

ИССЛЕДОВАНИЕ API GOOGLE CLOUD SPEECH-TO-TEXT И YANDEX SPEECHKIT ДЛЯ МОБИЛЬНЫХ ПРОГРАММНЫХ СРЕДСТВ ПОД ОПЕРАЦИОННУЮ СИСТЕМУ ANDROID

Силенкова Д.С., Зацепина М.И.

Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь

Научный руководитель: Писарчик А.Ю. – ст. преподаватель кафедры ПИКС

Аннотация. В работе представлен сравнительный анализ технологий преобразования речи в текст (*Speech-to-Text*) на примере API *Google Cloud Speech-to-Text* и *Yandex SpeechKit*. Основное внимание уделено ключевым параметрам, таким как точность распознавания, скорость обработки запросов, поддержка различных языков и акцентов, а также удобство интеграции в *Android*-приложения. На основе полученных данных предложены практические рекомендации по выбору оптимального решения для различных сценариев использования.

Ключевые слова: распознавание речи, преобразование речи в текст, сравнительный анализ, *Google Cloud Speech-to-Text*, *Yandex SpeechKit*, интеграция, *Android*-приложения.

Введение. Технологии преобразования речи в текст (*Speech-to-Text*, *STT*) становятся неотъемлемой частью цифровой трансформации, находя применение в таких сферах, как образование, СМИ и бизнес. Студенты используют их для записи лекций и генерации конспектов, распознавание речи позволяет обрабатывать голосовые сообщения клиентов для автоматического составления расписания и преобразовать интервью в текст для создания субтитров. Лидерами в этой области являются *Google* и *Яндекс*, предлагающие решения *Google Cloud Speech-to-Text* и *Yandex SpeechKit*, которые активно интегрируются в мобильные приложения, включая *Android*.

Основная часть. Технологии преобразования речи в текст (*Speech-to-Text*, *STT*) основаны на сложных алгоритмах, которые включают несколько этапов обработки звукового сигнала. Существуют два основных направления обработки речи: распознавание загруженных аудио файлов и потоковое распознавание в режиме реального времени. При этом сама обработка поступающих данных происходит по одному и тому же алгоритму [1].

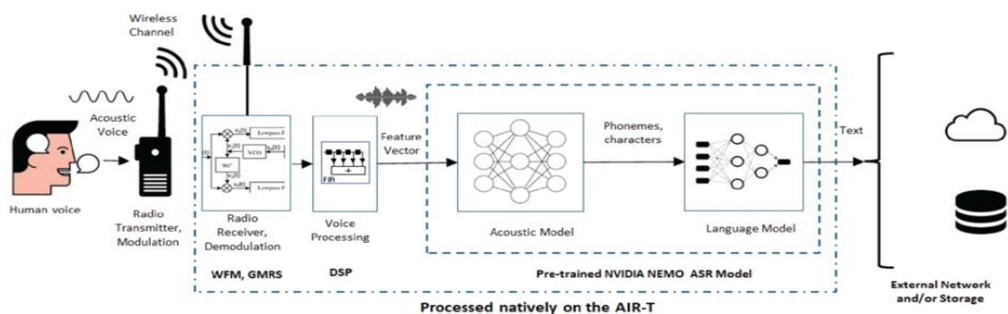


Рисунок 1 – Алгоритм преобразования речи в текст

На первом этапе происходит предобработка аудиоматериалов: устранение шумов, нормализация громкости и разделение звукового потока на фреймы. Затем система переходит к акустическому моделированию, где звуковые фрагменты сопоставляются с фонемами – минимальными единицами звука. Для этого используются методы машинного обучения, такие как *Hidden Markov Models (HMM)* или *Deep Neural Networks (DNN)*. На этапе языкового моделирования определяется наиболее вероятная последовательность слов на основе контекста. Завершающий этап – обработка ошибок, добавление пунктуации и форматирование

текста. Эти этапы позволяют системам *STT* достигать высокой точности распознавания даже в условиях шума или при наличии акцентов.

Google Cloud Speech-to-Text [2] является одним из лидеров в области распознавания речи, его ключевой особенностью является режим потокового распознавания, который позволяет обрабатывать аудиоматериалы в реальном времени. Кроме того, *Google* предлагает модель с повышенной точностью распознавания в сложных акустических условиях, которая адаптируется к шумам. Интеграция с другими сервисами *Google Cloud*, такими как *Translate* и *Natural Language API*, расширяет возможности использования этой технологии. Однако стоит учитывать, что *Google Speech-to-Text* требует наличия *Google Play Services*, что может быть ограничением для персонализированных сборок *Android*.

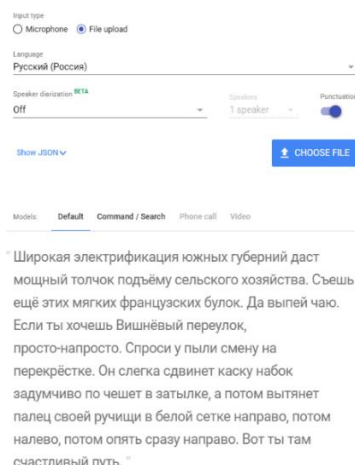


Рисунок 2 – Преобразование аудио файлов с помощью *Google Speech-to-Text*

Yandex SpeechKit [3], в свою очередь, оптимизирован для работы с русским языком и славянскими акцентами, что делает его предпочтительным выбором для русскоязычных проектов. Этот *API* отличается низкой задержкой – до 300 мс для коротких запросов, что обеспечивает быстрое взаимодействие с пользователем. *Yandex SpeechKit* поддерживает *REST API*, что упрощает интеграцию в мобильные программные средства. Несмотря на ограниченную поддержку языков в *Yandex SpeechKit*, для проектов, ориентированных на русскоязычную аудиторию, таких как голосовые помощники, приложения для образования или медиа, эта система демонстрирует высокую точность и доступную стоимость.

Ключевые требования к *STT*-системам включают точность распознавания, скорость обработки, поддержку языков, устойчивость к шумам и энергопотребление. Точность распознавания оценивается по формуле *WER* (*Word Error Rate*), которая учитывает замены, пропуски и вставки слов [4]. Скорость обработки измеряется как время от отправки аудио до получения текста. Поддержка языков определяет, насколько универсальным является решение, а устойчивость к шумам – его способность работать в сложных акустических условиях. Значения критериев для рассматриваемых технологий приведены в таблице 1.

Таблица 1 – Сравнительный анализ технологий

Критерий	<i>Google Cloud Speech-to-Text</i>	<i>Yandex SpeechKit</i>
Точность (WER)	5.8% (английский), 10.2% (русский)	8.5% (английский), 4.7% (русский)
Скорость обработки	1200 мс	800 мс
Поддержка языков	125	16
Устойчивость к шуму	высокая	средняя
Стоимость	\$0.006/мин (стандартная модель)	\$0.01/мин (русский язык)

Сравнительная характеристика *Google Speech-to-Text* и *Yandex SpeechKit* показывает, что каждая из технологий имеет свои сильные и слабые стороны. *Google Cloud Speech-to-Text* демонстрирует высокую точность для английского языка (*WER* ~5.8%) и поддерживает более 120 языков, что делает его идеальным для международных проектов. *Yandex SpeechKit*, напротив,

показывает лучшие результаты для русского языка ($WER \sim 4.7\%$) и имеет низкую задержку, что делает его предпочтительным выбором для русскоязычных приложений.

Интеграция *Google Speech-to-Text* начинается с создания проекта в *Google Cloud Console*, активации соответствующего API и генерации ключа. В Android-проекте также требуется реализовать компонент *SpeechRecognizer* с использованием *RecognitionListener* для обработки аудиоданных. Для *Yandex SpeechKit* необходима регистрация сервисного аккаунта в Яндекс.Облаке, получение IAM-токена и использование HTTP-клиента или SDK для передачи аудио в формате *LPCM*. В обоих случаях подключения требуется добавление разрешений на подключение к интернету и запись аудио файлов в манифест проекта.

Тестирование *STT*-технологий в Android-приложениях можно производить несколькими методами. Точность распознавания оценивается, записывая аудио с микрофона и сравнивая результат с эталонным текстом. Скорость обработки – это время от начала записи до получения текста. Энергопотребление отслеживается с помощью инструментов *Android Profiler*.

Продукты от *Google* и *Yandex*, пусть и являются лидерами в этой области, но это далеко не означает что они будут оптимальным выбором для всех технических решений. Например, *IBM Watson Speech-to-Text* лучше подходит для гибридных облачно-локальных решений, а *AWS Transcribe* оптимален для интеграции с инфраструктурой *Amazon*. Open Source решения, такие как *DeepSpeech*, могут быть полезны для проектов с ограниченным бюджетом и необходимостью кастомизации.

Заключение. Проведенный анализ технологий преобразования речи в текст (*Speech-to-Text*) показал, что *Google Cloud Speech-to-Text* и *Yandex SpeechKit* имеют свои уникальные преимущества. *Google* выделяется поддержкой более 120 языков и режимом потокового распознавания, что делает его идеальным для международных проектов. Однако для русскоязычных приложений *Yandex SpeechKit* демонстрирует более высокую точность ($WER \sim 4.7\%$) и низкую задержку, что делает его предпочтительным выбором. В будущем развитие *STT*-технологий будет направлено на улучшение устойчивости к шумам, расширение языковой поддержки и интеграцию с другими AI-решениями, что откроет новые возможности для их применения.

Список литературы

1. Что такое распознавание речи и нужно ли оно бизнесу [Электронный ресурс]. – Режим доступа: <https://www.mtt.ru/support/blog/chto-takoe-raspoznavanie-rechi-i-nuzhno-li-ono-biznesu/>. – Дата доступа: 09.03.2025.
2. *Speech-to-Text documentation* [Электронный ресурс]. – Режим доступа: <https://cloud.google.com/speech-to-text/docs>. – Дата доступа: 10.03.2025.
3. Обзор технологий *Yandex SpeechKit* [Электронный ресурс]. – Режим доступа: <https://yandex.cloud/ru/docs/speechkit/overview>. – Дата доступа: 12.03.2025.
4. Blokdyk, G. *Speech-to-Text. Second Edition*. – Toronto, Canada, 2022.

UDC 004.934.1

GOOGLE CLOUD SPEECH-TO-TEXT API AND YANDEX SPEECHKIT OVERVIEW FOR ANDROID OPERATING SYSTEM MOBILE SOFTWARE

Silenkova D.S., Zatsepina M.I.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Pisarchik A.Y. – Senior lecturer of the department of ICSD

Annotation. This paper presents a comparative analysis of speech-to-text conversion technologies using the example of Google Cloud Speech-to-Text and Yandex SpeechKit APIs. The main focus is on key parameters such as recognition accuracy, request processing speed, support for various languages and accents, and ease of integration into Android applications. Based on the obtained data, practical recommendations are provided for selecting the optimal solution for various use cases.

Keywords: speech recognition, speech-to-text conversion, comparative analysis, Google Cloud Speech-to-Text, Yandex SpeechKit, integration, Android applications.