Настройка и адаптация модуля распознавания речи CMU Sphinx на основе скрытых марковских моделей

В. В. Гончарова, email: vlntngoncharova@gmail.com
В.В. Гаршина, email: garshina@cs.vsu.ru

Воронежский государственный университет

Аннотация. *в данной работе рассматриваются основные методы оценки качества распознавания речи, а так же проводится адаптация системы распознавания речи речи CMU Sphinx на основе скрытых марковских моделей*

Ключевые слова: *оценка качества распознавания речи, CMU Sphinx , Word Recognition Rate, Word Error Rate, Word Correctly Recognized, адаптация, настройка.*

Введение

Технологии распознавания речи все уверенней входят в нашу жизнь. Однако существует проблема, с которой приходится столкнуться при эксплуатации большинства систем распознавания речи. Это недостаточная точность распознавания.

Из-за этого фиксируются частые случаи неправильно распознанных команд, ложных срабатываний (команда не была произнесена, но система распознала окружающий шум как команду) или игнорирование команды. С таким качеством распознавания речи конечному пользователю работать недопустимо.

Такое поведение системы объясняется акцентом или дефектами речи диктора, окружающим шумом при записи речи, некачественным микрофоном и многими другими факторами. Чтобы сократить процент ошибок распознавания систему чаще всего настраивают под конкретного пользователя или группу пользователей.

В данной работе будет проведена настройка системы распознавания речи, встроенная в программную платформу для решений, технологий и устройств вспомогательной и дополненной реальности. Так как платформа предназначена для умных очков, то управление ею происходит с помощью голосовых команд на русском языке. Поэтому в качестве системы распознавания была выбрана платформа CMU Sphinx, поддерживающая русский язык и имеющая режим распознавания отдельных команд. В ходе первичного тестирования программной платформы было выявлено, что качество

распознавания является неудовлетворительным и для дальнейшей работы его необходимо повысить.

1. Экспериментальная оценка качества распознавании речи с неадаптированной акустической моделью

Прежде чем приступать к настройке, необходимо зафиксировать текущее состояние системы.

Для оценки качества распознавания были взяты голосовые команды для управления описанной выше программной платформой. Выборка состояла из всех используемых команд, их количество равно 39, количество слов во всех командах – 57. Основным условием проведения эксперимента является окружающий уровень шума при записи распознаваемой речи. Данный эксперимент проходил в условиях малой зашумленности. В эксперименте участвовали 3 диктора.

В данной работе были использованы два способа вычисления точности системы [1]:

1. Количество правильно распознанных слов (WRR — Word Recognition Rate) или противоположное значение – количество ложно распознанных слов (WER — Word Error Rate).

Метод ложно распознанных слов состоит в подгонке двух строк. Первая – строка, полученная в результате распознавания. Вторая – исходная строка. Данный метод основан на использовании расстоянии Левенштейна, которое представляет оценку редактирования данных – минимальное количество операций вставки, удаления и редактирования одного символа для выравнивания двух строк текста. Показатель WER вычисляется по формуле

$$WER = \frac{S + D + I}{T} \quad (1)$$

где S – количество операций замены слов; D – количество операций удалений слов; I – количество операций вставки слов.

Соответственно показатель для правильно распознанных слов WRR рассчитывается по формуле

$$WRR = 1 - WER \quad (2)$$

Для оценки данным методом системе был предложен текст, состоящий из 39 команд. Результаты эксперимента приведены в табл. 1.

Таблица 1

Оценка качества распознавания WRR и WER

	Диктор 1	**Диктор 2**	**Диктор 3**
S	3	4	3
D	2	3	1
I	1	0	1
T	39	39	39
WER	15%	19%	13%
WRR	85%	81%	87%

2. Процент корректно распознанных слов (WCR — Word Correctly Recognized)

Данный показатель вычисляется по формуле

$$WER = \frac{H}{T}100\% \quad (3)$$

где H – количество верно распознанных слов; T – количество слов в исходной фразе.

Для вычисления показателя WCR был выбран альтернативный способ тестирования. В этом случае распознавание происходило в контексте мобильного приложения, то есть дикторы предъявляли речевые образцы команд модулю распознавания речи, взаимодействуя с ним через мобильное приложение. За единицу распознавания принимается отдельное слово из команды.

Результаты тестирования представлены в табл. 2.

Данные о ложных срабатываниях и игнорировании команд вынесены как отдельные показатели, так как данная оценка точности их не учитывает, но для оценки качества системы они важны.

Таблица 2

Оценка качества распознавания

	Диктор 1	**Диктор 2**	**Диктор 3**
T	57	57	57
H	49	51	52
Ложные срабатывания	7	4	5
Игнорирование команды	2	0	3
WCR	86%	89%	91%

2. Настройка и адаптация модуля распознавания речи

В данной работе для системы распознавания используются ненастроенные ресурсы, предоставляемые CMU Sphinx для русского языка. Так как ресурсы представляют из себя параметры скрытых марковских моделей, то с помощью обучения (адаптации) СММ можно добиться повышения качества распознавания речи.

Подготовка данных для адаптации

Первый этап – это создание корпуса адаптационных данных

- аудиозаписи речи;
- файл с транскрипцией надиктованной речи (*.transcription);
- порядковый список используемых аудиофайлов (*.fileids).

Основной материал, необходимый для адаптации – звуковая запись речи. Обычно используют большие речевые корпуса с несколькими десятками или сотнями часов надиктованной речи. Так как в данной работе настраивается система для распознавания ограниченного набора команд, то целесообразней будет производить настройку именно на записи этого ограниченного списка. Для повышения гибкости распознавания речи запись должна осуществляться несколькими дикторами в различных условиях зашумленности.

В качестве речевых образцов были использованы записи произнесения полного списка голосовых команд. Для адаптации были выбраны четыре диктора и два места записи с разным уровнем фонового звукового шума. Аудиофайлы должны обладать следующими качествами:

- количество каналов записи – mono;
- частота дискретизации – 16kHz;
- количество бит в сэмпле – 16bit;
- оцифровка – PCM.

Так же для адаптации необходимо подготовить транскрипции (текстовое представление) записей речи и специальный файл формата *.fileids с порядком следования аудиофайлов.

Обучение акустической модели

Обучение происходит на основе уже существующей акустической модели.

С помощью инструмента sphinx_fe библиотеки SphinxBase осуществляется генерация набора файлов характеристик (мел-частотных кепстральных коэффициентов) используемых аудио (листинг 1).

Листинг 1

Генерация набора файлов характеристик

```
sphinx_fe
-argfile en-us/feat.params \
-samprate 16000
-c arctic20.fileids \
-di .
-do .
-ei wav
-eo mfc
-mswav yes
```

Следующим шагом в адаптации является сбор статистики из данных адаптации. Делается это с помощью bw программы от SphinxTrain (листинг 2).

Листинг 2

Сбор статистики из данных адаптации

```
.
./bw \
-hmmdir en-us \
-moddeffn en-us/mdef.txt \
-ts2cbfn .ptm. \
-feat 1s_c_d_dd \
-svspec 0-12/13-25/26-38 \
-cmn current \
-agc none \
-dictfn cmudict-en-us.dict \
-ctlfn arctic20.fileids \
-lsnfn arctic20.transcription \
-accumdir .
```

Затем на основе максимального апостериорного критерия обновляется каждый параметр акустической модели. Это действие выполняется с помощью программы map_adapt (листинг 3).

Листинг 3

Обновление каждого параметра акустической модели

```
./map_adapt \
-moddeffn en-us/mdef.txt \
-ts2cbfn .ptm. \
-meanfn en-us/means \
-varfn en-us/variances \
-mixwfn en-us/mixture_weights \
-tmatfn en-us/transition_matrices \
-accumdir . \
-mapmeanfn en-us-adapt/means \
-mapvarfn en-us-adapt/variances \
-mapmixwfn en-us-adapt/mixture_weights \
```

```
-maptmatfn en-us-adapt/transition_matrices
```

После этого получается готовая к использованию адаптированная акустическая модель.

Заключительный этап – оценка качества адаптации. Этот пункт выполнялся экспериментальным путем.

С точки зрения работы с элементами речи в процессе обучения СММ можно выделить несколько этапов:

1. Вычисление мел-частотных кепстральных коэффициентов (MFCC) речевых образцов.
2. Акустическое моделирование. Обучение с использованием отдельных фонем – контекстно-независимые.
3. Отдельные фонемы объединяются с соседними фонемами по одной с каждой стороны – обучение модели трифонов.
4. Так как количество трифонов в языке может достигать нескольких тысяч, то необходимо объединить близкие по звучанию трифоны с помощью построения дерева фонем.
5. Кластеризация трифонов.
6. Обучение модели с кластерами трифонов.
7. Декодирование тестовой выборки речевых образцов и оценка качества распознавания (Word Error Rate) [2].

3. Экспериментальная оценка качества распознавания речи с адаптированной акустической моделью

Оценка качества производилась методами, описанными выше. Эксперимент проводился с соблюдением условий оценки качества неадаптированной моделью.

1. Количество правильно распознанных слов (WRR — Word Recognition Rate) или противоположное значение – количество ложно распознанных слов (WER — Word Error Rate) (табл. 3)

Таблица 3

Оценка качества распознавания WRR и WER

	Диктор 1	**Диктор 2**	**Диктор 3**
S	2	1	1
D	0	1	0
I	0	0	0
T	39	39	39
WER	5%	5%	2,5%
WRR	95%	95%	97,5%

2. Процент корректно распознанных слов (WCR — Word Correctly Recognized) (табл. 4)

Таблица 4

Оценка качества распознавания WCR

	Диктор 1	**Диктор 2**	**Диктор 3**
Т	57	57	57
Н	55	55	57
Ложные срабатывания	0	0	0
Игнорирование команды	1	2	1
WCR	96%	96%	100%

Сравнительный анализ оценки качества распознавания речи адаптированной и неадаптированной моделями представлен на диаграммах ниже (рис. 1 и 2).

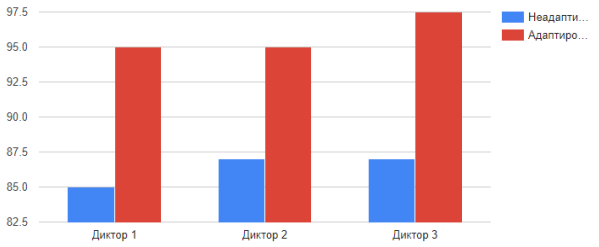

Рис. 1. Сравнительная диаграмма WRR

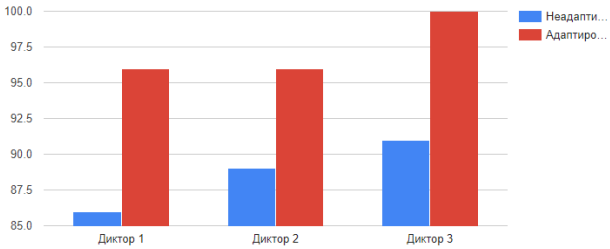

Рис. 2. Сравнительная диаграмма WCR

Анализируя сравнительные диаграммы можно сделать следующие выводы:

- количество правильно распознанных слов WRR возросло на 11%;
- количество ложно распознанных слов WER сократилось на 11,5 %;
- процент корректно распознанных слов WCR возрос на 8,6 %;

Заключение

В рамках данной статьи были рассмотрены и применены основные методы оценки качества распознавания речи: Word Recognition Rate, Word Error Rate и Word Correctly Recognized. Так же был проведен сбор и обработка речевых материалов для адаптации акустической модели. На основе собранного речевого корпуса проведена адаптация системы распознавания речи CMU Sphinx, используемая в программной платформе для решений, технологий и устройств вспомогательной и дополненной реальности. По различным методам оценки качества распознавания речи точность адаптированной модели возросла на 11% (WRR), а количество ошибок уменьшилось на 11,5 % (WER). В результате проделанной работы имеем настроенную систему распознавания речи с высокими показателями оценки качества.

Список литературы

1. Карпов А. А. Методология оценивания работы систем автоматического распознавания речи [Электронный ресурс] : научная статья по специальности «Компьютерные и информационные науки». - Режим доступа : http://pribor.ifmo.ru/file/article/5994.pdf

2. Adapting the default acoustic model [Электронный ресурс]. - Режим доступа : https://cmusphinx.github.io/wiki/tutorialadapt/