Исследовательская фонетическая база данных для Интернет “Региональная вариативность русской звучащей речи”

П.А.Скрелин, Т.Ю.Шерстинова

Санкт-Петербургский государственный университет,

paul@phonet.lang.pu.ru, tanya@ts4306.spb.edu

В настоящее время во многих странах мира развернулась работа по организации филологического материала в виде баз данных (текстовых или акустических). Большинство разрабатываемых акустических систем можно условно разделить на две группы: 1) предназначенные для целей архивного хранения материала, и 2) направленные на обеспечение решения конкретных прикладных задач (в области речевых технологий и разнообразных систем обработки естественного языка).

Особенностью всех филологических баз данных, разрабатываемых на кафедре фонетики Санкт-Петербургского государственного университета, является их ориентация на поддержку научных и культурологических исследований. Нашей целью является разработка такой системы, которая позволяла бы специалистам самых разных областей проводить изучение филологического материала как на звуковом, так и текстовом уровне непосредственно в среде базы данных.

Для обеспечения этой задачи в первую очередь требуется особое представление в базе материала исследования - звукового сигнала (или текста). Так, и текстовые и звуковые файлы должны сопровождаться информацией о разноуровневой сегментации, которая позволит пользователю-исследователю обращаться к любому интересующему его сегменту. Каждый значимый сегмент базы данных мыслится как потенциальный объект исследования и сопровождается таблицей признаков описания, отражающих присущие ему характеристики, причем этот список не является закрытым.

Научный проект "Региональные варианты русской звучащей речи в Интернет", поддерживаемый фондом РГНФ (грант N 99.04-12015в), имеет своей целью разработку интерактивной экспертно-исследовательской базы данных, доступ к которой планируется осуществлять через Интернет. Основное назначение базы данных - исследование региональной вариативности звуковой формы русской речи. Для прослушивания звукозаписей через сеть пользователю необходим компьютер со звуковой картой типа SoundBlaster и соответствующий Plug-in броузера или системный плейер.

В режиме on-line пользователю базы данных предоставляются следующие возможности:

1) указав интересующий его вариант произношения перейти к орфографической и транскрипционной записи текста и прослушать текст целиком или любой его фрагмент (для этого необходимо задать начальный и конечный элемент сегмента);

2) просмотреть базу данных сегментных и супрасегментных фонетических признаков конкретного варианта регионального произнесения и получить звуковые иллюстрации этих особенностей;

3) указав интересующие его фонетические особенности, найти региональный вариант или варианты, в которых они представлены.

4) при выделении для анализа/прослушивания какого-либо фрагмента текста (от слова или последовательности слов до синтагмы или последовательности фраз), пользователь получает из базы данных всю имеющуюся информацию о региональных особенностях реализации сегментных и супрасегментных единиц, попавших в область выделения.

Профессиональная версия разрабатываемой системы будет распространяться на цифровых компакт-дисках и будет предоставлять возможность пользователю создавать собственные комментарии и вносить дополнения (изменения) во все информационные поля (таблицы) базы данных.

Логическими компонентами базы данных являются:

речевой материал (звуковые файлы); вспомогательная БД информации о сегментации речевого материала; БД фонетических признаков; HTML-интерфейс пользователя; обслуживающие CGI-программы.

Для того, чтобы пользователь мог прослушать любой фрагмент конкретной записи разрабатывается специальная программа, основанная на использовании вспомогательной БД информации о сегментации. Звуковые реализации текстов оцифровываются и с помощью специальной программы акустической обработки сигналов сегментируются на слова, синтагмы и фразы. Программа помещает информацию о сегментации в файле, имеющем стандартный табличный вид (что позволяет легко его конвертировать во вспомогательную БД информации о сегментации). БД должна состоять из такого количества базовых таблиц, сколько значимых сегментов (объектов исследования) предполагается исследовать. В настоящее время это таблица сегментов-слогов, сегментов-слов и сегментов-синтагм. Эти таблицы впоследствии могут быть расширены собственно фонетической информацией (на каждом сегментном уровне могут быть заданы поля описания, которые будут необходимы для работы конкретных пользователей в зависимости от специфики их задач).

Текст для озвучивания представлен в виде HTML–формы. Минимальным звуковым сегментом для анализа/прослушивания в настоящее время является слово (в перспективе – слог). Пользователь отмечает флажком начало и конец интересующего его сегмента и пересылает запрос на сервер. Далее CGI-программа обрабатывает входные данные формы

и передает запрос пользователя о начале и конце сегмента вспомогательной БД информации о сегментации, а программа обработки звукового сигнала вычленяет по заданным отсчетам соответствующий сегмент, после чего происходит передача его по сети пользователю.

Помимо самих звукозаписей мы располагаем их подробными описаниями, составленными экспертами-фонетистами и представленных в виде таблиц фонетических признаков, характерных для каждого региона. Для того, чтобы пользователь получил доступ к доступ к этим таблицам, они также формализуются и представляются в виде компьютерной базы данных фонетических признаков, основными таблицами которой являются: 1) особенности реализации гласных звуков, 2) особенности реализации согласных, 3) интонационные особенности. Каждая таблица содержит указание на звуковые примеры, озвучивание которых реализуется в виде гиперссылки на соответствующие звуковые сегменты (слова, словосочетания, синтагмы, фразы).

Для того, чтобы пользователь по заданным фонетическим особенностям мог найти региональный вариант или варианты, в которых эти особенности были отмечены, необходима поисковая система по признакам модификации как на сегментном, так и интонационном уровнях. Такая опция системы обеспечивается благодаря системе запросов и обмена данными с БД фонетических признаков.

Вспомогательная БД информации о сегментации может быть дополнена полями-комментариями, в которые заносится фонетическая информация (а потенциально и любая другая в зависимости от переориентации экспертной системы). Это позволит придать системе “объектно-ориентированный” характер и сделает работу с ней еще более гибкой. Тогда одновременно с озвучиванием произвольного фрагмента пользователь сможет получить всю имеющуюся в базе данных содержательную (в частности, фонетическую) информацию.

Разрабатываемая версия должна продемонстрировать возможности баз данных для проведения сопоставительных исследований, озвучивания любого фрагмента звукового материала - от слова до предложения или всей звукозаписи, классификации материала по выбранным признакам. И хотя главным объектом хранения и описания ее информации является звуковой сигнал, основные принципы подготовки и представления филологического материала в той же степени пригодны и для разработки баз данных, ориентированных на материал, представленный в текстовой форме. Более того, некоторые процедуры (как например, процесс автоматической сегментации материала) в полнотекстовой базе данных значительно упрощаются. Таблицы признаков описания сегментов могут быть расширены по мере необходимости введения в базу данных новых параметров описания, а также для обеспечения работы с базой пользователей самых разных филологических, культурологических и смежных специальностей. Результаты исследований вместе с примерами (текстовыми или акустическими) станут доступными через сеть Интернет исследователям всего мира.

Представление в виде описанной БД уникального звукового материала, собранного специалистами кафедры фонетики в 70-80-е годы на территории бывшего СССР, который

представляет собой чтение фонетически сбалансированного текста коренными жителями 51 города, является единственной возможностью ввести его широкий научный оборот. Более того, организация самого звукового материала и его фонетического описания в виде БД, аналогичной TIMIT, позволит использовать его и в прикладных целях, например, для тренировки систем автоматического распознавания речи.