ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

DOI: 10.18137/RNU.V9187.23.04.P.87 УДК 004.93

Д.А. Елизаров

РАЗРАБОТКА СИСТЕМЫ ТРАНСКРИБАЦИИ АУДИО- И ВИДЕОКОНТЕНТА

Аннотация. Проведен анализ существующих систем транскрибации аудио- и видеоконтента. Разработка системы транскрибирования позволит повысить эффективность работы сотрудников, обеспечить текстовыми данными и визуализацией информации. Существующие системы на данном этапе не позволяют с высокой точностью проводить распознавание аудио- и видеофайлов, особенно при работе с файлами с посторонними шумами.

Ключевые слова: транскрибация, модель, Python, система, архитектура.

D.A. Elizarov

DEVELOPMENT OF A TRANSCRIPTION SYSTEM FOR AUDIO AND VIDEOCONTENT

Abstract. The article provides an analysis of existing systems for transcribing audio and video content. The development of a transcription system will improve the efficiency of employees, provide text data and information visualization. The existing systems at this stage do not allow high-accuracy recognition of audio and video files, especially when working with files with extraneous noise.

Keywords: transcription, model, Python, system, architecture.

Введение

В современном мире технологий, когда необходимо быстро получить и обработать информацию, порой не хватает времени, чтобы прослушать длинное голосовое сообщение или просмотреть видеоролик; гораздо проще прочитать текстовое сообщение и выделить для себя важные моменты из прочитанного. Текстовый формат представления информации предпочтительнее, например, когда находишься в общественных местах без наушников, на работе, на учебе и др. Стоит также отметить, что видео с субтитрами и текстовым описанием набирают больше просмотров, так как удобны для пользователей [1].

Текстовая расшифровка аудио- и видеозаписей абсолютно необходима для специалистов, любителей и обычных пользователей аудио- и видеоконтента. Наиболее широкопризнанным видом транскрипции является преобразование источника разговорного языка в текст, например, в компьютерную запись, подходящую для печати в виде документа, отчета [2].

Можно переводить, записывая слова вручную, слушая текст, останавливать запись, перематывать назад, а можно запись преобразовать в текст автоматически, при этом важно сохранить смысл и авторскую подачу. Ручная транскрипция – это процесс набора произносимых слов так, как они слышатся. Хотя этот метод может занять много времени и быть

[1] Исследование выполнено за счет гранта РНФ (проект № 22-28-20359), https://www.rscf.ru/project/22-28-20359/.

Елизаров Дмитрий Александрович
кандидат технических наук, доцент, доцент кафедры автоматики и систем управления. Омский государственный университет путей сообщения, город Омск. Сфера научных интересов: разработка и верификация программного обеспечения, вопросы обеспечения информационной безопасности. Автор более 60 опубликованных научных работ. ORCID: 0000-0001-7326-9674, AuthorID:78813, SPIN-код: 9739-5327.
Электронный адрес: elizarovdaib@gmail.com

неточным, он предлагает некоторые преимущества, такие как доступность и точность. Автоматическая транскрипция использует программное обеспечение для преобразования аудиофайлов в текстовый формат и предлагает несколько преимуществ по сравнению с ручной транскрипцией, включая скорость, точность и экономичность [3].

Задача программ по преобразованию звука в текст – грамотно перевести и отредактировать аудио- или видеофайл в текстовый формат: убрать слова-паразиты, неправильные фразы, паузы, шум [4]. Готовый текст должен состоять из логичных фраз, построенных по правилам русского языка. Контент для Интернета все чаще представлен в виде видео и подкастов, поэтому возможность аудиотранскрипции становится все более важной и актуальной задачей [5].

Разработка системы транскрибирования позволит повысить эффективность работы сотрудников, обеспечить текстовыми данными и визуализацией информации. Существующие системы на данном этапе не позволяют с высокой точностью проводить распознавание аудио- и видеофайлов. Развитие данной области может привести к новым значимым открытиям.

Анализ существующих сервисов транскрибации

Существует достаточное количество сервисов и сайтов для перевода звука в текст. Рассмотрим самые популярные из них, определим их достоинства и недостатки.

Онлайн-сервис GoogleDocs предназначен для работы с текстом и данными. Внутри платформы можно включить микрофон, который поможет перевести речь в текстовый формат. Чтобы текст перевелся качественно, нужно говорить четко и громко, иначе при медленной или тихой диктовке сервис не распознает речь. GoogleDocs плохо справляется с тихой и шумной диктофонной записью. Главный недостаток – сервис работает только в активном окне GoogleDocs, то есть наговорить что-то с другой вкладки или включить запись на компьютере не получится. Чтобы добавить в текст знаки препинания, нужно произносить их словами, что не совсем удобно для пользователя. Ошибки при распознавании необходимо править вручную.

Онлайн-инструмент Speechpad позволяет расшифровать голосовую запись. В десктопной версии голосовой ввод возможен только в браузере Chrome, также разработаны приложения под Android и iOS. Расшифровка возможна при наличии хорошего микрофона. Можно загрузить свои файлы и воспользоваться функцией «Транскрибация», можно расшифровать видео с видеохостинга YouTube. Запись без фоновых шумов высокого качества распознается достаточно хорошо, плохая запись – крайне посредственно [6]. Бесплатно Speechpad обрабатывает только 15 минут записи.

Платный сервис Real Speaker для перевода аудиофайлов в текстовый контент работает исключительно с готовыми файлами, поэтому использовать микрофон не получится. Сервис поддерживает 38 языков (в том числе русский), прост в использовании, достаточно выбрать язык и загрузить файл для расшифровки. Получившийся файл можно скачать и редактировать вручную. Так как сервис платный, то длинные файлы будут доступны для транскрибации только после оплаты.

Бесплатный онлайн-сервис Dictation позволяет распознавать запись с микрофона. Встроенный редактор позволяет форматировать текст, оформлять списки. Для добавления нового абзаца или знака препинания необходимо воспользоваться голосовой командой. После расшифровки сервис сохраняет преобразованный текст в браузере локально, данные больше никуда не загружаются [7].

AmazonTranscribe – сервис на основе искусственного интеллекта, позволяющий преобразовывать речь в текст. Имеется возможность добавления субтитров по требованию пользователя к трансляции, расшифровки телефонных обращений в службу поддержки; также можно производить анализ аудио- и видеоконтента. Качество записи или разговора влияет на точность распознавания сервиса. Одно из достоинств – постоянное обновление сервиса для улучшения распознавания, главный недостаток – платный сервис [8].

Бесплатный сервис Dragon Dictation предназначен для обладателей устройств Apple и создан для улучшения функции Siri. Поддерживает функцию диктовки и автоматического распознавания качественной речи в текст. Главное достоинство сервиса – полное голосовое управление: можно записать текст, создать сообщение и отправить его по нужной социальной сети, используя только голосовые команды [9]. Для качественного использования разработки необходимо обеспечить быстрое и бесперебойное подключение к сети Интернет.

Сервис Otter.ai для записи видеоконференций и создания на основе их заметок имеет возможность запоминать голос диктора, что позволяет разделять реплики в записях с несколькими собеседниками. Сервис позволяет осуществлять запись и расшифровку звонка в реальном времени, расшифровку сохраненной записи [10].

Разработка системы транскрибации

В результате анализа работы популярных сервисов для транскрибации и после выявления общих недостатков (платность использования, плохая расшифровка при некачественной записи) можно утверждать, что существующие системы на данном этапе не позволяют с высокой точностью проводить распознавание аудио- и видеофайлов, и необходимо разработать собственное программное обеспечение для развития данной области. Для решения многих из выявленных минусов представленных выше систем было принято решение о разработке собственного программного обеспечения на языке программирования Python с удобным интерфейсом для пользователя, которое расшифровывает аудио- или видеофайлы в текст. Python широко используется в системах распознавания речи благодаря своей простоте и гибкости. Возможности языка Python в области обработки аудио- и видеоданных делают его одним из наиболее подходящих языков программирования для разработки подобных систем. В ходе разработки системы было реализовано несколько моделей, которые можно использовать в зависимости от своих целей и задач.

Первая модель достаточно быстро транскрибирует аудио- и видеофайлы в текст, но из-за быстроты обработки теряется качество полученного текста. Модель работает следующим образом: быстро «пробегает» по файлу и выдает текстовый документ. При идеаль-

ном звучании обрабатываемого файла читабельность конечного файла может быть приемлемой, но даже при таких идеальных условиях могут возникать ошибки транскрибации. Процент ошибок резко возрастает при шумах, наложенных на аудиодорожку (плохой микрофон, посторонние звуки и др.), а также при присутствии фоновой музыки. При реализации модели была использована библиотека wave, ffmpeg и SpeechRecognition.

Wave – библиотека для работы с .waw-файлами – позволяет извлечь канальность аудиодорожки, фреймрейт и другие полезные характеристики для последующего улучшения обработки аудиофайла [11].

Ffmpeg – библиотека-«процессор» для множества других библиотек-надстроек. Она является исполнителем операций, производимых над всякого рода звуковыми файлами. По сути, все используемые библиотеки являются надстройками (врапперами) для обеспечения удобства использования этого FFmpeg [12].

Библиотека SpeechRecognition – это инструмент для передачи речевых API от компаний (Google, Microsoft, Soundhound, Ibm, а также Pocketsphinx), который, в отличие от остальных, имеет возможность работы офлайн [13].

Вторая модель на порядок медленнее первой, но процент ошибок у нее значительно меньше. Длительность транскрибирования растет экспоненциально (примерно) по мере увеличения длительности обрабатываемой аудиодорожки. Модель более углубленно обрабатывает аудиодорожку, но, несмотря на это, присутствуют значительные проблемы с транскрибированием файлов, в которых присутствуют шумы, фоновая музыка или посторонние звуки. При реализации модели была использована библиотека Vosk.

Vosk – библиотека-модель для распознавания речи, она поддерживает около 20 языков и различных диалектов, имеет небольшой вес (около 50 МБ), работает в автономном режиме (без использования Интернета). Библиотека также не умеет автоматически определять язык, поэтому необходимо вручную ее настраивать перед запуском транскрибации [14].

Описанные выше модели не умеют расставлять знаки препинания, и для этого необходимо подключать дополнительную библиотеку, специально предназначенную для этого, однако справляется со своей задачей данная библиотека достаточно посредственно [15].

Третья модель обладает низким процентом ошибок по сравнению с первыми двумя моделями. При реализации модели была использована библиотека Pywhisper для работы с Whisper-моделью. Для транскрибации аудио- или видеофайла используется функция Transcribe() этой библиотеки. Внутренне эта функция считывает весь файл и обрабатывает аудио с помощью скользящего 30-секундного окна, выполняя авторегрессионные предсказания последовательности в последовательности в каждом окне. Модель обучается на большом наборе данных с разнообразным аудио, а также является многозадачной моделью, которая может выполнять многоязычное распознавание речи, а также перевод речи и идентификацию языка [16]. В этой модели есть пять типоразмеров модели, и каждый типоразмер подразделяется еще на два вида модели – общую и языковую. Первая подходит для любого языка (модель сама определяет используемый язык, но она намного больше по размеру, чем вторая), а вторая используется для конкретного языка. От выбранного типоразмера зависит качество обработки. При обработке аудиофайлов (с шумом, без шумов, с фоновой музыкой, с посторонними звуками и др.) модель обеспечивает низкий процент ошибок. Также данная модель сама расставляет знаки препинания. Единственным минусом данной модели является медленное транскрибирование.

Whisper – это система автоматического распознавания речи (ASR), обученная на 680 000 часов многоязычных и многозадачных контролируемых данных, собранных из Интернета. Она показывает, что использование такого большого и разнообразного набора данных приводит к повышению устойчивости к акцентам, фоновому шуму и техническому языку. Архитектура Whisper представляет собой простой сквозной подход, реализованный в виде преобразователя кодер-декодер. Входной звук разбивается на 30-секундные фрагменты, преобразуется в спектрограмму Log-Mel, а затем передается в кодировщик. Декодер обучен предсказывать соответствующий текстовый заголовок, смешанный со специальными токенами, которые направляют единую модель для выполнения таких задач, как идентификация языка, временные метки на уровне фраз, транскрипция многоязычной речи и перевод речи на английский язык.

Модель преобразования последовательности в последовательность обучается множеству различных задач обработки речи, включая распознавание многоязычной речи, перевод речи, идентификацию разговорного языка и обнаружение голосовой активности [17]. Все эти задачи совместно представлены в виде последовательности токенов, которые должны быть предсказаны декодером, что позволяет одной модели заменить множество различных этапов традиционного конвейера обработки речи. Формат многозадачного обучения использует набор специальных токенов, которые служат спецификаторами задач или целями классификации.

В результате программа формирует текстовый документ, содержащий транскрибацию аудиофайла. На Рисунке 1 отображено сравнение исходного распознавания программы Whisper с ручным распознаванием.

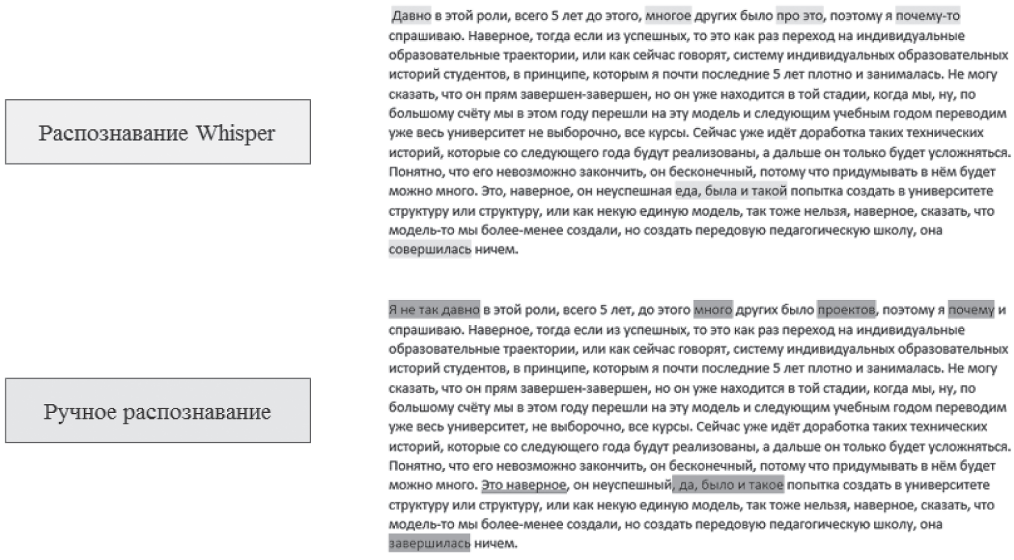

Рисунок 1. Сравнение исходного распознавания программы Whisper с ручным распознаванием
Источник: рисунок выполнен автором.

На Рисунке 2 представлены результаты обработки аудио- и видеофайлов разработанной системы распознавания речи. В качестве источников были проанализированы интервью длительностью от 30 минут до 2 часов. Файлы интервью были представлены в аудио-

и видеоформатах. При анализе было определено количество слов, некорректно обработанных тремя предложенными моделями. Процент ошибок рассчитывался как отношение общего количества слов в интервью к количеству некорректно обработанных слов. Третья модель показала наименьший процент ошибок.

Из экспериментальных результатов было отмечено несколько ограничений и областей для будущей работы.

Улучшенные стратегии декодирования. Когда был масштабирован Whisper, было замечено, что более крупные модели добились устойчивого и надежного прогресса в уменьшении ошибок, связанных с восприятием, таких как путаница похожих по звучанию слов. Более точная настройка моделей Whisper на высококачественном контролируемом наборе данных и/или использование обучения с подкреплением для более непосредственной оптимизации производительности декодирования поможет еще больше уменьшить эти ошибки.

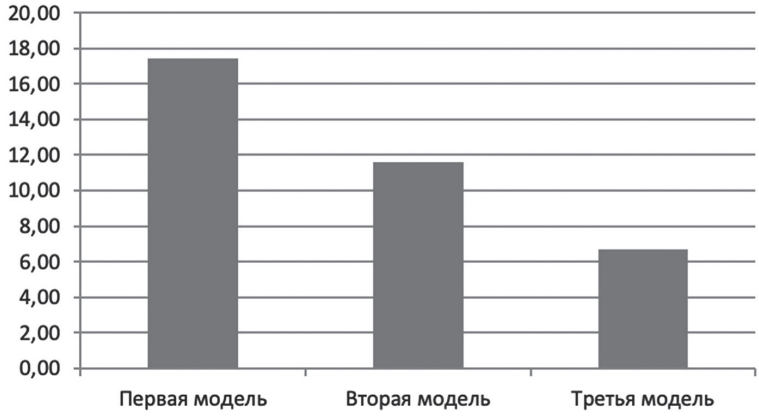

Рисунок 2. Результаты работы системы распознавания речи
Источник: диаграмма построена автором.

Надежность Whisper частично обусловлена его сильным декодером, который представляет собой условную звуковую языковую модель. В настоящее время неясно, в какой степени преимущества Whisper связаны с обучением его кодировщика, декодера или и того, и другого. Это можно изучить либо путем удаления различных компонентов модели Whisper, таких как обучение модели CTC без декодера, либо путем изучения того, как производительность существующих кодировщиков распознавания речи, таких как wav2vec 2.0, изменяется при использовании вместе с языковой моделью.

Добавление дополнительных целей при обучении Whisper заметно отличается от самых последних современных систем распознавания речи из-за отсутствия неконтролируемого предварительного обучения или методов самообучения. Хотя в них нет необходимости для достижения хорошей производительности, возможно, что результаты могут быть улучшены за счет их включения.

Заключение

В работе проведен анализ существующих систем для перевода аудио- и видеозаписей в текстовый формат. В результате анализа и выявления общих недостатков существующих сервисов и решений для повышения качества распознавания необходимо разрабаты-

вать собственное программное обеспечение. В ходе разработки системы было реализовано три модели, которыми можно воспользоваться в зависимости от своих целей и задач. Результатом распознавания является формирование текстового документа, содержащего транскрибацию аудио- или видеофайла. Третья модель с использованием библиотеки Pywhisper, обладает низким процентом ошибок по сравнению с первыми двумя моделями, расставляет знаки препинания, работает с файлами с посторонними шумами, однако медленно транскрибирует.

Литература

1. *Елизаров Д.А., Колпакова П.Е.* Применение системы транскрибирования // Сборник материалов Всероссийской научно-практической конференции с международным участием, Биробиджан, 15 декабря 2022 г. / Под науч. ред. В.М. Козина. Биробиджан : Приамурский государственный университет им. Шолом-Алейхема, 2023. С. 13–16. EDN BXAZZQ.
2. *Девяткина Е.* Способы перевода видео в текст, автоматическое транскрибирование // Yagla. URL: https://yagla.ru/blog/marketing/6-sposobov-perevesti-audio-i-video-v-tekst--2110m94955/ (дата обращения: 03.08.2023).
3. *Ибушева М.* Перевод аудио и видео в текст: способы транскрибации // SEOnews. 2021. 1 августа. URL: https://www.seonews.ru/analytics/7-sposobov-perevoda-video-v-tekst/ (дата обращения: 03.08.2023).
4. 5 ways to transcribe audio to text // MyNewsdesk. 2019. April 18. URL: https://www.mynewsdesk.com/en/blog/5-ways-to-transcribe-audio-to-text/ (дата обращения: 03.08.2023).
5. *McMullin C.* Transcription and Qualitative Methods: Implications for Third Sector // Voluntas. 2023. Vol. 34. No. 1. Pp. 140–153. DOI: 10.1007%2Fs11266-021-00400-3
6. Блокнот для речевого ввода. URL: https://speechpad.ru (дата обращения: 03.08.2023).
7. Voice Dictation – Online Speech Recognition. URL: https://dictation.io (дата обращения: 03.08.2023).
8. Вопросы и ответы по AmazonTranscribe // AmazonWebServices (AWS). URL: https://aws.amazon.com/ru/transcribe/faqs/ (дата обращения: 03.08.2023).
9. *CloudExpert.* Dragon Dictation – распознавание голоса в текст // IaaSSaaSPaaS.ru: Обзоры облачных сервисов. 2022. 22 декабря. URL: https://iaassaaspaas.ru/servisy/dragon-dictation-raspoznavanie-golosa-v-tekst (дата обращения: 03.08.2023).
10. Otter.ai – сервис для стенографирования видеоконференций и звонков // Startpack. URL: https://startpack.ru/application/otter-ai (дата обращения: 03.08.2023).
11. Wave – чтение и запись WAV-файлов // Digitology.tech. 09.08.2023. URL: https://digitology.tech/docs/python_3/library/wave.html (дата обращения: 03.08.2023).
12. Как обработать аудио с помощью ffmpeg? // VC.RU. NewTechAudit. 03.02.2020. URL: https://vc.ru/newtechaudit/110160-kak-obrabotat-audio-s-pomoshchyu-ffmpeg (дата обращения: 03.08.2023).
13. NewTechAudit. Распознавание и анализ речи с помощью библиотеки SPEECH RECOGNITION, PYAUDIO и LIBROSA // Хабр. 14.09.2021. URL: https://habr.com/ru/post/577806/ (дата обращения: 03.08.2023).
14. Офлайн-распознавание речи. Библиотека Vosk // VC.RU. NewTechAudit. 18.05.2021. URL: https://vc.ru/dev/247450-oflayn-raspoznavanie-rechi-biblioteka-vosk (дата обращения: 03.08.2023).
15. Мы опубликовали модель, расставляющую знаки препинания и заглавные буквы в тексте на четырех языках // Хабр. 06.10.2021. URL: https://habr.com/ru/post/581946/ (дата обращения: 03.08.2023).

16. Топ-10 библиотек Python для Data Science // DataStart. URL: https://datastart.ru/blog/read/top-10-bibliotek-python-dlya-data-science (дата обращения: 03.08.2023).
17. Elsayed O.S., Petrov S.N. (2020) Speech and voice recognition system based on machine learning methods // Современные средства связи: материалы XXV Международной научно-технической конференции (Минск, 22–23 октября 2020 г.). Минск : Белорусская государственная академия связи, 2020. С. 222–223. EDN MZYGIJ.

References

1. Elizarov D.A., Kolpakova P.E. (2023) Application of the transcribing system. In: *Matematicheskoe i jeksperimental'noe modelirovanie fizicheskih processov: Vserossijskaja nauchno-prakticheskaja konferencija s mezhdunarodnym uchastiem (Birobidzhan, 15 dekabrja 2022 g.)* [Mathematical and Experimental Modeling of Physical Processes] : Collection of materials of the All-Russian scientific and practical conference with international participation, Birobidzhan, December 15, 2022. Ed. by V.M. Kozin. Birobidzhan : Sholom-Aleichem Priamur State University, 2023. Pp. 13–16. EDN BXAZZQ. (In Russian).
2. Devyatkina E. Methods of Translating Video into Text, Automatic Transcription. *Yagla*. URL: https://yagla.ru/blog/marketing/6-sposobov-perevesti-audio-i-video-v-tekst--2110m94955/ (accessed 08.03.2023) (In Russian).
3. Ibusheva M. (2021) Translation of audio and video into text: methods of transcription // *SEOnews.* August 1. URL: https://www.seonews.ru/analytics/7-sposobov-perevoda-video-v-tekst/ (accessed 08.03.2023) (In Russian).
4. 5 ways to transcribe audio to text. *MyNewsdesk.* 2019. April 18. URL: https://www.mynewsdesk.com/en/blog/5-ways-to-transcribe-audio-to-text/ (accessed 08.03.2023).
5. McMullin C. (2023) Transcription and Qualitative Methods: Implications for Third Sector. *Voluntas.* Vol. 34. No. 1. Pp. 140–153. DOI: 10.1007%2Fs11266-021-00400-3
6. *Voice notepad.* URL: https://speechpad.ru (accessed 08.03.2023) (In Russian).
7. *Voice Dictation – Online Speech Recognition*. URL: https://dictation.io (accessed: 08.03.2023) (in Russian).
8. Amazon Transcribe FAQs. *Amazon Web Services (AWS)*. URL: https://aws.amazon.com/transcribe/faqs/?nc1=h_ls (accessed 08.03.2023).
9. CloudExpert (2022) Dragon Dictation – Apple Service for Transcribing Voice into Text. *IaaSSaaSPaaS.ru: Obzory oblachnykh servisov*. December 22. URL: https://iaassaaspaas.ru/servisy/dragon-dictation-raspoznavanie-golosa-v-tekst (accessed 08.03.2023) (In Russian).
10. Otter.ai – service for video conference and call transcription. *Startpack*. URL: https://startpack.ru/application/otter-ai (accessed 08.03.2023) (In Russian).
11. Wave – Reading and Writing WAV files. *Digitology.tech.* 09.08.2023. URL: https://digitology.tech/docs/python_3 /library/wave.html (accessed: 08.03.2023) (In Russian).
12. NewTechAudit (2020) How to Process Audio Using Ffmpeg? *VC.RU.* 03.02. URL: https://vc.ru/newtechaudit/110160-kak-obrabotat-audio-s-pomoshchyu-ffmpeg (accessed 08.03.2023) (In Russian).
13. NewTechAudit (2021) Speech Recognition and Analysis Using the Speech Recognition, Pyaudio and Librosa Libraries. *Habr.* 14.09. URL: https://habr.com/ru/post/577806/ (accessed 08.03.2023). (In Russian).
14. NewTechAudit (2021) Offline Speech Recognition. Library Vosk. *VC.RU.* 03.02. URL: https://vc.ru/dev/247450-oflayn-raspoznavanie-rechi-biblioteka-vosk (accessed 08.03.2023). (In Russian).

15. We Published a Model that Places Punctuation and Capitalization in Text in Four Languages. *Habr.* 06.10. URL: https://habr.com/ru/post/581946/ (accessed 08.03.2023). (In Russian).
16. Top 10 Python Libraries for Data Science. *DataStart.* URL: https://datastart.ru/blog/read/top-10-bibliotek-python-dlya-data-science (accessed 08.03.2023) (In Russian).
17. Elsayed O.S., Petrov S.N. (2020) Speech and Voice Recognition System Based on Machine Learning Methods. In: *Sovremennye sredstva svjazi* [Modern Means of Communication: Proc. XXV Int. Sci. and Tech. Conf., Minsk, October 22–23, 2020). Minsk : Belarusian State Academy of Communications. Pp. 222–223.