Квазиречевой видеонавигатор для слепых

Б.М. Лобанов,
доктор технических наук
О.Г. Сизонов,
соискатель

Описывается структура электронного «видящего и говорящего» устройства-поводыря для слепых, основанного на принципах преобразования видеоизображения в речеподобный сигнал. Приводятся предварительные результаты экспериментального исследования эффективности предложенного метода.

Abctract

A structure of the electronic «seeing and speaking» guide-device for blind people, based on the principles of conversion of a videoimage into a speech-similar signal is described. Preliminary results of an experimental investigation of effectiveness of the proposed method are given.

Введение

Как известно, незрячий человек испытывает огромные трудности при самостоятельном передвижении в пространстве квартиры, улицы, приусадебного участка, города. В настоящее время он использует для навигации специальную трость или сопровождающего (человека или собаку). В данной работе рассматривается возможность создания в помощь слепому принципиально нового электронного «видящего и говорящего» устройства-поводыря на принципах преобразования видеоизображения в речеподобный сигнал.

До недавнего времени задача распознавания зрительных образов не рассматривалась в направлении помощи слепым. Начало было положено в 2001 году, когда в Стенфордском университете США был начат проект под названием «Blind Navigator» («Навигатор для слепых»), который ставит целью создать для слепого человека телеробот-шапку [1]. Голова, шея, туловище человека с такой шапкой направляют взор двух телекамер, закреплённых на голове, в нужную точку пространства впереди идущего. Миниатюрный компьютер, вмонтированный в «шапку», может распознавать около ста часто встречающихся в комнате предметов. После этого синтезатор речи сообщает идущему слепому информацию о том, что за объекты

встречаются на его пути. Точно так же, как курсор мышки компьютера управляется рукой и скользит по экрану, аналогично курсор взора двух телекамер будет управляться головой и скользить по реальному трёхмерному физическому пространству.

Существенные результаты по этому проекту, пригодные для внедрения, до сих пор отсутствуют. Это связано, прежде всего, с огромными трудностями в решении проблемы машинного распознавания трёхмерных зрительных образов и видеосцен.

1. Основная идея и конечная цель

Очевидно, что человеческий мозг справляется с задачей распознавания образов намного лучше любой существующей сегодня искусственной системы. Поэтому, если реализовать механизм адекватного преобразования визуальной информации в акустическую, а когнитивную работу оставить человеку, можно добиться ощутимых результатов и избежать решения проблемы машинного распознавания зрительных образов. Для этого предлагается выделить информативные параметры изображения и преобразовать их в адекватный звуковой сигнал. Для того чтобы способствовать лучшему восприятию и запоминанию звука человеком, принято решение приблизить его к звучанию человеческой речи, т.е. на основе поступающей видеоинформации синтезировать речеподобный сигнал. Здесь можно провести отдалённую аналогию с любым человеческим языком, в том смысле, что реальные объекты описываются набором звуков, который человек может запомнить и распознать.

Основная идея заключается в следующем. Система содержит в своём составе электронную фотокамеру, которая совершает одномоментные снимки пространства впереди незрячего человека. Снимок цветного формата в форме электронного файла поступает для обработки на вход процессора. Из пространственного видеокадра специальная программа формирует последовательный временной ряд сегментов, которые последовательно озвучиваются, образуя псевдослова. Разные сцены впереди незрячего будут давать различные видеокадры, а разные кадры будут порождать для слепого разные псевдослова. В процессе тренировки слепой человек, согласно нашей гипотезе, сможет на слух распознать набор из нескольких сотен типовых псевдослов. После нескольких сеансов обучения с поводырём, двигаясь в привычном пространстве, он сможет узнавать впереди себя знакомые предметы и сцены, ориентируясь на слуховое восприятие псевдослов.

Конечной целью является создание технической системы «Квазиречевой видеонавигатор» для инвалидов по зрению на базе мобильного телефона, оснащённого телекамерой и специальным программным обеспечением для навигации в пространстве. Реализованная на базе мобильного телефона система может оказаться, по нашему мнению, недорогой и простой для тиражирования.

К настоящему времени в лаборатории распознавания и синтеза речи ОИПИ НАНБ накоплен значительный опыт синтеза речевых, в том числе и речеподобных, сигналов [2]. В перспективе, при реализации достаточно

эффективных алгоритмов распознавания и анализа изображений, вместо речеподобного сигнала планируется использовать встроенную в мобильный телефон систему генерации навигационного текста и речи.

2. Алгоритмы реализации

Исходное изображение сцены (***рис. 1***) делится на заданное число «временных срезов». Для преобразования каждого среза в речеподобный сигнал определяется значение усреднённого цвета и контраста. Каждый срез делится на 32 части и определяется значением яркости на каждом участке. Речеподобный сигнал формируется методами спектрально-полосного синтеза путём последовательного во времени сканирования картинки. При этом оси X картинки соответствуют временные отсчёты сигнала, а оси Y — его частотный спектр.

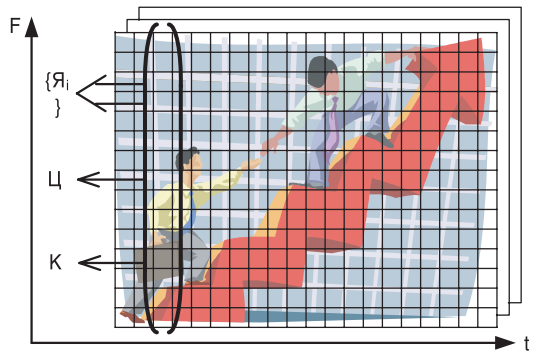

Рис. 1. Исходное изображение сцены

Выходной речеподобный сигнал формируется в соответствии со схемой, представленной на ***рис. 2***. Яркость среза определяет входные значения для используемых цифровых фильтров. Средний цвет — входные значения для генераторов тона, шума. Контраст — соотношение шума и тона.

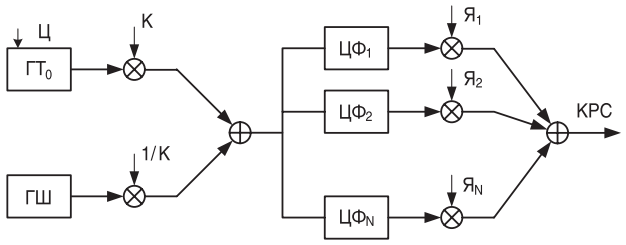

Рис. 2. Преобразователь изображения в речеподобный сигнал

Битовая матрица рисунка приводится к цветовой схеме HSV [3]. Цвет в цветовой схеме HSV изменяется в диапазоне 0÷360°. Яркость изменяется в диапазоне 0÷1. При минимальной яркости сигнал имеет нулевую амплитуду, при максимальной яркости амплитуда приравнивается к единице. Под контрастностью здесь принимается разница между максимальным и минимальным значениями яркости участка. Чем выше средняя контрастность среза — тем сильнее сигнал тона, чем контрастность ниже — тем больше зашумлённость.

На практике цветные снимки дают при синтезе достаточно сложно воспринимаемый звук, поэтому введена возможность приближения картинки к чёрно-белой палитре. Это может быть исполнено двумя способами. Первый из них предлагает вручную задавать значение порогов максимальной и минимальной яркости. При втором способе автоматически вычисляется дисперсия яркости, которая и определяет пороги белого и чёрного.

ГШ — генератор шума.
$ГТ_0$ — генератор основного тона.
ЦФ — цифровой фильтр.
КРС — квазиречевой сигнал.
К — контраст среза.
$\{Я_i\}$ — вектор яркости среза.
Ц — средний цвет среза.

В синтезаторе квазиречевого сигнала используется 32 полосовых фильтра Баттерворда второго порядка с полосами пропускания, определяемыми по шкале Мела в диапазоне 200–4850 Гц. Источником основного тона служит генератор пилообразного сигнала, источником шумового сигнала — генератор белого шума.

Частота основного тона звука задаётся значениями среднего цвета среза:

$$F_0 = \pi/[F_{min} + (F_{max} - F_{min})]^{*}\text{Ц}.$$

Согласно схеме (рис. 2) сигнал основного тона умножается на значение контрастности, а сигнал шума — на обратную величину. Тем самым задаётся соотношение тон/шум в синтезированном сигнале.

3. Программная модель навигатора

По описанным выше алгоритмам разработана программная модель навигатора. Программа позволяет загружать список файлов изображений. При выборе элемента этого списка выводится соответствующее изображение. При этом происходит обработка битовой матрицы, разбиение её по вертикали и горизонтали, синтез квазиречевого сигнала, который можно прослушать и сохранить в файл. Главный интерфейс программы представлен на ***рис. 3***, а интерфейс настроек — на ***рис. 4***.

Для обработки изображений предусмотрены два режима: «ручной» — с возможностью установки порогов чёрного и белого цветов; «автоматический» —

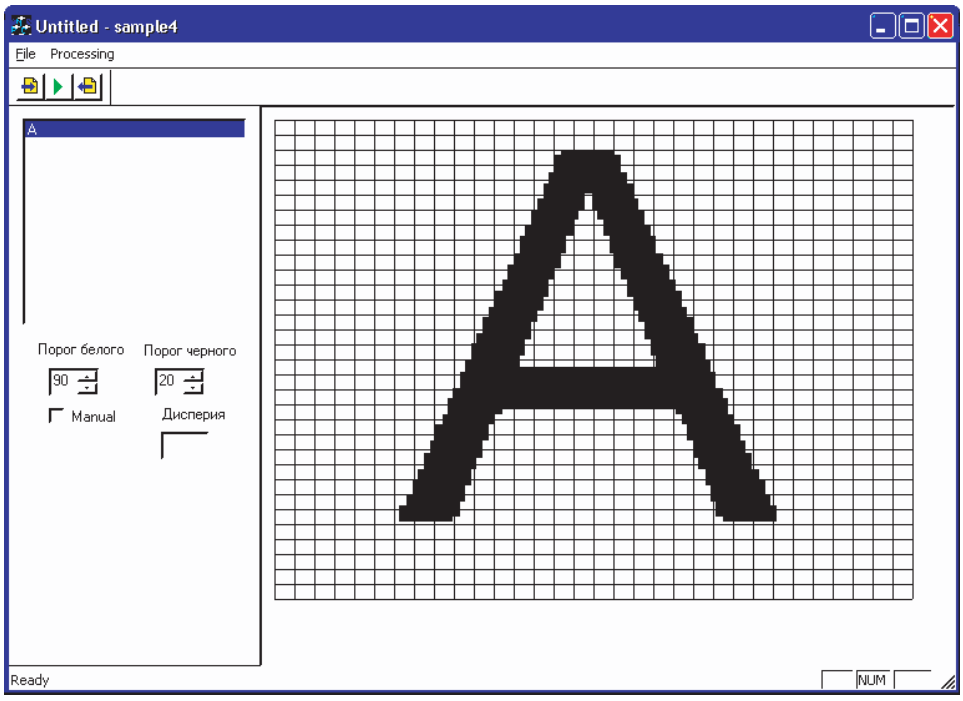

Рис. 3. Главный интерфейс программы

по автоматически вычисленной дисперсии. Для изменения характеристик синтезируемого звукового сигнала используются следующие переменные:

- количество делений изображения по горизонтали — 8, 16, 32, 64;
- Fmax — максимальная частота основного тона в Гц;
- Fmin — минимальная частота основного тона в Гц;
- относительная длительность импульсов сигнала основного тона.

Для проверки качества преобразования изображения в речеподобный сигнал на вход системы подавались различные картинки, а для выходного сигнала рассчитывались соответствующие им динамические спектрограммы (сонограммы). Примеры сонограмм, полученных для изображений цифр, приведены на ***рис. 5***.

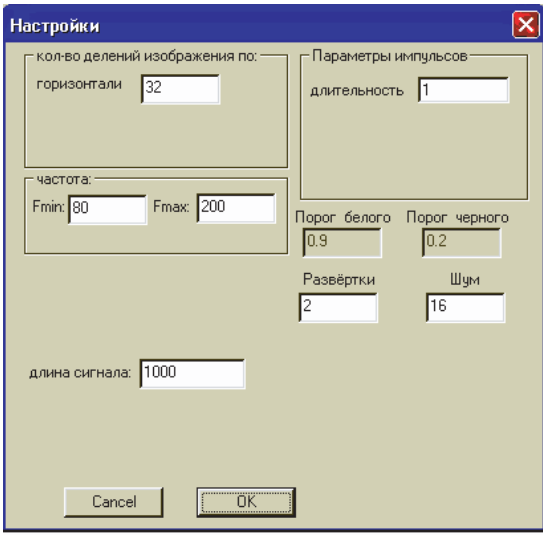

Рис. 4. Интерфейс настроек

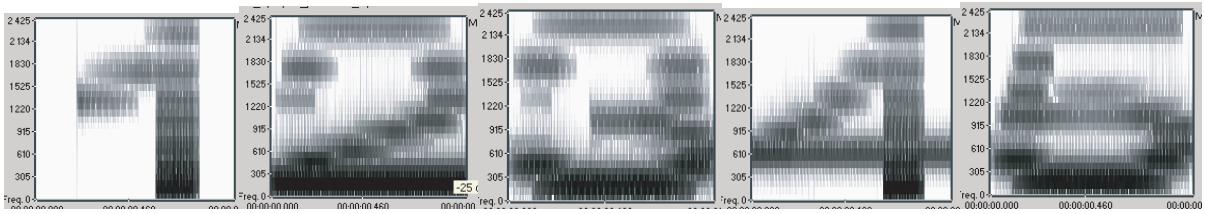

Рис. 5. Примеры сонограмм для изображений цифр

Для увеличения информативности звукового сигнала, кроме синтеза по основному направлению (слева направо), применяется изменение направления сверху вниз и справа налево. При этом звуковой сигнал, синтезированный на дополнительных развёртках, последовательно добавляется к основному. Такой подход предоставляет дополнительные пространственные составляющие и обеспечивает большую однозначность звуковым образам. В системе предусмотрена возможность задания длительности развёртки кадра (в мсек.) и числа развёрток (от 1 до 3). На ***рис. 6*** представлены сонограммы сигналов с двойной развёрткой.

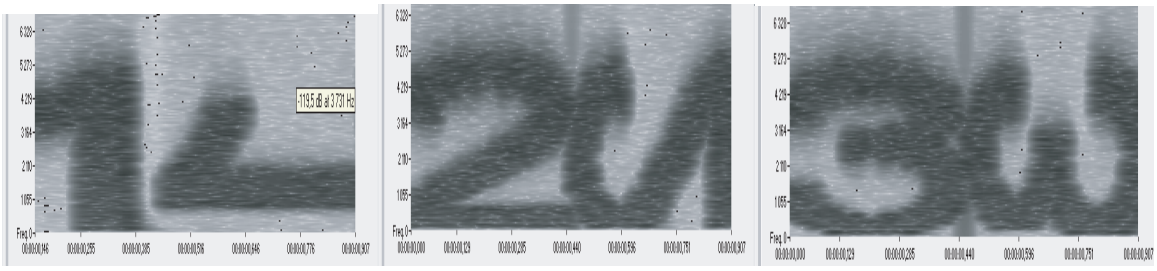

Рис. 6. Примеры сонограмм сигналов с двойной развёрткой

4. Экспериментальное исследование метода

Цель исследования — изучить и оценить трудоёмкость обучения звуковым образам изображений, эффективность их слухового распознавания незрячим человеком, определить направления развития описанной модели синтеза в задачах навигации.

Эксперимент 1. Оценка обучаемости и эффективности распознавания на изображениях цифр.

Методика: Испытуемому, тотально слепому человеку К.Л., предоставляется возможность самостоятельно изучить звуковые образы цифр от 0 до 9. Для этого автоматически озвучивается название каждой выбранной пользователем цифры и воспроизводится её синтезированный звуковой образ. Обучаемый сам определяет, когда он закончил обучение, т.е., по его мнению, готов узнавать предъявляемые цифры. Далее компьютерной программой ему предъявляются в случайном порядке звуковые образы, а испытуемый должен назвать, каким цифрам они соответствуют. Общее количество предъявлений — 100.

Результаты эксперимента (матрицы спутывания) представлены в ***таблицах 1 и 2*** для одинарной и двойной развёрток при длительности стимулов 0,5 и 1 сек. соответственно.

Таблица 1

Матрица спутывания для одинарной развёртки

Предъявлено \ Распознано	0	1	2	3	4	5	6	7	8	9
0	8						2			
1		9			1					
2			10							
3				10						
4					10					
5						8			2	
6	3						6		1	
7								10		
8						5			5	
9										10

Общая эффективность узнавания при одной развёртке — 86%, при двойной развёртке — 95%.

Эксперимент 2. Изучение восприятия сходства и различия изображений на фотографиях прямых и пересекающихся пешеходных дорожек.

Таблица 2

Матрица спутывания для двойной развёртки

		Распознано									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Предъявлено	**0**	9						1			
	1		10								
	2			10							
	3				9					1	
	4					10					
	5						9	1			
	6							9			1
	7								10		
	8							1		9	
	9										10

Методика: Испытуемому предъявляются пары синтезированных звуковых образов, полученные на основе фотографий. Задача испытуемого — определить на слух, на каких фотографиях отображены сходные объекты, а на каких — различные.

На ***рис. 7*** представлены примеры фотографий, воспринятых как схожие изображения прямых пешеходных дорожек, а на ***рис. 8*** — пересекающихся.

Рис. 7. Фотографии прямых пешеходных дорожек

Рис. 8. Фотографии пересекающихся пешеходных дорожек

В результате эксперимента испытуемый осуществил на слух правильную классификацию предъявленных фотографий прямых и пересекающихся пешеходных дорожек с надёжностью 93%.

Заключение

Проведённые эксперименты показали достаточно высокий уровень обучаемости слепого человека и способностей к распознаванию звуковых образов предъявляемых изображений. Остаются проблемы обработки изображений для избавления от таких артефактов, как блики, тени и пр.

В целом, однако, предложенный подход можно характеризовать как многообещающий для создания достаточно недорогого и эффективного нового средства навигации слепых в пространстве.

Цель настоящей статьи была бы достигнута в полной мере, если бы она послужила стимулом к привлечению необходимого финансирования для реализации действующей системы на базе мобильного телефона.

Литература

1. Материалы Стэндфордской конференции, 2003 г. // [Электронный ресурс] http://mediax.stanford.edu/news/conference_nov03/dave_grossman.pdf.

2. *Воробьев В.И., Давыдов А.Г., Лобанов Б.М.* Синтез речеподобных сигналов с использованием аллофонов // Сб. трудов XIII сессии Российского акустического общества, М.: «Геос», 2003, с. 110–114.

Лобанов Борис Мефодьевич —

Почётный радист СССР (1981), обладатель серебряной и бронзовой медалей ВДНХ СССР (1983), главного приза международного конкурса фирмы HEWLETT-PACKARD за работу «Распознавание голоса» (1992). С 1987 — член Международного акустического общества, с 1994 — координатор Белорусского отделения Европейской сети по компьютерной лингвистике и речи, с 1995 — член Европейской ассоциации речевых исследований, с 2001 — эксперт Европейской сети языковых технологий. Член докторских советов по защите диссертаций (ОИПИ НАН Беларуси, БГУИР, БГУ, МГЛУ). С 1998 — профессор БГУИР, с 2003 — профессор Университета в Белостоке.

О.Г. Сизонов —

Окончил факультет информационных технологий Белорусского Национального технического университета. Соискатель учёной степени кандидата технических наук. С 2007 года — младший научный сотрудник лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси. Область научных интересов — методы лингвистической обработки русского текста в синтезе речи по тексту, синтез и обработка речевых и квазиречевых сигналов, применение синтеза речи в системах реабилитации инвалидов по зрению и слуху.