**Анализ и распознавание голоса человека**

И.В. Луковкин

Научный руководитель – Устюков Д.И., канд. техн. наук, старший преподаватель

В современном мире повсеместно используется технология распознавания речи: люди общаются с голосовыми ассистентами (например, “Алиса” от компании “Яндекс”) и получают ответы от искусственного интеллекта на интересующие их вопросы; в разговоре по мобильной сети нередко диалог происходит не с человеком по ту сторону линии передачи, а со специальной программой, которая помогает пользователю в случае обращения (наглядный тому пример - голосовой помощник Олег от компании “Тинькофф”) и т.д.

Вышеперечисленные продукты являются примерами использования технологии распознавания речи (от англ. Speech-to-text), которая появилась еще в конце XX века, но широкое применения получила лишь в начале 2000-х годов в ходе совершенствования вычислительной техники и искусственного интеллекта.

Принцип анализа речи представляет собой обширный алгоритм, состоящий из множества процессов обработки исходных данных. Через записывающее устройство голос сохраняется в памяти ЭВМ. Следом при использовании специальной нейросети запись проходит анализ, в результате которого исходный голос разделяется на фрагменты - фонемы. Далее каждый участок сравнивается нейросетью на предмет схожести с предварительно загруженными шаблонами. Так происходит процесс распознавания в материале букв, слогов и слов. Если некоторые фонемы были разобраны некорректно, искусственный интеллект обращается к библиотеке, состоящей из обширного словаря. Здесь идет поиск слов, которые наиболее точно совпадают с обрабатываемым участком, и найденные объекты заменяются в исходном тексте. После того, как все фонемы были обработаны, ИИ приступает к следующему этапу: контекстному преобразованию - степень обученности нейросети позволяет преобразовать отдельно распознанные фразы в словосочетания, предложения и, в конечном итоге, текст. Эта стадия работы отличается от предыдущей лишь используемыми библиотеками. Здесь сравнение данных происходит с библиотеками, состоящими уже не из слов, а предложений, текстов. В процессе сопоставления исходный текст меняет свой вид: слова меняют свои окончания в соответствии с правилами грамматики.

В настоящее время уже существует такой ИИ, который способен подбирать по контексту не только фонемы, но и целые слова, исходя из того, какие фразы находятся перед неизвестным словом и после него. Кроме того, современные нейросети способны в распознанном тексте грамотно расставить знаки препинания. В результате обработки речи мы получаем не просто набор высказываний, а целостный текст, состоящий из точек, запятой, тире, двоеточий и т.д.

Особенность каждой нейросети заключается в ее постоянном совершенствовании в процессе работы. Каждый раз, распознавая новое высказывание, в библиотеку шаблонов заносятся новые примеры фонем и словосочетаний. Так при последующем обращении к машине искусственный интеллект анализирует данные, схожие с предшествующими быстрее.

Преимущество использования в рассматриваемой технологии такой нейросети состоит и в том, что в ее библиотеки загружаются не только слова и тексты книжного стиля, но и высказывания, которые чаще всего используются в разговорной речи. Таким образом технология способна подстроиться под каждого пользователя индивидуально, поэтому работа с обладателем такого ИИ с каждым разом будет эффективнее.

Метод распознавания речи на момент написания статьи уже используется во многих сферах человеческой деятельности. Во многих программах межъязыкового перевода текста интегрировано использование технологии распознавания речи для коммуникации людей, имеющих языковой барьер; вышеуказанный голосовой помощник позволяет людям с ограниченными возможностями пользоваться современными инновациями в сфере информационных технологий при использовании одного лишь источника ввода информации - звукозаписывающего устройства и т.д.

Широкое распространение метод распознавания человеческой речи получил и в процессе идентификации человека. Наряду с формой и характерными чертами лица и отпечатками пальцев человеческий голос достаточно уникален, чтобы провести распознавание личности.

При первоначальной настройке идентификации в память компьютера сохраняются индивидуальные характеристики записанного голоса: его высота, громкость и скорость речи. Эти характеристики в свою очередь формируют набор признаков, по которому происходит идентификация. Так при дальнейшем использовании такого метода входной сигнал сравнивается нейросетью с первоначальной записью на предмет совпадения характерный черт.

Безусловно, такой метод имеет явные недостатки. Голос человека может подвергаться изменениям из-за болезней, эмоциональных факторов и возраста, поэтому возникают случаи, когда актуальность записанных первоначальных данных пропадает, и вместе с тем использование этой технологии для подтверждения личности становится нецелесообразным. Кроме того существует вероятность того, что запись речи может быть осуществлена на технику ненадлежащего качества, из-за чего качество обрабатываемой записи может быть хуже, что в будущем явно отразится на процессе идентификации.

Помимо указанных факторов риска необходимо учесть и то, что современные нейросети способны преобразовывать пользовательский голос: менять его тембр, скорость и громкость. Такие технологии позволяют сторонним лицам изменять свой голос и создавать речь, похожую на ту, которая используется в идентификации, и получать доступ к конфиденциальным данным.

Для повышения безопасности и уменьшения рисков многие сервисы при регистрации просят отвечать на заранее подготовленные вопросы, чтобы усложнить процесс взлома злоумышленникам. Помимо этого просят ввести надежный пароль и письменный ответ на заданные вопросы. Это позволяет системе спросить у пользователя текстовые данные в случае непрохождения процесса распознавания по голосу.

Как и все другие нейросети, искусственный интеллект, применяемый в процессе распознавания человеческой речи и идентификации человека по голосу, совершенствуется с каждым днем, и вместе с тем улучшается взаимодействие человека со средствами вычислительной техники.

Одним из бесплатных API для распознавания голоса человека является Google Cloud Speech-to-Text API. Это API идеально подходит для обработки аудиофайлов, содержащих речь, и его можно использовать для распознавания речи на более чем 120 языках и диалектах.

Одной из главных особенностей этого API является то, что оно достаточно точно распознает голос на фоне шума или других раздражающих факторов. Более того, Google Speech-to-Text API имеет возможности для обработки больших объемов данных, что делает его идеальным для использования в индустриальном масштабе.

Для использования Google Cloud Speech-to-Text API потребуется зарегистрироваться в Google Cloud Platform и создать проект, после чего можно будет сгенерировать ключ API и начать работу. Работа с Google Speech-to-Text API является достаточно простой и требует всего лишь отправки POST-запросов на API с помощью различных языков программирования.

Преимущества Google Cloud Speech-to-Text API:

1. Высокая точность распознавания голоса. Google Speech-to-Text API использует алгоритмы машинного обучения и нейросетей, что позволяет достичь точности распознавания высокого уровня.

2. Поддержка большого количества языков и диалектов. Google Speech-to-Text API поддерживает распознавание речи на более чем 120 языках и диалектах.

3. Хорошая работа на фоне шума. Благодаря применению алгоритмов машинного обучения и нейросетей, Google Speech-to-Text API может эффективно обрабатывать аудиофайлы с фоновым шумом.

4. Удобная работа с большими объемами данных. Google Speech-to-Text API может обрабатывать большие объемы данных, что делает его идеальным для использования в промышленном и коммерческом масштабе.

Недостатки Google Cloud Speech-to-Text API:

1. Ограниченный бесплатный доступ. Хотя API бесплатен при небольших объемах обработки данных, при работе с большими объемами данных потребуется оплачивать использование.

2. Проблемы с интеграцией сторонних приложений. Некоторые разработчики отмечают некоторые сложности при интеграции Google Speech-to-Text API с собственными приложениями.

3. Требуется наличие стабильного интернет-соединения. Работа с Google Speech-to-Text API требует наличия стабильного интернет-соединения, что может быть проблематично при работе в условиях ограниченной доступности интернета.

Еще одним бесплатным API для распознавания голоса на основе open source является Mozilla DeepSpeech. Это библиотека глубокого обучения, разработанная Mozilla, которая позволяет выполнять распознавание речи на основе моделей нейронных сетей. Она доступна для использования бесплатно и поддерживает несколько языков. Это решение позволяет обработать как аудио файлы, так и потоковое аудио. Для запуска DeepSpeech необходимо установить соответствующий пакет на свой компьютер или сервер и подключить к API. При этом разработчикам доступен полный исходный код библиотеки, что позволяет модифицировать ее под нужды конкретного проекта. Также следует учитывать, что точность распознавания голоса с DeepSpeech несколько меньше, чем у коммерческих решений, но совершенствование сетей может улучшить этот показатель. В целом, Mozilla DeepSpeech - это мощное и гибкое решение для распознавания голоса, которое может быть использовано бесплатно в проектах с открытым исходным кодом.

Одним из главных преимуществ Mozilla DeepSpeech является то, что это решение open source, то есть пользователи могут бесплатно скачать и использовать его, модифицируя под себя. Благодаря этому, библиотека имеет большое сообщество разработчиков, которые постоянно работают над ее улучшением. Кроме того, Mozilla DeepSpeech позволяет работать с аудиофайлами и потоковым аудио, а также поддерживает несколько языков, что обеспечивает большую гибкость. Точность распознавания речи Mozilla DeepSpeech находится на весьма высоком уровне, так что среди недостатков можно указать только некоторое снижение точности по сравнению с другими коммерческими решениями. Кроме того, запуск Mozilla DeepSpeech на компьютере может требовать дополнительных вычислительных ресурсов, но этот недостаток можно компенсировать есть возможность использовать облачные вычисления. В целом, Mozilla DeepSpeech представляет собой гибкое и бесплатное open source решение для распознавания голоса, которое можно использовать в самых разных проектах.

*Библиографический список*

1. Как работает распознавание речи и где его можно использовать: [Электронный ресурс] // URL:<https://www.mango-office.ru/newsletter/kak-rabotaet-raspoznavanie-rechi>
2. Распознавание речи: очень краткий вводный курс: [Электронный ресурс] // URL:<https://habr.com/ru/company/toshibarus/blog/490732/>
3. Идентификация по голосу: [Электронный ресурс] // URL:<https://www.biolink.ru/technology/voice.php>
4. Биометрия: [Электронный ресурс] // URL:<https://cs.hse.ru/mirror/pubs/share/473209293.pdf>