**УДК 004.62**

# РЕШЕНИЕ ЗАДАЧИ STT С ПРИМЕНЕНИЕМ НЕЙРОННЫХ СЕТЕЙ

**Курочкин Ю.И.** (Университет ИТМО, Санкт-Петербург), **Крапивин С.Р.** (Университет ИТМО, Санкт-Петербург)
**Научный руководитель – кандидат культурологии, доцент Пучковская А.А.**
(Университет ИТМО, Санкт-Петербург)

В данной работе представлены ASR-алгоритмы, целью которых является решение задачи STT. Объектом исследования являются видео и аудио данные, которые впоследствии трансформируются в текстовый формат, для дальнейших манипуляций. Реализованные решения направлены на упрощение анализа текстовых данных.

**Введение.** В 21 веке мы живём в реальности, в которой мы всё чаще получаем информацию из аудио и видеоисточников. Такая тенденция вполне объяснима, ведь появление Интернета дало человеку возможность потреблять объёмы информации, ограниченные лишь возможностями самого человека. Логичным способом преодоления таких ограничений стало донесение информации в аудио и видео форматах. Таким образом, экономилось время в виду простоты восприятия вышеуказанных форматов в сравнении с текстовыми. В тоже время росли возможности машинного анализа данных, а поскольку на этот момент значительная часть данных уже была представлена в аудио и видео форматах, то встал вопрос о трансформации таких данных в текстовый формат для решения различных задач, связанных с анализом текстов. Актуальность данной работы заключается в применении передовых технологий машинного обучения для непосредственной реализации алгоритма, который был бы способен выполнять задачу STT, а именно извлекать текстовую информацию из исходного видеофайла.

**Основная часть.** Целью работы является реализация алгоритмов машинного обучения, которые были бы применимы к задаче преобразования видео и аудио информации в текст. В первую очередь возникла необходимость в детальном изучении сферы NLP, а именно задачи извлечения текстовой информации из аудио (STT), в результате чего встал вопрос о выборе конкретной технологии. Выбор пал на библиотеку Vosk, которая успешно реализована на многих языках, а в частности на Python. Vosk является автономным инструментом для распознавания речи с открытым исходным кодом. Он обеспечивает возможность распознавания речи для более, чем 20 языков и диалектов. Для непосредственной работы выбранной модели распознавания речи «vosk-model-ru-0.10» необходимо указать путь к исходному видеофайлу с вызовом класса VideoFileClip библиотеки moviepy.editor для дальнейшей конвертации.

Был реализован конвертер исходного видеофайла в аудио формат при помощи библиотеки moviepy.editor, класса AudioFileClip и метода write_audiofile, который возвращает аудиофайл в формате моноканального wav при помощи параметра ffmpeg_params=["-ac", "1"].

Далее, для работы программы, необходимо скачать и распаковать модель, которая будет содержать каталоги am, conf, graph и другие, и указать на неё путь. Стоит упомянуть, что для локальной распаковки модели (для запуска на локальной машине) реализован небольшой скрипт при помощи модуля zipfile и метода ZipFile.extractall(), которые извлекает все элементы архива в текущий рабочий каталог.

Далее происходит открытие аудиофайла, которые появляется в результате работы реализованного конвертера и проверка соответствия формату моноканального wav.

Следующим этапом является загрузка модели и последующий запуск распознавания текстовой информации в аудиофайле.

Для вывода конечного ответа был реализован скрипт, который извлекает распознанный текст из финального отчёта работы модели, который генерируется методом FinalResult и формируется результатирующий файл с названием «result.txt», который сохраняется в той же директории, в которой находится сама модель.

В качестве ещё одного алгоритма для решения задачи STT была выбрана библиотека Speech Recognition и модель, обученная Google. Данная библиотека является пакетом, позволяющим строить различные сценарии для обработки аудиоданных. По сути это лишь оболочка для вызова существующих речевых API. Именно данный принцип работы даёт пакету ряд преимуществ, таких, как гибкость, скорость и простота. Далее, важно уделить вниманию классу Recognizer. Именно этот класс отвечает за распознавание текста из аудиоданных и определяет API, который мы будем использовать в нашем случае. Вышеописанный класс имеет 7 методов для анализа. Итак, в силу простоты эффективности и простоты использования, выбор был сделан в пользу метода recognize_google() на основе Google Web Speech API.
Также, в качестве третьего алгоритма, задача которого перекликается с основной (STT), был выбран Tesseract OCR, суть которого заключается в распознавании вшитых в видео субтитров и преобразовании в текст. Результатом является алгоритм, который может быть успешно использован для обработки видеоинформации с целью получения её текстового представления, которое в дальнейшем может быть использовано для решения широкого спектра задач, а в частности nlp.

**Выводы.** Итоги работы включают в себя успешную реализацию конвертеров видео в аудио, а также применение алгоритмов машинного обучения для автоматического распознавания речи и извлечения текста из видеофайлов.

Курочкин Ю.И. (автор) Подпись

Пучковская А.А. (научный руководитель) Подпись