ПРОЕКТ «РАСПОЗНАВАНИЕ РЕЧИ И ПРЕВРАЩЕНИЕ ЕЁ

В ТЕКСТ БЕЗ ПОДКЛЮЧЕНИЯ К СЕТИ ИНТЕРНЕТ»

Автор: А.А. Круглей

Класс: 9В

Учебное заведение: ГУО «Гимназия №10 г. Минска»

2025

Содержание

[Введение 3](#_Toc42075205)

[1 Основная часть 4](#_Toc42075206)

[1.1 Обзор предметной области и мотивация 4](#_Toc42075207)

[1.2 Архитектура и реализация решения 4](#_Toc42075208)

[1.2.1 Захват аудиопотока 4](#_Toc42075208)

[1.2.2 Постобработка и ресемплинг 4](#_Toc42075208)

[1.2.3 Оффлайн-распознование 5](#_Toc42075208)

[1.2.4 Тестирование и результаты 5](#_Toc42075208)

[2 Заключение](#_Toc42075209) 6

[3 Список использованных источников](#_Toc42075212) 7

[4 Приложения](#_Toc42075218) 8

# Введение

Современные интерфейсы взаимодействия с электронными системами всё активнее используют голос как способ управления и общения. Если научиться надёжно распознавать речь и превращать её в текст в оффлайн-режиме, это открывает широкие возможности: создание голосовых помощников, систем управления приложениями голосом, а также применение в игровой индустрии — где игроки смогут управлять персонажами и взаимодействовать с миром через голосовые команды. Подобные технологии расширяют функциональные возможности программных продуктов, повышают их ценность на рынке и могут приносить экономические преимущества.

Цель данного исследования — разработать и протестировать прототип программного решения, выполняющего оффлайн распознавание речи (speech-to-text) на платформе Windows 10, с последующим анализом точности распознавания и описанием архитектуры решения.

Задачи исследования:

1. Проанализировать подходящие оффлайн-методы и библиотеки для распознавания речи.
2. Реализовать прототип захвата аудиопотока и его обработки.
3. Интегрировать оффлайн-распознавание (на основе выбранной модели) и провести тестирование.

4. Произвести оценку качества распознавания и описать возможные направления улучшения.

1 Основная часть

* 1. Обзор предметной области и мотивация

Распознавание речи в реальном времени традиционно опирается на методы статистического анализа аудиопотока.

Для оффлайн-решений важны два фактора:

* 1. локальная доступность моделей
  2. скорость обработки на целевом оборудовании.

В рамках проекта выбран путь использования открытых/бесплатных компонентов, обеспечивающих работу без постоянного подключения к сети.

1.2 Архитектура и реализация решения

Прототип реализован как настольное приложение для Windows 10 на C#/.NET8 с использованием кросс-платформенного UI-фреймворка «AvaloniaUI». Основные компоненты:

1. Захват аудиопотока с выбранного микрофона с помощью бесплатного фреймворка «NAudio».
2. Буферизация и преобразование аудиоформата.
3. Оффлайн-распознавание речи с использованием библиотеки «Vosk» и предварительно подготовленными речевыми оффлайн-моделями.
4. Обработка результатов, распознавание речи, отображение текста в интерфейсе.

1.2.1 Захват аудиопотока

Захват реализован через API Windows (NAudio WasapiCapture), позволяющий работать с устройствами захвата и получать поток данных в реальном времени. Для совместимости с выбранной моделью поток переводится в моно и ресемплится до частоты 16000 Гц.

1.2.2 Постобработка и ресемплинг

Поскольку многие микрофоны выдают 44100 или 48000 Гц и/или 32-bit float, входной поток преобразуется через «WdlResamplingSampleProvider» в формат, ожидаемый распознавателем, т.е. библиотекой «Vosk», а именно: 16 kHz, 16-bit PCM, моно. Такой формат является стандартным для многих моделей распознавания.

1.2.3 Оффлайн-распознование

В качестве движка распознавания выбран «Vosk» — он поддерживает оффлайн-режим и предоставляет готовые языковые модели (в проекте проводились тесты с русским и английским языками). Модель хранится в каталоге приложения; при инициализации создаётся экземпляр «Vosk.Model» и «VoskRecognizer». Для минимальной задержки используется потоковая подача фреймов и чтение «PartialResult()» для «on-the-fly» вывода промежуточного текста.

1.2.4 Тестирование и результаты

В прототипе присутствует «заготовка» для реализации функционала синтеза распознанной речи, однако изначально проект нацелен исключительно на распознавание аудиопотока с микрофона и преобразование речи в текст. Результаты распознавания зависят от качества модели и свойств микрофона. В эксперименте наблюдаются колебания точности в диапазоне ~50–90% для «сжатых» моделей.

Метод тестирования: запись фраз и фрагментов речи, чтение коротких фраз в спокойной обстановке.

Результаты:

* прототип корректно распознаёт простые фразы и отдельные слова; точность зависит от модели и условий записи.
* при использовании «малых» (компактных) моделей наблюдалось распознавание с точностью порядка 50–90% в зависимости от дикции, уровня шума и акустики.
* время распознавания частично варьировалось: от почти мгновенного вывода (благодаря частичным результатам) до задержек порядка 1.0–1.5 с при длительных фразах (это связано с размером блоков и внутренней обработкой модели).

Выводы о надёжности:

* для задач командной активации и управления прототип демонстрирует приемлемую скорость и точность.
* для применения в продуктах с высокими требованиями к точности и низкой латентности рекомендуется использовать более крупные/качественные модели и выполнить оптимизацию разработанного решения.

2 Заключение

Практическая значимость:

Решение может служить основой для оффлайн-ассистентов, систем голосового управления приложениями на локальных машинах, а также встраиваться в игровые и образовательные продукты, где недопустимо хранение или передача пользовательской речи в сеть.

Рекомендации по улучшению:

* использование более крупных и качественных языковых моделей для повышения точности;
* аппаратное ускорение (использование GPU) для снижения задержек при использовании тяжелых нейросетей;
* подключение систем адаптации к голосу пользователя (fine-tuning или адаптивная фильтрация шума).

Ограничения и риски:

* требования к объёму и скорости обработки растут с качеством модели;
* на слабых ПК большие модели могут работать медленно — требуются компромиссы между скоростью и качеством.

В ходе работы разработан прототип оффлайн-системы распознавания речи, выполняющей потоковое преобразование аудиосигнала в текст на платформе Windows 10. Продемонстрирована возможность получения частичных и окончательных результатов распознавания в реальном времени при использовании локальных языковых моделей. Прототип подтверждает практическую применимость подхода для задач голосового управления и взаимодействия, однако его точность и латентность зависят от выбранной модели и оборудования. Дальнейшие исследования целесообразно направить на интеграцию более совершенных моделей, оптимизацию обработки аудиопотока и адаптацию к условиям реального шума.

3 Список использованных источников

1. Vosk — оффлайн распознавание речи. Официальный репозиторий и модели. https://alphacephei.com/vosk (дата обращения: xx.07.2025).
2. eSpeak NG — оффлайн TTS (использовался для тестирования звука). https://github.com/espeak-ng/espeak-ng (дата обращения: xx.07.2025).
3. NAudio — библиотека для работы со звуком в .NET. https://github.com/naudio/NAudio (дата обращения: xx.07.2025).
4. Документация AvaloniaUI — UI-фреймворк. https://avaloniaui.net (дата обращения: xx.06.2025).
5. VB-Audio Virtual Cable — виртуальный аудиокабель (при необходимости создания виртуального микрофона). https://vb-audio.com/Cable/ (дата обращения: xx.06.2025).
6. Общие статьи и руководства по оффлайн STT и TTS (обзорные материалы, статьи 2018–2024 гг.).

4 Приложения

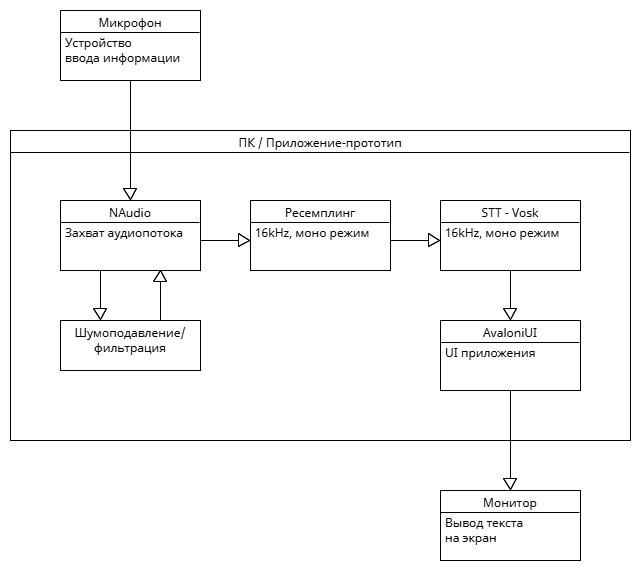


Рисунок 1. Схема работы приложения

Ссылка на проект: <https://github.com/Akrumax/ArteKrugMicApp>