КОМПЬЮТЕРНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ СОЗДАНИЯ РЕЧЕВОГО КОРПУСА[1]

Широкова А. М. (anna_a@stel.ru),
Главатских И. А. (glavatskih_ia@stel.ru),
Платонова Т. С. (platonova_ts@stel.ru),
Рогожина В. С. (rogozhina_vs@stel.ru),
Смолина А. А. (smolina_aa@stel.ru)
Котов М. А. (kotov@stel.ru)

«Стэл — Компьютерные Системы», Москва, Россия

Ключевые слова: речевой корпус, речевая база, речевые данные, транскрипции, обработка речи, распознавание речи

SOFTWARE TOOLS FOR SPEECH CORPORA CREATING[2]

Shirokova A. M. (anna_a@stel.ru),
Glavatskih I. A. (glavatskih_ia@stel.ru),
Platonova T. S. (platonova_ts@stel.ru),
Rogozhina V. S. (rogozhina_vs@stel.ru),
Smolina A. A. (smolina_aa@stel.ru)
Kotov M. A. (kotov@stel.ru)

Stel—Computer systems ltd., Moscow, Russia

Speech database is an organized collection of speech fragments and their text transcripts provided with computer software which allows users to access them. Speech databases are an essential component of automatic speech processing, in particular, of speech recognition systems. This

---

[1] Настоящие исследования проводятся ООО «Стэл КС» при финансовой поддержке Министерства образования и науки Российской Федерации (соглашение №14.579.21.0058). Уникальный ID прикладных научных исследований (проекта) RFMEFI57914X0058. Ответственность за представленные данные, результаты и выводы несут исключительно авторы исследования.

[2] Research is conducted by Stel—Computer systems ltd. with support of the Ministry of Education and Science of the Russian Federation (Contract№14.579.21.0058). Unique ID for Applied Scientific Research (project) RFMEFI57914X0058. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

paper focuses on the software products designed by the company “Stel—Computer systems” for the purposes of optimization and simplification of database creation process. *Speech_to_text* is a program used to collect and organize speech data from certain speakers, as well as to automate the creation of spoken corpora on the basis of specific text categories. It enables simultaneous recording via a number of different microphones. *Speech_utility* simplifies and speeds up the build-up of databases for given speech recordings, containing their orthographic transcripts as well as the descriptions of some signal and speaker characteristics. Automatic recognition tools can be applied for preliminary processing to sort out the data valid for human analysis. *Audioprotocol* is a web-based application which interacts with media sources (open web content and TV channels) and automatically stores audio and video data. The collected speech data is categorized in order to create mass media speech databases for various languages. Multilingual speech recognition engine is used to automatically generate orthographic transcripts of this data.

Key words: speech corpora, speech database, speech data, transcript, speech processing, speech recognition

1. Введение

Речевой корпус — это структурированная совокупность речевых фрагментов, которая обеспечена программными средствами доступа к ним. Речевой фрагмент как базовая единица корпуса представляет собой оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определенного типа (типов) (Кривнова 2001). Как отмечено в (Кривнова 2004), формирование речевых корпусов представляет собой определенный технологический процесс, в котором можно выделить следующие этапы:

- подготовка фонетического обеспечения для формирования речевого корпуса;
- разработка стандартов для транскрипции речевого сигнала на разных уровнях;
- подготовка текстового материала;
- разработка программного обеспечения для формирования речевого корпуса;
- подбор дикторского состава;
- запись речевых фрагментов, произнесенных дикторами;
- проверка качества записи речевых фрагментов;
- создание детальных инструкций по разметке и фонетической интерпретации (транскрипции) речевых сигналов;
- фонетическая верификация речевых фрагментов и их разметка;
- верификация разметки и аннотаций речевого материала, полученных автоматически;
- обработка результатов верификации.

В настоящей статье обсуждаются применяемые нами способы оптимизации процесса создания речевых корпусов, суммируется опыт разработки и использования собственного программного обеспечения для автоматизации

и технологического упрощения ряда перечисленных выше этапов. В частности, запись речевых фрагментов осуществляется в программе *Speech_to_text*, создание орфографических и фонетических транскрипций речевых фрагментов выполняется в программе *Speech_utility*. Программный комплекс *Аудиопротокол* отвечает за формирование речевого корпуса в целом в автоматическом режиме, включая сбор, сохранение и текстовое аннотирование речевого материала.

Сразу оговоримся, что наши речевые базы имеют в первую очередь технологическое назначение и собираются для нужд конкретного проекта по автоматической обработке речи. Целевое использование определяет специфику речевого материала. Так, в проектах по распознаванию слитной речи перед нами стоит задача отекстовки множества образцов спонтанной речи, поскольку для улучшения точности системы распознавания требуется провести дополнительное обучение акустических моделей на реальных речевых данных, предоставленных заказчиком. Полученные речевые базы используются в дальнейших разработках, а также представляют самостоятельную ценность, в том числе для задач исследования особенностей слитной речи.

В силу особенностей области применения все описываемые программные средства являются многофункциональными и включают различные модули автоматической обработки речевых данных. В рамках этой статьи обсуждаются только те технические возможности наших компьютерных инструментов, которые непосредственно участвуют в реализации этапов формирования речевого корпуса, остальные же, по возможности, будут опущены.

2. Speech_to_text

Программный продукт *Speech_to_text* создавался как многофункциональная программа-клиент для взаимодействия пользователя с системой распознавания речи, установленной на удаленном сервере. Здесь рассматривается часть функционала, позволяющая автоматизировать процесс записи речевых фрагментов.

Рассмотрим подробнее структуру и характеристики программного обеспечения *Speech_to_text*, которое имеет графический интерфейс, представленный на рис. 1.

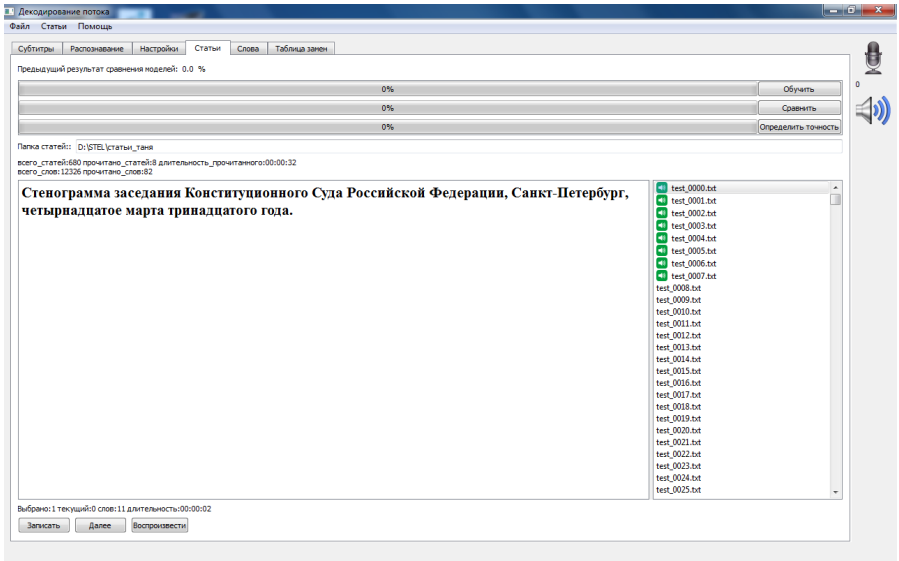

Рис. 1. Интерфейс программы *Speech_to_text*, процесс записи

2.1. Входные и выходные данные

Входными данными программного обеспечения являются:

- текстовые файлы для записи в формате *.txt*. Тексты на произвольную или определенную тематику подготавливаются заранее, затем они автоматически разбиваются на отдельные предложения.
- поток аудиоданных, поступающий с микрофона и/или видеокамеры, подключенных к персональному компьютеру пользователя;

Выходными данными являются:

- звуковые файлы в формате *.wav* со всех активных записывающих устройств, подключенных к персональному компьютеру (запись с каждого из них сохраняется в отдельный файл), названные в соответствии с именем файла в формате *.txt*.

2.2. Функциональные возможности

Для удобства записи в программе реализованы следующие возможности:

- отображение текста в графическом интерфейсе записывающей программы;
- возможность прослушивания и перезаписи любого записанного ранее файла текущей папки;
- автоматическое сохранение файлов в исходной папке.

Рассмотрим подробнее принцип работы ПО *Speech_to_text*. В процессе подготовки звуковых данных для последующего обучения, адаптации или тестирования модели диктора используются специально подготовленные текстовые данные (статьи), которые пользователь загружает в программу. Затем производится запись речи целевого диктора, читающего эти статьи в микрофон. Для удобства статьи предварительно автоматически разбиваются на предложения так, что во время записи в окне программы содержится текст только одного предложения; после его записи пользователь переходит к следующему, и так далее до конца текста. Кроме того, доступна перезапись текста.

Следует особо подчеркнуть, что ПО *Speech_to_text* позволяет осуществлять запись одного диктора одновременно со всех активных микрофонов, подключенных к персональному компьютеру, сохраняя запись с каждого из них в отдельный файл, название которого получает вид X_Y, где X — имя соответствующего текстового файла, а Y — код записывающего устройства.

В результате программа создает набор звуковых файлов (сохраняются автоматически в папку с текстовыми файлами, которые использовались для диктовки), содержащих речь целевого диктора. Таким образом, автоматически выстраивается файловая архитектура речевого корпуса, обеспечивается структурированность и порядок хранения речевого материала.

Важно отметить, что оператор, работая в полуавтоматическом режиме, имеет возможность активно контролировать процесс записи. Прямой доступ к созданным аудиофайлам из окна программы позволяет осуществлять верификацию их качества и перезапись (при необходимости) в неотложном режиме.

2.3. Применение

Исходя из вышесказанного, ПО *Speech_to_text* можно применять для решения широкого спектра задач в области корпусной лингвистики. Одной из базовых функций программного средства является запись потока речи, непосредственно поступающего от диктора на вход ПО с микрофона. Данный продукт предоставляет возможность быстрого и удобного сбора и структурирования речевых данных конкретных дикторов, а также автоматизации создания речевых корпусов на основе специально подготовленных текстов, в том числе определенной тематики. Кроме того, многофункциональный инструментарий программы обеспечивает доступ к этим данным и обучение на них моделей распознавания речи различных типов, а также выбор оптимальной модели распознавания в рамках поставленных задач.

Задействование одновременно нескольких микрофонов решает задачу сбора речевой базы, включающей звуковые файлы, одинаковые по содержанию, но с разными техническими характеристиками, при однократной записи каждого диктора. Такая специфика данных позволяет, в частности, адаптировать модель распознавания для разных типов микрофонов, сравнить качественные характеристики используемых микрофонов, а главное, обучить обобщенную акустическую модель, независимую от типа микрофона.

С помощью ПО *Speech_to_text* экспертами нашей компании созданы корпуса речи для различных языков (русского, английского, немецкого, испанского). Собраны узкоспециализированные базы данных, например, база данных читаемых текстов судебно-правовой тематики подготавливалась в рамках проекта по внедрению системы распознавания речи для протоколирования судебных заседаний. Продолжающийся научно-исследовательский проект — разработка системы автоматической верификации дикторов по случайно сгенерированной парольной фразе — использует речевой материал, записанный через *Speech_to_text* с использованием четырех активных микрофонов с разными техническими характеристиками.

3. Speech_Utility

Создание различных типов аннотаций к звуковым файлам — наиболее трудоемкая и затратная по времени экспертная работа, традиционно выполняемая вручную[3]. Для оптимизации временных затрат и облегчения кропотливого ручного труда в рамках проектов по созданию речевых корпусов часто разрабатывается соответствующий инструментарий, позволяющий существенно уменьшить долю чисто технических, несодержательных операций, выполняемых человеком. Так, в ИСА РАН была создана открытая, гибкая и многофункциональная программная оболочка для проведения работ в области исследования речи, сопровождения речевых баз данных, простой визуализации и обработки речевых сигналов. Эта же оболочка используется для обеспечения процесса создания речевых баз данных, который включает в себя накопление звуковых файлов и разметку этих файлов экспертами на фонетически и акустически значимые сегменты (Арлазаров 1998). Как отмечается в (Богданова 2009), широкие возможности для ручной экспертной обработки речевых данных имеют зарубежные системы, открыто доступные для некоммерческого использования. В частности, программное средство Praat (Boersma 2001) предназначено для лингвистов, исследующих звучащую речь. Его функционал позволяет осуществлять детальный акустический и перцептивный анализ фонетических явлений, а также выполнять сегментацию звукового потока. Компьютерный инструмент ELAN (Wittenburg 2006) является средой профессионального лингвистического аннотирования аудио- и видеоматериалов, которая поддерживает неограниченное количество уровней аннотации, различные шрифты и кодировки данных, сложные иерархические связи между данными, экспорт и импорт аннотаций в основные форматы представления данных. Нашей группой разработано собственное программное средство, имеющее аналогичное назначение, но обладающее дополнительными технологическими возможностями.

---

[3] Возможно, это в меньшей степени относится к орфографической записи, однако, учитывая обработку больших объемов аудиоданных, является вполне справедливым.

3.1. Описание системы

Speech_Utility — это комплекс программ пакетной и единичной специальной обработки звуковых файлов, предназначенный для ручной и автоматической подготовки массивов оцифрованных звуковых данных, содержащих речевую информацию.

- В его состав входят следующие компоненты:
- основная программа *Speech_utility*;
- независимо компилируемые программные модули, динамически подключаемые к основной программе.

Основная программа *Speech_Utility* имеет графический интерфейс, представленный на рисунке 2, при помощи которого оператор может вызывать те или иные функции комплекса.

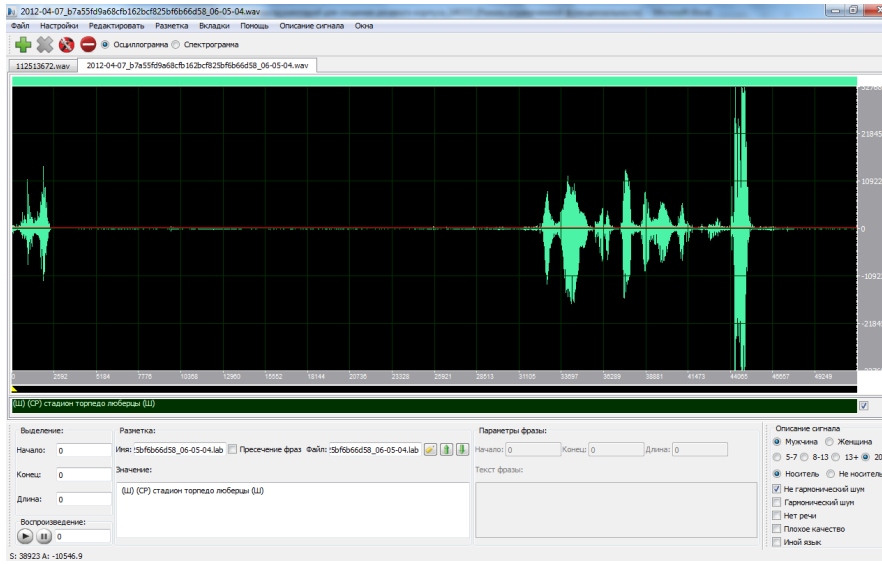

Рис. 2. Интерфейс программы *Speech_Utility*

Базовой задачей ПО *Speech_Utility* является автоматизация процесса создания аннотаций речевых фрагментов и обеспечение удобства работы лингвиста при их формировании и редактировании.

3.2. Входные и выходные данные

Для начала работы со *Speech_Utility* звуковые файлы должны быть приведены к следующему формату:

- расширение файла: *.wav*;
- метод оцифровки сигнала: PCM;
- частота дискретизации: 8 кГц;

- разрядность: 16 бит;
- количество аудиоканалов: один (моно).

Выходными данными являются:
- файлы разметки с расширением *.lab*
- файлы формата *.json*, содержащие информацию о дикторе и свойствах сигнала.

3.3. Функциональные возможности

Для удобства работы с акустическим сигналом в главном окне можно выбрать тип его визуального отображения (осциллограмма/спектрограмма) и формат измерения времени (отсчеты/секунды).

Результатом обработки звукового файла является файл разметки, т. е. массив данных, содержащий информацию об одном или нескольких выделенных в нем речевых фрагментах. Пример файла такого типа приведен на рис. 3. Информация о конкретном фрагменте речи указывается в отдельной строке разметки. Строка разметки содержит следующие поля:
- временная метка начала фрагмента;
- временная метка окончания фрагмента;
- текст фрагмента[4].

```
468750 197657500 ЖС_ (Д) до старта олимпиады в сочи ровно сто дней седьмого февраля в
двадцать четырнадцать весь мир увидит церемонию открытия (ШД) обратный отсчёт времени
сегодня начнётся во многих городах запланированы %праздничные мероприятия (ШД) но главные
конечно же в сочи (Д) большинство объектов столицы зимних игр уже готовы (Д) а наши
спортсмены усиленно тренируются чпобы выступить как можно лучше (Д) репортаж ивана
прозорова
197657500 230907500 (КР)
230907500 307852500 МК_ на кубке россии по бобслею почти олимпийское напряжение сейчас
каждый заезд для спортсменов как экзамен на попадание в сборную страны экипаж анастасии
307852500 360676250 МК_ (РМ тамбовцевой на пьедестале на первом месте) (Д) (РМ но даже
лидеры пока не знают кто поедет на олимпиаду)
360676250 404500000 ЖЗ_ %очень хочется попасть на олимпийские игры выступить достойно и
стоять так же на первом месте
404500000 559246250 МК_ в сочи бобслеисты провели почти месяц десятки раз проезжали по
трассе запоминали виражи и испытывали новое оборудование женские экипажи теперь называют
принцессами на бобах (Д) за поистине сказочные результаты стали быстрее почти на секунду
(Д) ускориться помогут и родные стены уверены спортсменки
559246250 675183750 ЖЗ_ не каждый спортсмен своей страны может почувствовать что у него
дома %будут проходить олимпийские игры то есть (Д) это поддержка зрителей будут русские
люди которые будут поддерживать болеть и вот ну как бы ещё больше хочется выступить
675183750 680065625 (НП)
```

Рис. 3. Файл разметки формата *.lab*, включающей множество речевых фрагментов

В программе *Speech_Utility* существует возможность создавать несколько уровней разметки и автоматически сохранять полученные разметки в файлы *.lab* в папку, указанную пользователем, или по умолчанию в исходную папку

---

[4] В текстовой аннотации используются специальные пометы для различных типов речевых и неречевых событий, например, (Д) — дыхание. Эти пометы делаются экспертами в соответствии с определёнными правилами аннотирования, которые составляются с учётом конкретных задач и типа речевого материала. В данной статье эти правила не рассматриваются.

со звуковыми файлами. Если пользователь хочет создать и затем сохранить несколько уровней разметки одного и того же файла, в этом случае формат сохраняемых файлов разметки по умолчанию будет следующим:

- уровень 1 разметки:имя_файла.*lab*
- уровень 2 разметки:имя_файла.*lab1*
  и т. д.

Непосредственно в программе *Speech_Utility* предусмотрена возможность фиксирования сведений о дикторе и свойствах сигнала. В правом нижнем углу окна программы находится меню «Описание сигнала», в котором оператор может выбрать нужное значение из предложенных характеристик диктора и установить соответствующие параметры сигнала:

- мужчина/женщина/ребенок;
- возраст (от 5 до 7 лет, от 8 до 13 лет, от 13 до 20 лет, от 20 лет и старше);
- носитель/не носитель;
- негармонический/гармонический шум;
- нет речи;
- плохое качество;
- иной язык.

Вся эта информация автоматически сохраняется в файл формата *.json* в папку, указанную пользователем, или по умолчанию в исходную папку со звуковыми файлами. В дальнейшем эти данные позволяют группировать базы данных по различным признакам: пол диктора, возраст, диалектные особенности, определенные типы шумов и т. д.

3.4. Индивидуальные настройки

В *Speech_Utility* можно установить следующие настройки:

- внешний вид графического интерфейса программы;
- количество плагинов программы;
- горячие клавиши (сочетание клавиш для вызова выбранной функции).

Все изменения сохраняются на компьютере пользователя, поэтому каждый оператор может настроить программу индивидуально.

3.5. Подключаемые модули

Как отмечено выше, в *Speech_utility* доступно использование отдельных модулей, предназначенных для автоматического распознавания акустических объектов разной природы и размерности (фонем, различных типов шумов и т. д.). Применение модуля *Plugin_VAD* позволяет в автоматическом режиме предварительно отсеивать сильно зашумленные и некачественные записи,

содержащие данные, не пригодные для ручной обработки, а также создавать первичную разметку звукового файла на речевые и неречевые участки. Модули *Plugin_PhonemeRecognition* и *Plugin_PhonemePlacer* предназначены для автоматизации создания фонетической транскрипции с временными метками. Первый модуль производит пофонемное распознавание речи и расставляет границы фонем, распознанных в речевом фрагменте, опираясь только на обученные акустические модели. Второй модуль работает по имеющейся фонетической транскрипции речевого фрагмента: он соотносит известную последовательность фонем и акустический сигнал и определяет границы звучания каждой фонемы. При использовании данных функций задача лингвиста сводится к ручной корректировке результатов автоматической разметки.

3.6. Применение

Итак, функционал *Speech_Utility* позволяет одновременно загружать большой массив аудиоданных и иметь несколько различных уровней разметки, отображаемых в одном окне и синхронизированных с осциллограммой (или спектрограммой) аудиосигнала. Возможность активации горячих клавиш, настройки автосохранения, быстрого выбора из заданных свойств сигнала делает *Speech_Utility* мощным инструментом для создания не только аннотированных, но и структурированных определенным образом баз данных. По нашим приблизительным оценкам, использование этого инструмента ведет к троекратному увеличению скорости обработки аудиоданных. В общей сложности с его помощью наши специалисты подготовили аудиоматериал, насчитывающий более 300 часов чистой речи на разных языках и взятый из разных источников, и сформировали орфографические транскрипции для различных типов речевых записей (теле- и радиовещание, телефонные разговоры, заседания и т. д.).

4. Аудиопротокол

Речевые корпусы имеют принципиальное значение в области автоматической обработки речи. Современный подход к распознаванию речи, базирующийся на статистическом моделировании ее акустических характеристик, предполагает обучение системы на больших объемах речевых данных.

На сегодняшний день в области речевых технологий уже достигнуты ощутимые результаты, предлагаются готовые, работающие решения. Нашими специалистами создана система распознавания русской речи, интегрированная в коммерческий программный продукт *Аудиопротокол*[5][6].

---

[5] http://speech.stel.ru:8080 (данные для онлайн-регистрации: логин guest, пароль 1).

[6] http://speech.stel.ru:8080 (online access: login guest, password 1).

4.1. Описание системы

Данное программное обеспечение (далее ПО) объединяет в себе несколько функционально разнородных возможностей. В первую очередь, обеспечивается автоматизированное взаимодействие с медийным пространством (открытыми интернет-ресурсами и телеканалами), автоматическое получение аудио- и видеоинформации, а также предварительная автоматическая рубрикация данных в зависимости от источника (новости, фильмы и т. д.). Второй компонент ПО позволяет с довольно высокой точностью распознавать и документировать речь, звучащую в полученных аудио- и видеоматериалах.

Остановимся более подробно на структуре и функциональных характеристиках системы *Аудиопротокол*, обеспечивающей автоматический захват и распознавание речи в аудио- и/или видеоданных, опубликованных цифровым информационным источником открытого доступа.

4.2. Входные и выходные данные

Входными данными ПО *Аудиопротокол* являются:

- аудио- и видеоданные, опубликованные на поддерживаемых открытых источниках;
- файлы формата *.wav*, *.mp3*, *.flv*, *.mp4*, *.webm*, содержащие аудио- и видеоданные;
- поток аудио- и/или видеоданных, поступающий с микрофона или камеры, подключенных к персональному компьютеру пользователя.

В ПО *Аудиопротокол* реализована возможность ввода аудио- и видеоданных следующими способами:

- мониторинг канала цифрового телевидения;
- мониторинг сайта в сети Интернет;
- разовая загрузка данных по внешней ссылке;
- загрузка данных по локальной информационной сети из одного или нескольких файлов, указанных пользователем;
- загрузка потока данных с периферийного устройства, подключенного к компьютеру пользователя.

Выходными данными ПО *Аудиопротокол* являются файлы следующих форматов:

- *.mp3* (содержат аудиоданные), *.mp4* (содержат видеоданные);
- *.docx*, *.txt* (содержат результирующую орфографическую транскрипцию с временной разметкой на слова).

4.3. Принципиальное устройство и структурная схема

В состав ПО *Аудиопротокол* включены программы, перечисленные ниже в таблице 1.

Таблица 1. Перечень программ в составе ПО *Аудиопротокол*

Название	Назначение
Программа захвата данных	Обеспечивает мониторинг и захват данных на входных разъемах ТВ-тюнера и с сайтов в сети Интернет
Программа обработки данных	Выполняет сохранение данных, распознавание речи и сохранение результатов

На рисунке 4 представлен процесс мониторинга целевого сайта и обработки полученных данных.

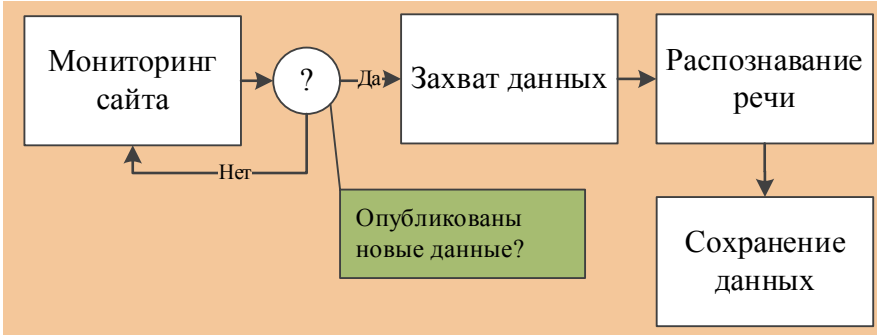

Рис. 4. Процесс мониторинга целевого сайта и обработки полученных данных

В ПО *Аудиопротокол* обработка данных выполняется в несколько этапов, перечисленных в таблице 2.

Таблица 2. Этапы обработки данных

Этап	Результат	Пояснение
1. Ввод данных	Данные введены	Выполняется ввод данных тем или иным методом. Оператор устанавливает настройки выполнения следующих этапов 2–4.
2. Распознавание речи и ключевых слов	Данные разделены на фрагменты, для каждого фрагмента создана разметка на слова	Этап выполняется автоматически в несколько действий: а) Разделение данных на фрагменты тем или иным методом. б) Распознавание речи во фрагментах данных. в) Создание разметки на словах для каждого фрагмента.

Этап	Результат	Пояснение
3. Создание разметки данных	Для данных создана разметка на слова	Разметки всех фрагментов данных включаются в состав единой разметки.
4. Сохранение результатов	Данные сохранены	Этап выполняется автоматически в несколько действий: а) введенные на этапе 1 данные сохраняются в файл *.mp3* или *.mp4*; б) создается карточка файла, в которой размещаются: • текстовая разметка данных, созданная на этапе 3; • настройки, по которым были выполнены этапы 2–4.

4.4. Функциональные характеристики

В ПО *Аудиопротокол* реализованы следующие возможности:

- мониторинг и захват данных, опубликованных открытым источником информации в сети Интернет и цифровым телевизионным каналом;
- автоматическое распознавание речи и ключевых слов на русском языке в аудио- и видеоданных;
- воспроизведение аудио- и видеоданных с синхронным графическим маркированием слов в разметке;
- автоматическое формирование статистики встречаемости слов в аудио- и видеоданных, загруженных в ПО *Аудиопротокол*;
- отбор данных по ключевому слову или по дате загрузки;
- экспорт аудио- и видеоданных в файлы типа *.wav*, *.mp3*, *.flv*, *.mp4*, *.webm*;
- экспорт разметки в файл *MS Word* или *.txt*;
- редактирование разметки данных.

4.5. Технологические параметры

Основные технические требования:

- скорость: локальная сеть должна обеспечивать передачу пакетов данных по протоколу *Ethernet* со скоростью не менее 1 Гбит/с;
- программная совместимость: рекомендуется использовать браузер *Google Chrome* версии 29.0.1547 и выше.

Важные технологические характеристики:

- быстродействие: в десять раз быстрее времени звучания;
- точность: до 85 % правильно распознанных слов.

Приведем несколько замечаний об условиях успешной работы системы. Как видно из таблицы 3, средняя точность составляет 75–80 % правильно распознанных слов, однако она существенно падает при распознавании беглой речи, при невнятном, неотчетливом произнесении (ср. результаты распознавания речи диктора 2Ж). Более высокая точность распознавания может быть достигнута путем адаптации акустической модели на речевых данных конкретного диктора (ср. результаты распознавания речи диктора 5М, для которого была проведена адаптация).

Таблица 3. Результаты работы системы распознавания, указан процент ошибочно распознанных слов. Тестовый материал: 5 текстов, озвученных 5 дикторами, в скобках указан пол диктора

	Диктор 1 (Ж)	**Диктор 2 (Ж)**	**Диктор 3 (М)**	**Диктор 4 (М)**	**Диктор 5 (М)**
Текст 1	27,7	32,4	21,5	23,5	14,9
Текст 2	28,9	41,4	22,8	25,1	19,6
Текст 3	24,4	37,6	20,0	20,0	12,6
Текст 4	23,6	34,4	22,7	22,4	14,5
Текст 5	24,4	41,1	16,5	21,8	16,9
Среднее	25,8	37,38	20,5	22,56	15,7

4.6. Возможности использования и перспективы развития

Итак, согласно приведенному выше описанию, программный продукт *Аудиопротокол* можно использовать как средство автоматической генерации речевого корпуса, имеющего заданный стандарт иерархической организации и хранения аудио- и видеоданных и соответствующих файлов разметки. Он включает в себя многофункциональный инструментарий, обеспечивающий доступ к этим данным и их обработку. Фактически, множество перечисленных выше этапов формирования речевого корпуса, требующих участия человека, сводится к следующим:

- проверка качества записи речевых фрагментов;
- фонетическая верификация речевых фрагментов;
- верификация разметки и аннотаций речевого материала, полученных автоматически;
- обработка результатов верификации.

Существует принципиальная возможность создания в автоматическом режиме приведенных ниже типов аннотаций:

- орфографическая запись;
- фонемная / фонетическая транскрипция;
- акустико-фонетическая разметка сигнала.

Потенциально это решается встраиванием периферийных функций в состав ПО и не требует изменения его исходной архитектуры. В настоящей версии результат распознавания фиксируется в виде орфографической транскрипции. Благодаря высокой скорости обработки распознанная последовательность лексических единиц выдается пользователю синхронно с воспроизводимой речью, при этом звучащий в данный момент времени фрагмент речи (слово или группа слов) выделяется другим цветом шрифта. Следует отметить, что орфографическая запись генерируется без знаков препинания, восстановление которых потребовало бы дополнительного использования инструментов извлечения синтаксической и просодической информации. На сегодняшний день, по материалам научных источников, разработаны статистические методы анализа синтаксической структуры текста, однако возможность их интегрирования в нашу систему пока не рассматривалась. Также ведутся исследования по моделированию просодических характеристик речи, однако, по известным нам данным, технологии использования высокоуровневых признаков находятся еще в стадии формирования и не имеют широкого практического применения.

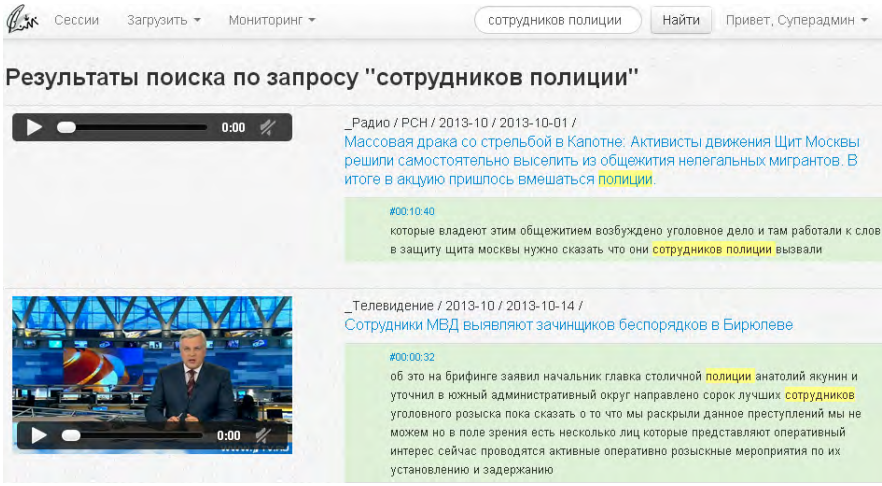

Рис. 5. Интерфейс системы *Аудиопротокол*, функция поиска

Функция поиска слов в орфографических транскрипциях, имеющая программный интерфейс, представленный на рис. 5, обеспечивает автоматический переход к соответствующему фрагменту звукового сигнала, делая возможным целевое прослушивание. Следует отметить, что удобство организации работы с образцами звучащей речи может служить подспорьем в проведении специальных фонетических исследований, требующих изучения большого количества фонетических контекстов, например, при анализе допустимой вариативности произнесения в речи носителей.

5. Заключение

Подводя итог, обобщим функциональные преимущества компьютерного инструментария, разработанного нами для оптимизации процесса формирования речевых корпусов. Еще раз подчеркнем, что мы работаем с речевыми данными различных типов (записи спонтанной речи, новостей, поисковых запросов) и ориентируемся, в первую очередь, на технологическое применение конечных корпусных решений.

Программное средство *Speech_to_text* позволяет автоматизировать процесс записи речевых фрагментов, предоставляя легкий доступ к текстовым и звуковым файлам и возможность их актуальной верификации и перезаписи. Также здесь реализована функция одновременной записи с множества микрофонов.

Специализированный программный продукт *Speech_Utility* является удобным инструментом, упрощающим и ускоряющим процесс формирования текстовых и иных аннотаций звуковых записей. Кроме того, доступны возможности автоматической предобработки речевых данных и сохранения информации о дикторе и свойствах сигнала. Конкретные разработки аналогичного типа отличаются целевой ориентацией, уровнем сложности и числом поддерживаемых функций, степенью продуманности интерфейса. Однако следует признать, что это лишь разные способы автоматизации ручной экспертной обработки, но ключевая роль самого эксперта остается при этом неизменной.

Программное обеспечение *Аудиопротокол* можно рассматривать как готовый инструмент для автоматического сбора больших объемов аудио- и видеоинформации, размещенной в открытом доступе и транслируемой открытыми источниками в сети Интернет и каналами цифрового ТВ. При этом значительная часть работы по структурному формированию и содержательному наполнению базы осуществляется автоматически, в частности, за создание текстовых аннотаций для аудио- и видеоданных отвечает система автоматического распознавания речи. Кроме того, предусмотрена автоматическая сортировка данных по источнику и по ключевым словам.

Литература

1. *Арлазаров В. Л., Богданов Д. С., Паклин М. Л., Розанов А. О., Финкельштейн Ю. Л.* (1998) Инструментальная система для исследования и обработки речевых сигналов и создания речевых баз данных // В сб. Интеллектуальные технологии ввода и обработки информации М., Эдиториал УРСС.
2. *Кривнова О. Ф., Богданов Д. С., Подрабинович А. Я.* (2004) Современный инструментарий для разработки речевых технологий // Информационные технологии и вычислительные системы. N2. М., УРСС.
3. *Кривнова О. Ф.* (2006) Области применения речевых корпусов и опыт их разработки // Труды XVIII сессии Российского акустического общества РАО. Таганрог.

4. *Богданова Н. В., Асиновский А. С., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* (2009) Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 8 (15). По материалам ежегодной международной конференции «Диалог 2009», Москва.
5. *Boersma P., Weenink D.* (2001). Praat, a system for doing phonetics by computer. Glot International, 5 (9/10).
6. *Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H.* (2006). ELAN: a Professional Framework for Multimodality Research. In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

References

1. *Arlazarov V. L., Bogdanov D. S., Paklin M. L., Rozanov A. O., Finkelshtein Yu. L.* (1998) Instrumental system for speech signal research and processing and speech database creation [Instrumental'naya Sistema dlya issledovanya i obrabotki rechevykh signalov i sozdaniya rechevykh baz dannykh], Intelligent technologies of information input and processing [Intellectual'nyye tekhnologii vvoda i obrabotki informatsii], Editorial URSS, Moscow.
2. *Krivnova O. F.* (2006) Speech corpora development and application [Oblasti primeneniya rechevykh korpusov i opyt ikh razrabotki], Proc. 18th Russian acoustic society sess. [Trudy XVIII sessii Rossiyskogo akusticheskogo obschestva RAO], Taganrog.
3. *Krivnova O. F., Bogdanov D. S., Podrabinovich A. Ya.* (2004) State-of-the-art tools for speech technologies development [Sovremenny instrumentariy dlya razrabotki rechevykh tekhnologiy], Information technologies and computation systems [Informatsionnyye tekhnologii i vychislitel'nyye sistemy], Vol. 2, URSS, Moscow.
4. *Bogdanova N. V., Asinovsky A. S., Rusakova M. V., Ryko A. I., Stepanova S. B., Sherstinova T. Yu.* (2009) A speech corpus as a tool for monitoring and fixation of various forms of natural language [Zvukovoy korpus kak sposob monitoringa i fiksatsii raznykh form yestestvennogo yazyka] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" / [Komp'yuternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoy conferentsii "Dialog 2009"], Moscow.
5. *Boersma P., Weenink D.* (2001). Praat, a system for doing phonetics by computer. Glot International, 5 (9/10).
6. *Wittenburg P., Brugman H., Russel A., Klassmann A.,Sloetjes H.* (2006). ELAN: a Professional Framework for Multimodality Research. In Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.