МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ ИДЕНТИФИКАЦИИ ДИКТОРА В СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 - Прикладная информатика

механико-математического факультета
Власова Владислава Андреевича

Научный руководитель
д. ф.-м. н. _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н. _____ С. П. Сидоров

Саратов 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Речь существенный элемент человеческой деятельности, позволяющий человеку познавать окружающий мир, передавать свои знания и опыт другим людям. Устная составляющая речи проявляется в виде голосовых высказываний, которые возможны благодаря голосовому аппарату человека. Каждый человек имеет индивидуальные голосовые характеристики, которые определяются особенностями строения его голосовых органов. В процессе общения люди способны на подсознательном уровне различать голоса других людей, однако для вычислительной техники данная задача является нетривиальной.

В настоящее время наблюдается рост интереса к технологиям, связанным с распознаванием речи. Это задачи управления устройствами с помощью голосовых команд, задачи справочной службы, которая предоставляют информацию после запроса в более естественной форме - с помощью голоса. Такая возможность удобна в некоторых ситуациях, например, когда не хочется искать информацию в смартфоне или если необходимо обратиться к некоторой функции устройства на расстоянии (если, например, нужно сделать фотографию). Для создания такой технологии необходимо было решить такую задачу, как распознавание речи. Исследования возможностей перевода человеческой речи в текст ведутся еще с середины XX века, и на сегодняшний день существует несколько основных подходов. Качество распознавание речи зависит от количества слов, которые необходимо будет распознан, каков будет тип распознаваемой речи, какого назначение (системы диктовки или командные системы) и многие другие характеристики. Системы распознавания могут быть разделены на текстозависимые и текстонезависимые. При текстозависимом распознавании могут использоваться как фиксированные фразы, так и фразы, сгенерированные системой и предложенные пользователю. Текстонезависимые системы предназначены обрабатывать произвольную речь. Также системы распознавания различаются методами и алгоритмами, на которых они построены. Это могут быть, например, скрытые Марковские модели или нейронные сети.

В последние годы набирают популярность алгоритмы, использующие системы идентификации диктора по голосу для выбора оптимальных акусти-

ческих моделей для распознавания речи. Задача распознавания личности по голосу была поставлена более 40 лет назад, но исследования в этой области продолжаются и в настоящее время. Известно много современных систем, которые пытаются решать данную задачу достаточно эффективно, однако точность подобных систем не всегда соответствует достаточному уровню для их реального применения.

На точность современных систем распознавания речи накладывается довольно много ограничений. Сюда относят проблемы, связанные с несоответствием условий обучения и распознавания диктора, проблемы различных акустических условий, в том числе наличия посторонних шумов и помех, проблемы отличия в спектральных составляющих записей голоса из-за применения различных микрофонов. Все это, в дополнение к несовершенству моделей и методов, применяемых для идентификации диктора, ведет к уменьшению точности распознавания. Поэтому интерес исследователей привлекают возможности использования адаптированных акустических моделей дикторов в системах распознавания речи.

В рамках магистерской работы была реализована система распознавания речи с использованием адаптированных акустических моделей дикторов выбор которых будет осуществляться на результатах работы реализованной системы идентификации диктора по голосу.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции «Актуальные проблемы математики и механики» которую проводил механико-математический факультет СГУ в апреле 2019 года, в секции «Анализ данных» и в VII Международной молодежной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2018 года.

1 Идентификация диктора по голосу

Работа систем распознавания содержит два основных этапа: регистрация пользователей в системе и сам процесс распознавания (идентификации личности). Пользователи предварительно регистрируются в системе, записав свои голоса. Образец голоса каждого диктора обрабатывается с целью извлечения признаков, которые могут быть использованы для распознавания. На основе извлеченных признаков строятся модели (в некоторых случаях более подходящим термином является шаблон) пользователей. Модель представляет собой некоторую структуру, позволяющую при данных признаках оценить степень подобия либо сразу принять решение.

Во время процесса идентификации происходит извлечение признаков из предъявленного образца (входного сигнала), которые затем сравниваются со всеми имеющимися в системе моделями.

Общую схему работы системы распознавания диктора можно описать следующим образом:

1. Обработка сигнала. Из сигнала выделяются признаки, существенные для задачи распознавания. Речевой сигнал представляется в виде набора векторов признаков.
2. Моделирование. Может заключаться как в простом копировании векторов признаков, так и в построении вероятностных моделей. После этого становится возможным при данных признаках вычислить степень подобия между признаками и сохраненной моделью.
3. Принятие решений.

Обработка сигнала в данных приложениях имеет целью выделить в речевом сигнале информацию, релевантную для задачи распознавания по голосу, то есть информацию, представляющую индивидуальные особенности голоса человека, или признаки. Выделенные признаки будут использованы для формирования шаблона или для сравнения с зарегистрированными шаблонами. Априори невозможно оценить, какие признаки более подходят для распознавания. Процесс определения подходящих признаков заключается в переборе возможных вариантов признаков с последующей экспериментальной оценкой. Выделяют два вида признаков: низкоуровневые (обусловленные анатомическим строением речевого аппарата) и высокоуровневые (приобре-

тенные, связанные с манерой произношения). Сложившийся подход к процедуре обработки речевого сигнала состоит в использовании кратковременного анализа. То есть сигнал разбивается на временные окна фиксированного размера, на которых, как предполагается, параметры сигнала не меняются. Для речевого сигнала размер окна обычно выбирается в пределах 10 - 30 мс. Для более точного представления сигнала между окнами делают перекрытие, равное половине длины окна. Затем к каждому окну применяются алгоритмы извлечения признаков, такие как спектральный анализ, метод линейного предсказания или другие. Методы извлечения я признаков предназначены для выделения характеристик на небольшом участке. Для того чтобы сохранить информацию о динамике речи, применяют подход, заключающийся в объединении векторов признаков с их первыми и, возможно, вторыми производными. Такие производные получили название Δ- и $\Delta-\Delta$- коэффициентов (дельта- и дельта-дельта-коэффициентов).

Последовательность векторов, полученная после этапа обработки сигнала, используется для построения шаблона/модели диктора или для осуществления сравнения с уже построенными шаблонами. Процесс определения диктора, зарегистрированного в системе, по входному речевому сигналу во всех рассматриваемых методах состоит в поиске наиболее подходящей сохраненной модели на основе каких-либо критериев.

В рамках магистерской работы был изучен метод на основе гауссовых смесей для идентификации диктора по голосу.

Модель гауссовых смесей широко используется в области распознавания дикторов. Представляет собой взвешенную сумму M компонент и может быть описана выражением:

$$P(\overline{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\overline{x}) \tag{1}$$

где λ - модель диктора, M - количество компонентов модели, $\overline{x}$ - D-мерный вектор случайных величин, p_i - веса компонентов модели, $b_i(\overline{x})$ - функции плотности распределения составляющих модели:

$$b_i(\overline{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \mid \sum_i \mid^{1/2}} \exp\left\{ -\frac{1}{2}(\overline{x} - \overline{\mu_i})^T \sum (\overline{x} - \overline{\mu_i}) \right\} \tag{2}$$

где $\overline{\mu_i}$ - вектор математического ожидания и $\sum$ – ковариационная матрица. При этом веса смеси удовлетворяют условию:

$$\sum_{i=1}^{M} p_i = 1 \tag{3}$$

Полностью модель гауссовой смеси определяется векторами математического ожидания, ковариационными матрицами и весами смесей для каждого компонента модели:

$$\lambda = \left\{ p_i, \overline{\mu_i}, \sum_i \right\}, i = 1, .., M \tag{4}$$

При использовании данного метода каждый диктор представляется моделью гауссовых смесей λ.

Чаще всего в системах, реализующих данную модель, используется диагональная матрица ковариации. Возможно также использование одной матрицы ковариации для всех компонентов модели диктора или одной матрицы для всех моделей.

Таким образом, для построения модели диктора необходимо определить векторы средних, матрицы ковариации и веса компонентов. Данную задачу решают с помощью ЕМ-алгоритма. На вход подается обучающая последовательность векторов $X = x_1, ..., x_T$. Параметры модели инициализируются начальными значениями и затем на каждой итерации алгоритма происходит переоценка параметров.

Для определения начальных параметров обычно используют алгоритм кластеризации такой, как алгоритм К-средних.

Построив разбиение множества обучающих векторов на М кластеров, параметры модели могут быть инициализированы следующим образом. Начальные значения μ_i совпадают с центрами кластеров, матрицы ковариации рассчитываются на основе попавших в данный кластер векторов, веса компонентов определяются долей векторов данного кластера среди общего количества обучающих векторов.

Для оценки систем идентификации в большинстве случаев ограничива-

ются замкнутым множеством пользователей, то есть все пользователи, проходящие попытку идентификации, зарегистрированы в системе. Результат зависит от количества зарегистрированных пользователей и от размера возвращаемого списка (чаще всего используют только один идентификатор) или от порога включения в список. Вероятность идентификации (истинноположительной идентификации) оценивают как долю попыток идентификации, в результате которых был возвращён список кандидатов, содержащий верный идентификатор.

2 Распознавание произвольной речи

Распознавание речи — это процесс преобразования речевого сигнала в текстовый поток. Системы искусственного интеллекта, распознающие речь, прошли большой путь развития от появления в 1970-х годах до наших дней. Прогресс связан не только с тем, что возникли новые технологии, но и с тем, что появились большие вычислительные мощности и качественные речевые корпусы.

В настоящее время на рынке существует большое количество готовых систем распознавания человеческой речи которые применяются в широком спектре систем - от приложений на смартфонах до систем «Умный дом». Рассмотрим самые популярные из них.

Подавляющее большинство работающих систем являются проприетарными продуктами, т.е. пользователь или потенциальный разработчик не имеет доступа к их исходному коду. Это негативно сказывается на возможности интеграции систем распознавания речи в проекты с открытым кодом. Также не существует какого либо централизованного источника данных, описывающего положительные и отрицательные стороны систем распознавания речи с открытым кодом. В результате возникает проблема выбора оптимальной системы распознавания речи для решения поставленной задачи.

По точности системы сравниваются по наиболее распространенным метрикам: Word Recognition Rate (WRR), Word Error Rate (WER), которые вычисляются по следующим формулам:

$$WER = \frac{S + I + D}{T} \tag{5}$$

$$WRR = 1 - WER \tag{6}$$

где S - число операций замены слов, I - число операций вставки слов, D - число операций удаления слов из распознанной фразы для получения исходной фразы, а - число слов в исходной фразе и измеряется в процентах. По скорости распознавания сравнение было проведено с использованием Real Time Factor - показателя отношения времени распознавания к длительности распознаваемого сигнала, также известного как Speed Factor (SF). Данный

показатель можно рассчитать используя формулу:

$$SF = \frac{T}{T} \tag{7}$$

где T - время распознавания сигнала, T - его длительность и измеряется в долях от реального времени.

Все системы проверялись с применением речевого корпуса WSJ1 (Wall Street Journal 1), содержащего около 160 часов тренировочных данных и 10 часов тестовых данных, представляющих собой отрывки из газеты Wall Street Journal. Данный речевой корпус включает в себя записи дикторов обоих полов на английском языке.

Таблица 1 – Результаты сравнения систем распознавания речи по точности и скорости

Система	WER, %	WRR, %	SF
HTK	19.8	80.2	1.4
CMU Sphinx	21.4	78.6	0.5
Kaldi	6.5	93.5	0.6
Julius	23.1	76.9	1.3
iAtros	16.1	83.9	2.1
RWTH ASR	15.5	84.5	3.8

В рамках магистерской работы использовалась система CMU Sphinx для реализации системы распознавания речи. Такой выбор обусловлен следующими свойствами системы:

— Достаточно высокая точность распознавания
— Высокая скорость распознавания
— Открытый исходный код, лицензией BSD
— Возможность адаптировать акустические модели встроенными инструментами
— Простота интеграции с системами написанными на Python

CMU Sphinx — это дикторонезависимый распознаватель непрерывной речи, который использует Скрытые Марковские модели и n-граммную статистическую языковую модель. Sphinx имеет возможности распознавания продолжительной речи, дикторонезависимый огромный словарь распознавания. Sphinx4 полный и переписанный речевой движок Sphinx, главная цель ко-

торого обеспечить гибкий каркас для исследования в распознавании речи. Sphinx4 написан полностью на языке программирования Java.

CMU Sphinx показывает посредственную точность распознавания и лучшую скорость распознавания из всех рассмотренных. Нужно заметить, что наибольшая скорость распознавания достигается при использовании декодера pocketsphinx, написанного на C.

CMU Sphinx в настоящее время является крупнейшим проектом по распознаванию человеческой речи. В инструментарий входят следующие программы и библиотеки:

- Pocketsphinx — небольшая программа, которая принимает на вход произвольные акустические модели, грамматики и словари, а также звуковой поток(либо звуковой файл, либо сам берет поток с микрофона). На выходе получается распознанный текст. Написана на C, работает быстро
- Sphinxbase — библиотека необходимая для работы Pocketsphinx
- Sphinx4 — гибкая библиотека для распознавания, написана на Java
- Sphinxtrain — программа для обучения акустических моделей

3 Применение системы идентификации диктора в решении задачи распознавания речи

В рамках магистерской работы была реализована система распознавания речи с использованием адаптированных акустических моделей на базе CMU Sphinx, при этом для выбора адаптированной модели в режиме реального времени использовать систему идентификации диктора.

Общие требования к системе:

— Система предоставляет API для идентификации диктора по произвольной речи

— В случае успешной идентификации диктора система предоставляет идентификатор адаптированной акустической модели для идентифицированного диктора, в противном случае предоставляется идентификатор базовой модели

— Система позволяет идентифицировать диктора и распознать текст во входящем звуковом сигнале за $SF \leq 1.2$

— Система предоставляет API по регистрации нового пользователя и адаптации акустической модели

В качестве инструмента реализации системы был выбран язык программирования Python.

Python - высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. Синтаксис ядра Python минималистичен. В то же время стандартная библиотека включает большой объем полезных функций.

Python поддерживает структурное, объектно-ориентированное, функциональное, императивное и аспектно-ориентированное программирование. Основные архитектурные черты — динамическая типизация, автоматическое управление памятью, полная интроспекция, механизм обработки исключений, поддержка многопоточных вычислений, высокоуровневые структуры данных. Поддерживается разбиение программ на модули, которые, в свою очередь, могут объединяться в пакеты.

Python — активно развивающийся язык программирования, новые версии с добавлением/изменением языковых свойств выходят примерно раз в два с половиной года. Язык не подвергался официальной стандартизации, роль

стандарта де-факто выполняет C Python, разрабатываемый под контролем автора языка. Разработчики языка Python придерживаются определённой философии программирования, называемой «The Zen of Python». Её текст выдается интерпретатором Python по команде import this (работает один раз за сессию). Автором этой философии считается Тим Петере (Tim Peters).

Для работы с большими использовалась библиотека pandas. В экосистеме Python, pandas является наиболее продвинутой и быстроразвивающейся библиотекой для обработки и анализа данных. Чтобы эффективно работать с pandas, необходимо освоить самые главные структуры данных библиотеки: DataFrame и Series. Структура/объект Series представляет из себя объект, похожий на одномерный массив (питоновский список, например), но отличительной его чертой является наличие ассоциированных меток, т.н. индексов, вдоль каждого элемента из списка. Такая особенность превращает его в ассоциативный массив или словарь в Python. Объект DataFrame лучше всего представлять себе в виде обычной таблицы и это правильно, ведь DataFrame является табличной структурой данных. В любой таблице всегда присутствуют строки и столбцы. Столбцами в объекте DataFrame выступают объекты Series, строки которых являются их непосредственными элементами.

При необходимости использовать различные математические функции и операции, использовалась высокопроизводительная библиотека NumPy. Библиотека NumPy предоставляет реализации вычислительных алгоритмов (в виде функций и операторов), оптимизированные для работы с многомерными массивами. В результате любой алгоритм, который может быть выражен в виде последовательности операций над массивами (матрицами) и реализованный с использованием ***NumPy***, работает так же быстро, как эквивалентный код, выполняемый в MATLAB.

Адаптированные акустические модели в CMU Sphinx можно использовать как для адаптации базовой модели под особенности произношения диктора так и под особенности микрофона и окружения. Процесс адаптации использует описанные входные данные для улучшения уже существующей модели. Для улучшения модели под особенности произношения диктора рекомендуется использовать не менее 5 минут записанной речи.

Модуль *recognition* разрабатываемой системы позволяет использовать

адаптированные акустические модели дикторов для улучшения качества распознавания речи.

Все имеющиеся в системе модели загружаются в рабочую память при инициализации системы. Каждой модели присваивается уникальный идентификационный номер совпадающий с идентификационным номером диктора. Ниже представлен код инициализации модуля Pocketsphinx с адаптированной акустической моделью.

Каждой модели присваивается уникальный идентификационный номер совпадающий с идентификационным номером диктора. На вход модуль `recognition` принимает обработанный звуковой сигнал и идентификатор. На первом этапе ищется совпадение полученного идентификатора с идентификаторами загруженных моделей. Если совпадение найдено, выбирается модель диктора. Если совпадений нет, выбирается базовая акустическая модель языка. На втором этапе производится распознавание фонем в полученном звуковом сигнале с использованием выбранной акустической модели.

ЗАКЛЮЧЕНИЕ

Самое быстрое и эффективное взаимодействие между людьми происходит посредством устной речи. С помощью речи могут быть переданы различные чувства и эмоции, а главное — полезная информация. Необходимость создания компьютерных интерфейсов звукового ввода-вывода не вызывает сомнений, поскольку их эффективность основана на практически неограниченных возможностях формулировки в самых различных областях человеческой деятельности. Можно утверждать, что уже в скором времени голосовые интерфейсы практически не будут отличаться по надежности от классических способов ввода.

Распознавание речи — одна из самых интересных и сложных задач искусственного интеллекта. Здесь задействованы достижения весьма различных областей: от компьютерной лингвистики до цифровой обработки сигналов. Самыми известными и эффективными алгоритмами распознавания речи являются алгоритмы, построенные на использовании скрытых марковских моделей и нейронные сети. Технологии речевого распознавания нашли свое применение в различных областях. Однако в данной области множество проблем все еще остаются не решенными, многие идеи требуют дальнейшего развития.

В рамках данной работы исследовалась возможность применения систем идентификации диктора и адаптированных акустических моделей для повышения качества распознавания речи. Была реализована система распознавания речи по адаптированным моделям дикторов. Для выбора модели диктора использовалась реализованная в рамках данной работы система распознавания диктора по голосу.