

Создание устно-речевой базы данных (УРБД) спонтанных речевых сообщений (на материале русского языка)

Абрамов Ю.В.

Потапова Р.К.

Хитина М.В.

Маслов А.В.

Бобров Н.В.

Московский государственный лингвистический университет.
Россия, 119034, Москва, ул. Остоженка, д. 38.
E-mail: RKPotapova@yandex.ru

Для создания УРБД спонтанных речевых сообщений на материале русского языка использовались различные виды речевой деятельности (чтение и говорение: монолог, диалог, полилог). Формирование УРБД многоцелевого назначения применительно к различным языкам мира является одной из приоритетных задач современного речеведения (Потапова, 2009). Подавляющее большинство конструируемых сегодня автоматизированных систем, работающих со звучащей речью, так или иначе используют устно-речевые базы данных.

В частности, УРБД находят применение там, где используются вероятностные и статистические методы анализа и синтеза речевого сигнала. В первую очередь здесь следует упомянуть системы автоматического распознавания и синтеза речи, идентификации и верификации говорящего по голосу и речи, идентификации психофизического и эмоционального состояния говорящего по речи, а также обучающие системы. Далее, УРБД составляют основу автоматизированных систем, в задачи которых входит сбор и хранение речевых сообщений, поиск и выдача записанных речевых сообщений по запросу (например, автоматизированные системы приёма голосовых сообщений в колл-центрах, комплексы для тестирования трактов связи). В ряде других случаев использование УРБД, не будучи строго необходимым технически, оказывается разумной альтернативой разработке сложных процедурных решений.

Как правило, УРБД содержат большие объёмы численной информации, трудно поддающейся автоматическому структурированию и сжатию. В то же время, в силу специфики систем, в которых применяются УРБД, в большинстве случаев эта информация должна быть доступна для обработки в режимах, близких к режиму реального времени, поэтому структура УРБД должна обеспечивать максимальное быстродействие системы при разумной ресурсоёмкости. По причине большого объёма информации изменение, а следовательно, и оптимизация структуры действующей УРБД, обычно является технически трудновыполнимой и крайне нежелательной операцией. С учётом многообразия задач, для решения которых применяются УРБД, это означает, что её структура должна быть универсальной и, как следствие, максимально простой.

При разработке УРБД неминуемо встаёт проблема выбора системы управления базами данных (СУБД). Здесь возможны следующие варианты: выбрать существующую, хорошо зарекомендовавшую себя СУБД из числа присутствующих на рынке информационных технологий или разработать свою СУБД специально для этой задачи. Желательно также предусмотреть средства, облегчающие (автоматизирующие) пакетный запуск алгоритмов различных видов обработки речевых данных.

УРБД разрабатываются для решения конкретной задачи. Круг возможных применений велик, однако конкретная задача задаёт непосредственные характеристики базы.

В качестве дикторов отбирались носители русского языка с учётом факторов возраста и гендера. В эксперименте участвовали дикторы трёх возрастных групп: до 25 лет, 25-40 лет, старше 40 лет. Группы испытуемых формировались как однородными по полу, так и состоящими из дикторов разного пола. К дикторам предъявлялись следующие требования: владение русским литературным языком; общеценностными знаниями; уровнем общей речевой культуры не ниже среднего; положительной модальностью при проведении записи (т.е. желанием и готовностью диктора принять участие в эксперименте).

До начала записи проводилось анкетирование дикторов, в которой учитывались антропологические параметры дикторов и их социальные и личностные характеристики. В случае необходимости фиксировались также особые условия записи (в виде примечания). В общей сложности оценивались не менее 34 параметров. Все данные, занесённые в анкету, были затем включены в качестве компонента в БД.

Следующим этапом подготовки к записи являлась настройка аппаратуры для каждого диктора (уровень интенсивности записи). Для этого проводилась пробная запись.

Непосредственно перед записью дикторов инструктировали, разъясняя им, какие сведения они должны сообщить в начале записи. Те же данные фиксировались в именах файлов, что позволило в дальнейшем находить нужную фонограмму без специального прослушивания.

В процессе проведения записи оператор должен был учитывать время говорения каждого диктора (в монологе, диалоге, полилоге); средний темп речи диктора (чтобы избежать необоснованного ускорения или замедления темпа); наличие небольших перерывов (2—3 мин) между звукозаписью; недопустимость обсуждения дикторами темы, развития диалога и полилога перед звукозаписью и в перерывах между звукозаписями; нежелательность присутствия в процессе записи посторонних лиц, не участвующих в эксперименте. Особое внимание обращалось на то, чтобы речь дикторов (за исключением чтения) была неподготовленной, спонтанной.

Для звукозаписи был использован аппаратно-программный комплекс, включающий аппаратное и программное обеспечение, проводился мониторинг хода звукозаписи и контролировалось их качество. Звукозапись проводилась в безэховой камере.

Проведённый предварительный лингвистический анализ позволил выявить совокупность признаков ($n=31$), относящихся к синтаксической и структурно-семантической организации спонтанной речи. Некоторые из них могли быть определены на основании экспертного прослушивания, другие требовали привлечения специальных методов анализа.

Для дальнейшего исследования было отобрано ограниченное число признаков, которые использовались для последующего аннотирования подготовленной УРБД. Среди них:

- наличие дефектов речи;
- характеристика голоса (мужской/женский/подростковый; высокий/средний/низкий);
- супрасегментные признаки артикуляции и фонации (громкость — высокая/средняя/низкая, темп — быстрый/средний/медленный, мелодика — монотон /«изрезанная»/комбинированный тип);
- фонетические средства выделения смыслового центра (ремы) — за счёт выделения слога (мелодическими, динамическими, средствами (громкость) / темпоральными средствами);
- нарушение порядка слов в предложении (инверсия, дистантное расположение связанных по смыслу слов и словосочетаний);
- наличие синтаксического эллипсиса;
- вводных синтаксических конструкций и слов-паразитов;
- простых нераспространённых предложений;
- простых распространённых предложений;
- сложных предложений (сложносочинённых, сложноподчинённых, бессоюзных сложных, предложений с сочинением и подчинением);
- паузация (синтактико-смысловые пауза и паузы хезитации).

Для проверки обоснованности выбора данного перечня признаков было проведено контрольное прослушивание отобранного для перцептивно-слухового анализа материала ведущими экспертами, мнение которых подтвердило состоятельность использования вышеперечисленных признаков. Несколько видоизменённая анкета использовалась для последующего аннотирования всего звучащего материала (за исключением чтения).

Для реализации базы данных фонограмм спонтанных речевых сообщений на материале русского языка (чтение, монолог, диалог, полилог) использована СУБД MS Access, поскольку она позволяла достаточно быстро разработать схему данных и логику работы. База данных состоит из следующих таблиц:

T2 — таблица, содержащая анкетные данные по каждому диктору (N=72)

F2 — таблица, содержащая сведения о физиоантропологических характеристиках дикторов (рост, вес и т.д.). Таблица F2 связана с таблицей T2 отношением «один к одному», поэтому в случае возникновения технологической необходимости может быть присоединена к ней.

S2 — таблица, в которую сведены данные о файлах фонограмм. Поля этой таблицы содержат следующие сведения: имя файла, уникальный код фонограммы, дату и время создания файла, а также размер файла.

L3 — таблица, содержащая подробную информацию о характеристиках речи каждого диктора, чей голос присутствует в фонограммах. Каждому диктору в каждой фонограмме в этой таблице соответствует отдельная запись. Таким образом, фонограммам монологов в таблице L3 соответствуют единичные строки, фонограммам диалогов — группы из 2 строк, фонограммам полилогов — группы из большего числа строк. Таблица L3 связана с таблицами F2 и S2 отношением «многие к одному».

Общее число единиц БД, представляющих информацию о фонограммах речи каждого из 72 дикторов, составляет 1067.

Для работы с БД можно использовать формы «Диктор», «Фонограмма» и «Речевое высказывание», вызываемые нажатием соответствующих кнопок из главной кнопочной формы (открывается автоматически при запуске БД). В этих формах реализованы возможности поиска информации по любому полю, просмотра списка фонограмм для каждого участника, данных лингвистического паспорта фонограммы, а также прослушивания фонограмм.

Перед началом работы с БД следует указать в соответствующих полях главной кнопочной формы пути к программе-проигрывателю (например, Windows Media Player) и к каталогу размещения аудиофайлов БД в системе, на которой предполагается вести работу с БД.

Для корректной работы интерфейса БД необходимо включить поддержку макросов (в СУБД Microsoft Access она отключена по умолчанию).

Поиск по фамилиям (именам) дикторов осуществляется путём выбора фамилии из выпадающего списка вверху формы Диктор.

Поиск по значениям других полей ведётся в следующем порядке: установить курсор в интересующее поле; нажать сочетание клавиш Ctrl-F; ввести искомое значение параметра; нажать Enter (или клавишу ОК на экране).

Воспроизведение фонограммы осуществляется путём выбора фонограммы из списка в нижней части формы Диктор (в этом списке отображаются все фонограммы, содержащие голос диктора, информация о котором представлена в полях данной формы) двойным щелчком левой кнопки мыши.

В дальнейшем возможна доработка базы данных в следующих направлениях: поиск информации по различным критериям; отображение различных поднаборов данных; графический анализ данных; визуализация фонограмм и их свойств.

Помимо результатов аудитивной экспертной оценки БД содержит релевантную информацию о каждом из дикторов, включая как стандартный набор анкетных сведений, так и информацию обличного характера, а также самооценку эмоционального и физического состояния дикторов перед началом записи.

Качество представленных в БД фонограмм и система паспортизации позволяют анализировать материал с учётом всех языковых уровней (фонетико-фонологического, лексического, синтаксического, семантического) и индивидуальных особенностей дикторов применительно к русской звучащей спонтанной речи.

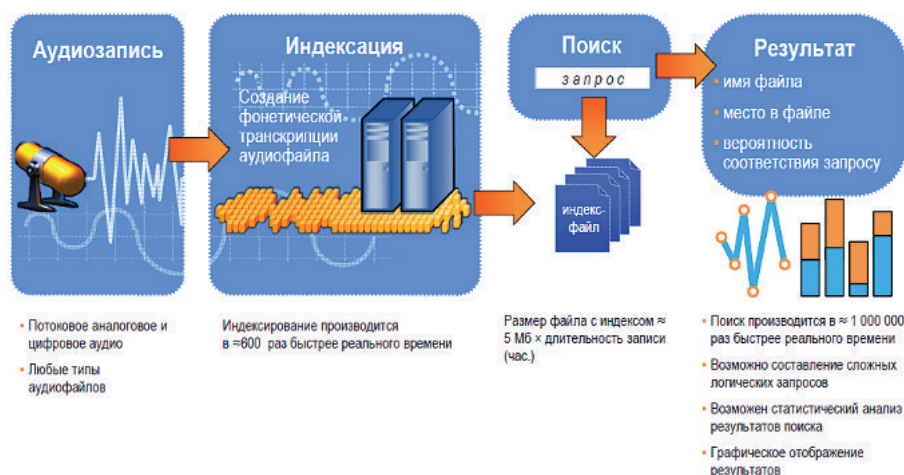
Конечный продукт реализован на высоком научно-исследовательском и научно-техническом уровне с учётом мировых достижений в области корпусной лингвистики. Разработанное приложение на базе СУБД Microsoft Access не имеет аналогов как в России, так и за рубежом и может быть успешно использовано для решения приоритетных задач специального назначения (например, автоматизированной идентификации личности на базе индивидуального вербально-смыслового оформления звучащей речи).

Дальнейшее развитие данного направления включает углубление и расширение информации, содержащей индивидуальные смыслонесущие признаки не только на базе вербалики, но также пара- и экстравербалики.

ЛИТЕРАТУРА

1. Потапова Р.К. Тенденции развития многоязычной корпусной лингвистики. Речевые технологии. № 3–4. М., 2009.

Информационное пространство, в котором функционируют современные организации, существенно изменилось за последние годы. Крайне высокими темпами растут объёмы создаваемой и хранимой информации, так в 2006 году объём хранимой в мире в цифровом виде информации составлял 161 экзбайт, в то время как в 2010 году, по экспертным оценкам, эта цифра составила уже 988 экзбайт, причём большую часть накопленной в мире информации — около 85% — составляют неструктурированные данные, среди которых наиболее быстрыми темпами растут объёмы аудиовизуальных данных.



В связи с этим за последнее время бурное развитие, как в фундаментальном, так и в прикладном плане получают технологии, позволяющие производить поиск и анализ неструктурированных, и в первую очередь, аудиовизуальных данных.

Научно-технический центр «ПОИСК-ИТ» находится на переднем крае таких исследований и представляет мощное, легко масштабируемое решение для анализа речи, позволяющее за короткое время извлекать и анализировать значимую информацию непосредственно из массивов аудио данных. В данном решении реализована непрерывно совершенствуемая технология фонемного поиска, позволяющая:

- производить поиск любых слов и выражений непосредственно в речевой составляющей файла;
- создавать логически сложные поисковые запросы, позволяющие максимально точно находить нужные сведения;
- проводить статистический анализ результатов поиска, извлекая таким образом новые знания из массивов аудио-данных.

Принципиальной особенностью системы является её способность производить высококачественный поиск по материалам низкого качества (зашумленным записям, записям речи по телефону и пр.).



Немаловажной характеристикой системы является её быстрдействие. Так, например, аудио-архив общей продолжительностью звучания 600 часов может быть доступен для поиска менее чем за один час. Непосредственно поиск выполняется со скоростью в 1 000 000 раз превышающей реальное время звучания.

Применение фоновго поиска позволяет значительно повысить эффективность поиска по аудио-данным и открывает пользователям богатейший источник информации.

Предлагаемая технология находит своё применение в различных областях человеческой деятельности, начиная от служб безопасности объектов разного уровня и заканчивая контакт-центрами, обладающими аудио-архивами наибольших объёмов. Применение предлагаемого решения, реализующего технологию фонетического поиска, в контакт-центрах позволяет значительно повысить их эффективность и расширить услуги, предлагаемые заказчикам и абонентам, за счёт возможности содержательного анализа записанных телефонных разговоров, а комбинация такого анализа с традиционными статистическими методами и встроенные аналитические отчёты позволяют поднять на новый уровень качество управления контакт-центром.