Б.М. Лобанов, Л.И. Цирульник

СИСТЕМА ПЕРСОНАЛИЗИРОВАННОГО АУДИОВИЗУАЛЬНОГО СИНТЕЗА РЕЧИ И ЕЁ ВОЗМОЖНЫЕ ПРИЛОЖЕНИЯ

Описывается модель системы аудиовизуального синтеза речи, позволяющая осуществлять персонализированный синтез речи русской по тексту, сопровождающийся речевой мимикой персонализированной «говорящей головы». Рассмотрены некоторые возможные практические приложения разработанной системы.

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 может также стать полезным дополнением при обучении и реабилитации артикуляции звуков русской речи. Число исследований в области аудиовизуального распознавания и синтеза речи постоянно увеличивается. Разработкой бимодальных систем синтеза речи занимаются, в частности, научные коллективы Кэмбриджского университета, Великобритания; Политехнического университета, Монз, Бельгия; Университета Крита, Греция; Университета в Загребе, Хорватия; Университета Западной Богемии, Чехия.

Для решения задачи создания систем аудиовизуального синтеза речи (часто называемых «Говорящая голова») существует два подхода: имитационный, при котором создаётся 2D или 3D модель лица и настраиваются управляющие параметры для передачи мимики, выражения лица и движения губ при говорении [2], и компиляционный, при котором «говорящая голова» формируется путём выбора соответствующих видеофрагментов или изображений из визуальной БД конкретного диктора [3].

Преимуществом первого подхода является меньший физический объём 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», связанные с неизбежной схематичностью отображения речедвижений (рис. 1).

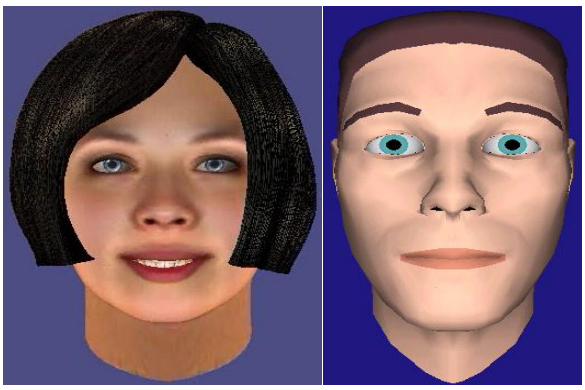


Рис. 1. Примеры реализации компьютерной «говорящей головы»

При компиляционном подходе требуемый объём 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 [4].

1. Компиляционная модель синтеза видеоизображений речи

В основу компиляционного метода положен метод плавной сшивки отдельных кадров изображения, предложенный ранее для плавной сшивки звуковых волн в микроволновом синтезаторе речи [5]. 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 определённая визема.

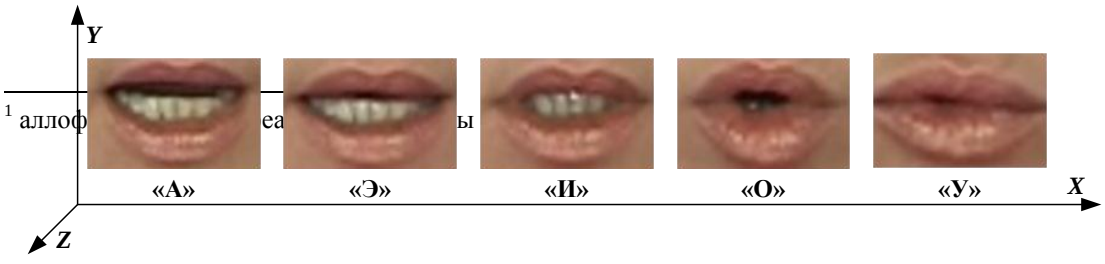
2. Для каждой виземы устанавливается её относительная длительность N_v, определяемая числом кадров показа изображения виземы:

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

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

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

Определение необходимого и достаточного набора визем осуществляется на основе известной классификации фонем и аллофонов русской речи по артикуляторным признакам способа и места образования с учётом эффектов коартикуляции и редукции. Как показали исследования [6], для русской речи практически полностью скрытой остаётся динамика движения тела, кончика и боковинок языка, нёбной занавески, голосовых связок. Обозрению доступны лишь движения губ и нижней челюсти. Наиболее яркие различия в



[1] аллоф ea ы

виземах связаны с изображениями губ говорящего (рис. 2).

Рис. 2. Виземы гласных фонем

Исходя из описанных выше явлений, выбран необходимый и достаточный набор визем русской речи, представленный в таблице 1.

Таблица 1. Соответствие «фонема-визема» для аудиовизуального синтеза речи

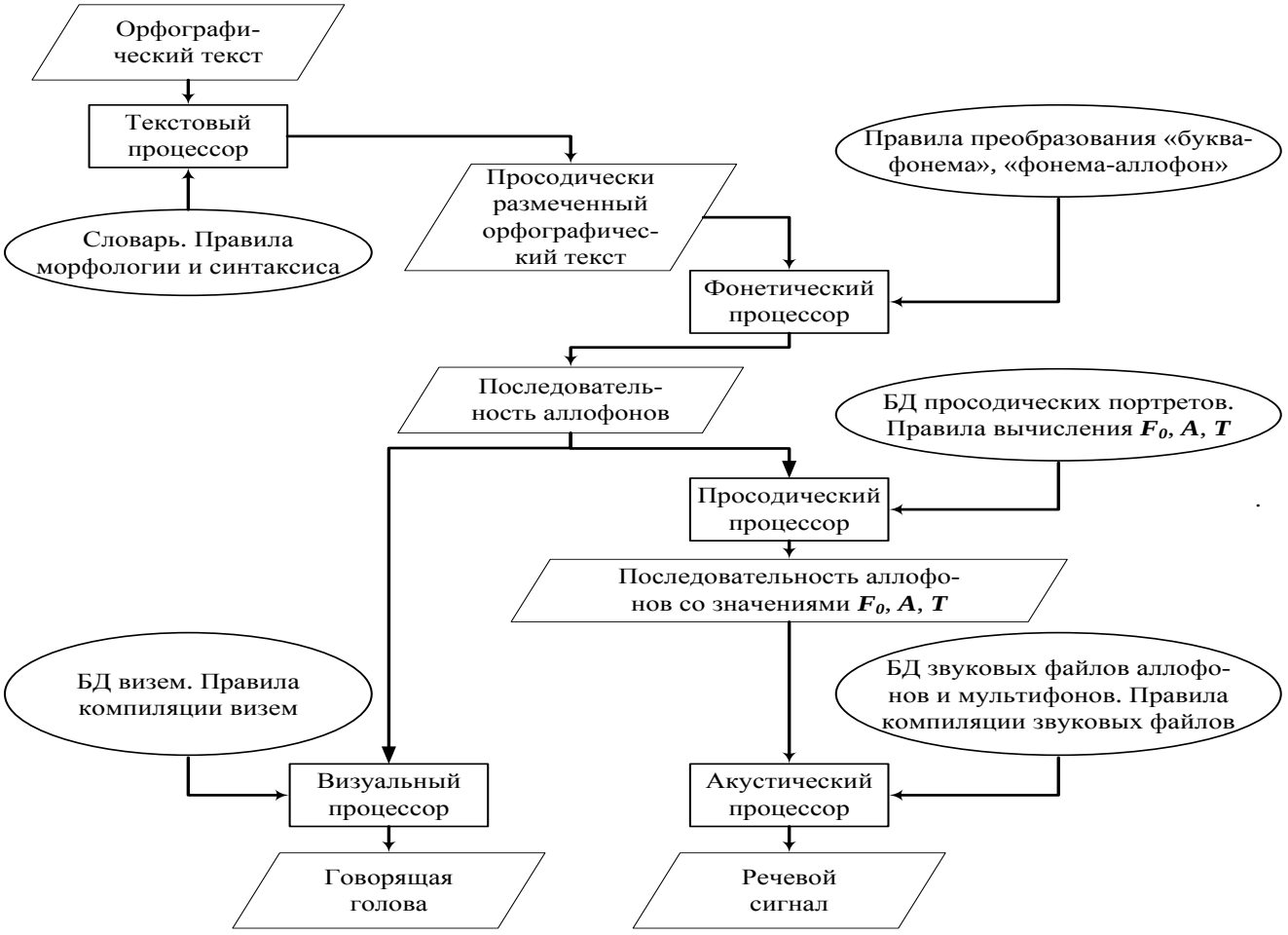
Визема	**Аллофоны фонем**	**Визема**	**Аллофоны фонем**
V_0	пауза	V_9	Ф, В
V_1	$А_0$, $А_1$	V_{10}	Ф’, В’
V_2	$Е_0$, $Е_1$	V_{11}	Ц, С, З, Д, Т, Н
V_3	$И_0$, $И_1$, $И_2$, $И_3$	V_{12}	С’, З’, Д', Т’, Н‘
V_4	$О_0$, $О_1$	V_{13}	Г, К, Х
V_5	$У_0$, $У_1$, $У_2$, $У_3$	V_{14}	Г’, К’, Х’, Й’
V_6	$Ы_0$, $Ы_1$, $Ы_2$, $Ы_3$, $А_2$, $А_3$, $Е_2$, $Е_3$	V_{15}	Ч', Ш’, Ш, Ж
V_7	Б, П, М	V_{16}	Л, Р
V_8	Б’, П’, М’	V_{17}	Л’, Р’

Процесс создания изображений визем может быть выполнен двумя способами: с использованием моментальных фотографий лица диктора, который имитирует произношение того или иного звука, соответствующего каждой виземе, и с использованием видеосъёмки диктора, произносящего фонетически сбалансированный текстовый корпус. Опыт создания визем показал, что второй способ является более предпочтительным с точки зрения соответствия полученных визем положению артикуляторных органов в процессе говорения, хотя и является более трудоёмким.

2. Система аудиовизуального синтеза речи по тексту

Общая структура системы аудиовизуального синтеза речи по тексту представлена на рисунке 3.

Рис. 3. Общая структура системы аудиовизуального синтеза речи по тексту



Входной орфографический текст последовательно подвергается преобразованиям, осуществляемым несколькими процессорами: текстовым, фонетическим, просодическим, акустическим и визуальным. Особенностью реализации системы аудиовизуального синтеза речи по тексту является совместная работа визуального и акустического процессоров.

3. Перспективы практического использования системы

Созданная система может быть встроена и использоваться как компонент в комплексных системах дистанционного обучения с персональным виртуальным учителем; в информационных киосках; в авиационных тренажёрах; в мобильных телефонах для озвучивания SMS-сообщений; в говорящих рекламных роликах, а также в составе тренажёра для постинсультной реабилитации устной речи.

В качестве примера на рисунке 4 представлены примеры 7-ми визем для различных звуков речи, полученных на основе киносъёмки лица профессионального логопеда, специализирующегося на проблемах восстановления речевой функции при афазии (постинсультной реабилитации речи). Аудиовизуальный синтезатор речи, в основу которого положены персонализированные наборы визем и аллофонов профессионального логопеда, планируется в дальнейшем использовать в составе компьютерного

тренажёра для постинсультной реабилитации устной речи.

04. – 478 p.
2. Tekalp A. M., Ostermann J. Face and 2-D mesh animation in MPEG-4 // Signal Processing: Image Communication, Special Issue on MPEG-4. – 2000. – V. 15. – P. 387-421.
3. Cosatto E., Graf H.P. Photo-realistic talking-heads from image samples // IEEE Transactions on Multimedia. – Sept. 2000. – V. 2. – P. 152-163.
4. Лобанов Б.М. Компьютерное клонирование персонального голоса и речи // Новости искусственного интеллекта – 2002. – № 5(55). – С. 35-39.
5. Лобанов Б.М. Микроволновой синтез речи по тексту // Анализ и синтез речи: сб. науч. тр. – Минск: Ин-т техн. кибернетики, 1991. – С. 21-38.
6. Karpov A., Tsirulnik L., Železný M., Krňoul Z., Ronzhin A., Lobanov B. Study of Audio-Visual Asynchrony of Russian Speech for Improvement of Talking Head Naturalness // Speech and Computer: proc. of International conference SPECOM'2009. – S.-Petersburg, June 21-25, 2009. In print.