Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра Информационно-Аналитических Систем

Михайлов Дмитрий Алексеевич

Анализ эффективности системы CMUSphinx

Бакалаврская работа

Научный руководитель:
ст. преп. Ярыгина А. С.

Рецензент:
ст. преп. Литвинов Ю. В.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Sub-Department of Analytical Information Systems

Mikhailov Dmitrii

Analysis of CMUSphinx system performance

Bachelor's Thesis

Scientific supervisor:
Senior Assistant Prof. Yarygina Anna

Reviewer:
Senior Assistant Prof. Litvinov Yurii

Saint-Petersburg
2016

Оглавление

Введение

Автоматическое распознавание речи представляет собой актуальную задачу, связанную с множеством различных приложений, таких как, например, голосовое управление, автоматическая генерация субтитров к видеоматериалам, перевод аудиозаписи в текст и т.п. При этом качество аудиозаписи может сильно разниться в зависимотсти от формата. Записи спонтанной речи часто сопровождаются шумом от окружающей обстановки и звукозаписывающей аппаратуры. Например, записи с различного рода конференций содержат не только речь докладчика, но и звуки разговоров на фоне, звуки передвигаемой мебели, иногда помехи микрофона и т.п. Кроме того говорящий может запинаться, менять темп речи. Так как эти факторы влияют на восприятие речи человеком, естественно, что они же будут влиять и на автоматическое распознавание.

В данной работе под качеством аудиозаписи подразумеваются:

- наличие/отсутствие шума на аудиодорожке и его уровень,
- речевые особенности говорящих.

Шум может быть связан с аппаратурой, используемой при записи. У микрофонов есть собственный уровень шума, кроме него качество записи звука уменьшается при отсутствии защиты микрофона: поп-фильтров, звукозаглушающих решёток,– так как кроме речи записывается звук дыхания. Шумы могут исходить от окружающей обстановки. Например, если одновременно говорят несколько человек, понять речь хотя бы одного довольно сложно. Также распознавание речи затрудняют речевые дефекты (сигматизмы, ротацизм), акцент говорящего. Наиболее распространёнными оказываются записи, сочетающие несколько из вышеперечисленных факторов.

Системы распознавания речи нередко обучают и тестируют на данных с минимальным уровнем шума, как, например, HUB-4 [1, 6] или TIMIT [2, 3], и дикторским произношением (HUB-4 или AN4 [5]). При этом естественно возникает вопрос: каково качество распознавания данных более низкого качества у таких систем? В данной работе будет исследоваться система распознавания речи CMUSphinx [4, 5], развивающаяся уже несколько десятилетий. Необходимо исследовать, насколько эффективно данная система будет работать с данными, более приближенными к реальной жизни: зашумлённые записи, спонтанная речь. Под эффективностью будет пониматься точность распознавания речи (WER - Word Error Rate).

1. Основные понятия

1.1. Структура речи

Речь представляет собой непрерывный аудиопоток, пребывающий в различных постоянно изменяющихся состояниях. Среди этих состояний выделяются наиболее схожие классы, фонемы. При этом свойства звуковой волны, соответствующей данной фонеме, зависят от многих факторов, таких как контекст фонемы, речевые особенности говорящего и т.д. Более того, так как переходы между звуками более информативны, чем устойчивые отрезки, исследователи выделяют дифоны(diphones) – участки аудиосигнала между двумя последовательными фонемами.

Часто фонемы рассматриваются в некотором контексте [7,8]. Такие фонемы называют трифоны(triphones), квинфоны(quinphones). То есть звучание конкретной фонемы может отличаться в зависимости от окружающих её фонем. В отличие от дифонов, трифонам соответствуют те же отрезки звуковой волны, что и фонемам (на рис. 1 отмечены участки звуковой волны, соответствующие трифонам и дифонам).

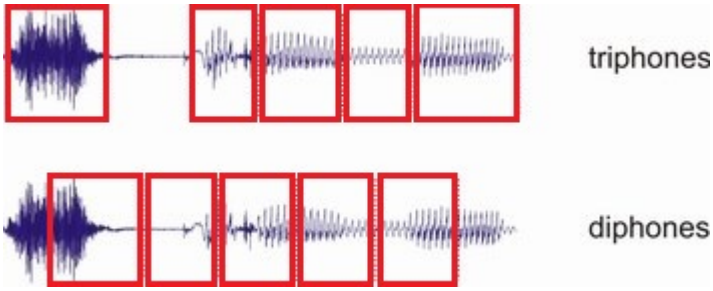

Рис. 1: Отрезки аудиосигнала, соответствующие трифонам и дифонам [40]

Для удобства проведения вычислений бывает полезно выявлять части трифонов, а не их целиком. Например, можно построить детектор для начала некого трифона и использовать его в дальнейшем для многих других трифонов с таким же началом. То есть можно создать множество таких детекторов для очень коротких звуков. В CMUSphinx такие детекторы называют сеноны (senones). Зависимость от контекста для сенонов может быть сложнее, чем только правый и левый контексты. Объём контекста пока является областью изучения [9].

Из фонем формируются подслова (subwords), такие как слоги. Иногда их определяют как устойчивые к редукции сущности (reduction-stable entities). Так как при ускорении речи фонемы могут изменяться, слоги остаются неизменными. Существуют также другие способы выделения подслов [10], но они не относятся к данной работе. Различные комбинации подслов формируют слова. Слова и нелингвистические звуки, например, вздохи, мычание, кашель, формируют высказывания - участки аудиозаписи между паузами.

1.2. Распознавание

Обычно процесс распознавания проходит следующим образом: звуковая волна разбивается по участкам тишины. В зависимости от системы и используемых данных устанавливается определённый минимум громкости, например, 10 Дб, и время поддержания этого уровня, например, 100 мс, или 10 фреймов; также может рассматриваться понижение уровня громкости на определённое количество Дб [6]. Также существуют другие методы выделения участков аудиосигнала, на которых присутствует речь [11, 12]. После этого производятся попытки сопоставления возможных комбинаций слов с аудиорядом, из которых выбирается лучшая.

Аудиосигнал разбивается на фреймы (обычно, по 10 мс), и для каждого фрейма вычисляется вектор признаков. В системе CMUSphinx данный ветор состоит из 39 чисел. Их содержание зависит от метода извлечения признаков. Например, в данной работе первые 13 чисел – кепстр (power cepstrum [13]) сигнала, вторые – первая производная от кепстра, третьи – вторая производная. Поиск наиболее эффективного способа вычисления этих чисел всё ещё является предметом изучения [14, 15], однако в общих случаях вычисляется производная от спектра. Далее, получившиеся векторы рассматриваются в рамках определённых моделей – языковой, акустической и фонетического словаря – для подбора наиболее подходящей комбинации слов. Моделью распознавания паттернов в речи является скрытая марковская модель [25] (HMM, Hidden Markov model), поэтому в данном случае 'подходящей комбинацией' является та, которая является наиболее вероятной. В каждый момент времени поддерживаются текущие лучшие комбинации, которые с течением временя расширяются засчёт поиска комбинаций для следующего фрейма.

1.3. Модели, соответствующие структуре речи

В распознавании речи используют комбинацию следующих моделей:

1. **Акустическая модель**, которая содержит информацию о звуковых параметрах (признаках) каждого сенона; бывают как контекстно-зависимые, так и контекстно-свободные. Обозначим набор признаков как O, тогда для слова W вероятность встретить набор O – $P(O \mid W)$

2. **Фонетический словарь**, содержащий отображение из слов в фонемы. Можно заметить, что представление в виде словаря не является обязательным: это может быть некая функция, полученная в результате алгоритма машинного обучения. Обозначим произношение слова W как $Q(W)$

3. **Языковая модель**, которая ограничивает слова для поиска. Она определяет, какое слово может следовать за данным и с какой вероятностью, таким обра-

зом отсекая невозможные варианты. Наиболее часто используются n-граммные языковые модели. Обозначим вероятность встретить слово W как $P(W)$

Если обозначить исходный аудиосигнал за X, а множество всех слов за $\mathbb{W}$, то задача распознавания речи формализуется следующим образом.

$$W^* = \arg\max_{W \in \mathbb{W}} P(W \mid X) = \arg\max_{W \in \mathbb{W}} P(O \mid Q(W))P(W)$$

1.4. Используемая метрика

В данной работе для определения эффективности систем распознавания речи используется метрика Word Error Rate(WER) [28], основанная на подсчёте вставок, замен и удалений слов. Пусть в тексте, полученном в результате распознавания речи, N слов. Тогда

$$WER = \frac{S + D + I}{N}$$

где

- S – количество заменённых слов
- D – количество удалённых слов
- I – количество вставленных слов

То есть, чем выше WER, тем менее качественно произошло распознавание. Обычно выражается в процентах.

2. Обзор существующих систем

Существует несколько известных систем распознавания речи с открытым исходным кодом: CMUSphinx, HTK [16, 17], Kaldi [18, 19], Julius [20, 21]. CMUSphinx реализована на языке Java (некоторые библиотеки написаны на C), HTK и Julius – на C, Kaldi – на C++. Их работа основана на HMM. Различия составляют используемые алгоритмы и модели.

2.1. HTK

HTK (Hidden Markov Model Toolkit) является инструментом для построения и модификации HMM. HTK в основном используется в исследованиях в области распознавания речи; также данная система использовалась для синтеза речи, распознавания символов и цепочек ДНК. HTK состоит из множества модулей и инструментов, реализованных на языке C. Эти инструменты обеспечивают платформу для анализа речи, обучения и тестирования HMM, позволяют построение сложных систем, основанных на HMM.

2.2. CMUSphinx

CMUSphinx – система распознавания речи, объединяющая ряд инструментов для задач различной сложности. Эти инструменты разрабатывались специально для низкопроизводительных систем. В отличие от HTK, применяется больше в прикладных целях, а не в исследованиях. На данный момент реализована поддержка американского и британского английского, французского, мандаринского, немецкого, голландского и русского языков. Обладает довольно активным сообществом.

Выбор системы CMUSphinx для данной работы обусловлен большим объёмом научных работ на её основе [33–37], а также субъективным фактором – удобством использования.

2.3. Kaldi

Основное направления использования и развития Kaldi такое же, как у HTK – исследования в области распознавания речи. Отличием от большинства других систем – интеграция Finite State Transducers (FSTs) и широкое использование линейной алгебры. Поддержку FST обеспечивает библиотека с открытым исходным кодом OpenFST [38], линейной алгебры – библиотеки BLAS[1] (Basic Linear Algebra Subroutines) и LAPACK[2] (Linear Algebra PACKage). Как и вышеперечисленные системы, Kaldi

[1] http://www.netlib.org/blas/
[2] http://www.netlib.org/lapack/

предоставляет код в наиболее общей форме, что позволяет использовать его для большого спектра задач. Также может использоваться для создания сложных систем распознавания речи.

2.4. Julius

Система Julius использует двупроходную стратегию [39] (two-pass strategy) распознавания речи. Основана на n-граммных языковых моделях. Крайне низкие требования памяти: для триграммных моделей со словарём из 20 тысяч слов используется не более 64 мегабайт памяти. Изначально разрабатывалась японскими исследователями для японского языка, поэтому на данный момент существуют только простейшие (на несколько десятков слов) языковые модели для английского языка.

3. Система CMUSphinx

На данный момент актуальной является версия sphinx4-5prealpha. Sphinx-4 представляет собой модульный фреймворк(рис. 2). Модульная структура позволяет варьировать параметры системы исходя из требований конкретной задачи. Выделяются 3 основных модуля: FrontEnd, Decoder и Linguist. FrontEnd преобразует входные данные в вектор параметров. Linguist на основе выбранных языковой, акустической моделей и словаря строит SearchGraph. Наконец, подмодуль Decoder'a – SearchManager – использует вектор параметров и построенный граф для декодирования и выдаёт результат.

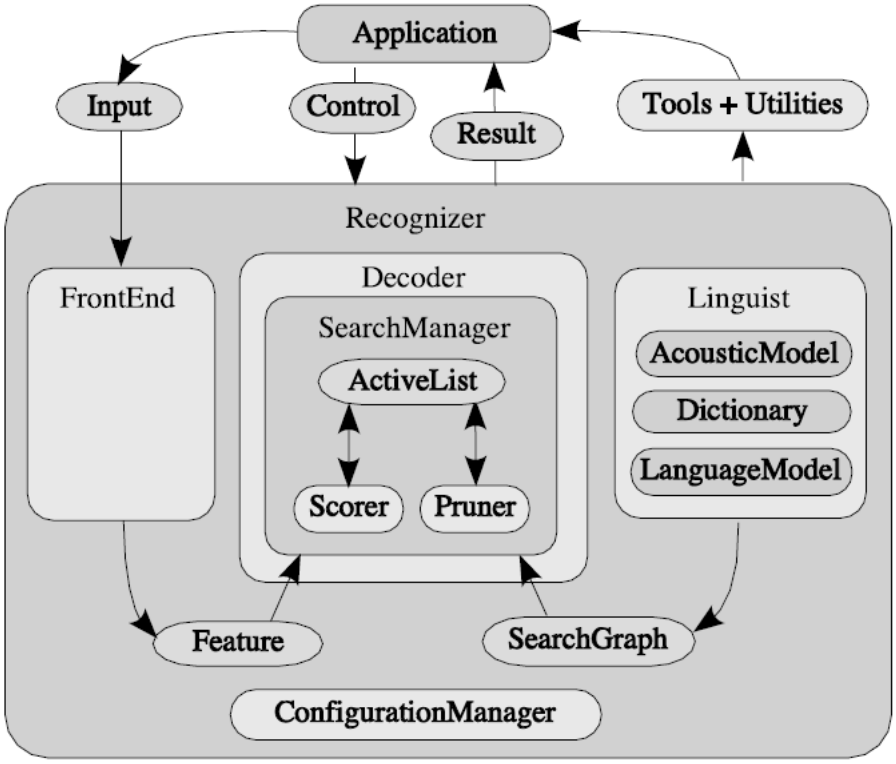

Рис. 2: Взаимодействие модулей системы CMUSphinx [5]

3.1. FrontEnd

FrontEnd объединяет в себе несколько цепочек из модулей сообщающихся обработчиков данных, каждый из которых способен вырабатывать различные параметры в результате вычислений. Наличие нескольких цепочек позволяет как производить вычисления для разных типов параметров, так и принимать несколько входных сигналов одновременно.

3.2. Linguist

Как упоминалось выше, построение SearchGraph'а в Linguist происходит на осовании данных о языке, получаемых из языковой и акустической модели. Каждая из них представляет собой HMM для элементарных звуковых единиц, используемых в конкретной системе. Словарь сопоставляет слова из языковой модели и комбинации элементов акустической модели.

- Языковая модель описывает структуру языка на уровне слов. Обычно используются следующие два вида моделей: графовые грамматтики и n-граммные модели. Графовые грамммматики представляют собой ориентированный граф, в котором вершинами являются слова, а каждому ребру соответствует вес, являющийся вероятностью перехода к следующему слову. В случае n-граммной модели имеется набор вероятностей встретить данное слово, если известны предыдущее $n-1$ слово. В данной работе использовалась триграммная языковая модель для американского английского.

- Словарь содержит варианты произношения для всех слов, встречающихся в данной языковой модели. Произношение слова разбито на наборы некоторых элементарных блоков. Например, abandon $\leftrightarrow$ AH B AE N D AH N

- Акустическая модель отображает элементы речи (в данном случае – трифоны) на HMM. Естественно, отображение может принимать информацию о контексте и положении в слове. Левый и правый контексты были рассмотрены выше, а информация о положении показывает, находится ли трифон в начале, середине или конце слова. Или же сам по себе является словом.

  В Linguist происходит разбиение каждого слова на контекстно-зависимые элементарные блоки (по информации из словаря), по которым затем по акустической модели строится множество графов HMM. Вершины этих графов – фонемы или трифоны, а рёбра имеют веса – вероятности перехода между фонемами). И на основе полученных графов и языковой моедли строится SearchGraph (рис. 3)

- SearchGraph – ориентированный граф, в котором каждая вершина представляет производящее (emitting) или непроизводящее (non-emitting) состояние. Производящим состояниям соответствуют рассматриваемые звуковые признаки, а непроизводящие состояния представляют более высокоуровневые языковые конструкции, такие как слова или фонемы, которые к вычислению признаков относятся опосредованно. Дуги графа представляют переходы между состояниями, каждой дуге соответствует определённая вероятность прохода по ней.

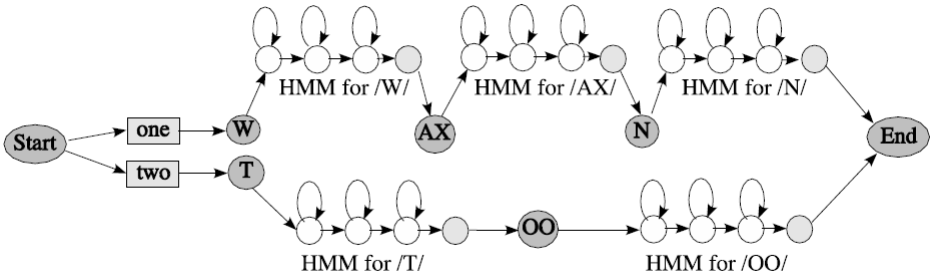

Рис. 3: Пример SearchGraph'а [5]

3.3. Decoder

Основная функция данного модуля – по вычисленным в FrontEnd признакам и построенному в Linguist'е SearchGraph'у получить множество гипотез. Decoder посылает подмодулю под названием SearchManager запрос на распознавание множества фреймов с указанными выше данными. На каждом этапе работы SearchManager строит все пути, которые достигают конечного непроизводящего состояния. SearchManager использует алгоритм передачи токенов (token passing algorithm [29]). Для используемого алгоритма SearchManager может, но не обязан, содержать множество активных токенов (ActiveList). Для упрощения вычислений подмодулем Pruner проводится сокращение множества токенов. Подмодуль Scorer по запросу вычисляет оценки плотности распределения для данных состояний в данные моменты времени.

Система CMUSphinx позволяет изменять код любого из модулей, если это необходимо в рамках определённых данных или задачи. Также встроенные средства позволяют адаптировать акустическую модель под речевые особенности конкретных говорящих: акценты, нарушения произношения и т.п..

4. Эксперименты

4.1. Экспериментальные данные

Для анализа эффективности системы CMUSphinx были выбраны следующие 2 речевых корпуса: Edinburgh University Speech Timing Archive and Corpus of English [22] и Santa Barbara Corpus of Spoken American English [23, 24].

4.1.1. EUSTACE

Первый корпус состоит из записей речи 6 человек – трёх мужчин и трёх женщин – и содержит в общей сложности 384 (от каждого говорящего по 64) записи длительностью от 25 секунд до минуты. Каждая запись является зачитыванием набора (10–20 штук) похожих однотипных фраз. Например, в одном из файлов записаны фразы вида 'Bob said he saw the fish again' – 'Bob said he saw the fisherman' – 'Bob said he saw the fissure'. Из шумов присутствует негромкий фон, связанный, скорее всего, с записывающим устройством. Шумов, исходящих от окружающей обстановки замечено не было. У разных говорящих замечены разные варианты произношения одинаковых слов. Формат аудиофайлов соответствует стандартному входному формату CMUSphinx (частота дискретизации 16000 Гц, моно)

4.1.2. Santa Barbara Corpus of Spoken American English

Santa Barbara Corpus of Spoken American English содержит 60 записей спонтанной речи. Это записи бесед нескольких человек, записи с собраний. Длительность варьируется в пределах от 20 до 40 минут. При разговорах использовано около 249000 слов. Уровень шума довольно высок. Присутствует как фон звукозаписывающего оборудования, так и шум от окружения: звон посуды, шум передвигаемой мебели, голоса на фоне и т.п. Кроме этого, так как речь спонтанна, люди могут говорить одновременно или перебивать друг друга, что увеличивает сложность для систем распознавания речи. Также формат аудиофайлов не соответствовал стандартному входному формату CMUSphinx: частота дискретизации составляла 20050 Гц, стерео. Форматирование к стандарту было произведено с помощью инструмента Audacity[3].

4.2. Эксперименты

Оба корпуса были полностью пропущены через систему CMUSphinx два раза. Первый проход осуществлялся с использованием стандартных языковой и акустической моделей и словаря (в словаре 134522 слова) для американского английского. Для второго прохода была осуществлена адаптация акустической модели.

[3]https://sourceforge.net/projects/audacity/

Адаптация проводилась с помощью двух методов: MLLR [30] (Maximum Likelihood Linear Regression) и MAP [31, 32] (Maximum A Posteriori).

В случае первого корпуса была осуществлена адаптация для каждого говорящего по отдельности. Было использовано 5% от суммарного количества зачитанных фраз каждого говорящего. Выбор предложений был произведён случайным образом. WER подсчитывался для каждого говорящего по всем его записям.

Для второго корпуса адаптация была осуществлена для каждого аудиофайла отдельно, потому что говорящие меняются от записи к записи. Использовано около 10% (5% + 5%) от длительности записи. При этом в каждом случае одна половина фраз была выбрана случайно, а другая вручную: при просмотре транскрипции прослушивался аудиофайл, и выбирались фразы или части фраз; кроме того, так как во всех записях данного корпуса было больше одного говорящего, выбирались отрезки для каждого из них. Это было сделано для того, чтобы покрыть как можно больше разнообразных особенностей речи говорящих. WER подсчитывался для каждой записи отдельно.

5. Результаты

Для первого корпуса при первом проходе (стандартные параметры) были получены следующие результаты WER(табл. 1).

первая женщина	39,3%
вторая женщина	31,95%
третья женщина	35,16%
первый мужчина	28,02%
второй мужчина	59,64%
третий мужчина	28,93%

Таблица 1: WER для EUSTACE до адаптации

Среднее значение WER составляет 36,58%. Как можно заметить, от среднего значения сильнее всего отличаются результаты распознавания речи второго мужчины. Это объясняется его произношением: наблюдается 'зажёвывание' некоторых слогов и вибрирующий звук. Возможно, это также связано с настройкой звукозаписывающего оборудования.

При втором проходе (после адаптации) для каждого из говорящих были улучшены результаты (табл. 2).

первая женщина	30,02%
вторая женщина	26,25%
третья женщина	24,05%
первый мужчина	21,84%
второй мужчина	45,51%
третий мужчина	20,42%

Таблица 2: WER для EUSTACE после адаптации

Среднее значение WER после адаптации составляет 28,01%(рис. 4). Улучшение составило 4-14% в зависимости от говорящего. Небольшие отличия в снижении WER может объясняться тем, что были зачитаны однотипные наборы фраз, и поэтому случайность выборов фрагментов для адаптации появлияла не слишком сильно.

Для второго корпуса среднее значение WER при стандартных параметрах составило 67,79%(рис. 5). После адаптации – 47,41%. Можно заметить, что присутствует несколько записей с выделяющимся WER (около 80-90%). Это записи с собраний большого числа людей, естественно, что от этого увеличилось количество постороннего шума и одновременных разговоров нескольких людей, что не распознаётся системой как речь. Также есть записи с изначально низким показателем WER – 30-40%. Это диалоги в форме интервью, в которых окружающие шумы сведены к минимуму.

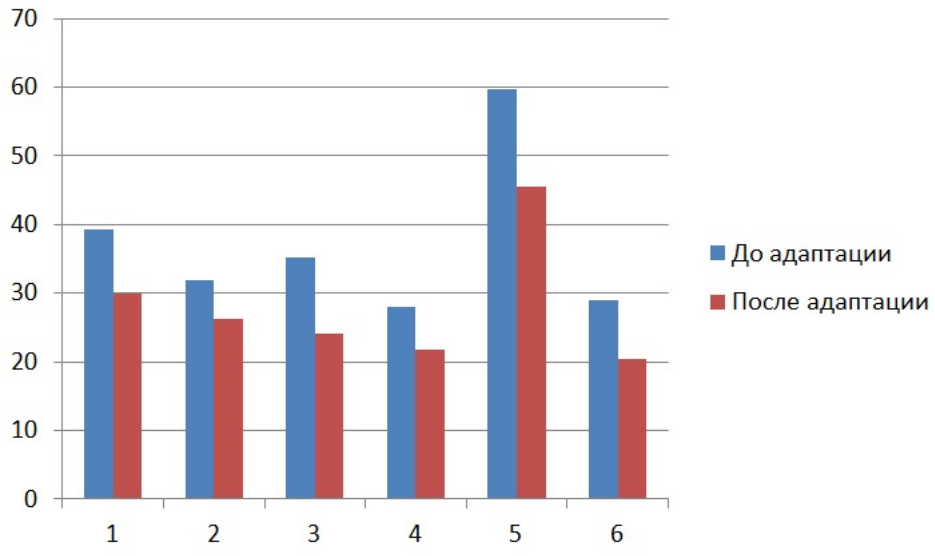

Рис. 4: Результаты WER для EUSTACE

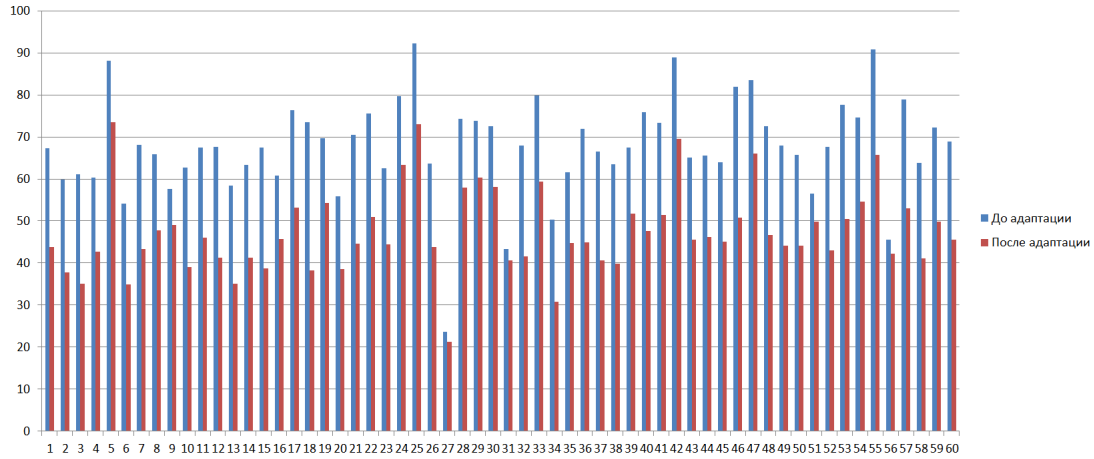

Рис. 5: Результаты WER для Snata Barbara Corpus of Spoken American English

Снижение WER после адаптации имеет неравномерный характер – в некоторых случаях снижение серьёзное (с 70% до 40%), в некоторых – нет (с 45% до 42%). Это может объясняться следующим образом: половина фрагментов для адаптации была выбрана случайно, оставшаяся половина – вручную; то есть субъективный взгляд на необходимые к покрытию особенности речи в каких-то случаях оказался более удачным, в других же – менее.

6. Заключение

В перспективе имеется приложение использовать тексты, полученные в результате распознавания речи какой-либо системой, для автоматической обработки (например, для поиска). Для этого система распознавания речи должна обладать достаточно высокой точностью. В данной работе была рассмотрена система CMUSphinx, обладающая высокими показателями точности на данных высокого качества (как показано в [5]). Показатели WER для слабозашумлённых данных (корпус EUSTACE) оказались близки к 30%, что является препятствием для эффективной автоматической обработки текста. Результаты WER для зашумлённых и сильнозашумлённых данных с разнообразными речевыми особенностями говорящих (Santa Barbara Corpus of Spoken American English) – почти 50% – являются неприемлемо высокими и делают автоматическую обработку полученных текстов бессмысленными. Таким образом, необходимо производить дополнительную обработку аудиосигнала для уменьшения уровня шума. При этом во всех аудиозаписях оказались улучшены результаты WER при адаптации акустической модели, что свидетельствует о состоятельности методов адаптаций применительно к данным различного качества.

Список литературы

[1] John S. Garofolo, Jonathan G. Fiscus, William M. Fisher. Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora. 1997

[2] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. 1993

[3] P. J. Moreno, R. M. Stern. Sources of degradation of speech recognition in the telephone network. 1994

[4] CMU Sphinx Project by Carnegie Mellon University. http://cmusphinx.sourceforge.net/

[5] Walker, Lamere, Kwok, Raj, Singh, Gouvea, Wolf, Woelfel. Sphinx-4: A Flexible Open Source Framework for Speech Recognition. 2004

[6] Placeway, Chen, Eskenazi, Jain, Parikh, Raj, Ravishankar, Rosenfeld, Seymore, Siegler, Stern, Thayer. The 1996 Hub-4 Sphinx-3 System

[7] K. -F. Lee. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. 1990

[8] George E. Dahl, Dong Yu, Li Deng, Alex Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. 2011

[9] Luciana Ferrer, Yun Lei, Mitchell McLaren, Nicolas Scheffer. Spoken language recognition based on senone posteriors. 2014

[10] C. -H. Lee, B. -H. Juang, F. K. Soong, L. R. Rabiner. Word recognition using whole word and subword models. 1989

[11] Sohn, Kim, Sung. A Statistical Model-Based Voice Activity Detection. 1999

[12] T. Hughes, K. Mierle. Recurrent neural networks for voice activity detection. 2013

[13] Alan V. Oppenheim, Ronald W. Schafer. From Frequency to Quefrency: A History of the Cepstrum. 2004

[14] Shreya Narang, Ms. Divya Gupta. Speech Feature Extraction Techniques: A Review. 2015

[15] Jing Dong, Dongsheng Zhou, Qiang Zhang. Robust Feature Extraction Based on Teager-Entropy and Half Power Spectrum Estimation for Speech Recognition. 2015

[16] HTK Speech Recognition Toolkit. http://htk.eng.cam.ac.uk/

[17] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kersha, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Anton Ragni, Valtcho Valtchev, Phil Woodland, Chao Zhang. The HTK Book (for HTK Version 3.5, documentation alpha version). 2015

[18] Kaldi ASR. http://kaldi-asr.org/

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlıcek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely. The Kaldi Speech Recognition Toolkit. 2011

[20] Open-Source Large Vocabulary CSR Engine Julius. http://julius.osdn.jp/en_index.php

[21] Akinobu Lee, Tatsuya Kawahara. Recent Development of Open-Source Speech Recognition Engine Julius. 2009

[22] Edinburgh University Speech Timing Archive and Corpus of English. http://www.cstr.ed.ac.uk/projects/eustace/index.html

[23] Santa Barbara Corpus of Spoken American English. http://www.linguistics.ucsb.edu/research/santa-barbara-corpus

[24] Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

[25] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. 1989

[26] http://musslap.zcu.cz/en/acoustic-speech-synthesis/

[27] C. J. Leggetter, P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. 1995

[28] Ye-Yi Wang, Alex Acero, Ciprian Chelba. Is word error rate a good indicator for spoken language understanding accuracy. 2003

[29] S. J. Young, N. H. Russell, J. H. S. Russell. Token passing: A simple conceptual model for connected speech recognition systems. 1989

[30] C. J. Leggetter, P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. 1995

[31] Michiel Bacchiani, Michael Riley, Brian Roark, Richard Sproat. MAP adaptation of stochastic grammars. 2006

[32] Liang Lu, Arnab Ghoshal, Steve Renals. Maximum a posteriori adaptation of subspace gaussian mixture models for cross-lingual speech recognition. 2006

[33] Ziad Al Bawab. An Analysis-by-Synthesis Approach to Vocal Tract Modeling for Robust Speech Recognition. 2009

[34] Xiang Li. Combination and Generation of Parallel Feature Streams for Improved Speech Recognition. 2005

[35] Jon P. Nedel. Duration Normalization for Robust Recognition of Spontaneous Speech via Missing Feature Methods. 2004

[36] Michael L. Seltzer. Microphone Array Processing for Robust Speech Recognition. 2003

[37] Balakrishnan Narayanaswamy. Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities. 2005

[38] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: a general and efficient weighted finite-state transducer library. 2007

[39] S. Matsunaga, H. Sakamoto. Two-pass strategy for continuous speech recognition with detection and transcription of unknown words. 1996

[40] http://musslap.zcu.cz/en/acoustic-speech-synthesis/