ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

Мартынова Александра Сергеевна

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ В БЕСЕРМЯНСКОМ ДИАЛЕКТЕ УДМУРТСКОГО ЯЗЫКА

Выпускная квалификационная работа
студента 4 курса бакалавриата группы БКЛ131

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю. А. Ландер
_____
«___» _____ 2017 г.

Научный руководитель
канд. филологических наук, доц.
Т. А. Архангельский
_____
«___» _____ 2017 г.

Москва 2017

Оглавление

Автоматическое распознавание речи в бесермянском диалекте удмуртского языка

Александра Мартынова

1. Введение

Речь является основным средством общения между людьми. Она несет в себе одновременно очень большое количество данных. Это как лингвистические данные (например, сообщение и язык говорящего), так и динамические (эмоции, региональные отличия говорящего и физиологические характеристики его речевого аппарата). Несмотря на то, что такая информация кодируется в сложной форме, люди могут относительно легко расшифровать её большую часть. Эта человеческая способность вдохновила многих исследователей на разработку систем, которые бы смогли приблизиться к возможностям человека в этой области. Среди всех речевых задач автоматическое распознавание речи (ASR) было в центре внимания исследователей на протяжении нескольких десятилетий.

Автоматическое распознавание речи — это одно из направлений в области разработки искусственного интеллекта, которое позволяет автоматически преобразовывать речь в текст. Главная проблема распознавания речи — сложность естественного языка. Во время речевой коммуникации слушающий использует не только уши, но и знания о говорящем и его окружении. Однако, в области автоматического распознавания речи у исследователей и разработчиков есть только речевой сигнал.

Если раньше распознавание речи было чем-то невероятным из области фантастики, то сейчас оно начинает широко распространятся в повседневной жизни людей. Поиск в интернете, голосовые помощники, голосовое управление интерфейсом — все эти функции доступны для обычного пользователя мобильного телефона. Однако ASR используется не только для облегчения жизни пользователей. Распознавание речи так же является инструментом для сокращения ручного труда. Например, многие крупные компании, имеющие колл-центр, начинают заменять людей, занимающихся обработкой обращений по голосовым каналам связи, на машины. Но системы распознавания речи могут облегчить работу не только в бизнесе, но и в науке.

Бесермяне — это малочисленный народ России, проживающий на северо-западе Удмуртии. По переписи на 2002 год, численность бесермян составляет 3100 человек. Бесермянский можно назвать наиболее полно описанным диалектом удмуртского языка. На протяжении 10 лет группы лингвистов проводят полевые исследования этого идиома. Одной из главных задач этих поездок является создание корпуса языка. Бесермянский — бесписьменный

язык, поэтому корпус основан на расшифровках устных текстов. Поскольку в Москве, где в основное время находятся исследователи этого идиома, нет носителей языка, все тексты необходимо расшифровывать совместно с носителями непосредственно во время экспедиции. Эта рутинная работа занимает очень много времени и сил, а самое главное — записанных аудио больше, чем расшифровок, так как записи появляются намного быстрее, чем экспедиционеры успевают расшифровывать тексты. Данная проблема является насущной не только для исследователей бесермянского языка, но и для всех ученых, занимающихся бесписьменными языками.

Цель данной работы — обзор существующих методов для создания систем распознавания речи и создание такой системы для бесермянсокого диалекта удмуртского языка.

2. История развития распознавания речи

Первые попытки создания системы для автоматического распознавания речи были предприняты ещё в 1950-х годах, когда многие исследователи пытались использовать для этого идеи акустической фонетики.

В 1952 году К. Дейвис, Р. Биддульф и С. Балашек в Bell Laboratories разработали систему «Audrey», которая была способна распознавать только цифры (Davis et al. 1952).

Первое время системы использовали для обучения так называемую «тренировку», во время которой говорящий начитывал текст или изолированные слова для системы. Система в свою очередь анализировала конкретный голос и использовала его для точной настройки распознавания речи этого человека, что сильно повышало точность распознавания. Такие системы называются дикторозависимыми.

Алгоритм работы первой системы распознавания речи кажется сейчас очень простым и примитивным, но в то время это был большой прорыв. Говорящий произносил выбранную цифру или несколько цифр, убедившись, что пауза между ними не меньше 350 миллисекунд. Дальше система, основываясь на измерениях спектральных резонансов в области гласных каждой цифры, сравнивала полученные данные с образцами, которые были сделаны заранее и хранились в аналоговой памяти. Важно отметить, что «Audrey» работала с высокой точностью (97%) только при изолированном произнесении цифр одним диктором.

В 1956 году в RCA Laboratories Олсон и Белар разработали систему, которая распознавала 10 слогов. Система также была дикторозависимая и основывалась на спектральном анализе (Olson & Belar 1956).

В 1959 году Фрай и Денес попытались создать систему распознавания фонем, чтобы декодировать четыре гласных и девять согласных (Fry 1959). Они использовали спектральный анализ и сравнение шаблонов. Нововведением в этой работе было использование статистической информации о допустимых последовательностях фонем в английском языке. Такая «фонемная модель» позволила улучшить общую точность распознавания слов, состоящих из двух и более фонем.

В 1960-е были начаты три ключевых исследовательских проекта, которые сильно повлияли на развитие распознавания речи в течение следующих двадцати лет. Первым из них была работа Т. Мартина и его коллег из RCA Laboratories. Они занимались решением проблем, связанных с неравномерностью временных отрезков в речевых сообщениях. Проще говоря, было неясно, как сравнивать шаблон и речевой отрезок, если они одинаковы по содержанию, но разные по длительности. Т. Мартин разработал набор простых методов нормализации речевых сообщений, основанный на способности точно обнаруживать начало и конец высказывания, что значительно уменьшило вариативность оценок распознавания (Martin et al. 1964). Примерно в то же время в Советском Союзе Т. Винцюк предложил использовать методы динамического программирования для выравнивания времени пары речевых высказываний (Винцюк 1968). Несмотря на то, что алгоритмы динамического программирования для распознавания слитной речи были реализованы в работе Т. Винцюка, его работа была неизвестна на западе вплоть до начала 1980-я годов. Это уже было после того, как другие исследователи предложили и реализовали подобные методы.

Заключительным достижением в 1960-х годах стало исследование Д. Редди в области распознавания слитной речи (т. е. не изолированные слова) путем динамического определения фонем (Reddy 1966). Исследования Д. Редди в конечном счете подарили жизнь весьма успешной лаборатории распознавания речи в Университете Карнеги-Меллона, которая до сих пор остается лидером в разработке систем распознавания речи.

В 1970-х годах область распознавания отдельных слов или дискретных высказываний стала жизнеспособной и практичной технологией, основанной на фундаментальных исследованиях В. Величко и Н. Загоруйко в России (Velichko & Zagoruyko 1970), Х. Сакоэ и С. Чиба в Японии (Sakoe & Chiba 1978) и Итакура в США (Itakura 1975). Исследования в России помогли применить идеи распознавания образов в распознавании речи; Японские исследователи показали, как можно успешно применять методы динамического программирования; а исследование Ф. Итакура, как идеи линейного предсказательного кодирования (LPC) могут приняться к системам распознавания речи.

До 70-х годов 20 века хорошие результаты показывали только системы с маленьким словарем (цифры, отдельные фонемы, ограниченное количество слов). Чем больше становился словарь, тем сильнее падала точность распознавания. В это время в IBM началась работа в области распознавания речи с большим словарем.

В 1980-е в центре внимания была проблема распознавания слитной речи. Перед исследователями стояла задача создать надежную систему, способную распознавать слитное речевое высказывание на основе его сопоставления с конкатенированным шаблоном отдельных слов. В то же время, речевые исследования в 1980-х можно охарактеризовать переходом от шаблонных технологий к методам статистического моделирования. Самым популярным среди них была скрытая марковская модель (HMM). К тому времени она уже активно применялась в нескольких лабораториях (прежде всего IBM, Institute for Defence Analyzes и Dragon Systems), но только после широкой публикации теории HMM (Ferguson 1980), этот

метод стал применяться практически в каждой исследовательской лаборатории распознавания речи в мире.

Ещё одной «новой» технологией, которая была повторно представлена в конце 1980-х, была идея применения искусственных нейронных сетей к проблемам распознавания речи. Впервые нейронные сети были представлены в 1950-х годах, но изначально они оказались бесполезными для задач распознавания речи, поскольку в то время ещё не было достаточно мощных компьютеров, на которых было бы возможно построить нейронную сеть, и у самой модели было много практических проблем. Однако в 80-е годы слабые и сильные стороны данной технологии были глубже изучены, а так же её связь с классическими методами классификации сигналов (Lippmann 1987; Waibel et al. 1989)

Использование Скрытой Марковской модели и искусственных нейронных сетей позволило значительно улучшить качество систем распознавания речи. Нынешние системы способны распознавать слитную речь с большим словарных запасом. Единственной большой проблемой оставались вычислительные мощности. Эта проблема решилась после одного важного события: появления приложения Google Voice Search для iPhone. У Google была возможность разгрузить процесс распознавания, используя свои облачные дата-центры, направив всю их мощь для крупномасштабного анализа данных, а именно, для поиска совпадений между словами пользователей и огромного числа образцов голосовых запросов, которые они получали.

3. Архитектура систем распознавания речи

Статистические методы для распознавания естественной слитной речи были созданы больше тридцати лет назад. Наиболее популярными статистическими методами в этой области являются Скрытые Марковские модели для акустической модели и N-граммы для языковой модели.

И акустическая, и лингвистическая модели являются важными частями современной системы распознавания речи.

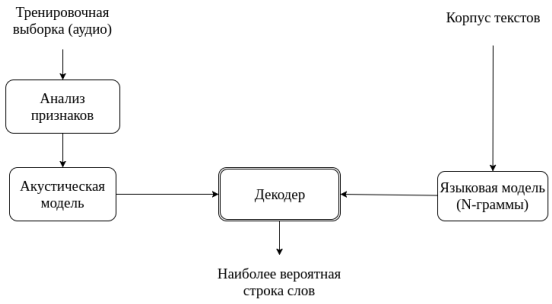

Рис. 1. Архитектура статистической системы анализа речи

Цель статистической системы распознавания речи — декодировать наиболее вероятную последовательность слов из полученного голосового сигнала. При распознавание речи

это эквивалентно распознаванию последовательности слов. Формально, мы ищем наиболее вероятную последовательность слов, учитывая акустические наблюдения.

3.1. Анализ речевого сигнала

Первым шагом в любом системе распознавания речи будет извлечение спектрально-акустических признаков. Другими словами, цель анализа речевого сигнала — это получение вектора признаков: для одинаковых фонем векторы будут максимально приближены друг к другу, а для разных фонем они будут максимально отличаться друг от друга.

Основные факторы, которые могут привести к сильной разнице между двумя речевыми сигналами с одинаковым содержанием:

- различия в произношении говорящего, его или её поле, высоте голоса и т.п.
- микрофон и другие свойства канала передачи звука
- фоновый шум, акустика помещения и т.п.

Обычные методы обработки сигналов, используемые при автоматическом распознавании речи, основаны на распределении Мел-частотных кепстральных коэффициентов (Mel Frequency Cepstral Coefficient — MFCC) и перцептуальных коэффициентах линейного предсказания (Perceptual Linear Prediction — PLP) (Suba & Bharathi 2014).

Оба этих распределения представляют собой вектор коэффициентов низкой размерности, который описывает спектр речи на небольшом временном интервале, как правило от 20 до 50 миллисекунд. Соответственно, из каждой аудиозаписи можно выделить множество подобных векторов, а их распределение моделирует особенности речи того или иного диктора.

Для разработки ASR для бесермянского диалекта удмуртского языка я буду использовать MFCC.

3.2. Акустическая модель

Акустическая модель — это сердце системы распознавания речи. АМ представляет описание последовательности векторов из акустического анализа с учетом последовательности слов. Это значит, что у акустической модели нет части информации, необходимой для её обучения, потому что соответствующая звуку текстовая транскрипция не синхронизирована с ним. Получение информации о времени начала и конца слова во фразе делает обучение акустической модели в разы сложнее. Современные средства распознавания речи для моделирования неопределенности между акустическими характеристиками и соответствующей им транскрипции используют Скрытую Марковскую модель.

Важно отметить, что из-за ограниченности данных, модель для отдельных слов а так же модель для целых предложений получается путем конкатенации акустических моделей подсловных единиц (отдельных фонем или, например, трифонов). Подсловные единицы, меньшие чем слово, позволяют распознавать слова, которые не встретились в тренировочной выборке.

Тип подсловных единиц, используемых в распознавании речи, зависит напрямую от количества данных в тренировочной выборке и сложности модели. Для системы с маленьким

словарем, например, для распознавания цифр, обычно применяются модели целых слов, а системы с большим словарем часто используют небольшие подсловные единицы, которые могут состоять из слогов или отдельных фонем или контекстно-зависимых фонем, иначе говоря, N-фонов. Контекстно-зависимые фонемные модели позволяют учитывать коартикуляцию, что значительно увеличивает точность распознавания.

3.3. Языковая модель

Языковая модель предоставляет информацию о вероятности последовательности слов. По своей сути главная цель ЯМ предоставить нашей системе информацию о синтаксисе, семантике и прагматике языка. Поскольку языковая модель не зависит от акустических наблюдений, её можно обучить на больших текстовых коллекциях, будь то корпус языка или сборник газет, журналов и художественной литературы. Для построения языковой модели в системах распознавании речи используются N-граммы.

3.4. Декодер

На конечном этапе, чтобы определить наиболее вероятную последовательность слов, в игру вступает декодер, который совмещает полученные данные от акустической и языковой модели, и после их объединения выдает конечный результат работы системы.

3.5. Оценка

Качество работы систем распознавания речи как правило оценивается при помощи метрики WER (Word Error Rate), которая так же используется в машинном переводе.

Эта метрика происходит от вычисления расстояния Левенштейна, однако работает она на уровне слов, а не фонем.

Метрика WER вычисляется по формуле $WER = \frac{S+D+I}{N}$, где

- S — число подстановок; например, в исходном тексте предложение: 'Что за чудесный день', а в результатах обучения: 'Что за ужасный день',
- D — количество удалений; например, в исходном тексте предложение: 'Что за чудесный день', а в результатах обучения: 'Что за день',
- I — число вставок; наоборот, в исходном тексте предложение: 'Что за день', а в результатах обучения: 'Что за чудесный день',
- N — количество слов в правильном варианте.

4. Основная часть

4.1. Данные

Для построения системы распознавания речи бесермянского диалекта удмуртского языка были использованы данные, собранные во время лингвистических экспедиций 2009-2017 годов.

Несмотря на наличие большого количество расшифрованных текстов, для данной работы подходят далеко не все расшифровки. Большая часть расшифровок бесермянского языка сделана в текстовом редакторе и не сопоставлена со звуком. Однако есть записи лингвистических экспериментов, которые были размечены и отсегментированы в программах для анализа и обработки звука Praat и ELAN. Помимо расшифровок экспериментов, были использованы записи изолированных слов для электронного словаря. Каждое слово повторяется диктором по три раза. Всего таких слов 2646. Общая длительность аудио данных составляет 4.3 часов от тринадцати носителей, среди которых семь женщин и шестеро мужчин.

Для языковой модели был использован бесермянский корпус, который насчитывает около 65000 словоупотреблений.

4.2. Kaldi

Для разработки системы распознавания речи я использовала Kaldi (Povey et al. 2011). Это инструмент с открытым исходным кодом для распознавания речи, написанный на C++ и распространяющийся под лицензией Apache v2.0.

Важные функции, которыми обладает этот инструмент:

- интеграция с библиотекой для конструирования и поиска взвешенных конеччных преобразователей (OpenFst)
- поддержка линейной алгебры
- открытая лицензия
- готовые рецепты для некоторых языков
- тщательное тестирование
- подробная документация (которая, однако, непонятна без предварительной подготовки)

4.3. Обработка данных

В любом проекте, связанном с большим корпусом устной речи, возникает проблема разнообразности форматов записей и их обработки. Из-за того, что бесермянские аудио-данные собирались в течение десяти лет, возникают следующие проблемы:

- разный формат аудиозаписей (.wav vs. .mp3)
- аудиозаписи собраны при помощи разных диктофонов (mono vs. stereo)
- расшифровки аудиозаписей представлены в разных форматах (ELAN vs. Praat)
- наличие переключения языкового кода

Для решения этих проблем был написан праат скрипт, соединял два канала аудиодорожки в один, а также проводил процесс шумоочистки. Для дальнейшей обработки данных было решено крнвертировать ELAN (.eaf) разметку в Praat формат (.TextGrid).

Kaldi для запуска требует предварительную ручную подготовку данных, которые должны быть представлены в соответствии со строгим форматом. Всего необходимо подготовить 6 файлов:

1. spk2gender (<speaker-id> <gender>)
   В этом файле находится информация о гендерной принадлежности дикторов.
2. wav.scp (<utterrance-id> <full-path-to-audio>)
3. text (<utterrance-id> <text-transcription>)
   Этот файл содержит соответствия каждого высказывания с его текстовой транскрипцией.
4. utt2spk (<utterrance-id> <speaker-id>)
   Этот файл говорит системе, какому диктору принадлежит то или иное высказывание
5. segments (<utterance-id> <recording-id> <segment-begin> <segment-end>)
   Этот файл не нужен, если для обучения используются маленькие файлы. Однако при наличие длинных отсегментированных файлов, этот файл оказывается очень полезен, так как не надо резать длинный аудиофайл на маленькие кусочки.
6. corpus.txt
   Здесь находится корпус текстов, на котором обучается языковая модель

Самым простым способом подготовки всех данных, на первый взгляд, кажется написание Praat скрипта, который легко взаимодействует со звуком и разметкой в формате .TextGrid, но оказалось, что при выделении расшифровки высказываний, Praat скрипт смотрит на всё, что стоит в кавычках. Если исследователь при расшифровке заключил часть текста в кавычки, то скрипт на этом месте обязательно сломается. К счастью, Kaldi выделяет спектрально-аккустические признаки и для этого не надо прибегать к praat скриптам. Поэтому для создания всех необходимых для Kaldi файлов, был написан скрипт на языке программирования Python, который анализировал только файлы с разметкой.

5. Заключение

Системы распознавания речи могут значительно облегчить работу полевых исследователей по анализу аудиозаписей. Если сейчас большинству лингвистов необходимо сидеть с информантом и разбирать текст, то системы распознавания речи позволяют сконцентрироваться и тратить время исключительно на сбор материала и его дальнейшую обработку.

Конечно, чтобы обучить хорошую и точную систему, необходимо как минимум 2 часа устной размеченной и расшифрованной речи, что уже является хорошим корпусом. Однако, многие лингвистические эксперименты основываются на спонтанной речи носителей и довольно часто после сбора материала, исследователям, хорошо знающим язык, необходимо потратить целый год только на расшифровку этих данных.

В рамках данного исследования была проделана вся необходимая и трудоёмкая работа с данными, которая необходима для дальнейшего построения системы распознавания речи, а именно:

- приведение всех аудиофайлов к единому формату, что так же включает в себя модуль шумоочистки
- приведение разметки файлов к единому формату

- удаление всех выражений, в которых встречается языковое переключение кодов
- автоматическое создание всех необходимых для Kaldi файлов, что позволит в дальнейшем легко добавлять к тренировочной выборке новые размеченные данные

Проблемы, с которыми я столкнулась:

- незнание C++
- Выделение признаков для обучения акустической модели при помощи метода MFCC, реализованного в Kaldi.

Созданная нами база данных и все вспомогательные скрипты доступны онлайн (`https://github.com/melanoya/beserman_ASR`) и могут использоваться для дальнейшей разработки и тестирования системы распознавания речи для бесермянского диалекта удмуртского языка.

Список литературы

KH Davis, R Biddulph, & Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

Jack Ferguson. Hidden markov models for speech. *IDA, Princeton, NJ*, 1980.

DB Fry. Theoretical aspects of mechanical speech recognition. *Journal of the British Institution of Radio Engineers*, 19(4):211–218, 1959.

Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, 1975.

Richard Lippmann. An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2): 4–22, 1987.

Thomas B Martin, AL Nelson, & HJ Zadell. Speech recognition by feature-abstraction techniques. Technical report, DTIC Document, 1964.

Harry F Olson & Herbert Belar. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6):1072–1081, 1956.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, & Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

D Raj Reddy. Approach to computer speech recognition by direct analysis of the speech wave. *The Journal of the Acoustical Society of America*, 40(5):1273–1273, 1966.

Hiroaki Sakoe & Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

P Suba & B Bharathi. Analysing the performance of speaker identification task using different short term and long term features. In *Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on*, pages 1451–1456. IEEE, 2014.

VM Velichko & NG Zagoruyko. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3):223–234, 1970.

Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, & Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

ТК Винцюк. Распознавание слов устной речи методами динамического программирования. *М.: Кибернетика*, (1):15–22, 1968.