Разметка разговорного речевого материала

Чучупал В.Я.,
кандидат физико-математических наук

Оценка параметров вероятностных моделей звуков в современных системах распознавания слитной речи осуществляется с использованием больших корпусов данных, которые включают, помимо собственно речевого сигнала, также его разметку. Переход к решению задач распознавания естественной речи делает неоптимальной разметку, которая использовалась при создании многих известных корпусов данных. В работе описан опыт формализации содержания разметки для корпуса данных русской разговорной речи и создания соответствующего программного инструментария.

The paper describes the issues that arise when markup of speech corpora of Russian telephone spontaneous speech is performed. The spontaneous nature of telephone conversation, speech disorders, accents, disfluences, use a variety of channels, places and environments should be represented in the annotation. The XML based annotation for telephone stereo conversations is described and DTD is presented.
The scheme utilize multilevel annotation (turn level, word level, phone level etc.) and the levels are liked both to text and signal representation of conversation. The software annotation tool is presented. The software is based on the open source multi-platform Tcl/Tk code with Snack and tDom packages. The graphic user interface lets possibility to work at the same time on the hypertext window and graphic representation of speech wave and markup elements.

Разметка речевого материала для решения задач моделирования и распознавания

Современные системы распознавания слитной речи основаны на использовании вероятностных моделей для описания звуков речи, слов и свойств языка. Такие модели могут содержать 10^6 и более параметров, поэтому объём обучающей выборки должен быть очень большим. Вследствие этого обстоятельства прогресс в области речевых технологий (уменьшение числа ошибок, возможность распознавания естественной речи, увеличение размера словаря) тесно связан с наличием адекватных, по размерам и условиям использования, обучающих выборок — корпусов речевых данных, на которых обучаются, настраиваются и проверяются системы распознавания речи.

Речевой корпус данных — это массив речевой информации, который записан в цифровой форме и снабжён текстовым описанием — разметкой (аннотацией) для того, чтобы обеспечить удобный доступ и автоматическую обработку данных с помощью машины. Разметка может включать общую информацию о речевом материале: о времени, месте, характеристиках (полоса частот, тип канала, микрофона и т. п.) записи и сведения о дикторах (пол, возраст и т. п.), а также орфографическое, фонетическое и т. п. представление записей речевого сигнала. Термины «разметка» и «аннотация» в данной работе являются синонимами.

Разметка — важнейший элемент речевых корпусов, поэтому вопросам содержания и способов представления разметки посвящено много публикаций [2–6]. Формат разметки определяется задачами, которые предполагается решать и общепринятых стандартных форматов аннотирования речевых баз данных пока нет. Существующие корпуса данных в основном создаются на основе локальных соглашений в рамках конкретных проектов. Среди существующих есть мощные и универсальные форматы разметки, которые поддерживают многоуровневое описание речевого сигнала и возможности поиска участков речевого сигнала по тем или иным комбинациям признаков [5–7]. Однако такие системы нацелены скорее на запросы лингвистов и фонетистов и практически редко используются в задачах оценивания параметров моделей или показателей эффективности, когда присутствуют огромные массивы информации, обработка которых имеет последовательный характер.

В речевой технологии корпуса данных обычно используются при обучении акустических моделей (например, для задания априорного соответствия между участками речевого сигнала и звуками речи, автоматической сегментации речевых высказываний с помощью процедуры Витерби, оценки параметров смесей нормальных распределений с использованием процедур прямого и обратного хода и Баума-Уэлча и т. д.). При оценивании эффективности распознавания наличие разметки тестового материала даёт возможность автоматического измерения вероятности пословных и фонемных ошибок распознавания, ошибок пропусков и вставок ключевых слов. Специфика этой области состоит в очень большом объёме данных (сотни часов звучания) в сочетании с последовательностью (запись за записью) их использования при обработке. Поэтому наличие произвольного доступа и поиска данных по запросам (хотя это выглядит привлекательно) не является определяющим при выборе форматов или способов обработки разметки и обычно уступает место более простым и удобным в данном случае форматам представления материала.

Один из первых и самых известных вариантов разметки был предложен и использован в корпусе данных TIMIT [12]. Это самая распространённая на сегодняшний день акустико-фонетическая база данных в мире. Название TIMIT образовано из имён двух организаций: TI (Texas Instruments) и MIT (Massachusetts Institute of Technology), создавших этот корпус данных по заказу правительственного агентства DARPA DARPA [9] в 1988 году.

TIMIT включает 6300 высказываний от 630 дикторов. Набор высказываний ограничен, словарный запас этой базы мал и её используют главным образом для обучения моделей звуков и проверки эффективности фонемных распознавателей. Разметка TIMIT включает орфографическую транскрипцию, фонемную произносительную транскрипцию и фонетическую сегментацию. Корпус организован по фразам: каждое высказывание записано в отдельный файл, которому соответствует несколько файлов разметки, для каждого типа разметки отдельный файл. Например:

- файл с расширением .txt, содержит орфографическую транскрипцию фразы;
- файл с расширением .wav, содержит оцифрованный речевой сигнал;
- файл с расширением .phn, содержит фонетическую транскрипцию фразы, с временами начал и концов звуков;
- файл с расширением .wrd, содержит список слов фразы с временами начал и концов.

Дополнительно к этому некоторая информация о дикторах, диалектах, принадлежности файлов к той или иной категории кодируется в самом имени файла.

«Пофразовая» организация оказалась довольно удобной и гибкой, позволяя организовать анализ и обработку данных в пакетном режиме.

Появившийся в 1989 году речевой корпус данных ATIS (Air Travel Information System) [12] также был собран в рамках проекта ARPA. Он состоит из записей речевых запросов информации в информационной системе воздушного транспорта. Записи получены путём симуляции присутствия системы распознавания речи на стороне сервера. Организация данных в ATIS также, как и в TIMIT, имеет пофразовый характер: каждое высказывание записано в отдельном wav-файле, которому соответствует несколько (точнее, 9) различных типов разметки — каждый в своём файле, например файлы, содержащие орфографическую транскрипцию фразы (так, как диктор намеревался её произнести), текст подсказки-приглашения, текст фразы в нормализованной форме, реально произнесённый текст с перечнем основных акустических событий (пропусков, вставок и т.п.), фонетическую транскрипцию высказывания, лог-файл сценария сессии записи, файл с дополнительной информацией о фразе и файл категории фразы.

Форматы разметки и организация данных TIMIT и ATIS были удобными на практике и оказали большое влияние на дизайн разметки последующих корпусов.

К настоящему времени созданы и доступны сотни корпусов речевых данных на разных языках, предназначенных для решения различных задач. Значительная часть их сосредоточена в трёх источниках: в Европе это Агентство по оценке и распространению языковых ресурсов, ELDA [10] и Европейское агентство языковых ресурсов ELRA [11]., а в США — Консорциум лингвистических данных, LDC [12].

В Российской Федерации практически все организации, которые осуществляют проекты в области речевых технологий, также располагают своими собственными корпусами данных. Возможно, наиболее представительные речевые корпуса русского языка были собраны по иностранным заказам. Так, фирма ОдиТек [13] создала целый ряд корпусов данных в рамках европейских проектов SpeechDat, SpeeCon, LC-star, в институте системного анализа РАН [14]

совместно с сотрудниками МГУ им. М.В. Ломоносова были разработаны известные корпуса данных ISABase и RuSpeech, а также мощный и удобный для проведения разметки программный комплекс DiffKit [3,4].

В ВЦ РАН (ВЦ АН СССР) накоплен опыт сбора, разработки и использования специального программного инструментария для речевых корпусов данных. Одна из первых работ по проблематике организации и технологии работы с речевыми корпусами принадлежит А.А. Кольцовой, которая реализовала корпус данных «Речь» на ЭВМ БЭСМ-6 с использованием СУБД «Компас» в конце 70-х годов [1].

Переход на персональные ЭВМ и появление недорогих массовых средств для ввода-вывода аудио придало импульс работам в этой области. В 1990 году в рамках работ по использованию нейронных сетей по распознаванию речи и обработкой речи в шумах С.В. Андреевым была создана система SLIRE-3 [2], которая включала все необходимые функции по вводу-выводу, анализу, визуализации и ручной разметке речевого материала. В начале 90-х годов по заказу фирмы KayElemetrics с использованием аппаратно-программного комплекса CSL (Computerised Speech Laboratory) и программного обеспечения MultiSpeech был собран корпус данных языков народов России, а в конце 90-х годов для задач распознавания телефонной речи для нужд работ ВЦ РАН был собран и аннотирован корпус данных TeCoRus. TeCoRus содержит двухканальные записи (синхронные высококачественные микрофонные и телефонные) от 130 дикторов, размечен в TIMIT-подобном формате и до сих пор является основным средством оценки параметров акустико-фонетических моделей в телефонных каналах в ВЦ РАН; этот корпус также лицензирован в исследовательском центре фирмы Motorola. Несмотря на некоторую «непрофильность», «вынужденные» работы, связанные с разметкой данных, продолжаются в ВЦ РАН практически всё время.

Очевидно, что разметка материала должна соответствовать тому кругу задач, для которых предполагается использовать этот материал. Поскольку круг таких задач в речевой технологии достаточно широк, более того, возникают новые и новые запросы, универсального варианта разметки может и не быть или такой формат будет слишком нетехнологичным в работе.

Известные и востребованные корпуса данных периодически подвергаются реинженерии для приспособления к решению возникающих вновь задач. Например, первая большая коллекция телефонной спонтанной речи — корпус Switchboard-1, собранный Texas Instruments в 1990 г. в последующие годы подвергался реинженерии разметки различными другими коллективами как минимум шесть раз [8]. Добавлялись временные границы фраз, делалась разметка нарушений речевого потока, добавлялись тэги частей речи, разметка для дискурс-анализа, фонетические транскрипции и проводилась полная ресегментация.

Желательно иметь некоторый достаточно гибкий формат разметки, который при необходимости несложно конвертировать, редактировать, пополнять и использовать в процедурах обучения моделей и тестирования систем. Поскольку для подобных целей был создан и ныне повсеместно используется язык разметки документов XML, естественно, что разметка речевых данных может быть реализована средствами этого языка.

Цель настоящей статьи заключается в том, чтобы кратко представить накопленный опыт работы с XML-разметкой разговорной речи, привести формализацию такой разметки в виде DTD, а также описать спецификации программного обеспечения (далее, ПО) для выполнения разметки. Возможно, будет сформировано более-менее общепринятое описание документа разметки естественно диалоговой речи и использование речевых корпусов, импортированных из других организаций, не будет связано с трудоёмкой реинженерией разметки и ПО.

Разметка разговорной речи

Разметка речевого материала, применённая в TIMIT, ATIS, RuSpeech и других распространённых ресурсах, естественна и удобна в работе в тех направлениях, для которых готовились эти корпуса. Это, в основном, анализ и распознавание грамматически правильной, контролирумой речи, которая записана в хороших условиях, с хорошим качеством.

В последние годы нам пришлось столкнуться с необходимостью решения задач распознавания и анализа естественной разговорной речи. Работа с таким речевым материалом, записанным также в «естественных» условиях, подразумевает и соответствующие изменения в разметке. Поскольку используемые акустические модели вероятностные, разметка должна предоставить описания для всех, более-менее регулярно наблюдаемых акустических условий и событий. Также, поскольку эффективность работы автоматических систем в случае разговорной речи довольно низка, важно исследовать наиболее существенные факторы снижения этой эффективности и способов априорной оценки эффективности по интегральным характеристикам качества сигнала, например, таким, как полоса частот, отношение «сигнал-шум», относительное количество материала с тем или иным дефектом и т. п.

Например, в располагаемом нами разговорном материале можно было выделить следующие группы условий и событий, влияющих на результаты автоматической обработки:

- тип кодека (GSM) или метода компрессии сигнала;
- использование автоматического регулятора уровня сигнала и детектора речевой активности;
- наличие потерь пакетов с речевым сигналом;
- наличие сигналов состояния канала передачи (линии);
- акустические помехи — внешние шумы, посторонние голоса, эхо-сигналы и просачивания сигнала из смежных каналов, нелинейные искажения и крайние значения уровней сигнала;
- особенности речевого поведения говорящих, нарушения речевого потока (запинки, обрывки слов, смех, кашель, плач), заполненные паузы, неправильные акценты слов, неправильное использование телефонной гарнитуры, иноязычная речь и сленг;
- темп речи;
- чёткость артикуляции дикторов.

Разговорная речь весьма разнообразна. Например, даже в формате телефонных разговоров обмен репликами часто осуществляется между 3–5 собеседниками. Задача подготовки разметки такого материала не нова в мировой практике. В конце 1990-х годов Национальный институт стандартов США — NIST оценивал эффективность распознавания естественной речи на примере новостных передач. Для разметки тестового речевого материала был предложен универсальный формат UTF [15] на основе XML, который и был нами использован в качестве основы при создании своей собственной разметки.

Как и в UTF, автономными единицами корпуса являются разговор и его разметка. Формат UTF был изменён для более адекватного описания материала и соответствия решаемым задачам. В частности, он был пополнен сведениями о характеристиках речевого материала, влияющих на эффективность

распознавания — тип кодека, отношения «сигнал-шум», наличие и количество клиппированного сигнала и сигнала с низким уровнем, особенности речи диктора, включая нарушения речевого потока. Были структурированы помехи, которые оказались типичными, в частности, в случаях разговора в движении — звук шагов, ветер, работа радио и телевизора. Были добавлены элементы, которые описывают произносительные транскрипции, отдельные слова и звуки. В результате получилось описание формата разметки, DTD или Design Type Document, структура которого приведена ниже на **рис. 1**.

Основная единица разметки — это реплика или Turn в латинских обозначениях. Разговорная речь, даже в ограниченном формате телефонных переговоров, настолько многообразна, что иногда не сводится к очевидному чередованию реплик, во всяком случае на практике иногда было неясно, каким образом то или иное «многоголосие» представить в виде обмена репликами. Тем не менее, в основном, использование такого деления обычно не вызывало проблем при разметке.

Реплика впоследствии являлась и автономной единицей при обучении акустико-фонетических моделей.

Каждая реплика состоит из элемента «расширенного» текста, который, кроме орфографической записи фактически сказанных слов, содержит также элементы, которые обозначают нарушения речевого потока (обрывки слов, запинки) и другие акустические события. Границы реплики, а также элементов акустических событий «привязаны» (номер канала для многоканальной записи, начало и конец) к речевому сигналу. Кроме того, реплика включает отдельные элементы для описания сегментации высказывания на слова и фонемы, также привязанные к соответствующим фрагментам сигнала.

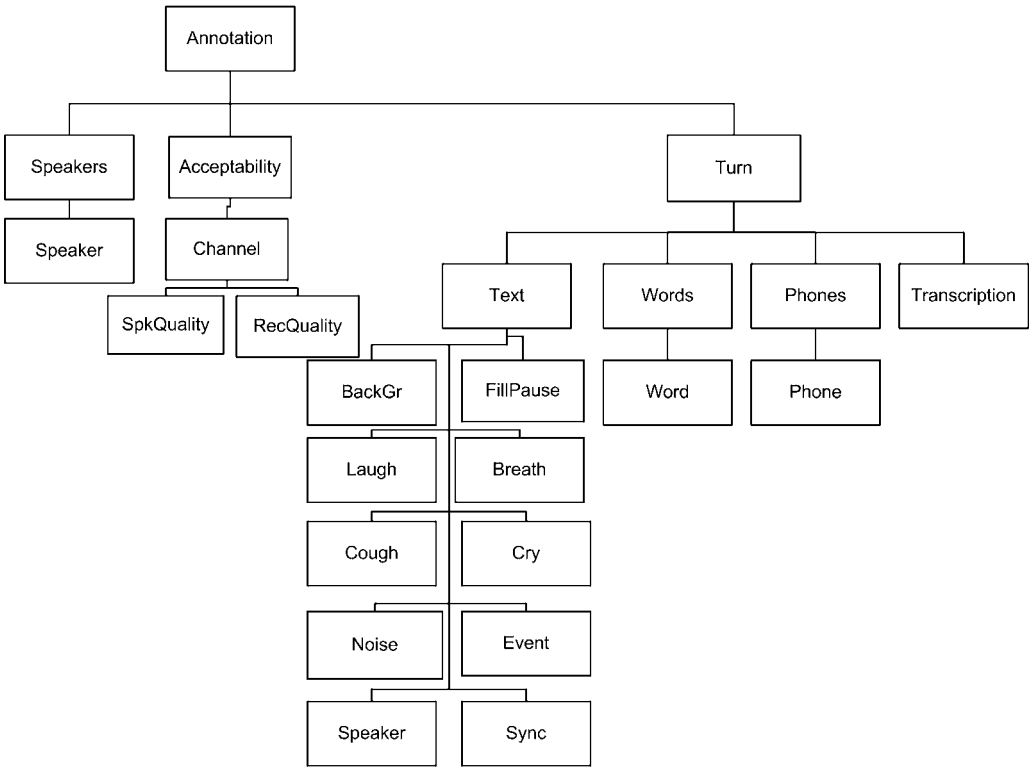

Рис. 1. Структура DTD разметки корпуса данных с разговорной речи

Формат DTD для разметки речевого материала представлен ниже, обозначения элементов и атрибутов объясняют их смысл:

```
<?xml encoding="UTF-8"?>
<!-- Общие сведения о записи -->
<!ELEMENT Annotation (Acceptability,Speakers,Topics,Section+)>
<!ATTLIST Annotation
 audio_filename     CDATA                       #IMPLIED
 version            NMTOKEN                     #IMPLIED
 version_date       CDATA                       #REQUIRED
 rate               CDATA                       #REQUIRED
 bps                CDATA                       #REQUIRED
 encoding           (psm,alaw,mulaw,g723,mp3)   #REQUIRED
 nChannels          CDATA                       #REQUIRED
 media              (wire|gsm|studio)           #REQUIRED

<!-- Интегральные показатели качества записи в каждом канале -->
<!ELEMENT Acceptability (Channel*)>
<!ATTLIST Acceptability>
<!ELEMENT Channel (RecQuality,SpkQuality,Comment) >
<!ATTLIST Channel
   id      CDATA          #IMPLIED
   name    (left|right)   #IMPLIED
>
<!-- качество сигнала (уровни, сигнал-шум, клиппирование) -->
<!ELEMENT ReqQuality EMPTY >
<!ATTLIST ReqQuality
   loc_SNR        CDATA          #IMPLIED
   magn_more_95   CDATA          #IMPLIED
   magn_less_5    CDATA          #IMPLIED
   mean_level     CDATA          #IMPLIED
   clipping_rate  CDATA          #IMPLIED
>
<!-- качество произношения для обучения, тестирования-->
<!ELEMENT SpkQuality EMPTY >
<!ATTLIST SpkQuality
   score   (trainig|test|poor)    #IMPLIED
>
<!ELEMENT Speakers (Speaker*)>
<!ATTLIST Speakers>
<!-- сведения о дикторе и особенностях его речи -->
<!ELEMENT Speaker EMPTY>
<!ATTLIST Speaker
   id             ID                              #REQUIRED
   name           NMTOKEN #IMPLIED
   gender         (male|female|child|unknown)     #REQUIRED
   dialect        (native|nonnative)              #IMPLIED
   accent         CDATA                           #IMPLIED
   age            CDATA                           #IMPLIED
   ROSpeech       CDATA                           #IMPLIED
   RODisflnce     CDATA                           #IMPLIED
>
<!-- тематика -->
<!ELEMENT Topics (Topic*)>
<!ATTLIST Topics>
   <!ELEMENT Topic EMPTY>
<!ATTLIST Topic
   id      ID             #REQUIRED
   name    NMTOKEN        #IMPLIED
>
```

```
<! – - Section – это то что относится к одной развиваемой теме – - >
<!ELEMENT Section (Turn)+ >
<!ATTLIST Section
   start   CDATA   #REQUIRED
   end     CDATA   #REQUIRED
   topic   CDATA   #IMPLIED
>
<!–- реплика -->
<!ELEMENT Turn (Text|Words|Transcription|Phonemes|Comment)*>
<!ATTLIST Turn
    speaker       IDREF       #REQUIRED
    start         CDATA       #REQUIRED
    end           CDATA       #REQUIRED
    channel       CDATA       #IMPLIED
>
<!–- текст реплики с акустическими событиями -->
<!ELEMENT Text (#PCDATA|FillPause|Laugh|Cough|Breath|BackGr|Speaker|Cry|Sync|Ev
ent)>
<!ATTLIST Text>
<!ELEMENT BackGr EMPTY>
<!ATTLIST BackGr
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!–- заполн.пауза, содержимое равно на слух «тексту» паузы -->
<!ELEMENT FillPause (#PCDATA)>
<!ATTLIST FillPause
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!ELEMENT Laugh EMPTY>
<!ATTLIST Laugh
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!ELEMENT Cough EMPTY>
<!ATTLIST Cough
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!ELEMENT Breath EMPTY>
<!ATTLIST Breath
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!ELEMENT Speaker EMPTY>
<!ATTLIST Speaker
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!ELEMENT Cry EMPTY>
<!ATTLIST Cry
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>

<!ELEMENT Sync EMPTY>
<!ATTLIST Sync
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
```

```
<!ELEMENT Event (#PCDATA)>
<!ATTLIST Event
   start          CDATA          #REQUIRED
   end            CDATA          #REQUIRED
>
<!-- список слов -->
<!ELEMENT Words (Word*)>
<!ATTLIST Words>
<!ELEMENT Word (#PCDATA)>
<!ATTLIST Word
   start           CDATA         #IMPLIED
   end             CDATA         #IMPLIED
>
<!-- транскрипция -->
<!ELEMENT Transcription (#PCDATA)>
<!ATTLIST Transcription>
<!-- список фонем -->
<!ELEMENT Phonemes (Phone*)>
<!ELEMENT Phone (#PCDATA))>
<!ATTLIST Phone
   start          CDATA          #IMPLIED
   end            CDATA          #IMPLIED
>
<!ELEMENT Comment EMPTY>
<!ATTLIST Comment
   desc           CDATA          #REQUIRED
>
```

Как пример ниже приведён фрагмент текста аннотации на XML, соответствующее графическое представление дано на **рис. 3 и 4**:

```
<Annotation audio_filename="9.wav" version="0.0" date="20:59:53
23.03.2010" rate="8000" bps="16000" encoding="psm" encoding_orig="g732"
nChannels="2" media="gsm">
  <Acceptability>
    <Channel id="1" name="left">
      <RecQuality loc_SNR=" " magn_more_95=" " magn_less_5=" " mean_
level=" " mean_magn=" "/>
      <SpkQuality score="training"/>
      <Comment/>
    </Channel>
    <Channel id="2" name="right">
      <RecQuality loc_SNR=" " magn_more_95=" " magn_less_5="
" mean_level=" " mean_magn=" "/>
      <SpkQuality score="training"/>
      <Comment/>
    </Channel>
  </Acceptability>
  <Speakers>
    <Speaker Id="1" Gender="male"/>
    <Speaker Id="2" Gender="female"/>
  </Speakers>
  <Turn Speaker="1" start="83.898" end="85.975">
    <Text>
      <Breath start="83.898" end="85.975"/>
    </Text>
  </Turn>
  <Turn Speaker="2" start="84.842" end="88.053">
    <Text>
      Отнимется скоро. У тебя какая, левая болит?
    </Text>
    <Words>
      <Word start="84.848" end="85.064">.</Word>
```

```
        <Word start="85.076" end="85.760">отнимется</Word>
        <Word start="85.772" end="86.132">скоро</Word>
        <Word start="86.144" end="86.252">у</Word>
        <Word start="86.264" end="86.408">тебя</Word>
        <Word start="86.420" end="86.768">какая</Word>
        <Word start="86.780" end="87.080">левая</Word>
        <Word start="87.092" end="87.536">болит</Word>
        <Word start="87.548" end="88.004">.</Word>
      </Words>
    </Turn>
    <Turn Speaker="1" start="88.399" end="89.626">
      <Text>
        Да
      </Text>
      <Phones>
        <Phone start="88.405" end="88.717">sil</Phone>
        <Phone start="88.717" end="88.753">d</Phone>
        <Phone start="88.753" end="89.113">a^</Phone>
        <Phone start="89.113" end="89.575">sil</Phone>
      </Phones>
      <Words>
        <Word start="88.405" end="88.705" scoreAv="-16713">.</Word>
        <Word start="88.717" end="89.101" scoreAv="-22118">да</Word>
        <Word start="89.113" end="89.581" scoreAv="-23811">.</Word>
      </Words>
    </Turn>
    <Turn Speaker="2" start="89.463" end="90.893">
      <Text>
         А у меня правая, блин
      </Text>
    </Turn>
    <Turn Speaker="1" start="90.667" end="92.890">
      <Text>
        Ну будем вдвоём на пару ходить <Unclear start="92.324" end="92.763"/>
      </Text>
    </Turn>
  </Annotation>
```

Программное обеспечение для разметки речевого материала

Разметку речевого материала в приведённом формате можно выполнить разными способами, например, работая одновременно с текстовым и аудиоредакторами (для расстановки меток синхронизации по времени). При больших объёмах материала наиболее эффективный способ — использование специального программного обеспечения, которое оптимизировано для эффективного выполнения разметки. Существует достаточно много удобных и мощных инструментальных средств, в том числе платформо-независимых и с открытым кодом для проведения разметки речевых корпусов данных. Тем не менее желание провести процедуру разметки с наибольшей эффективностью привело к созданию собственного ПО для разметки разговорной речи (в данном случае — телефонных переговоров) в формате на основе XML. Российских программных средств для работы с разметкой на XML не было, западное ПО потенциально казалось уязвимо, например, из-за кодировок кириллицы. Кроме того, к ПО портируются уже существующие процедуры автоматического распознавания, сегментации и скоринга и хочется иметь полный контроль над программным обеспечением.

В качестве языка программирования используется скриптовый язык Tcl/Tk [16], визуализация и некоторые операции с речевым сигналом выполняются с помощью пакета Snack [16],

XML документы представляются с использованием объектной модели документа — DOM, которая реализована пакетом TDOM [17]. Более сложная обработка и распознавание осуществляются отдельными библиотеками, написанными на языке C++. Мы используем Tcl/Tk с 2000 года, в основном для создания прототипов ПО, этот язык оказался очень удобен как язык сборки системы из отдельных компонент и написания собственных сценариев работы. Tcl/Tk ранее использовался при создании ПО разметки Transcriber [6]. Помимо открытости кода дополнительное удобство Tcl/Tk заключается в его кроссплатформенности.

Основные функции ПО разметки состоят в удобном для восприятия представлении XML-документа-разметки речевого материала, визуализации речевого сигнала и его параметрического описания, визуализации элементов разметки, возможности интерактивной установки и редактирования элементов и их содержания, прослушивания содержания, проведения части операций (например, сегментации) в автоматическом режиме с возможностью ручной коррекции результатов.

Проблема с восприятием XML разметки в её естественном, текстовом виде заключается в том, что текстовое представление «одномерно», лишено наглядности, его неудобно редактировать, приходится набирать много служебной информации, относящейся к форматам элементов и атрибутов XML, по тексту трудно проверить, что атрибуты синхронизации по времени правильные и т.п. Всё это удобнее сделать, если доступно графическое отображение речевого сигнала, которое непосредственно синхронизировано с разметкой. Эта функция реализуются с помощью простого графического интерфейса, схематически, вместе с блоками обработки, представленного на следующем **рис. 2**.

ПО читает XML-разметку из файла (либо создаёт новый документ), строит дерево DOM, после чего отображает содержимое DOM в нескольких окнах: гипертекстовое окно (**рис. 3**) отображает текст разговора с графическим представлением тегов (элементов) разметки, которые идентифицируют

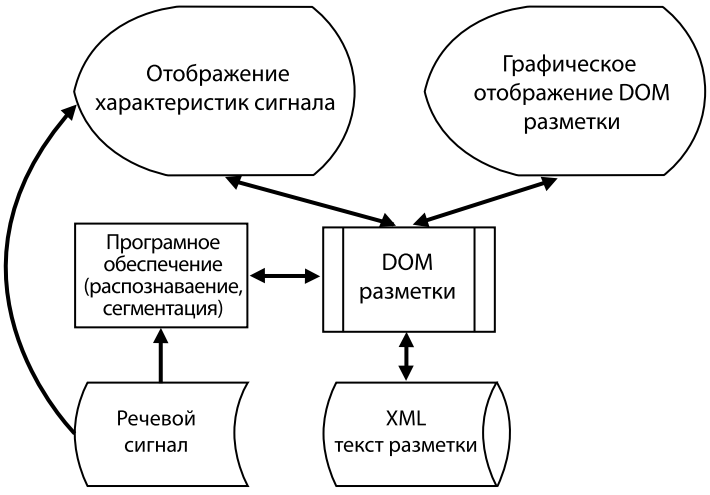

Рис. 2. Организация обработки и отображения документа аннотации

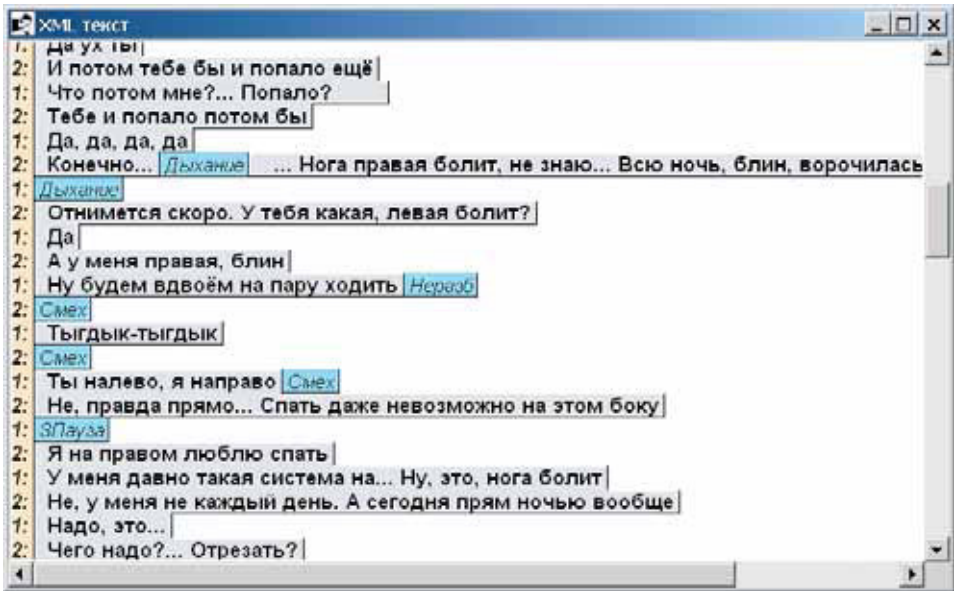

Рис. 3. Гипертекстовое окно для отображения документа разметки

того, кто произносит реплику, канал, а также тэги различных акустических событий и точки синхронизации. Цвет элементов показывает оператору, какие элементы не содержат полную информацию, например, не синхронизированы с сигналом. В графическом окне (**рис. 3**) отображены осциллограмма и сонограмма выбранного участка сигнала, синхронизированный с сигналом текст, границы и обозначения слов, звуков, акустических событий, а также маркеры синхронизации.

Конфигурационный файл определяет, какие элементы разметки отображаются, в каком из окон и в каком именно виде. Можно отображать элементы XML документа по имени, содержанию, атрибутам или по синонимам, что иногда удобно для наглядности, если,

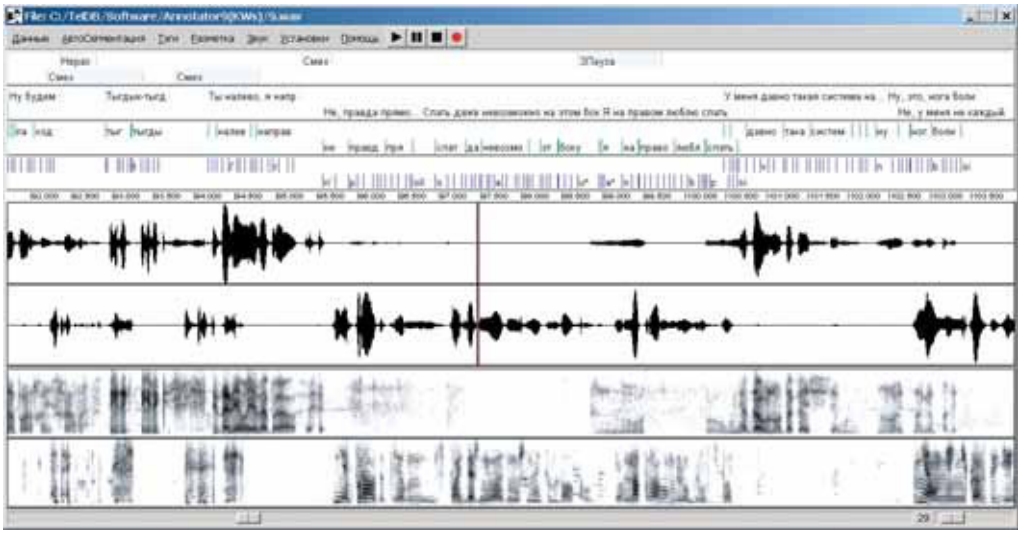

Рис. 4. Графическое окно документа разметки

например, заменить латинские обозначения тегов на их русскоязычные эквиваленты. Отображаются только те элементы разметки, которые нужны оператору (которые разрешаются конфигурационным файлом).

Программа позволяет выделять фрагмент текста и затем создавать элемент разметки с содержанием, равным выделенному тексту, ставить синхронизирующие текст и звук маркеры, интерактивно создавать все элементы, которые описаны в представленном выше DTD. Кроме того, можно в качестве разметки первоначально определять так называемый «расширенный» текст разговоров. Расширенный текстовый формат — это упрощённый формат записи разметки, который содержит почти все элементы разметки в более удобной для ввода с клавиатуры записи, из которой удалены атрибуты времени.

Помимо основной функции создания, представления и коррекции документа разметки, ПО выполняет дополнительные операции, которые существенно облегчают последующее использование разметки для обучения моделей и проверки эффективности. Такими дополнительными функциями являются транскрибирование и сегментирование указанных элементов разметки и создание скриптов для пакетного режима обучения. В последнем случае результатом операции является набор команд и файлов, который служит входным потоком для процедуры оценки параметров акустико-фонетических моделей звуков.

Заключение

В работе представлен формат для разметки разговорной речи, а также описано программное обеспечение с открытым кодом для создания, визуализации и коррекции разметки. Формат разметки и ПО был использован при создании разметки корпуса естественной разговорной русской речи.

Литература

1. *Емельянова Л.А., Кольцова А.А.* Организация базы данных проблемно-ориентированных диалоговых систем. М.: ВЦ АН СССР, 1980.

2. *Andreev S., Chuchupal V.* Workstation for Speech Analysis. Proc. of XIII Phonetical Congress, Aix En Provance, France, 1991.

3. *Арлазаров В.В, Богданов Д.С., Брухтий А.В., Подрабинович А.Я.* Программное обеспечение для формирования баз данных//Организаионное управление и искусственный интеллект: Сб. трудов Института системного анализа РАН, 2003, http://www.cognitive.ru/innovation/sbornic4.

4. *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных// Там же.

5. *Cassidy S., Harrington J.* Multi-level annotation in the Emu speech database management system. Speech Communication, vol. 33, pp. 61–77, 2001.

6. *McKelvie D., Isard A., Mengel A., Moller B.M., Grosse M., Klein M.* The MATE workbench — An annotation tool for XML coded speech corpora//Speech Communication, Vol. 33, Pp 97–112, 2001.

7. *Cassidy S., Bird S.*, Querying Databases of Annotated Speech//Proceedings of the 11 Australasian Database Conference, pp. 12–20, 2000.

8. *Graff D., Bird S.* Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies.

9. DARPA, Defence Advanced Research Program Adency. Сайт: http://www.darpa.gov/

10. Evaluations and Language resources Distribution Agency, ELDA. Сайт: http://www.elda.org/

11. ELRA, European Language Resource Association. Сайт: http://www.elra.info/

12. LDC Lingustic Data Consortium. Сайт: http://www.ldc.upenn.edu/

13. ООО AUDITECH. Сайт: http://www.auditech.ru/

14. Институт системного анализа РАН, ИСА РАН. Сайт: http://www.isa.ru/

15. 1998 DARPA/NIST Continuous Speech Recognition Broadcast News Hub-4 English Benchmark Test. Сайт: http://www.itl.nist.gov/iad/mig//tests/bnr/1999/

16. Tcl Developer Exchange, сайт: http://www.tcl.tk/The Snack Sound Toolkit. Сайт: http://www.speech.kth.se/snack/

17. tDOM. Сайт: http://www.tdom.org/

Чучупал Владимир Яковлевич —
закончил МГПИ им. В.И. Ленина в 1976 году по специальности «математика», в 1983 году закончил очную аспирантуру ВЦ АН СССР, с тех пор работает в ВЦ РАН. Основная область интересов: распознавание и обработка речевых сигналов. Кандидат физико-математических наук. Ведущий научный сотрудник Вычислительного центра им. А.А. Дородницына РАН.
Email: chuchu@ccas.ru