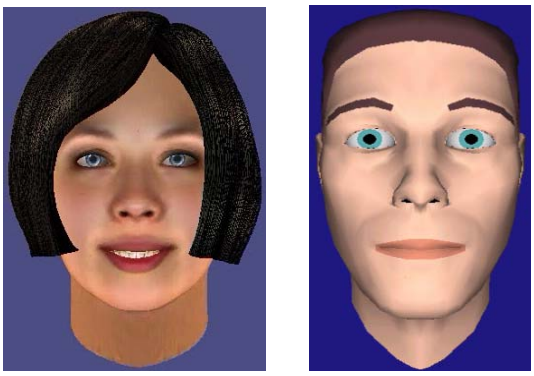


Рис. 1. Примеры реализации компьютерной «говорящей головы»

Важной задачей при разработке систем аудиовизуального синтеза речи по тексту, возникающей как при имитационном, так и при компиляционном подходах, является создание визуальных единиц речи – *визем* (изображений лица при произнесении различных фонем). При разработке классов визем необходимо учитывать положение и движение артикуляторных органов, доступных обозрению. В английской речи, например, видимыми артикуляторными органами являются зубы, движения губ, нижней челюсти, языка [10]. Для русской речи систематических исследований в данном направлении до сих пор не проводилось.

Еще одной задачей, решение которой необходимо для высококачественного аудиовизуального синтеза речи, является синхронизация звукового и визуального потоков. Степень синхронности потоков фонем и визем в процессе речеобразования для разных языков различна. Так, для японского языка движения губ и звуковой поток речи практически синхронны; английскому же языку (особенно американскому варианту) сопутствует достаточно богатая артикуляция, что зачастую вызывает определенные временные расхождения между потоками фонем и визем.

В настоящей работе исследуются оба метода реализации «говорящей головы» – компиляционный и имитационный. Особое внимание уделено разработке классов визем для русской речи, описана процедура синхронизации виземного и фонемного потоков, а также синтеза переходных участков видеоизображений речи.

1. Компиляционная модель синтеза видеоизображений речи

В основу компиляционной модели положен метод плавной сшивки отдельных кадров изображения, предложенный ранее для плавной сшивки звуковых волн в микроволновом синтезаторе речи [11, 12]. Основная идея компиляционного синтеза видеоизображений речи заключается в следующем:

1. Полное множество фонем русской речи $\{Ph\}$ разбивается на M подмножеств $\{Ph_i\}$, каждому из которых соответствует определенная визема.

2. Для каждой виземы устанавливается ее относительная длительность N_v, определяемая числом кадров показа изображения виземы:

$$N_v = k\frac{T_a}{n} + 1, \qquad (1)$$

где T_a – длительность текущего аллофона*, задаваемая синтезатором речи; k – коэффициент, изменяющийся в интервале [0..1], текущее значение которого определяется типом синтезируемого аллофона; n – число кадров в секунду, равное 24 согласно стандартам видеоформатов.

3. Для каждой пары визем устанавливается длительность перехода от одной виземы к другой, задаваемая таблично числом переходных кадров.

1.1. Создание визем русской речи

Определение необходимого и достаточного набора визем осуществляется на основе известной классификации фонем и аллофонов русской речи по артикуляторным признакам способа и места образования с учетом эффектов коартикуляции и редукции [13]. Как показали исследования, для русской речи практически полностью скрытой остается динамика движения тела, кончика и боковинок языка, небной занавески. Обозрению доступны лишь движения губ и нижней челюсти. Наиболее четко они проявляются при образовании гласных (рис. 2), а также губных согласных (рис. 3). Не столь сильное, однако достаточно заметное различие наблюдается между твердыми и мягкими согласными, а также между заднеязычными и другими негубными согласными (рис. 3).

*Аллофон – звуковая реализация фонемы в потоке речи.

Рис. 2. Виземы гласных фонем (положение губ и нижней челюсти)

Рис. 3. Виземы групп согласных фонем

Наиболее яркие различия в виземах связаны с изображениями губ говорящего (рис. 4). При этом проявляются три характерных координаты движения губ:

степень растягивания – координата X (наибольшее значение у гласной «И»);

степень раскрытия – координата Y (наибольшее значение у гласной «А»);

степень выпячивания – координата Z (наибольшее значение у гласной «У»).

Рис. 4. Степени растягивания, раскрытия и выпячивания губ, характерные для визем гласных фонем

Исходя из описанных выше явлений, выбран необходимый и достаточный набор визем русской речи (табл. 1). Индексы при гласных в таблице указывают на степень их позиционной редукции: 0 – полноударная гласная, 1 – частично ударная, 2 – предударная, 3 – заударная. Символ «'» после согласной обозначает ее мягкость.

Процесс создания изображений визем может быть выполнен двумя способами: с использованием видеосъемки диктора, произносящего фонетически сбалансированный текстовый корпус (например, мини-текст из приложения 1 в [14]); с использованием моментальных фотографий лица диктора, который имитирует произношение того или иного звука, соответствующего каждой виземе. Опыт создания визем показал, что второй способ является более предпочтительным как в плане меньшей трудоемкости создания визем, так и в плане возможности осу-

ществления более надежной фиксации стандартного положения головы диктора при создании изображений различных визем.

Таблица 1

Соответствие «фонема – визема» для аудиовизуального синтеза речи

Визема	Аллофоны фонем	Визема	Аллофоны фонем
V_1	$А_0$, $А_1$	V_8	Б', П', М'
V_2	$Е_0$, $Е_1$	V_9	Ф, В
V_3	$И_0$, $И_1$, $И_2$, $И_3$	V_{10}	Ф', В'
V_4	$О_0$, $О_1$	V_{11}	Ц, С, З, Ш, Ж, Д, Т, Л, Р, Н
V_5	$У_0$, $У_1$, $У_2$, $У_3$	V_{12}	С', З', Ч', Ш', Д', Т', Л', Р', Н'
V_6	$ЫІ_0$, $ЫІ_1$, $ЫІ_2$, $ЫІ_3$, $А_2$, $А_3$, $Е_2$, $Е_3$, пауза	V_{13}	Г, К, Х
V_7	Б, П, М	V_{14}	Г', К', Х', Й'

1.2. Установка длительностей показа изображения виземы и переходов между виземами

Синхронизация показа визем с синтезированным речевым сигналом осуществляется на основе информации о позиции моментов начала и конца каждого аллофона в текущем речевом потоке. На визуальном уровне необходимо задать три участка, суммарная длительность которых равна реальной длительности звучания каждого аллофона t_a: начальный переход t_b, стационарный участок t_m и конечный переход t_f. На рис. 5 схематично представлен динамический процесс отображения стационарных участков и переходов одного из параметров визем – степени раскрытия губ (координата Y в соответствии с введенными в п. 1.1 обозначениями). Динамическое визуальное отображение звучания i-го аллофона складывается из изображений последовательности кадров начального, стационарного и конечного участков. При этом длительность полного перехода t_c от $(i-1)$-го аллофона к i-му складывается из участков конечного и начального переходов.

Таким образом, для синхронизации видео- и аудиопотоков необходимо определить текущие длительности участков t_b и t_f (исчисляемые количеством кадров) на основе информации о позиции моментов начала и конца каждого аллофона и его длительности t_a. При этом длительность показа каждого кадра будет равна 40 мс (с учетом принятой частоты обновления видеоизображения, равной 25 кадров в секунду).

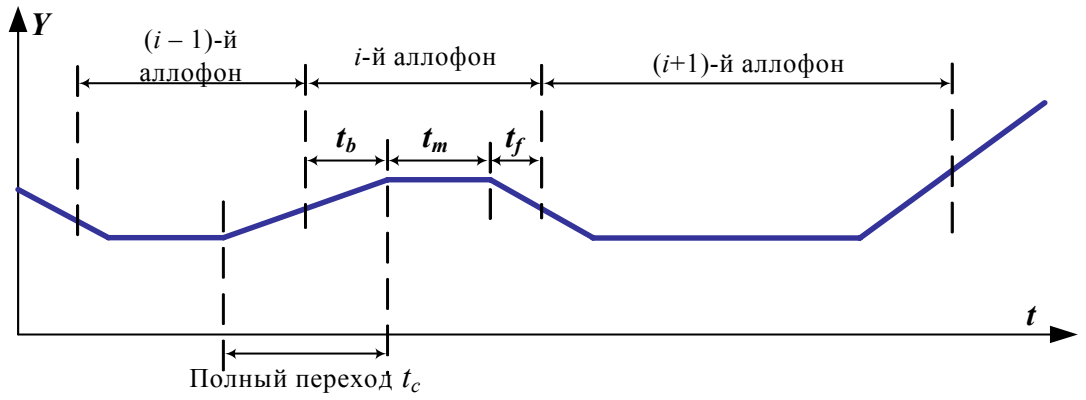


Рис. 5. Процесс отображения последовательности визем (Y – степень раскрытия губ, t – время)

Текущая длительность каждого аллофона t_a задается системой синтеза речи, исходя из собственной средней длительности аллофона и требуемого темпа речи. В табл. 2 приведены значения относительной длительности аллофонов (в %) при изменении темпа речи, а также их абсолютная длительность в миллисекундах (мс) и в числе видеокадров (кд).

Таблица 2

Относительные (%) и абсолютные (мс, кд) длительности звуков при изменении темпа речи

Тип звуковых единиц	Медленный темп (% – мс – кд)	Средний темп (% – мс – кд)	Быстрый темп (% – мс – кд)
Паузы	250 – 650 – 16	100 – 260 – 7	20 – 50 – 2
Ударные гласные	200 – 320 – 8	100 – 160 – 4	50 – 80 – 2
Предударные гласные	200 – 160 – 4	100 – 80 – 2	80 – 64 – 2
Заударные гласные	200 – 80 – 2	100 – 40 – 1	80 – 20 – 1
Сонанты	140 – 110 – 3	100 – 80 – 2	80 – 64 – 2
Звонкие взрывные и щелевые	120 – 120 – 3	100 – 100 – 3	80 – 80 – 2
Глухие взрывные	130 – 160 – 4	100 – 120 – 3	85 – 100 – 3
Глухие щелевые	130 – 180 – 4	100 – 140 – 4	85 – 120 – 3

Как видно из табл. 2, длительность самых коротких звуков может быть отражена только одним кадром, который должен соответствовать стационарному участку t_m, причем участки t_b и t_f отсутствуют. Это не означает, однако, что данные участки обязательно будут равны нулю на (i–1)-м и (i+1)-м аллофонах. Следовательно, переходный процесс может быть отображен даже при быстром темпе синтезируемой речи.

При определении длительности стационарного участка виземы (кроме рассмотренной выше информации о позиции моментов начала и конца каждого аллофона и его текущей длительности t_a) следует учитывать также хорошо изученное явление коартикуляции гласных и согласных фонем [13]. На визуальном уровне это явление проявляется в том, что в слогах типа «согласная – гласная» характерный артикуляционный уклад гласной фонемы устанавливается не только на самой гласной, но и на большей части согласного. При этом для различных комбинаций «согласная – гласная» проявление эффекта коартикуляции может быть различным. Для заднеязычных согласных «Х», «Г», «К» эффект коартикуляции проявляется в комбинации с любой из гласных, в то время как для остальных согласных – только в сочетании с губными гласными «У» и «О». Таким образом, показанный на рис. 5 процесс отображения последовательности визем и переходов отражает лишь случаи, когда эффект коартикуляции между фонемами отсутствует (например, в слове «Пенсильвания», в котором отсутствуют губные гласные и заднеязычные согласные).

На рис. 6 показано проявление эффектов коартикуляции на примере фонетического слова «В КАССУ», содержащего сочетания заднеязычной согласной «К» и гласной «А», согласной «С» и губной гласной «У». Стрелками указаны горизонтальные стационарные участки соответствующих визем, наклонными линиями – переходные участки, пунктирными линиями – позиции начала и конца произносимых фонем.

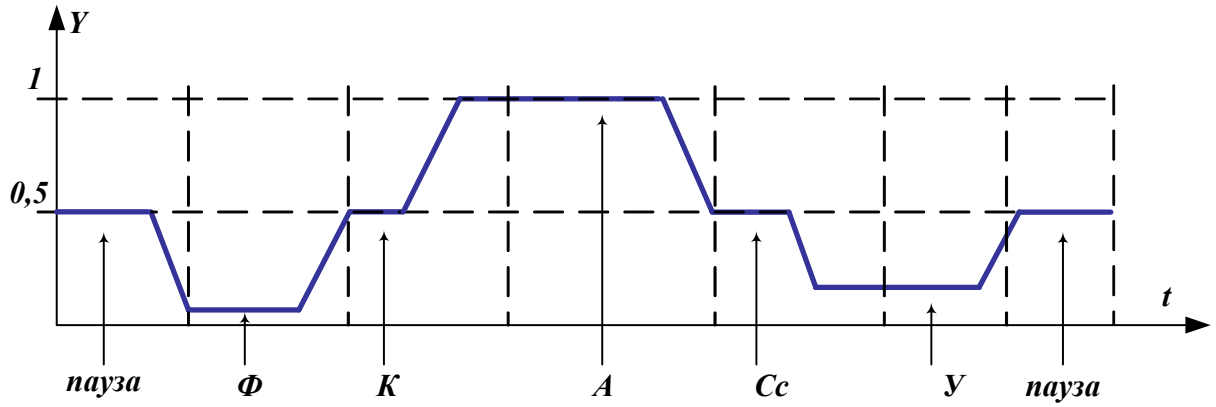


Рис. 6. Эффект коартикуляции (Y – степень раскрытия губ, t – время)

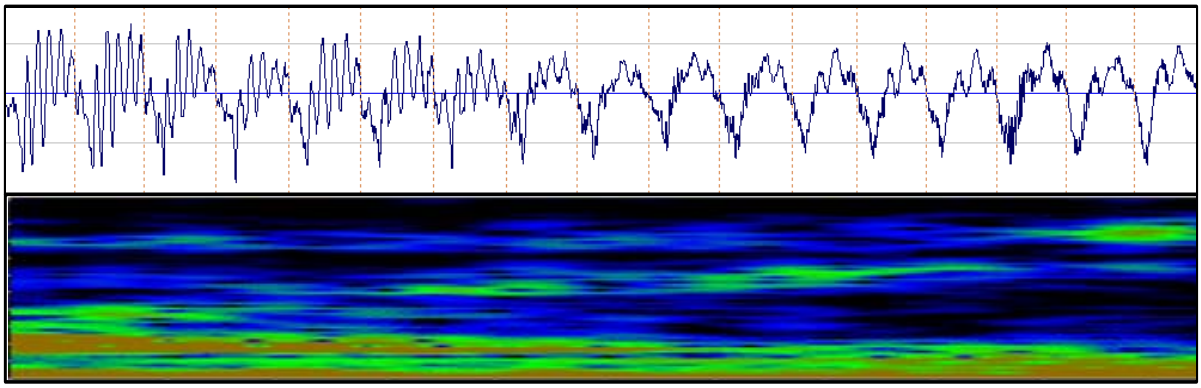
После определения длительности и локализации стационарного участка виземы с учетом эффектов коартикуляции необходимо задать некоторые определенные значения длительностей

переходных участков t_b и t_f. При среднем темпе речи, как показывает опыт, удовлетворительные результаты получаются при выборе значений $t_c = t_f + t_b$ = 70-90 мс (два кадра). В зависимости от требуемого темпа речи эти значения могут изменяться в большую или меньшую сторону.

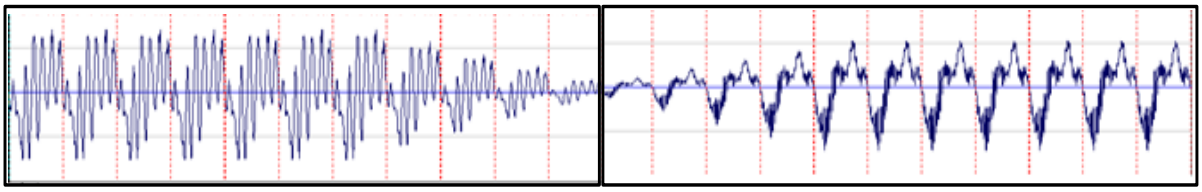
1.3. Синтез переходных участков звуковых и визуальных элементов речи

Для воссоздания непрерывного движения органов артикуляции по отдельным изображениям в [5] используются методы компьютерной анимации. В настоящей работе для создания эффекта непрерывного движения органов артикуляции предлагается другая, значительно более простая процедура – так называемая процедура мягкой сшивки (Soft Lacing, SL) отдельных кадров изображения, хорошо зарекомендовавшая себя при сшивке звуковых волн в микроволновом синтезаторе речи [11, 12].

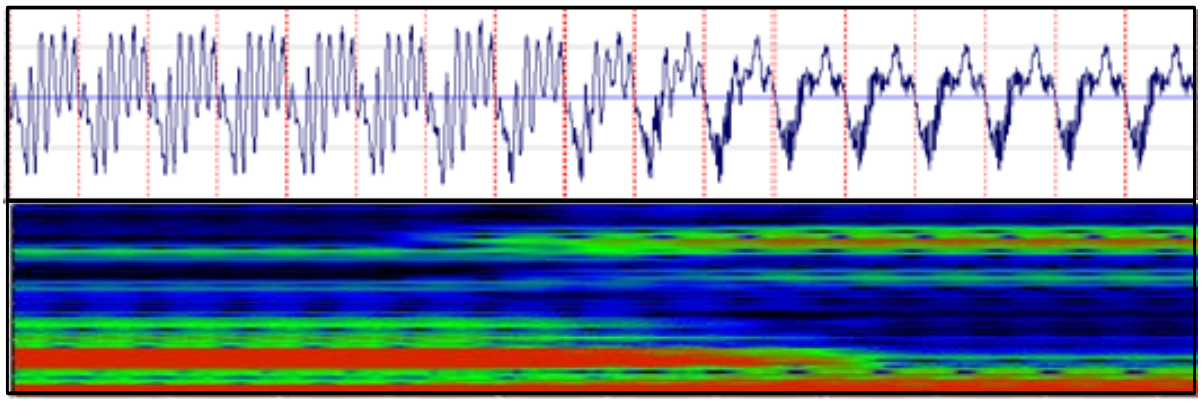
На рис. 7 демонстрируется применение SL-метода для сшивки звуковых волн на примере создания плавного звукового перехода от звука «А» к звуку «И» по двум опорным периодам колебаний, взятым из стационарных участков этих звуков.



а)



б)



в)

Рис. 7. Процедура SL-синтеза стационарных и переходных участков звукосочетания «АИ»: *а*) осциллограмма и спектрограмма естественного сигнала звукосочетания «АИ»; *б*) сформированные затухающий сигнал звука «А» (слева) и усиливающийся сигнал звука «И» (справа); *в*) осциллограмма и спектрограмма синтезированного сигнала звукосочетания «АИ»

Формирование переходных участков происходит путем суммирования затухающего первого сигнала и усиливающегося второго сигнала (рис. 8).

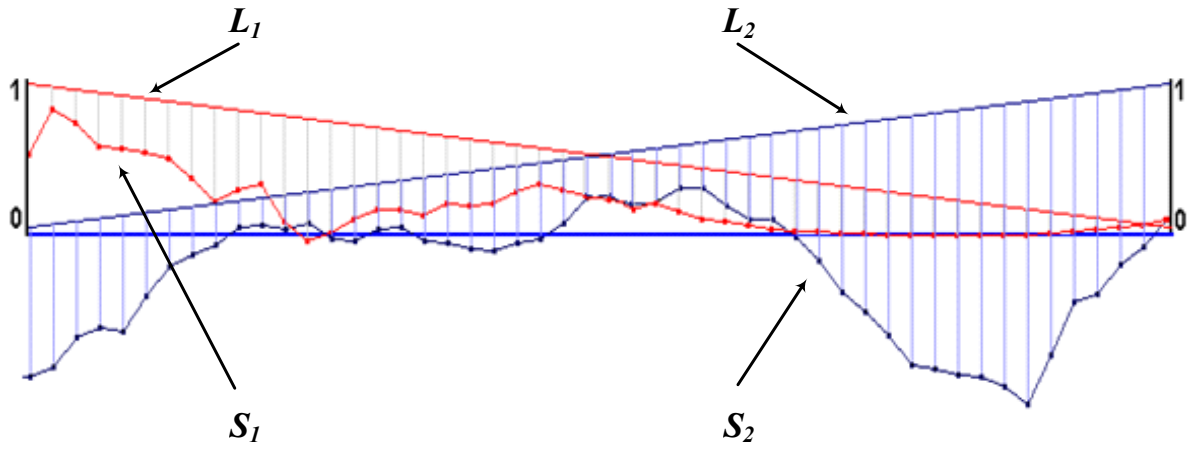


Рис. 8. Расчет участков плавного перехода

Затухание первого сигнала вычисляется по формуле

$$S_1'(n) = S_1(n)L_1(n),\ L_1(n) = 1 - \frac{n}{N},\ 0 \le n \le N, \tag{2}$$

где $S'_1(n)$ – модифицированный первый сигнал; $S_1(n)$ – исходный первый сигнал; N – размер окна плавного перехода.

Усиление второго сигнала выражается формулой

$$S_2'(n) = S_2(n)L_2(n),\ L_2(n) = \frac{n}{N},\ 0 \le n \le N, \tag{3}$$

где $S'_2(n)$ – модифицированный второй сигнал; $S_2(n)$ – исходный второй сигнал.

Вычисление суммарного сигнала переходного участка S' происходит следующим образом:

$$S'(n) = S_1'(n) + S_2'(n),\ 0 \le n \le N. \tag{4}$$

Описанная процедура попериодной плавной сшивки звуковых сигналов соответствует покадровой плавной сшивке визуальных изображений лица при произнесении этих звуков.

В области компьютерной графики процесс слияния двух изображений с целью создания эффекта частичной прозрачности известен как взвешенное наложение смешиваемых цветов или альфа-композиция (alpha-blending) [15]. Эффект прозрачности достигается путем смешивания значений цветов исходного и результирующего пикселей. Альфа-композиция последовательно применяется ко всем пикселям смешиваемых изображений. При этом если изображения представлены в стандарте RGB (red, green, blue) и каждый цвет задается одним байтом, то смешивание происходит отдельно по каждому из трех цветов согласно выражению

$$C = \frac{\alpha C_1}{255} + \frac{(255 - \alpha)C_2}{255},\ 0 \le \alpha \le 255, \tag{5}$$

где α – вес первого изображения относительно второго.

При формировании переходных участков визем значения α для каждой из RGB-составляющих одинаковы и вычисляются в соответствии с формулой

$$\alpha(n) = 255 - \left[255 \frac{n}{N} \right], \ 0 \le n \le N, \tag{6}$$

где n – номер текущего кадра переходного участка; N – количество кадров плавного перехода.

2. Имитационная модель синтеза видеоизображений речи

В основу имитационного метода синтеза видеоизображений речи положена параметрическая трехмерная модель головы человека [16], разработанная для чешского языка и адаптированная для русского.

Модель головы представляет собой набор точек – вершин виртуального пространства, которые соединены дугами, образуя треугольные поверхности, формирующие каркас трехмерной модели (рис. 9). Создание модели головы осуществляется полуавтоматическим методом в ходе видеозаписи и сканирования речи реального диктора с использованием проектора, видеокамеры и системы из четырех зеркал для создания эффекта стереоизображения на изображении от одной видеокамеры. Полученные данные обрабатываются и сохраняются в файле формата виртуальной реальности VRML в виде набора координат вершин каркаса, треугольных плоскостей и соответствующих текстур лица диктора. Полный каркас модели головы описывается несколькими десятками тысяч вершин (рис. 9, *а*), из которых лишь некоторые являются активными, т. е. могут управляться программой, имитируя движения лицевых мускулов. К активным вершинам относится набор точек области губ, управление которыми позволяет отображать визему.

В созданной системе синтеза видеоизображений речи используется не только общая модель головы, но и модели отдельных ее составляющих: глаз, языка, нижней и верхней челюстей, внутренних артикуляторных органов. Эти элементы описываются по тем же принципам, что и модель головы, однако созданы они не на основе обработки видеоизображений, а на базе знаний из антропологической физиологии. Каждая из моделей может управляться программой независимо от самой головы.

Благодаря созданию отдельных моделей артикуляторных органов и функций управления разработанную систему можно использовать для синтеза видеоизображений речи не только для чешского и русского, но и для других языков, в которых такие элементы, как боковинки и тело языка, являются видимыми артикуляторными органами.

а) *б)*

Рис. 9. Имитационная трехмерная модель головы диктора:
а) каркас модели; *б*) модель с наложением текстур

Еще одним преимуществом создания отдельных моделей для элементов «говорящей головы» является возможность управления морганием глаз, что создает иллюзию «живой» головы.

Набор визем для визуального синтеза чешской речи создан на основе анализа движений органов артикуляции нескольких реальных дикторов при произнесении речевого корпуса [16]. Визуальный синтез русской речи осуществляется с использованием функции соответствия ви-

зем для русского и чешского языков, которая учитывает место и способ образования соответствующих фонем.

Трехмерная реконструкция артикуляторных движений по последовательности визем осуществляется в рамках модели [17], использующей следующие признаки:

1) координаты губ и маркеры челюсти (27-мерный вектор);

2) ширину, высоту и степень выпячивания губ (трехмерный вектор);

3) сдвиг от центра: нижней и верхней губы, выпячивания и ширины губ (четырехмерный вектор);

4) четыре коэффициента анализа основных компонент PCA (четырехмерный вектор). Эти компоненты описывают форму губ: первая компонента – степень открытия рта; вторая – ширину и степень выпячивания губ; третья – степень поднятия верхней губы, четвертая – вертикальное положение нижней челюсти.

Программное управление трехмерной моделью головы осуществляется с помощью стандартных матричных преобразований средствами графической библиотеки OpenGL, которые приводят к изменениям положения головы, повороту головы, увеличению/уменьшению масштаба и др.

3. Базовая модель аудиовизуального синтеза речи по тексту и особенности ее реализации

В системе аудиовизуального синтеза речи по тексту входной орфографический текст последовательно подвергается преобразованиям, осуществляемым несколькими процессорами: текстовым, фонетическим, просодическим, акустическим и визуальным (рис. 10).

Текстовый процессор предназначен для преобразования входного орфографического текста в просодически размеченный текст и выполняет следующие задачи:

– разбиение текста на предложения;

– преобразование чисел, аббревиатур, сокращений и т. д.;

– разбиение предложений на просодические синтагмы;

– расстановку сильных и слабых словесных ударений;

– разбиение синтагм на акцентные единицы (АЕ);

– определение интонационного типа синтагм.

Просодически размеченный текст поступает в фонетический процессор, который выполняет следующие задачи:

– преобразование орфографического текста в последовательность фонем;

– генерацию последовательности аллофонов на основе последовательности фонем.

Сформированная аллофонная последовательность поступает на вход двух процессоров: просодического и визуального.

Функции просодического процессора:

– разбиение АЕ на элементы акцентных единиц: предъядро, ядро и заядро;

– вычисление целевых значений частоты основного тона (F_0), амплитуды (A) и длительности аллофонов (T) в соответствии с портретами акцентных единиц для каждой АЕ.

Акустический процессор использует информацию, поступающую от фонетического и просодического процессоров, для выполнения следующих операций:

– модификации просодических параметров звуковых волн аллофонов и мультифонов;

– конкатенации звуковых волн аллофонов и мультифонов в соответствующую последовательность.

Визуальный процессор использует информацию, поступающую от фонетического процессора, для выбора из БД требуемых визем и их конкатенации.

Особенностью реализации системы аудиовизуального синтеза речи по тексту является совместная работа визуального и акустического процессоров. Синхронизация аудио- и визуального потоков на программном уровне осуществляется с помощью стандартных средств обеспечения многозадачности операционной системы Windows.

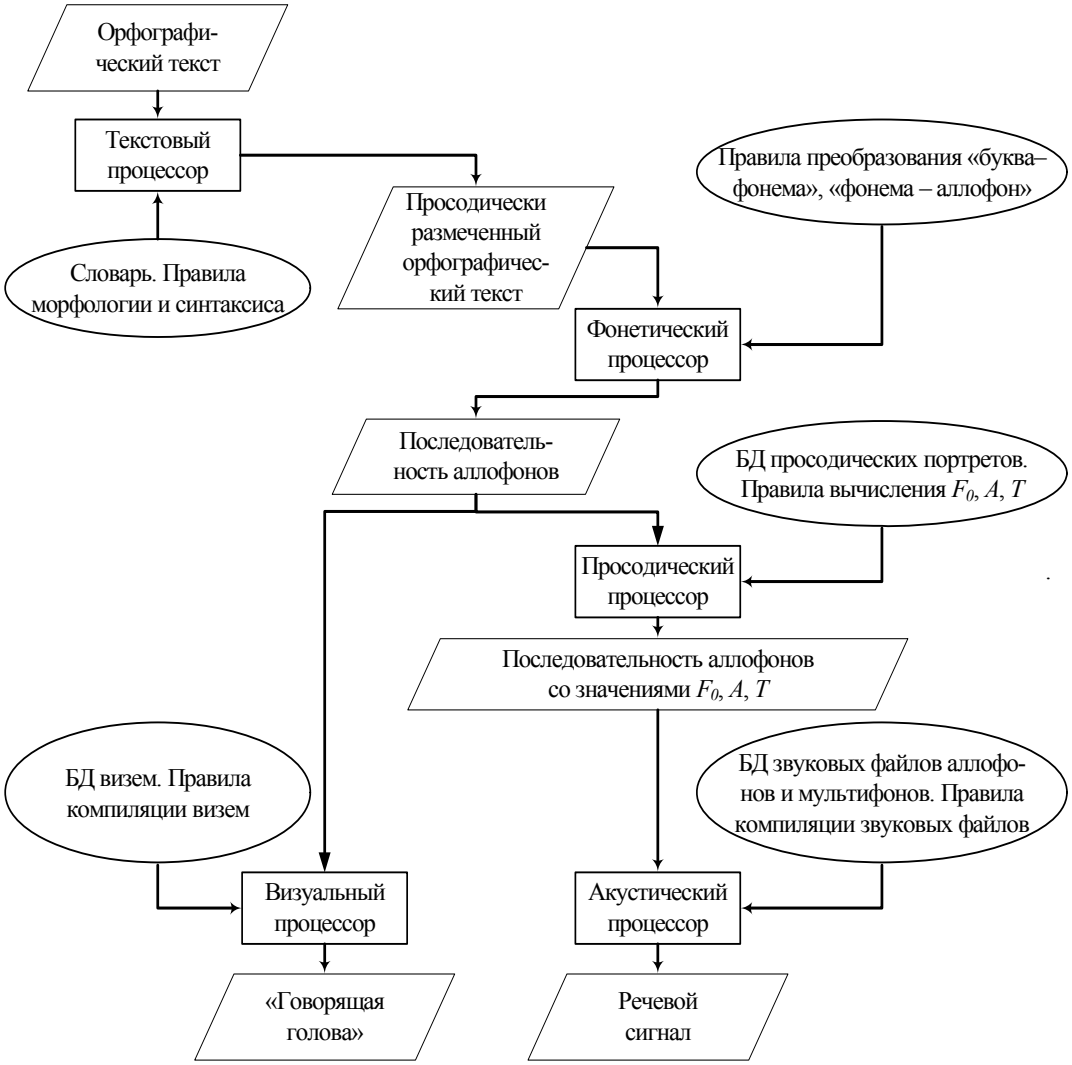


Рис. 10. Общая структура системы аудиовизуального синтеза речи по тексту

Заключение

Рассмотрены компиляционная и имитационная модели «говорящей головы». Представлена система аудиовизуального синтеза русской речи, которая позволяет использовать как одну, так и другую модель.

В описанной системе дикторозависимые и языкозависимые данные и правила преобразований организованы в виде специализированных БД, что позволяет использовать разработанную систему аудиовизуального синтеза речи по тексту как многодикторную и многоязыковую при добавлении соответствующих лингвистических, акустических и визуальных ресурсов.

За рамками данной работы осталось рассмотрение задачи автоматизации создания визем при обработке корпуса видеозаписей речи. На решение данной задачи будут направлены дальнейшие исследования авторов.

Исследование выполнено в рамках совместного российско-белорусского проекта «Модель аудиовизуального синтеза и распознавания речи для интеллектуальных устройств массового обслуживания», поддерживаемого грантами РФФИ и БРФФИ (№ 08-07-90002 и № Ф08Р-016).

Список литературы

1. Issues in Visual and Audio-Visual Speech Processing. – Cambridge : MIT Press, 2004. – 478 p.

2. Tekalp, A.M. Face and 2-D Mesh Animation in MPEG-4 / A.M. Tekalp, J. Ostermann // Signal Processing: Image Communication, Special Issue on MPEG-4. – 2000. – Vol. 15. – P. 387–421.

3. Animated Talking Head with Personalized 3D Head Model / L.S. Chen [et al.] // IEEE First Workshop on Multimedia Signal Processing. – 1997. – P. 274–279.

4. Video Rewrite: Driving Visual Speech with Audio / C. Bregler [et al.] // Proc. of 24 Int. conf. on Computer Graphics and Interactive Techniques «DIGGRAPH97». – Los Angeles, USA, 1997. – P. 353–360.

5. Cosatto, E. Photo-Realistic Talking-Heads from Image Samples / E. Cosatto, H.P. Graf // IEEE Transactions on Multimedia. – Sept. 2000. – Vol. 2. – P. 152–163.

6. Лобанов, Б.М. Компьютерное клонирование персонального голоса и речи / Б.М. Лобанов // Новости искусственного интеллекта. – 2002. – № 5(55). – С. 35–39.

7. Lobanov, B. TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian) / B. Lobanov, H. Karnevskaya // Phonetics and its Applications. – Stuttgart : Steiner, 2002. – P. 445–452.

8. Лобанов, Б.М. Персональные особенности синтагматического членения речи телеведущего Ю. Сенкевича / Б.М. Лобанов, Л.И. Цирульник // Тр. Междунар. конф. «Диалог-2004». – М. : Наука, 2004. – С. 419–423.

9. Lobanov, B. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS / B. Lobanov, L. Tsirulnik // Proc. of 9th Int. conf. «Speech and Computer» (SPECOM'2004). – SPb. : Anatolia, 2004. – P. 17–21.

10. Embodied Conversational Agents / Eds. : J. Cassell, J. Sullivan, S. Prevost, E. Churchill. – Cambridge : MIT Press, 2000. – 420 p.

11. Лобанов, Б.М. Микроволновой синтез речи по тексту / Б.М. Лобанов // Анализ и синтез речи : сб. науч. тр. – Минск : Ин-т техн. кибернетики АН БССР, 1991. – С. 21-38.

12. Lobanov, B. MW-Speech Synthesis from Text / B. Lobanov // Proc. of the XII International Congress of Phonetic Sciences ICPhS'91. – Aix-en-Provense, France, 1991. – P. 128–132.

13. Lobanov, B. Development of Multi-Voice and Multi-Language TTS Synthesizer (languages: Belarussian, Polish, Russian) / B. Lobanov, L. Tsirulnik // Proc. of 11th Int. conf. «Speech and Computer» (SPECOM'2006). – SPb. : Anatolia, 2006. – P. 274–283.

14. Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск : Белорусская наука, 2008. – 344 с.

15. Porter, Th. Compositing Digital Images / Th. Porter, T. Duff // Computer Graphics. – July 1984. – № 18(3). – P. 253–259.

16. Design, Implementation and Evaluation of the Czech Realistic Audio-Visual Speech Synthesis / M. Železný [et al.] // Signal Processing. – 2006. – № 86. – V. 12. – P. 3657–3673.

17. Krňoul Z. Innovations in Czech Audio-Visual Speech Synthesis for Precise Articulation / Z. Krňoul, M. Železný // Proc. of the workshop on Audio-Visual Speech Processing. – 2007. – P. 172–175.

Поступила 23.07.08

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6*
e-mail: lobanov@newman.bas-net.by, l.tsirulnik@newman.bas-net.by

[2]*University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Plzeň, Česká republika*
e-mail: zelezny@kky.zcu.cz, zdkrnoul@ kky.zcu.cz

[3]*Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, 14 Линия, 39*
e-mail: ronzhin@iias.spb.su, karpov@iias.spb.su

B. Lobanov, L. Tsirulnik, M. Železný, Z. Krňoul, A. Ronzhin, A. Karpov

AUDIO-VISUAL RUSSIAN SPEECH SYNTHESIS SYSTEM

The paper dealing with two models of Russian audio-visual speech synthesis: the simulation model and the compilative one. The advantages and disadvantages of each model as well as peculiarities of creation of the models for Russian language are considered. The audio-visual Russian speech synthesis system built on these models is described.