DOI 10.37882/2223–2966.2023.04–2.01

УЛУЧШЕННЫЙ МЕТОД РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ РАЗРАБОТКИ СИСТЕМЫ ГОЛОСОВОГО УПРАВЛЕНИЯ

IMPROVED SPEECH RECOGNITION METHOD FOR DEVELOPING A VOICE CONTROL SYSTEM

E. Amiraslanov
S. Saradgishvili
T. Leontieva

Summary. This paper discusses the issue of creating a voice control system, comparing modern recognition methods, and identifying the most effective and less resource-intensive approaches. Consideration of the issue of effective evaluation of these methods is one of the fundamental ones. The article also presents preliminary results of the developed system.

Keywords: speech recognition, hidden markov models, acoustic model, ASR.

Амирасланов Эльмар Гейдар Оглы
Аспирант, Санкт-Петербургский политехнический университет Петра Великого
elmar0131@gmail.com
Сараджишвили Сергей Эрикович
К.т.н., доцент, Санкт-Петербургский политехнический университет Петра Великого
ssaradg@yandex.ru
Леонтьева Татьяна Владимировна
К.т.н., доцент, Санкт-Петербургский политехнический университет Петра Великого
leontieva_tv@spbstu.ru

Аннотация. В данной работе рассматривается вопрос создания системы голосового управления, сравнение современных методов распознавания, и выявление наиболее эффективных и менее ресурсоемких подходов. Рассмотрение вопроса эффективного оценивания данных методов является одним из основополагающих. В статье так же представлены предварительные результаты разработанной системы.

Ключевые слова: распознавание речи, скрытые марковские модели, акустическая модель, ASR.

Введение

Ежедневно появляются новые возможности по улучшению жизни человека и внедрению в неё новых технологий и систем, которые позволяют сделать её более комфортной. С каждым появлением новых технологий, производители задаются вопросом о новшествах, и одним из актуальных является создание систем распознавания голосовых команд. Важнейшим направлением при реализации автоматизированных систем является добавление голосовых помощников. Эти системы уже пользуются спросом: голосовое управление мобильных телефонов, внедрение в систему умного дома, облегчение жизни людей с ограниченными возможностями и т.д.

Существующее направление активно развивается в настоящее время и с каждым годом выходят новые технологические решения. Отсчет начала появления голосовых помощников ведется с 2011 года. Компании Google и Amazon одними из первых презентовали широкой публике свои продукты: Google Home и Alexa. Они представляли собой небольшой аппарат в виде колонки, который реагирует на обширный диапазон голосовых команд, и исполняет роль посредника между пользователем и системой автоматизации домашних устройств.

Основная цель систем автоматического распознавания речи — преобразовать входной звуковой сигнал с определенной длинной в последовательность слов или символов. Акустические свойства формы сигнала, соответствующего фонеме, могут сильно различаться в зависимости от многих факторов. Рассмотрим наглядный пример работы системы распознавания речи: за основу берется звуковая дорожка, которая делится на семплы и затем вычисляется вектор признаков. Следующим этапом является анализ полученных векторов на уровне трех моделей — языкового, фонетического и акустического, для выявления набора наиболее подходящих комбинаций слов.

Наиболее распространенными методами распознавания речи являются: скрытые Марковские модели, нейронные сети, DTW алгоритмы. Задачи по распознаванию, решенные с помощью алгоритмов нейронных сетей, имеют огромное преимущество по сравнению с другими алгоритмами.

Процесс распознавания речи с использованием нейронных сетей выглядит следующим образом:

1. Акустический препроцессор обрабатывает входной речевой сигнал, и определяет последовательность векторов признаков, для каждого отрезка времени (семпл) и состоят они из спектральных или кепстральных коэффициентов, характеризующих отрезки речевого сигнала.
2. Далее полученные векторы подлежат сравнению с эталонными векторами, содержащимися в моделях слов.
3. С помощью метрик происходит временное выравнивание последовательностей векторов признаков с последовательностями эталонных векторов, образующими модели слов, а также вычисляется мера соответствия для компенсации изменений скорости произнесения и затем находится максимально соответствующее слово.

По мере увеличения размера словаря увеличивается только объем обучающего процесса, для этого нейронной сети придется тратить больше времени на обучение без изменения сложности этого процесса распознавания. Это преимущество позволяет использовать достаточно большое количество слов в словаре. Но у этого подхода есть и недостатки, одним из них является отсутствие внесения дополнений к словарю после окончания процесса обучения. Выход из этой ситуации — теория адаптивного резонанса, которая может быть использована для решения этой проблемы. То есть нейронные сети, построенные в рамках этой теории, позволяют нам сохранять гибкость при запоминании новых нейронных связей и в тоже время позволяют нам не затрагивать уже существующие связи.

Использование же скрытых Марковских моделей подходит для моделирования изменяющихся во времени спектральных векторных последовательностей. Большинство первых созданных систем распознавания речи использовали: СММ для моделирования состояния речи и смеси Гауссовских моделей для наблюдения состояния СММ. Это считалось прорывом в системах распознавания, прежде чем появились нейронные сети. В Марковских моделях для речевых распознаваний имеется два основных момента:

1. Речь может быть разделена на части, которые соответствуют состоянию в СММ, при этом характеристики речи в границах каждого фрагмента является постоянными.
2. Вероятность каждой части зависит только лишь от текущего состояния системы и никак не зависит от предыдущих состояний.

Так же необходимо создать кодовую книгу для реализации распознавания, на основе скрытых Марковских моделей, которая содержит большое количество наборов речевых характеристик. Для решения этой задачи записывают участки эталонной речи, делят на простые компоненты и далее выводятся параметры каждого из характерных признаков. Единому простому компоненту соответствует набор признаков среди большого количества предложенных вариантов в этом словаре.

Часть записанной речи делится на определенные сегменты, в которых речевые характеристики могут считаться неизменными. Затем вычисляют характеристики для всех речевых сегментов и далее выбирают запись кодовой книги с более оптимальными параметрами. Именно эти измерения данных записи формируют конкретную последовательность наблюдений для Марковской модели. Каждому из слов в словаре соответствует единственная последовательность.

Разработанная система

Имеется два подхода при распознавании речи. Первый из них основывается на онлайн распознавание голоса. Большие компании: Google, Amazon, Samsung, Apple, Яндекс развивают рассматриваемый подход, в надежде выпустить свой продукт раньше конкурентов и завоевать рынок. Они используют в своих технологиях нейронную сеть. Ее суть заключается в передаче полученной пользователем фразы на сервер, дальнейшей ее обработкой и отсылкой ответа пользователю. Недостаток этого подхода — наличие постоянного Интернет-соединения.

Другие же компании, которые не имеют тех финансовых возможностей, используют подход попроще, а именно запись ключевых фраз и дальнейшее сравнение с ней проговариваемых пользователем запросов. Основным минусом такого подхода является трудоемкость всех операций с добавлением запросов в базу. Система голосового распознавания также требует и систему голосового вывода, ввиду того что, возвращаемый ответ от системы следует куда-то выводить. Благодаря созданию ответных симплов для пользователя, система работает в разы быстрее и стабильнее.

В качестве движка для системы распознавания речи была использована Kaldi Speech Recognition. С точки зрения алгоритмов и структур данных, применяемых для распознавания речи, вышеупомянутая система предоставляет большое количество современных подходов, таких как использование нейронных сетей и Гауссовых моделей на этапе акустического моделирования и использование конечных автоматов на этапе языкового моделирования. Система имеет модульную

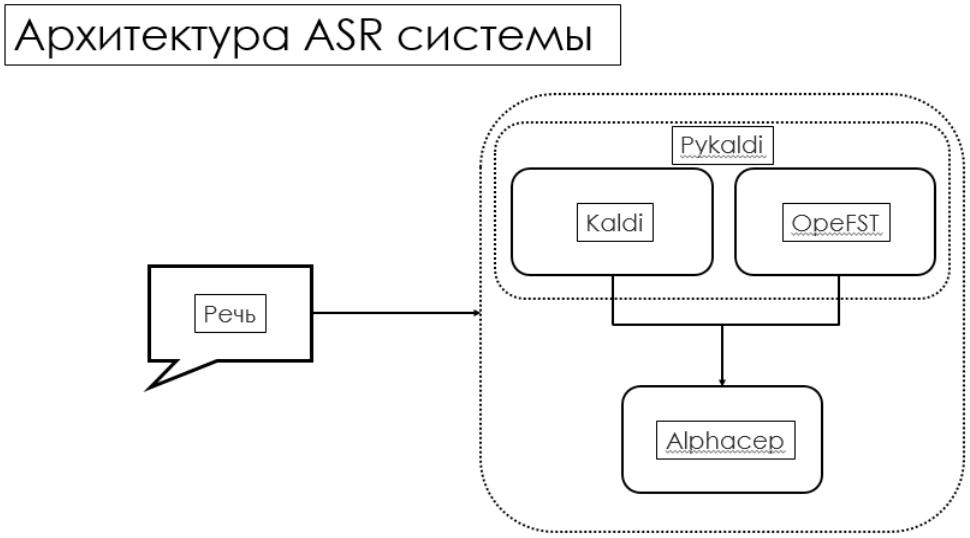

Рис. 1. Архитектура автоматизированной системы распознавания речи

структуру, что позволяет добавлять новый функционал без потери производительности.

Важнейшей задачей голосового модуля является управление системой не с помощью простых диалоговых команд, а использования тех команд, при которых алгоритм будет понимать повседневную речь пользователя. Для этого необходимо реализовать понимание естественного языка. При такой реализации система распознавания преобразовывает естественную речь в текстовый формат и уже в нем находит ключевые фразы.

Kaldi является технологией с открытым исходным кодом, а обеспечивает ей поддержку библиотека OpenFST. Вышеупомянутая библиотека позволяет технологии Kaldi работать эффективно, применяя комплексные решения.

Реализация программного обеспечения проходит на языке Python, в которой был использован скриптовый слой PyKaldi. В качестве распознавания речи при помощи использования акустической и языковой модели была использована русскоязычная система Alphacep. Сама система Kaldi хороша еще и тем, что в ней можно как прописывать отдельные фразовые команда, так и использовать нейронную сеть для улучшения распознавания речи.

Постоянное распознавание работает до тех пор, пока не произойдет соответствие произнесенной ключевой фразы с маркерами фраз, прописанными в базе данных, либо же пока пользователь не даст команду для прекращения работы голосового помощника. После того, как пользователь произнес фразу, голосовой модуль запускает процесс распознавания ключевой фразы и далее в текстовом формате передает ее в модуль выполнения команд. Архитектура системы распознавания представлена на Рис.1

Анализ результатов

Весь модуль голосового управления был подключен к комплектующим системы управления «Умного дома» в офисном помещении. Вся система в офисе состоит из нескольких подсистем: подсистема управления освещения, подсистема управления персональным компьютером, подсистема управления системой видеокамер. Управлять ими можно как с помощью отдельных команд, так и с помощью запуска сценариев, прописанных в системе. Так, для примера, в системе был реализован сценарий настройки рабочей среды. При произношение ключевой фразы, голосовой модуль передавал следующие команды системе: открытие в браузере часто используемых вкладок, снижение яркости экрана до среднего значение, отключение всех уведомление и открытие необходимых для работы программ.

По точности системы будут сравниваться по наиболее распространенным метрикам: Word Recognition Rate (правильно распознанные слова); Word Error Rate (неправильно распознанные слова); Speed Factor (Скорость распознавания) .

Вычисляются метрики по следующим формулам:

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

Таблица 1.

Название технологии	WER,%	WRR,%	SF
Google Speech Recognition	4,3	95,7	0,45
Yandex SpeechKit	8,3	91,7	0,51
My Speech Recognition	6,5	93,5	0,6

$$WRR = 1 - WER$$

где S — число операций замены слов, I — число операций вставки слов

D — число операций удаления слов из распознанной фразы

T — число слов в исходной фразе

$$SF = \frac{T_{\text{расп}}}{T}$$

где $T_{расп}$ — время распознавания сигнала, T — длительность сигнала

Подводя итоги, можно сказать, что все системы показывают результаты примерно на одном уровне (продемонстрированы в Таблице 1), и разработанный голосовой модуль не сильно уступал аналогам из больших компаний. Но при этом Yandex SpeechKit и Google Speech Recognition — являются закрытыми системами, которые работают на чужих серверах и недоступны для модификаций под собственные нужды пользователя. Тем временем созданную систему распознавания можно спокойно своими руками адаптировать под особенности решаемых задач.

Заключение

В результате данной работы была разработана система голосового управления. Описанная система в настоящее время используется в офисном помещение и полностью выполняет поставленные задачи.

Основной упор при реализации был сделан на дистанционное управление с помощью диалоговых команд. В системе реализована функция обратной связи с пользователем, и выполнение каждого запроса сопровождается звуковым подтверждением.

В дальнейшей перспективе предполагается развитие системы с целью ее усовершенствования. Будет вестись работа над созданием и внедрением пользовательского интерфейса в общее приложение управления системой Умного дома, а также над автоматизацией настройки голосового помощника под задачи каждого пользователя.

ЛИТЕРАТУРА

1. Voicehd: Hyperdimensional computing for efficient speech recognition / Imani Mohsen, Kong Dekian, Rosing Tajana // IEEE International Conference on Rebooting Computing.— c: IEEE, 2017.— С. 1–8.
2. Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury Deep Neural Networks for Acoustic Modeling in Speech Recognition — IEEE, Signal Processing Magazine, 2012
3. An Open Source Machine Learning Framework for Everyone // TensorFlow.— [Электронный ресурс] URL: https://www.tensorflow.org/
4. Graves A., Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks // Proceedings of International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013. P. 6645–6649
5. Беленко М.В., Балакшин П.В. Сравнительный анализ систем распознавания речи с открытым кодом // МНИЖ. 2017. № 4–4 (58)
6. Гапочкин, В.А. Нейронные сети в системах распознавания речи. / В.А. Гапочкин // "Science time".— 2014, № 1.— С. 29–36
7. Алимурадов, А.К. Обзор и классификация методов обработки речевых сигналов в системах распознавания речи. / А.К. Алимурадов и [др]. // Измерение. Мониторинг. Управление. Контроль.— 2015, № 2.— С. 27–35
8. Карпов Алексей Анатольевич, Кипяткова Ирина Сергеевна. Методология оценивания работы систем автоматического распознавания речи // Приборостроение. 2012. № 11
9. Документация системы распознавания речи Kaldi. [Электронный ресурс] URL: https://kaldi-asr.org/doc/
10. Использование MQTT протоколов и их предназначение [Электронный ресурс]: URL: https://ipc2u.ru
11. Вишнякова О.А., Лавров Д.Н. Применение преобразования Гильберта-хуанга к задаче сегментации речи // Математические структуры и моделирование. 2011. вып. 24. С. 12–18

Журнал «Современная наука: актуальные проблемы теории и практики»