

Разработка интерактивной системы Интернет-доступа к коллекциям звуковых архивов

П.А. Скрелин, А.О. Таланов, Т.Ю. Шерстинова, П.П. Щербаков

Санкт-Петербургский государственный университет

Адрес: Санкт-Петербург, 199034, Университетская наб. 11

Тел.: (812) 328-9565

Факс: (812) 312-2246

E-mail: paul@phonet.lang.pu.ru; tanya@ts4306.spb.edu; andre@AT4305.spb.edu; paul@icape.nw.ru

1. Общие характеристики системы

Основной задачей представляемого проекта является разработка такой системы, которая позволила бы пользователям сети Интернет проводить изучение коллекционного архивного материала как на звуковом, так и текстовом уровне непосредственно в режиме он-лайн.

Звуковой материал может представлять интерес для специалистов самых разных профилей: в первую очередь для фонетистов, акустиков, физиологов, работа которых непосредственно связана с обработкой и анализом звукового материала, а также для филологов общего профиля (например, текстологов), исследователей культуры, фольклористов, диалектологов, литературоведов, музыковедов, этнографов, социологов, психологов и специалистов других научных дисциплин, в том числе и негуманитарных. Следовательно, создавая систему доступа к звуковым коллекциям, мы должны учитывать разнообразие профессиональных интересов ее потенциальных пользователей.

Требования к разработке системы определяются главным образом характером архивного материала. Поскольку для одной группы пользователей звуковой сигнал является основным объектом исследования, а для многих других он является лишь иллюстрацией явлений, связанных со звуком лишь опосредовано, звуковой архив должен представлять собой не просто каталог звукозаписей с шифрами (адресами), а сложную логическую систему, состоящую из разных модулей. Так, помимо традиционных модулей – корпуса собственно звукозаписей и их каталога – система должна включать в себя модули многоаспектного описания материала, состоящие из таблиц признаков и даже баз данных, заполняемых специалистами разных научных направлений (например, фонетический, диалектологический, текстологический, музыковедческий модули и т.п.).

Далее, традиционный архивный каталог, пусть даже представленный в электронном виде, обладает с нашей точки зрения одним существенным

недостатком: даже если система и позволяет “озвучить” звукозапись непосредственно из каталога путем обращения к соответствующему звуковому файлу, пользователь в этом случае вынужден прослушать весь звуковой файл целиком, сколь не велика продолжительность его звучания (которая в архивных файлах может достигать нескольких часов).

Поэтому, прежде всего, мы выдвигаем требование обязательной текстологической расшифровки звукозаписи. Орфографическая запись текста, соответствующего звукозаписи, становится основным “посредническим” объектом, через который происходит обращение пользователя к любым фрагментам звукозаписей. Имея на экране орфографическую или транскрипционную запись озвученного текста, пользователь сможет прослушать (проанализировать) любой интересующий его фрагмент.

Техническая подготовка архивного звукового материала включает в себя оцифровку звукозаписей, их многоуровневую сегментацию (например, для фонетического модуля возможна сегментация на фразы, синтагмы, слова и даже отдельные звуки) и текстологическую расшифровку. В результате для каждой записи мы получаем звуковой файл с метками сегментации разных типов, информация о которых заносится во вспомогательную базу данных, а также текстовые файлы с орфографическим или транскрипционным представлением звукозаписи, размеченные на те же сегменты, что и звук.

Соотнесение расшифрованного текста и звучания становится одной из наших основных технических задач, которая решается путем построения и использования вспомогательных баз данных, содержащих информацию о сегментации. В результате пользователь может теперь прослушать звукозапись не только целиком, но и отдельными сегментами (фразы, синтагмы, слова), кроме того становится возможным и “озвучивание” произвольного фрагмента текста, ограниченного установленными метками сегментации.

Разумеется, чем более мелкими являются базовые сегменты членения (например, звуки), тем с большей точностью можно прослушивать интересующий пользователя фрагмент. Однако поскольку работа по сегментации довольно трудоемка, сегментацию на отдельные звуки имеет смысл производить лишь для экспертных фонетических коллекций, в то время как для большинства диалектных и фольклорных архивов достаточной является сегментация на слова или даже фразы.

Каждый значимый сегмент базы данных (фразы, синтагмы, отдельные слова) рассматривается как потенциальный объект исследования и сопровождается таблицей признаков описания, отражающих присущие ему характеристики. Таким образом, если пользователь выделяет для анализа/прослушивания один из значимых сегментов (конкретное слово или фразу), он должен получить из базы данных всю соответствующую ему информацию (комментарии, описания, транскрипцию и т.п.). В перспективе планируется предоставить пользователю возможность создавать собственные комментарии и описания непосредственно в режиме он-лайн, которые будут заноситься в единую базу данных.

На настоящий момент определяется техническое решение реализации отдельных логических компонентов системы (речевого корпуса; вспомогательной БД информации о сегментации речевого/музыкального материала; БД многоаспектных модулей описания звукозаписи и ее значимых сегментов; HTML-интерфейса пользователя и обслуживающих CGI-программ) и их интеграции в единую систему, производится разработка отдельных модулей.

Профессиональная версия 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, а благодаря Интернет-технологиям коллекционный материал должен стать доступным исследователям всего мира.

2. Специфика представления звуковых баз данных в Интернет

При подготовке системы для представления в Интернет возникает ряд новых проблем по сравнению с эксплуатацией ее в локальной сети или на отдельной машине, в частности – задача ее адекватного воссоздания на удаленных компьютерах, в том числе и на других платформах.

2.1. Проблема шрифтов и кодировки. Особенность разрабатываемой системы состоит в том, что одновременно на экране монитора пользователя сети могут быть открыты несколько фреймов: орфографическая, транскрипционная записи текстов, перевод текста на русский язык, комментарии.

Так, при создании Web-сайта для Фонограммархива Пушкинского Дома <http://www.speech.nw.ru/phonetics/homepage.html> нам пришлось описывать записи на коми и ненецком языках. Орфография текстов должна быть представлена на оригинальном алфавите для каждого из языков. Поскольку алфавит этих языков был разработан на основе кириллицы с добавлением символов из других алфавитов, то при определении кодировки для страниц на коми и ненецком мы столкнулись со значительными сложностями.

Одним из возможных способов решения проблемы является разработка нового шрифта, включающего в себя все необходимые для просмотра сайта символы. Недостаток такого подхода состоит в том, что пользователю придется сначала “скачать” этот шрифт с сервера, установить его на своей машине и перенастроить браузер. Сложность выполнения этой задачи определяется качеством описания требуемых технических процедур, которое должно быть обязательно представлено на справочной странице сайта, и сильно зависит от компьютерных навыков пользователя и опыта его работы с Интернет.

Другое решение состоит в использовании шрифтов Unicode (UTF-8), которые претендуют на охват всех существующих алфавитов мира. Наш же опыт показывает, что для одного из символов ненецкого языка так и не нашлось полного соответствия ни с одним из имеющихся на сегодняшний день символов Unicode. Как и в предыдущем случае, шрифт должен быть предварительно установлен на компьютере пользователя. Некоторые из шрифтов Unicode распространяются бесплатно в Интернет, однако как правило все они очень

громоздки. К примеру, установочный пакет шрифта UNICODE MS, разработанный фирмой Microsoft, насчитывает около 14 Мб, а в развернутом виде занимает примерно 23,5 Мб.

Для текстов, у которых вхождение “иностраных” символов в основной текст невелико, можно использовать графические вставки. Так мы поступали, например, для представления в Интернет фонетических комментариев, когда основной текст шел на русском языке и сопровождался примерами фонетической транскрипции, записанной символами Международного Фонетического Алфавита и представленной в виде инкорпорированных в текст графических файлов. Такой подход не требует от пользователя никаких предварительных установок шрифтов, однако едва ли может быть признан перспективным. Работа пользователя в сети несколько замедляется, поскольку кроме текста должны загрузиться и все картинки, а трудоемкость предварительной подготовки материалов для сайта значительно повышается.

Другой аспект проблемы состоит в том, что даже если символы разных алфавитов правильно отображаются на экране монитора пользователя, в настоящее время возможны проблемы с выводом страницы на печать, что зависит от типа принтера пользователя, операционной системы и ее настроек.

Поскольку нашей задачей является создание Интернет-системы, в разных фреймах которой одновременно могут быть представлены тексты на разных языках и другие символы (фонетическая транскрипция, нотная запись и т.п.), разрешение проблемы выбора шрифтов и кодировок (а также их доступности и универсальности) является крайне важным.

2.2. Проблема звукового формата. Выбор звукового формата для представления звукозаписей в Интернет определяется в первую очередь предназначением звуковой коллекции. Так, оцифрованный звуковой файл может быть представлен в Интернет как в исходном, так и в сжатом виде. Компрессия звукового файла позволяет существенно уменьшить его объем, что является достаточно важным для быстрой передачи его по сети Интернет, однако качество звучания падает пропорционально степени компрессии. Более того, получив по сети сжатый звуковой файл (например, в формате MP3 или RealAudio), пользователь уже не сможет вернуть его в декомпрессированное состояние. Этим компрессия звукового сигнала качественно отличается от архивации (ZIP, ARJ и все другие типы архивов при распаковке возвращают данные в исходное состояние, но сжатие звуковых файлов посредством стандартных архивов не эффективно, так как фактически не меняет объем файла).

Поэтому выставлять в Интернет компрессированный звук имеет смысл лишь для тех сайтов, которые имеют демонстрационный или общеобразовательный характер. Качественная работа со звуковым сигналом не может быть осуществлена на компрессированном материале, потому что в данном случае становится недоступной некоторая часть значимой информации, безвозвратно утраченная в результате компрессии. Поэтому в разрабатываемых нами сейчас сайтах “Региональные варианты русской звучащей речи” и “Web-коллекция эталонных звуковых образцов произношения для исследований в области синтеза

и распознавания речи” звук представлен в исходном WAV-формате, который является базовым для операционной системы WINDOWS.

Для прослушивания звукозаписей через сеть пользователю необходим компьютер со звуковой картой типа SoundBlaster и соответствующий Plug-in браузера или системный плеер. Если данные хранятся в сжатом формате, то возможно, что для их прослушивания пользователю придется установить соответствующий аудио-проигрыватель.

Таким образом, поскольку компрессия звукового сигнала для профессиональных систем Интернет-доступа нежелательна, для эффективной работы пользователя в режиме он-лайн необходимо подключение серверов звуковых коллекций к высокоскоростным сетям (например, vBNS), так как передача в Интернет оцифрованной речи в реальном времени становится возможной только при условии высоких скоростей и надежности каналов.

2.3. Проблема доступа и использования коллекций. Большинство коллекционных звукозаписей являются национальным достоянием России и не могут свободно распространяться через сеть Интернет. Мы предлагаем следующее решение этой проблемы: выставлять в Интернет со свободным доступом текстовые каталоги коллекций и лишь отдельные образцы и фрагменты звукозаписей. Для доступа к акустической базе данных в режиме он-лайн пользователю необходимо предварительно зарегистрироваться, заполнив стандартную регистрационную форму, и получить персональный пароль для входа в систему. Представленные образцы разрешается использовать исключительно в научных и культурологических целях. Несанкционированное копирование, тиражирование и любое их использование в коммерческих целях преследуется по закону.

Доступность звукового материала как можно большему числу потенциальных пользователей мы обеспечиваем посредством разработки простого и стандартизированного интерфейса, крайне необходимого для традиционных пользователей-филологов, большинство из которых до сих пор с недоверием и боязнью относятся к современным компьютерным технологиям.

3. Заключение

В процессе реализации проектов по представлению в Интернет звуковых коллекций и их каталогов выявилась острая необходимость в объединении усилий коллективов — исполнителей подобных проектов, заключающаяся в выработке единой концепции представления звукового материала, стандартизации формата хранения текстовой и звуковой информации каталогов коллекций, разработки единого и понятного пользователям-гуманитариям интерфейса взаимодействия с системой в режимах он- и офф-лайн.

Хотя главным объектом хранения и описания в разрабатываемой системе является звуковой сигнал, основные принципы подготовки и представления архивного материала в той же степени пригодны и для разработки баз данных, ориентированных на материал, представленный в текстовой или видео форме, а также для разнообразных мультимедийных приложений.

Представляемый доклад является обобщением работ, проводимых авторами в рамках проектов, поддерживаемых грантами РГНФ, № 99.04-12015в "Региональные варианты русской звучащей речи в Интернет" и РФФИ, № 00-07-90197 "Создание Web-коллекции эталонных звуковых образцов произношения для исследований в области синтеза и распознавания речи". Региональная и общенациональная значимость осуществляемых проектов состоит в сохранении национального наследия России, интеграции методов и концепций представления звуковых баз данных в Интернет, их стандартизации, а также в разработке модели пользовательского интерфейса, ориентированного на полноценную работу в сети пользователей-гуманитариев.

4. Литература

Бондарко Л.В., Вольская Н.Б., Скрелин П.А., Шерстинова Т.Ю., Кастров А.Ю. Концепция и структура Web-сайта "Каталог звукозаписей Санкт-Петербургских коллекций" // Труды I Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - Digital Libraries'99. Санкт-Петербург, СПбГУ, 19-22 октября 1999г. С. 216-218.

Бондарко Л.В., Вольская Н.Б., Скрелин П.А., Шерстинова Т.Ю. Электронные звуковые коллекции в Интернет // CD-ROM "Труды международной научной конференции "Управление электронными ресурсами библиотек"". Москва, Российская государственная библиотека, 18-19 апреля 2000.

Скрелин П.А., Шерстинова Т.Ю. Проект "Региональные варианты русской звучащей речи в Интернет" как модель интерактивной базы данных для филологических исследований // Тезисы II Всероссийской научно-методической конференции "Интернет и современное общество" - IMS'99. Санкт-Петербург, СПбГУ, 29 ноября - 3 декабря 1999г. С. 55-57.

Скрелин Павел Анатольевич - д.ф.н., доцент, зав. лабораторией экспериментальной фонетики СПбГУ, зам. декана по научной работе филологического факультета СПбГУ.

Шерстинова Татьяна Юрьевна - к.ф. н, дипл. спец. в обл. информационных систем и Интернет-технологий, ассистент кафедры фонетики СПбГУ

Таланов Андрей Олегович - программист, разработчик систем акустической обработки.

Щербаков Павел Петрович - к. физ-мат. н, спец. в обл. информационных систем и Интернет-технологий, доцент Междисциплинарного центра дополнительного профессионального образования СПбГУ.

Скрелин П.А., Таланов А.О., Шерстинова Т.Ю., Щербаков П.П. ©

