УДК 621.391

ОБРАБОТКА РЕЧЕВЫХ КОРПУСОВ ДЛЯ РАСПОЗНАВАНИЯ И СИНТЕЗА РУССКОЙ РЕЧИ В СИСТЕМАХ УПРАВЛЕНИЯ

Б.М. Лобанов[1], Л.И. Цирульник[1], И.В. Ли[2]
[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь;
[2]Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия

Описывается автоматизированная система разметки речевого корпуса, в основу которой положена идея «анализ через синтез». Рассматриваются процедуры фонетической обработки русских текстов и синтеза речевого сигнала. Обсуждаются особенности анализа и сопоставления синтезированного и естественного речевых сигналов в процессе разметки. Показана специфика использования размеченных речевых корпусов в системе многоголосого синтеза речи и системе дикторонезависимого распознавания речи.

Введение

На современном этапе развития средств вычислительной техники автоматическое распознавание и синтез речи по тексту являются, пожалуй, одними из наиболее востребованных функций систем управления, в частности, в чрезвычайных ситуациях. Действительно, подсистемы распознавания и синтеза речи, встроенные в систему управления, позволят сэкономить время, требуемое на ввод информации и оповещение абонентов о принятом решении, и, как следствие, предотвратить или, по крайней мере, уменьшить ущерб, приносимый чрезвычайной ситуацией.

Для создания как системы синтеза речи по тексту, так и системы распознавания необходимы речевые корпусы, размеченные на некоторые базовые единицы. Компиляционный синтезатор речи [1] использует звуковые волны аллофонов и мультифонов в процессе генерации речевого сигнала. Для системы распознавания, основанной на использовании скрытых марковских моделей (СММ) [2], необходимы звуковые эталоны фонем и морфем, на которых модель обучается.

Точность разметки речевой базы является весьма ответственной процедурой, так как от неё в конечном итоге зависит и качество синтезируемой речи, и точность распознавания.

Во многих системах процесс сегментации и маркировки речевого корпуса осуществляется вручную экспертом-фонетистом с использованием полуавтоматических средств просмотра осциллограмм и спектрограмм сигнала [3, 4]. «Ручной» метод сегментации и маркировки при высокой квалификации эксперта обеспечивает достаточно точную разметку речевого корпуса, но требует много времени и усилий.

В данной работе описывается система автоматизированной обработки речевых корпусов, в основу которой положена идея «анализ через синтез» [5]. Система создана в рамках продолжающихся исследований по клонированию голоса и дикции личности [6], созданию персонализированного синтезатора речи по тексту [7, 8] и дикторонезависимой системы распознавания русской речи [9], реализуемых в рамках проекта ИНТАС № 04 77 7404.

В разделе 1 описывается общая структура системы, раздел 2 посвящён фонетической обработке текста и синтезу речевого сигнала, процесс сопоставления синтезированного и естественного сигналов описан в разделе 3.

1. Структура системы сегментации и маркировки речевого корпуса

Общая структурная схема автоматизированной системы сегментации и маркировки речевого корпуса представлена на рис.1. На вход системы подается цифровая запись речевого сигнала и орфографический текст стенограммы записи. Орфографический текст поступает на вход модуля синтеза речи, осуществляющего фонетическую и акустическую обработку. На этапе фонетической обработки выполняется расстановка словесных ударений, маркировка границ морфем, преобразование графема-фонема и преобразование фонема-аллофон. Результат фонетических преобразований – последовательность аллофонов с маркерами границ морфем – поступает в блок обработки сигнала, осуществляющий выбор звуковых волн аллофонов из фонетико-акустической базы, их компиляцию и маркировку границ речевых единиц (РЕ) в генерируемом речевом сигнале.

Как синтезированный, так и естественный речевые сигналы поступают на вход модуля разметки и маркировки естественного речевого сигнала.

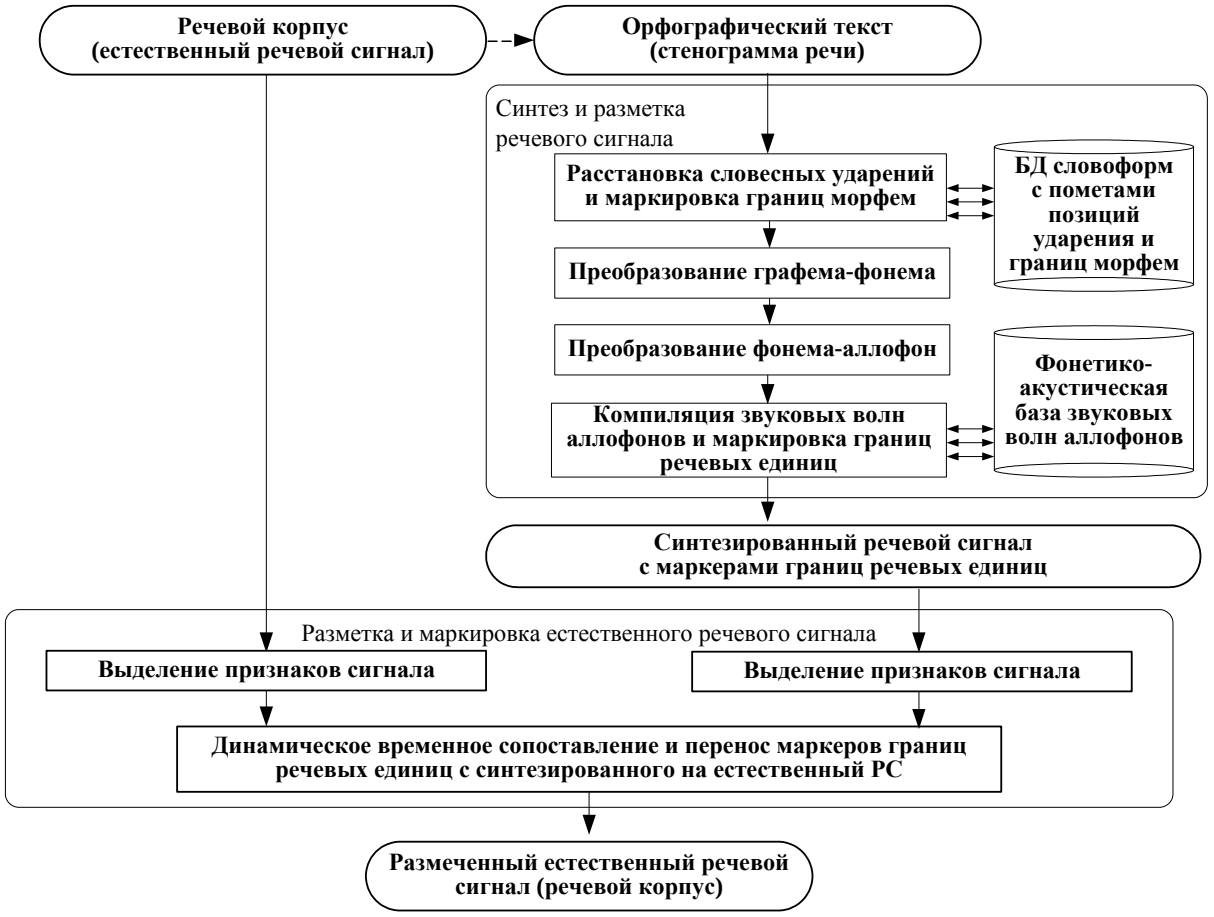

Рис. 1. Общая схема системы фонетической разметки речевых корпусов

Модуль разметки естественного сигнала включает в себя выделение признаков естественного и синтезированного сигналов и нелинейное временно́е сопоставление полученных признаков методом динамического программирования (ДП). По найденному соответствию производится перенос меток границ РЕ с синтезированного на естественный речевой сигнал.

2. Модуль фонетической обработки текста и синтеза речевого сигнала

Фонетическая обработка орфографического текста включает следующие этапы (см. рис. 1):

- расстановка словесных ударений и маркировка границ морфем;
- преобразование графема-фонема;
- преобразование фонема-аллофон.

Для расстановки словесных ударений используется база данных (БД) словоформ, созданная на основе грамматического словаря А.А. Зализняка. БД содержит около двух миллионов единиц, включающих как знаменательные, так и служебные слова. Для каждого знаменательного слова указана позиция ударения, для каждого служебного (произносимого, как правило, без ударения и присоединяемого в речи к ближайшему слову) – тип его присоединения: к последующему или к предыдущему слову. Кроме того, БД содержит пометы границ морфем для каждой словоформы.

Результатом работы первого блока является орфографический текст, в котором служебные слова присоединены к соответствующим знаменательным, позиция ударения для каждого знаменательного слова помечена знаком «+» после ударной гласной, границы морфем помечены символом «>».

Блок преобразования графема-фонема использует контекстно-зависимые правила отображения букв в последовательность фонем, учитывающие как внутрисловные, так и межсловные фонетические явления, исследованные в [10].

Блок преобразования фонема-аллофон генерирует позиционные и комбинаторные аллофоны фонем. Позиционный фактор учитывает позицию данной фонемы относительно словесного ударения. Комбинаторный фактор учитывает ближайшее фонемное окружение.

Для гласных фонем генерируются следующие позиционные аллофоны: полноударный, частично ударный, первый предударный, не первый предударный или заударный. С учётом левого контекста генерируются следующие комбинаторные аллофоны гласных: после синтагматической паузы, после твёрдых губных, передне- и среднеязычных, после твёрдых заднеязычных и гласных и после мягких. С учётом правого контекста генерируются следующие комбинаторные аллофоны гласных: перед синтагматической паузой, перед переднеязычными и заднеязычными твёрдыми согласными и гласными, перед губными согласными, перед мягкими согласными и гласной. Итого, при этих условиях обеспечивается генерация $N_v = 4*5*4*6$(гласных) = 480 аллофонов гласных. С учётом того, что многие комбинаторные и позиционные ситуации вообще не встречаются в речи, практически число генерируемых аллофонов гласных равно 300.

Аллофоны согласных генерируются только с учётом правого контекста, причём группы правых контекстов различны для разных по способу образования согласных фонем. В частности, для большинства глухих выделяется только два правых контекста: пауза и любая фонема, а для большинства сонорных и щелевых определены следующие группы правых контекстов: пауза, глухие согласные, звонкие согласные, гласные. Общее количество генерируемых аллофонов гласных равно 200.

Результирующая последовательность аллофонов с маркерами границ морфем является входным данным блока обработки сигнала, использующего для синтеза речи фонетико-акустическую базу звуковых волн аллофонов. Для каждого аллофона входной последовательности из базы выбирается соответствующая звуковая реализация, выбранные звуковые волны аллофонов компилируются в непрерывный речевой сигнал, в котором указываются границы и имена аллофонов и морфем.

Синтезированный размеченный речевой сигнал поступает в модуль разметки и маркировки естественного речевого сигнала.

3. Разметка и маркировка естественного речевого сигнала

Модуль разметки и маркировки состоит из следующих блоков (см. рис. 1):

- блок выделения признаков сигнала;
- блок ДП-сопоставления признаков и переноса маркеров границ РЕ с синтезированного на естественный сигнал.

Первый из блоков использует в качестве признаков сигнала спектр и усреднённые конечные разности спектра по времени. Для получения указанных признаков сигнал пропускается через набор 18 полосовых фильтров Баттерворта. Полосы пропускания фильтров представлены в таблице 1.

Таблица 1

Полосы пропускания фильтров: F_{min} – минимальная частота фильтра, F_{max} – максимальная частота фильтра

№ фильтра	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
F_{min}, Гц	100	200	300	400	500	600	708	814	1000
F_{max}, Гц	400	500	600	708	814	1000	1190	1415	1682
№ фильтра	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
F_{min}, Гц	1190	1415	1682	2000	2379	2829	3364	4000	4757
F_{max}, Гц	2000	2379	2829	3364	4000	4757	5657	6727	8000

Затем для каждого из 18 каналов вычисляется среднеквадратичное значение сигнала на каждом участке длительностью 1 мс и осуществляется вычисление конечных разностей спектра по формуле

$$\Delta S(i) = \frac{1}{K}\sum_{k=0}^{K-1}\left[S(i+k) - S(i-1-k)\right],$$

где $\Delta S(i)$ — усредненная конечная разность спектра по времени; K — интервал усреднения конечных разностей спектра; $S(i)$ — i-й спектральный отсчёт по времени.

Вычисленные признаки синтезированного и естественного сигнала подаются на блок ДП-сопоставления признаков и переноса меток границ аллофонов. Начальным этапом сопоставления является нахождение матрицы локальных расстояний между векторами признаков синтезированного и естественного сигналов. Затем вычисляется матрица интегральных расстояний между признаками, которая служит для нахождения временно□го соответствия между синтезированным и естественным сигналами.

Алгоритмы работы блока подробно описаны в [11]. Особенностью реализации является ограничение интервала поиска траектории соответствия между сравниваемыми сигналами, при этом ширина начала и ширина конца допустимого интервала – переменные величины, задаваемые в виде доли от длины синтезированного сигнала. Это позволяет увеличивать размер допустимого интервала для длинных фраз, где могут происходить значительные искажения временны□х шкал, и уменьшать его для коротких фраз.

По найденному соответствию маркеры границ аллофонов и морфем, а также их имена переносятся с синтезированного на естественный речевой сигнал.

Заключение

Разработанная система применяется для создания размеченных речевых корпусов, используемых как при дикторонезависимом распознавании речи, так и при многоголосом синтезе речи по тексту.

Для создания минимальной фонетико-акустической базы голоса, применяемой при персонализированном синтезе речи по тексту, используется речевой корпус, включающий набор русских слов, содержащих все аллофоны по описанной в п.2 классификации. Для пополнения фонетико-акустической базы голоса мультифонами, а также для создания звуковых эталонов морфем, применяемых при распознавании речи, используется фонетически репрезентативный речевой корпус, включающий набор из 500 фраз [12], в которых статистические характеристики распределения фонем и других фонетических единиц близки к характеристикам, получаемым на больших выборках.

Список литературы

1. Lobanov B., Tsirulnik L. Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) // Proc. of the XI International Conference SPECOM'2006, S.-Petersburg, Russia, June 25-29, 2006. – S.-Petersburg: Anatolya, 2006. – P. 274-283.
2. Ronzhin A., Karpov A., Li I. Russian Speech Recognition for Telecommunications // Proc. of the X International Conference SPECOM'2005, Patras, Greece, October 17-19, 2005. – P. 491-494.
3. Богданов Д. С., Кривнова О. Ф., Подрабинович А. Я., Фарсобина В. В. База речевых фрагментов русского языка «ISABASE» // Интеллектуальные технологии ввода и вывода информации. М., 1998. С. 20-23.
4. Вольская Н., Коваль А. и др. Синтезатор русской речи по тексту нового поколения. // Мат. Междунар. конф. «Диалог'2005», 1-7 июня, Звенигород, Россия. – М.: Наука, 2005. – С. 84-85.
5. Лобанов Б.М., Давыдов А.Г., Киселев В.В., Цирульник Л.И. Система сегментации речевого сигнала методом анализа через синтез. // Известия Белорусской инженерной академии № 1 (17)/1' 2004, С. 112-115.
6. Лобанов Б. М. Компьютерное «клонирование» персонального голоса и речи // Новости искусственного интеллекта. – 2002. – № 5(55). – С. 35–39.
7. Lobanov B., Karnevskaya E. TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian) // Phonetics and its Applications. – Stuttgart: Franz Steiner Verlag, 2002. – P. 445–452.
8. Lobanov B., Tsirulnik L. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS // Proc. of the IX International Conference SPECOM'2004, St.-Petersburg, Russia, September 20-22, 2004. – P. 17 – 21
9. Карпов А.А., Ронжин А.Л., Ли И.В. SIRIUS - система дикторонезависимого распознавания слитной русской речи. Известия ТРТУ, № 10, 2005, С. 44-53.
10. Лобанов Б.М., Цирульник Л.И. Внутрисловные и межсловные правила обработки текста для полного и разговорного стилей речи // Функциональные стили звучащей речи: сб. науч. тр. – М.: Макс-Пресс, 2006. – С. 21-30.
11. Лобанов Б.М., Слуцкер Г.С., Тизик А.П. Автоматическое распознавание звукосочетаний в текущем речевом сигнале. //Труды НИИР. Вып. 4. – М., 1969. – С. 67-75.
12. ГОСТ 16600-72. Передача речи по трактам радиотелефонной связи. Москва, 1973. Приложение 9. Тесты фраз и команд.