АНАЛИЗ МЕТОДОВ РАСПОЗНАВАНИЯ ГОЛОСА В ГОЛОСОВЫХ ПОМОЩНИКАХ

Ангапов В.Д.

Ангапов Василий Данилович – старший системный архитектор,
Digital IQ,
г. Москва

Аннотация: *на сегодняшний день формируется целый комплекс задач, решение которых возможно только при использовании средств информационных технологий. Одной из актуальных задач является распознавание голоса в различных голосовых помощниках. Цель текущей статьи состоит в анализе основных методов распознавания голоса. Научная ценность работы заключается в предпринимаемой попытке систематизации знаний относительно вопроса выбора и принципов работы методов распознавания голоса. Материалы статьи могут быть полезны при практической реализации различных голосовых помощников во время задачи выбора определенного метода распознавания голоса.*
Ключевые слова: *информационные технологии, распознавание голоса, ASR, NLP, идентификация речи, естественный язык.*

ANALYSIS OF VOICE RECOGNITION METHODS IN VOICE ASSISTANTS

Angapov V.D.

Angapov Vasily Danilovich – senior system architect,
DIGITAL IQ,
MOSCOW

Abstract: *to date, a whole set of tasks is being form, the solution of which is possible only with the use of information technology tools. One of the urgent tasks is voice recognition in various voice assistants. The purpose of the current article is to analyze the main methods of voice recognition. The scientific value of the work lies in the attempt to systematize knowledge regarding the choice and principles of voice recognition methods. The materials of the article can be useful in the practical implementation of various voice assistants during the task of choosing a certain method of voice recognition.*
Keywords: *information technology, voice recognition, ASR, NLP, speech identification, natural language.*

УДК 004.52

В современном мире наблюдаются устойчивые тенденции, связанные с развитием и интеграцией в различных бытовых и профессиональных сферах жизнедеятельности технологий распознавания голоса. На основе данных инструментов представляется возможность взаимодействия между человеком и цифровым устройством исключительно на основе использования голоса. Распознавание голоса - это процесс определения и идентификации голоса человека или любого звукового сигнала с использованием компьютерных технологий и алгоритмов. Эта технология позволяет преобразовывать аудио-сигналы в текст или другие формы данных, которые могут быть анализированы и использованы для различных целей, таких как автоматическое управление, системы безопасности, биометрическая идентификация и другие [1].

Использование методов распознавания голоса является основой работы современных голосовых помощников, способных упростить и качественно улучшить жизнь современного человека. Наиболее распространенными из примеров использования данных технологий является чтение книг, воспроизведение содержания документов и ряд иных задач, при выполнении которых происходит экономия временных ресурсов человека. Таким образом, технология распознавания голоса широко применяется на сегодняшний день в различных сферах, к примеру:

- Аутентификация. Распознавание голоса используется для идентификации личности в системах аутентификации, таких как системы доступа к компьютерам или мобильным устройствам, банковские системы и прочие системы, где требуется подтверждение личности;
- Телефония. Распознавание голоса применяется в голосовых помощниках на мобильных устройствах и системах голосовой почты. Оно позволяет пользователю установить голосовые команды для взаимодействия с устройством или выполнения определенных функций;
- Клиентский сервис. Распознавание голоса используется в системах автоматического ответа (IVR) для переадресации вызовов и выполнения простых операций, таких как проверка баланса или информации о заказах;
- Медицина. Распознавание голоса может быть использовано для диктования и транскрипции медицинских отчетов или документации;

- Транспорт. Распознавание голоса может быть реализовано в автомобилях и других транспортных средствах для управления различными функциями, такими как навигация, управление мультимедиа и даже управление телефонными вызовами;

- Конференцсвязь. Распознавание голоса используется для автоматического распознавания и транскрибирования речи во время конференцсвязи, что помогает повысить эффективность и качество коммуникации.

При этом представлены не все области, где применяется распознавание голоса. С каждым годом его применение становится все более широким. Помимо этого, в зависимости от используемых методов распознавания голоса, активно разрабатываются интеллектуальные системы обучения, способные принимать экзамен и помогать в изучении иностранных языков. При этом использование голосовых помощников получает свое развитие и в решении наиболее сложных задач из профессиональных сфер человеческой жизнедеятельности. Примерами данного применения являются – идентификация личности, судебная экспертиза, помощь людям с ограниченными возможностями и ряд иных задач [2].

Работа голосовых помощников возможна только на основе использования методов распознавания голоса. На сегодняшний день существует два основных направления развития по распознаванию речи – автоматическое распознание речи (Automatic Speech Recognition, ASR) и обработка естественного языка (Natural Language Processing, NLP). При этом необходимо отметить, что каждый из данных методов имеет уникальные особенности, преимущества и недостатки [3].

Метод автоматического распознавания речи представляет собой совокупность компьютерного оборудования и программных технологий, выполняющих прямую идентификацию и обработку человеческого голоса. Принцип работы данной технологии может быть определена в качестве автоматической транскрипции разговорного языка в читаемый текст. При этом распознавание голоса происходит в режиме реального времени на основе заранее заданных звуковых шаблонов. Таким образом, метод ASR представляет компьютеру возможность выявить слова из человеческой речи и перевести их в электронный текст [4].

На рис. 1 представлена блок-схема работы данного метода на примере голосового навигатора:

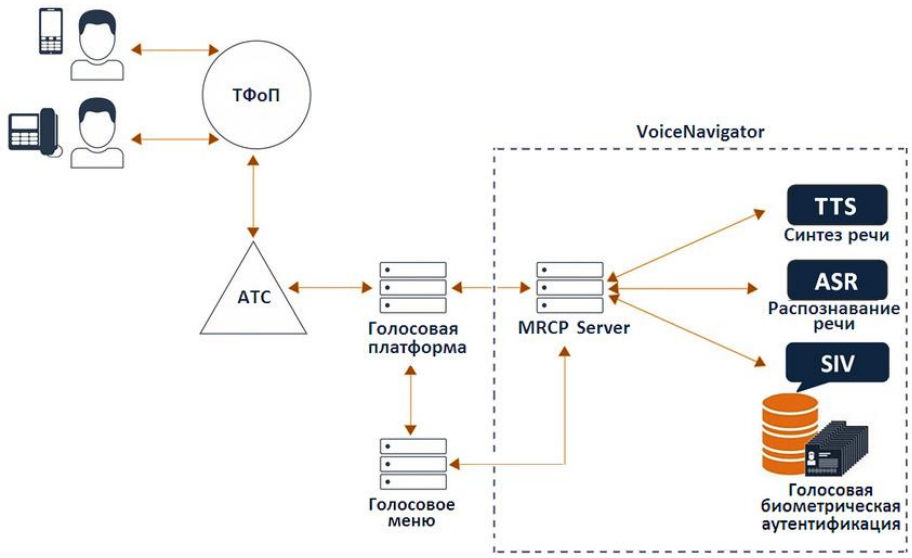

Рис. 1. Блок-схема работы VoiceNavigator на основе ASR.

В процессе работы ASR системы обычно выполняются следующие шаги:

1. Захват звукового сигнала. Сначала аудио-данные, включающие речевой сигнал, захватываются при помощи микрофона или другого устройства записи звука;

2. Предобработка сигнала. Захваченный звуковой сигнал может содержать шумы, эхо, искажения и другие нежелательные факторы. Для улучшения качества сигнала производится его предварительная обработка, которая может включать в себя шумоподавление, фильтрацию и нормализацию аудио;

3. Извлечение признаков. Затем из предобработанного звукового сигнала извлекаются акустические признаки. Это может быть выполнено с использованием алгоритмов, таких как Mel-Frequency Cepstral Coefficients (MFCC) или другие методов фурье-преобразования;

4. Распознавание фонем или слов. С полученными признаками происходит процесс распознавания речи. Возможны два основных подхода: основанный на фонемах и основанный на словах. При подходе,

основанном на фонемах, ASR система пытается распознать фонемы, минимальные звуковые единицы речи. При подходе, основанном на словах, система пытается распознать целые слова;

5. Применение языковой модели. Для улучшения точности распознавания ASR система применяет языковую модель, которая учитывает вероятность сочетания слов для определенного языка или предметной области. Языковая модель помогает системе выбрать наиболее вероятные последовательности слов;

6. Генерация текста. Наконец, ASR система генерирует текстовый выход на основе распознанных слов и языковой модели [5].

Основное использование данного метода наблюдается в задачах распознавания слов для идентификации речи человека. Автоматическое преобразование произнесенных слов в электронный вид позволяет применять данную технологию в не только различных задачах распознавания, но и в качестве голосового помощника для людей с ограниченными возможностями. Технологи работы данного метода основывается на предварительно настроенных или же сохраненных в компьютерной программе речи, образцов и словаря [6].

Особое внимание заслуживает инновационный метод распознавания голоса NLP – обработка естественного языка. Данный метод представляет собой актуальное направление из области машинного обучения, в котором обработка голоса производится на основе интеллектуальных алгоритмов. При этом один из вариантов процесса распознавания голоса в голосовом помощнике на основе NLP может выглядеть таким образом: запись речи человека; преобразование машиной слов из аудио в электронный текст; разбор текста на основные составляющие для понимания контекста беседы и целей человека; по результатам работы система определяет команду на выполнение [7].

На рис. 2 представлена схема работы распознавания голоса методами обработки естественного языка:

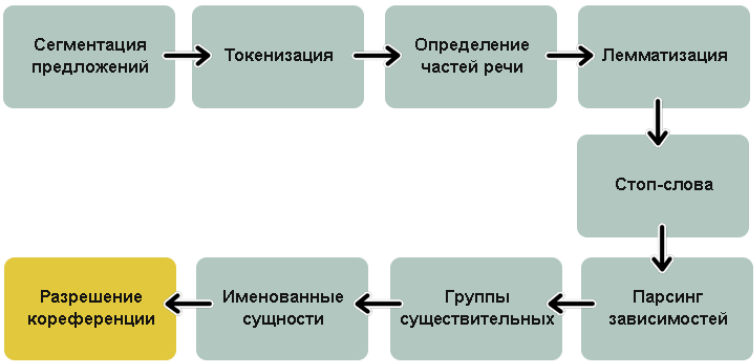

Рис. 2. Схема обработки естественного языка.

Процесс работы NLP включает в себя несколько шагов:

1. Токенизация. Первым шагом является разбиение текста на отдельные слова, фразы или другие единицы смысла, называемые токенами.

2. Лемматизация и стемминг. Затем происходит преобразование слов к их базовым формам. Лемматизация сводит слова к их леммам (например, «бежать» становится «бежать»), тогда как стемминг отсекает окончания слов (например, «бежал» становится «бежа»).

3. Часть-речная разметка. После преобразования слова каждый токен помечается соответствующей частью речи (существительное, глагол, прилагательное и т. д.).

4. Синтаксический анализ. Затем происходит анализ синтаксиса предложений, чтобы определить связи между словами и их ролями в предложении.

5. Смысловая аналитика. Последний шаг заключается в понимании смысла текста. Это может включать анализ семантики (значения слов и фраз), анафорических связей (замены местоимений на соответствующие существительные) и реализацию знаний о мире для интерпретации содержания текста [8].

Для реализации этих шагов NLP использует различные методы и алгоритмы, включая машинное обучение, глубокое обучение и статистический анализ. Это позволяет компьютеру понимать, интерпретировать и отвечать на текстовую информацию с помощью комплексной обработки. Сейчас NLP применяется в различных областях, таких как машинный перевод, распознавание речи, чат-боты, анализ социальных медиа и других [9].

Необходимо отметить, что до появления данного метода алгоритмы распознавания речи имели набор действий только на определенные конструкции из слов. Это в большей степени является не распознаванием речи, а реагированием на определенный набор символов. В свою очередь, обработка естественного языка имеет совершенно иной подход. Алгоритмы NLP имеют возможность обучения словам, их значениям, а также структуре фраз и общей логики внутри языка. Именно в результате работы данных алго-

ритмов появляется возможность понимания контекста и более качественного распознавания голоса человека. Данный метод получает свое активное развитие повсеместно. Так, к примеру, алгоритмы NLP уже используются в поиске Google или Яндекс, в чат-ботах, виртуальных ассистентах и других направлениях. В частности, использование рассматриваемого метода наблюдается в науке и решении коммерческих задач на основе создания «умных» систем, обрабатывающих естественный человеческий язык [10].

В табл. 1 сведены основные особенности, сферы использования, преимущества и недостатки рассмотренных методов распознавания голоса:

Таблица 1. Сравнение методов ASR и NLP.

	Автоматическое распознавание речи	**Обработка естественного языка**
Преимущества	- снижение затрат за счет автоматизации - высокая точность распознавания слов - распознавание в режиме реального времени	- возможность понимания контекста - возможность повсеместного использования - обработка естественного человеческого языка
Недостатки	- узкий круг решения задач - невозможность обработки естественного языка	- долгое распознавание и обработка голоса - дороговизна использования систем
Особенности	Имеет узкий круг применения относительно других методов, однако показывает максимальную точность распознавания голоса	Требует проработку с целью создания более быстрых алгоритмов распознавания и анализа голоса
Области использования	Применим к использованию в технологии чат-бота, помощника для людей с ограниченными возможностями, идентификация человека	Извлечение и анализ информации в задачах государственного уровня и бизнеса, виртуальные ассистенты, умные системы поиска

Automatic Speech Recognition и Natural Language Processing - это разные методы обработки речи. ASR используется для распознавания голоса и преобразования его в текст. Он обычно основывается на различных алгоритмах и моделях машинного обучения, которые обрабатывают акустический сигнал и пытаются правильно распознать произнесенные слова.

NLP, с другой стороны, предназначен для обработки и понимания естественного языка. Он может включать в себя различные задачи, такие как семантический анализ, синтаксический анализ, анализ тональности и другие. NLP позволяет понимать и интерпретировать текст, включая текст, полученный из распознавания голоса ASR. Так, ASR и NLP - это взаимосвязанные методы, где ASR используется для преобразования голоса в текст, а NLP - для понимания этого текста и выполнения задач анализа естественного языка. Оба метода являются важными и эффективными в своих сферах применения [11].

Необходимо отметить, что основной характеристикой данных решений является точность распознавания. Для рассматриваемых инструментов данная величина исходит из метрики Word Error Rate (WER) - процент ошибок в распознанном тексте по сравнению с эталонным текстом.

На сегодняшний день имеется целый ряд основных игроков на рынке по поставке программного обеспечения для распознавания голоса. Выделяются как программные, так и облачные решения. Так, некоторыми ключевыми программными решениями являются:

- Kaldi. Это открытая платформа для распознавания речи, которая предлагает широкий спектр инструментов и алгоритмов для обучения моделей распознавания речи. Она используется как академическими исследователями, так и индустрией. WER – 5-10%;

- Yandex SpeechKit. Это онлайн-сервис звукового анализа от компании Яндекс для реализации распознавания речи на основе программных алгоритмов машинного обучения в любых бизнес-приложениях. WER – 5-15%;

- IBM Watson Speech to Text. Это сервис распознавания речи от IBM, который можно использовать для преобразования речи в текст. Поддерживает несколько языков и может быть интегрирован в различные приложения и системы. WER – 10-15%.

Из облачных сервисов наиболее распространенными являются:

- Google Cloud Speech-to-Text. Это развитое API распознавания речи от Google, которое предоставляет возможность преобразования речи в текст. Оно может обрабатывать большие объемы аудиоданных и поддерживает несколько языков. WER 15-20%;

- Microsoft Azure Speech to Text. Это платформа для распознавания речи от Microsoft, которая позволяет преобразовывать аудиофайлы в текст. API поддерживает множество языков и предоставляет возможность настраивать модель для достижения более высокой точности. WER – 10-15%.

Как видно, наилучший результат распознавания имеет программное решение Kaldi. Точность распознавания речи с использованием Kaldi может быть высокой, но зависит от различных факторов, таких как качество аудиозаписи, язык речи, размер обучающего набора данных, использование аккуратной акустической и языковой моделей, а также от настройки параметров алгоритма распознавания речи. В ряде задач Kaldi достигает высокой точности распознавания, сравнимой с другими современными системами. Однако, для достижения наилучших результатов, возможно потребуется провести настройку и оптимизацию модели для конкретного случая.

Таким образом, основной целью представленной статьи являлось рассмотрение и анализ методов распознавания голоса в голосовых помощниках. По результатам работы определена актуальность развития методов распознавания голоса и роль использования голосовых помощников. Также представлены результаты анализа и сравнения таких основных методов распознавания голоса, как автоматическое распознавание речи (ASR) и обработка естественного языка (NLP).

В заключение необходимо отметить, что каждый из рассмотренных методов имеет индивидуальные особенности, а также ряд преимуществ и недостатков своего использования в голосовых помощниках. Так, несмотря на видимое отставание в эффективности и качестве обработки голоса, метод ASR имеет высокую значимость в задачах по разработке голосовых помощников для людей с ограниченными возможностями. При этом интеллектуальные методы распознавания находят свое применение в решении задач государственного и частного секторов бизнеса, однако нуждаются в продолжении своего развития и создании более быстрых методов распознавания голоса [12].

Список литературы / References

1. *Хеин М.З.* Современное состояние проблемы обработки, анализа и синтеза речевых сигналов // Computational nanotechnology. 2018.
2. *Хлопенкова А.Ю., Белов Ю.С.* Методы обработки естественного языка в виртуальных голосовых помощниках // E-Scio. 2019.
3. *Морозова А.А.* Речевой портрет голосового помощника «Алиса» // Вестник ЧелГУ. 2021.
4. *Малиновкин В.А., Валуйских Н.В., Шведов Н.Н., Кенин С.Л., Гребенникова Н.И.* Сравнительный анализ средств голосового интерфейса и технологий распознавания речи // Вестник ВГТУ. 2022.
5. *Abougarair A.J.* Design and implementation of smart voice assistant and recognizing academic words // International Journal of Robotics and Automation. 2022.
6. *Ермоленко Т.В., Пикалёв Я. С.* Система автоматического распознавания слитной русской речи на основе глубоких нейросетей // Речевые технологии/Speech Technologies. 2021.
7. *Цитульский А.М., Иванников А.В., Рогов И.С.* NLP - обработка естественных языков // StudNet. 2020.
8. *Валуйцева И.И., Филатов И.Е.* Разработка программы ASR для распознавания вариантов русского языка // Полилингвиальность и транскультурные практики. 2021.
9. *Khurana D., Koli A., Khatter k., Singh S.* Natural language processing: state of the art, current trends and challenges // Multimedia Tools and Applications. 2022.
10. *Пикалёв Я.С.* Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов // Проблемы искусственного интеллекта. 2020.
11. *Кудрявцев Н.Д., Семенов Д.С., Кожихина Д.Д., Владзимирский А.В.* Технология распознавания речи: результаты опроса врачей-рентгенологов Московского референс-центра лучевой диагностики // ОРГЗДРАВ: Новости. Мнения. Обучение. Вестник ВШОУЗ. 2022.
12. *Плешакова Е.С., Гатауллин С.Т., Осипов А.В., Коротеев М.В., Ушакова Ю.В.* Распознавание эмоций человека по голосу в борьбе с телефонным мошенничеством // Национальная безопасность / nota bene. 2022.