СВОБОДНЫЕ РЕЧЕВЫЕ БАЗЫ ДАННЫХ VOXFORGE.ORG
VOXFORGE.ORG FREE SPEECH CORPUS

Шмырёв Н.В. *(nshmyrev@yandex.ru)*
НИИСИ РАН

В докладе обсуждаются проблемы, связанные с созданием свободных баз для систем синтеза и распознавания речи. Будут рассмотрены источники свободной речи, способы обработки информации и возникающие проблемы.

В связи с развитием устройств хранения и коммуникации современное оборудование позволяет накапливать и обрабатывать огромные массивы данных. Базы речи применяются при построении систем синтеза и распознавания речи, для оценки различных методик при тестировании приложений. Большинство современных баз собрано вручную, значительные ресурсы затрачены на их создание. Остро стоит вопрос автоматизации процесса сбора и обработки данных, вовлечения носителей языка в процесс записи. С одной стороны, значительные исследования посвящены попыткам организовать обработку с минимальным вмешательством исследователя, разработаны многие алгоритмы, позволяющие обойтись только незначительной предварительной обработкой. Например, анализ текстов [1] позволяет составить морфологическую модель языка только на основе словаря, без какой-либо дополнительной информации. С успехом подобные методы применяются и в области распознавания изображений. К сожалению, полностью автоматизировать процесс не всегда возможно, часто вмешательство человека все же требуется. Более того, на наш взгляд, невозможно собрать и обработать значительный объем речевой информации без привлечения большого числа пользователей. Наша задача – заинтересовать носителя языка, вовлечь его в процесс сбора данных, использовать его опыт и время при создании и проверке записей. Поэтому, актуальной является проблема организации процесса сбора данных, рассмотренная в данном докладе.

В этом докладе мы опишем методы, используемые для сбора речевых баз данных для систем распознавания в рамках проекта VoxForge.org [2], расскажем о текущих проблемах. Проект VoxForge.org посвящен сбору речевых данных для использования в системах распознавания речи. Мы собираем речевую базу на нескольких языках - английском, русском, немецком, итальянском, голландском. База распространяется свободно по лицензии GPL и содержит записи, разбитые на небольшие речевые отрезки в оригинальном формате записи и транскрипцию. Некоторые части содержат дополнительную разметку, например, разметку интонации, точную сегментацию и так далее, но это скорее исключение. Для каждого диктора в базе сохраняется возраст, пол. Диктор указывает свой диалект, которому, к сожалению, не стоит доверять. Объем собранной и обработанной речи для наиболее активных языков: английский – 40 часов речи более ста дикторов, русский – 10 часов речи более двухсот дикторов, немецкий и голландский языки – по 10 часов речи. В перспективе мы надеемся собрать значительно больший объем данных – до 140 речи часов для каждого языка. Скорость наполнения базы значительна – база английской речи пополняется примерно на 5 часов в месяц, более того, мы полагаем, что скорость пополнения будет расти. Растет и число поддерживаемых языков, в апреле 2007 года база содержала записи только на английском, в 2008 году – уже на 8 языках.

Проект распространяет фонетические словари, приложения для обработки данных, разметку речи на речевые сегменты, акустические модели для систем распознавания CMU Sphinx, HTK и Julius. Для Julius английская акустическая модель VoxForge.org является основной. Акустическая базы служат не только для создания систем распознавания, например, база русской речи используется в русском голосе msu_ru_nsh_clunits для синтезатора Festival [3]. Подобные проекты развиваются и в других областях, например, стоит отметить проект Freesound [4], посвященный сбору коллекции звуков.

Свободно распространяемая база имеет ряд преимуществ. Наиболее ценна она для свободных приложений. К сожалению, выбор свободных приложений в области речевых технологий невелик. Несмотря на наличие нескольких пакетов распознавания речи, в настоящий момент отсутствуют реализации некоторых необходимых компонентов речевых интерфейсов, таких как система управления диалогом. Мы надеемся, что наша работа стимулирует развитие свободных приложений. Свободная лицензия позволяет нам также снять и технические ограничения. Распределенное хранилище позволяет неограниченно расширять объем базы, не заботясь ни о сохранности данных, ни о производительности системы.

К счастью необходимо отметить, что для некоторых наиболее популярных языков источников речевых данных достаточно. Основной объем данных, собранных на ресурсе был прислан от обычных посетителей ресур-

са, но, в последнее время, появляются и другие источники данных. Посетитель ресурса может прислать свою запись следующими способами: по обычному или IP-телефону, записав данные на домашнем компьютере и прислав запись, посетив сайт и записав речь прямо с него. Последняя возможность наиболее важна для нас, как показывает практика запись с компьютера и телефона сложна для посетителей ресурса. Необходимость установки и настройки приложений отпугивает многих посетителей. Возможность записи речи прямо из web-браузера значительно увеличивает вклад посетителей в создание базы. Java-апплет не требует установки и настройки и позволяет прослушать записанную речь и загрузить речь в исходном формате на сервер.

При записи посетителю предлагается произнести текст, состоящий из случайно выбранных предложений из большого текстового корпуса, составленного из свободных текстов. Запись пользователя проверяется с помощью уже существующей речевой модели языка и, если проверка прошла успешно, сохраняется в базу.

По собранной базе периодически рассчитываются и обновляются акустические модели. Подготовка оптимизированных речевых моделей – сложный многоступенчатый процесс. На данный момент используется пакет HTK для расчета моделей для Julius и пакет SphinxTrain для расчета моделей для семейства приложений CMU Sphinx. В настоящий момент модели не оптимизируются, используются параметры расчета модели по умолчанию. Качество их, тем не менее удовлетворительное. Например, точность распознавания модели русского языка со словарем в 30 тысяч слов – порядка 70%. К сожалению, частота обновления ограничена скоростью расчета модели.

Нами используются и другие источники информации, которых последнее время становится все больше. Например, мы сотрудничаем с проектом по записи аудиокниг Librivox, поставляющем нам книги в оригинальном несжатом формате. Поступают предложения использовать фильмы с субтитрами, возможно даже использование записей официальных переговоров. Власти города Мизула в США в рамках программы OpenGoverment позволяют использовать запись обсуждений в городском совете. Эти записи уже содержат транскрипцию и могут быть использованы без значительных затруднений.

К сожалению, без дополнительного стимулирования не всегда возможно собрать требуемый объем речи. Мы используем и различные призы для участников, например, компания Voice2type предлагает приз наиболее активному пользователю нашего ресурса, приславшему запись речи с мобильного телефона. Тем не менее, проблема стимулирования остается открытой. Одной из самых интересных работ в направлении организации сбора данных и проведения вычислений с использованием человека является работа [5], рассматривающая использование человеческих ресурсов в сети. Остроумное использование web-технологий и соревнования между участниками процесса позволяет эффективно организовать распределенное вычисление, задействовать ресурсы пользователей и обработать огромный объем информации для дальнейших исследований в области распознавания изображений. Схема, позволяющая заинтересовать пользователя в записи, пока еще не разработана. В разработке находится модуль, позволяющий использовать собственную запись для адаптации акустической модели для пользователя, при этом будет необходимо записать некоторый небольшой текст. Такие записи мы будем использовать для дальнейшего улучшения общей модели. Нужно надеяться, таким образом удастся собрать действительно значительный объем речевых данных.

Другой, не менее важной проблемой является унификация и обработка собранных данных. Остановимся на проблеме первичной обработки данных. Тут возникают несколько проблем. Во-первых, формат данных не всегда соответствует текущим требованиям систем расчета моделей. Часто звук закодирован с потерей данных, например, в формат ogg или mp3. Возникает вопрос, возможно ли использовать такие данные для системы распознавания? К сожалению, внятных ответов на него пока не получено. Известно, что звук, декодированный из mp3 отрицательно сказывается на качестве акустической модели, в тоже время некоторые исследователи в устной беседе утверждают, что с помощью моделей на основе речи, сжатой в формате mp3 можно отлично декодировать mp3 речь. Возможно ли смешивать данные с различной степенью сжатия, и как это сказывается на качестве распознавания, пока неизвестно. Нужно надеяться, современные методы извлечения параметров речи при расчете моделей позволят решить эту проблему.

Во-вторых, интересным направлением исследования является проблема выделения речевых отрезков и проверки большой записи, например, разбиения аудиокниги на небольшие куски, пригодные для расчета параметров акустической модели. Задача разбиения большого звукового файла имеет и самостоятельное значение. Она находит применение в системах обучения языку, где временные метки используются приложением для отображения синхронного перевода. Необходимо заметить, что в последнее время значительное количество работ позволило добиться успеха в этой области. Например, система [6] позволяет обработать аудиокнигу и получить базу речевых отрезков и транскрипцию. В настоящее время мы применяем разбиение с помощью выравнивания по существующий модели и тексту. Наиболее сложной проблемой является наполнение словаря неизвестными словами из текста. Для наполнения словаря используется система автоматической транскрипции, обученная по

существующим словарям. Наиболее точные системы, основанные на мультиграммных поточных моделях [7] обеспечивают точность транскрипции порядка 70%. Остальные 30% приходится корректировать вручную.

К сожалению, присутствуют и более серьезные проблемы, порожденные самим способом сбора данных. К сожалению, несмотря на значительные объемы хранимой информации, тяжело получить хорошо сбалансированную базу данных. Объем данных из некоторых источников доходит до 10 часов, часто это речь всего одного диктора. Конечно, таким базам тоже можно найти применение. Например, их можно использовать для создания систем высококачественного синтезатора речи. Не ясно, насколько такая несбалансированная база будет полезна в исследованиях по распознаванию речи. Мы надеемся, что возможно будет выделить некоторую сбалансированную часть базы.

Стилевое наполнение базы тоже не оптимально. Для некоторых задач, например, задачи поиска в звуковых файлах или в задаче транскрибирования видео более полезны записи повествовательной речи. Важная задача создания речевого интерфейса требует базы данных совсем другого стиля, базы, составленной в основном из диалогов и спонтанной речи. Увеличить наполнение диалоговой составляющей базы – задача на ближайшее будущее.

Мы описали работу, которая ведется в направлении создания системы распознавания речи на основе свободных речевых данных. Нужно признать, значительные аспекты еще не проработаны. Не исследованы проблемы построения акустической модели нескольких языков, создания оптимального текста и условий для записи. К счастью, круг проблем еще очень широк.

Список литературы

1. Mathias Creutz, Krista Lagus. Unsupervised Models for Morpheme Segmentation and Morphology Learning // ACM Transactions on Speech and Language Processing, Volume 4, Issue 1, Article 3, January 2007.

2. VoxForge project // http://voxforge.org

3. Русский голос для Festival // http://festlang.berlios.de/russian.html

4. FreeSound project // http://freesound.iua.upf.edu

5. Luis von Ahn, Laura Dabbish, Labeling images with a computer game. // Proceedings of the SIGCHI conference on Human factors in computing systems, 2004.

6. Prahallad K., Toth A., Black A. Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases // Interspeech 2007, Antwerp, Belgium.

7. Bisany M., Ney H. Joint-sequence models for grapheme-to-phoneme conversion. // Speech Communication,Volume 50, Issue 5, May 2008.