МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Министерство цифрового развития, связи и массовых коммуникаций
Ордена трудового Красного Знамени федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский технический университет связи и информатики»

Кафедра Математической кибернетики и информационных технологий

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА НА ТЕМУ

«Разработка эффективного метода определения нецензурной лексики в режиме реального времени»

Подготовил студент группы БВТ1901
Лапин Виктор Андреевич

Научный руководитель:
Мкртчян Грач Маратович

Москва 2023

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Описание предметной области

Производство прямых трансляций

Предотвращение конфликтных и опасных ситуаций

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Существующие решения

Intel Bleep

- Находится в разработке
- Нет русского языка

Soniox

- Выполняет полную транскрибацию, высокая задержка
- Нет русского языка

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Существующие решения

Yandex SpeechKit

- Ограничение по длительности
- Только облачный сервис

Существуют и другие продукты, но они:

- Не работают в реальном времени
- Имеют ограниченную длительность
- Не распознают нецензурные выражения автоматически

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

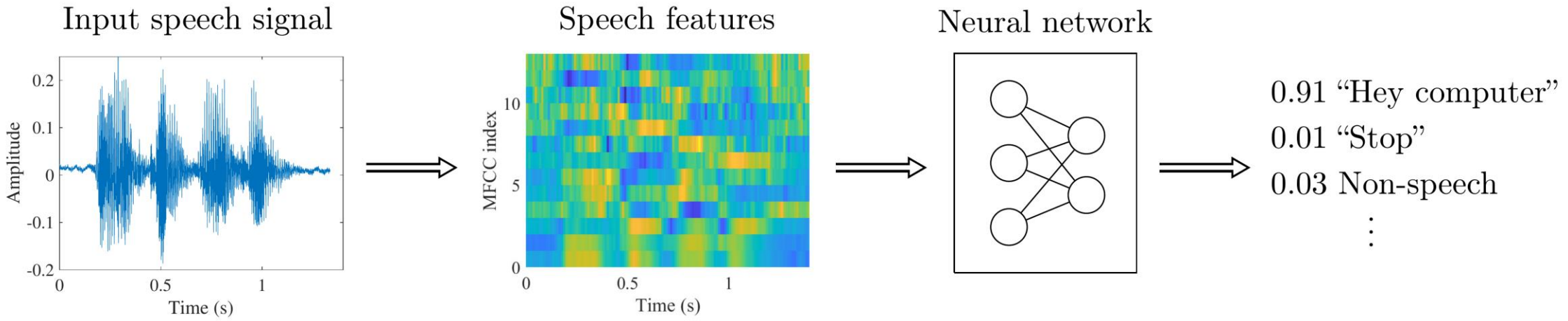
Автоматическое распознавание речи

- Модель состоит из нескольких компонентов
- Содержит словарь с множеством слов
- Выполняет полную транскрибацию аудио
- Имеет задержку в несколько секунд

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Распознавание ключевых слов

- Модель содержит меньше компонентов
- Находит только необходимые слова
- Имеет более низкую задержку



МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Сбор датасета

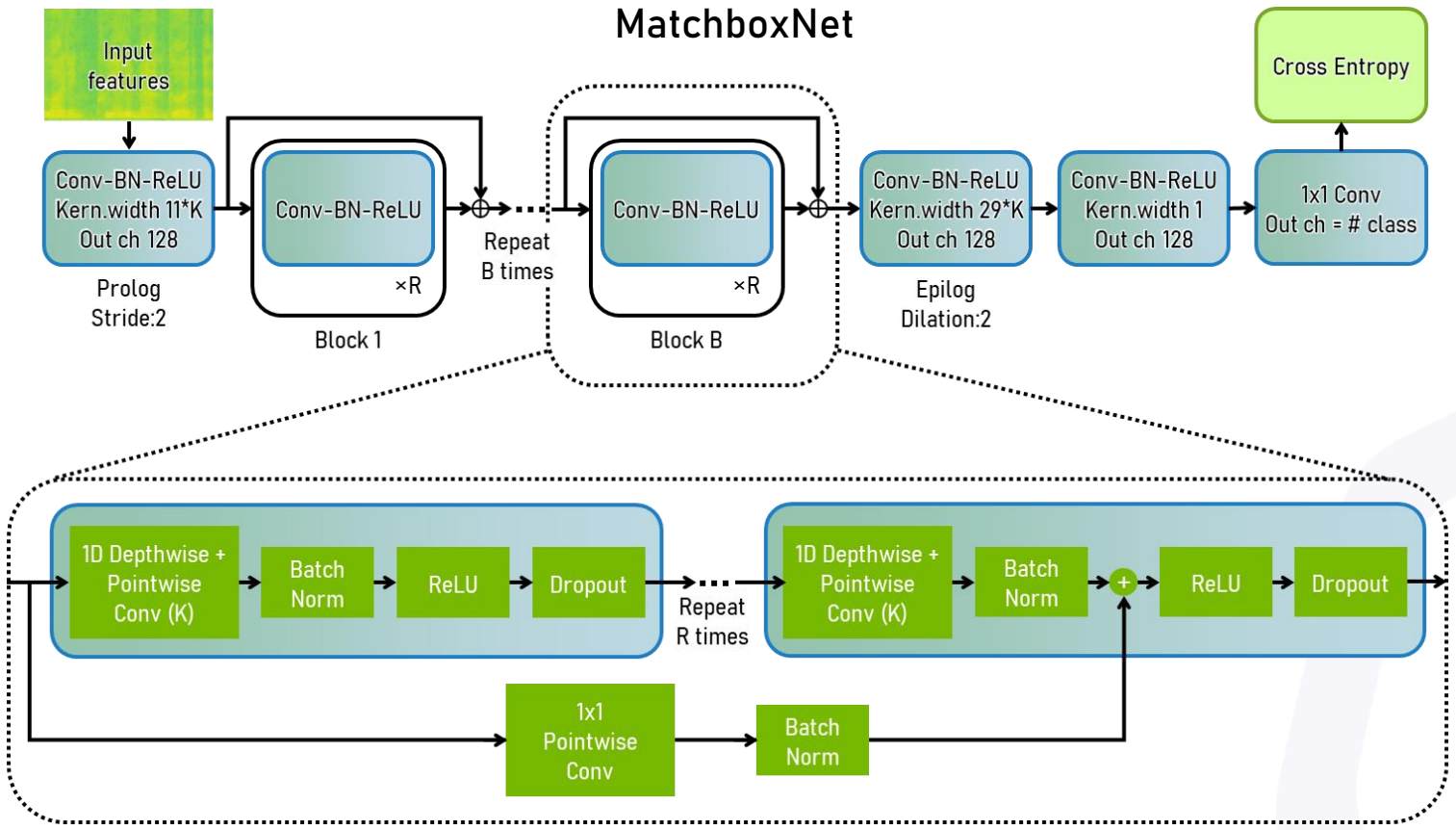
- Собран из открытых источников и размечен с использованием распознавания речи
- Содержит специально собранные данные

Описание датасета:

- 10 классов с 200 примерами в каждом
- 1 класс *unknown* с 20 000 примерами
- Длительность примеров до 1 секунды
- Содержит в основном корни нецензурных слов

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Выбор модели



МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Параметры обучения

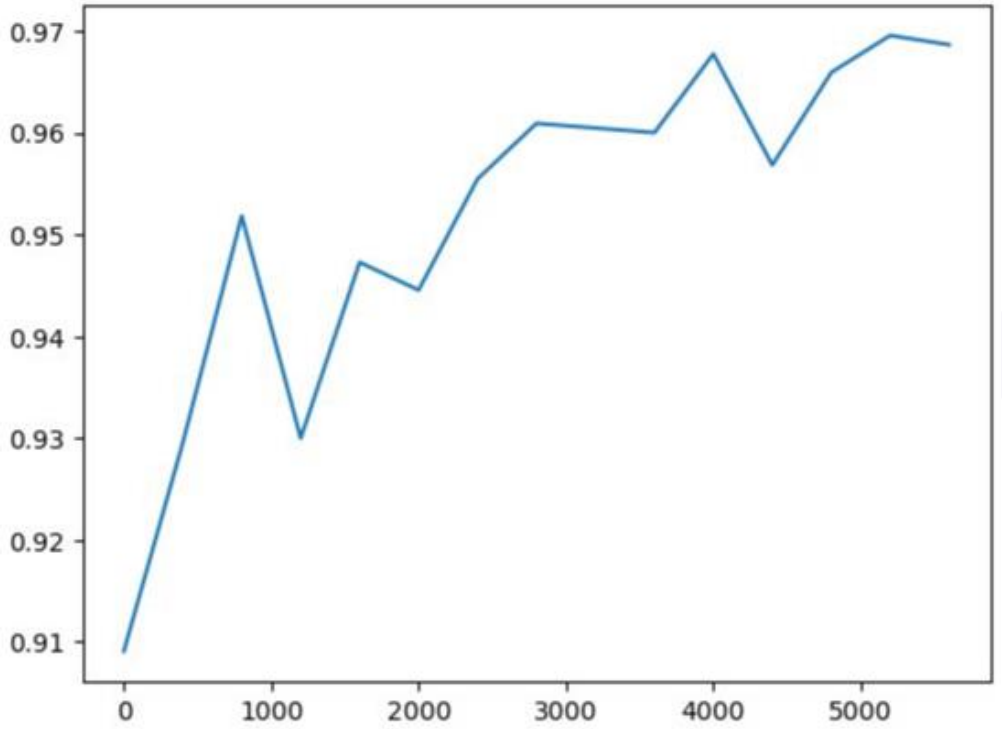
Размеры выборок

- **Train:** 1600 + *16000 unknown*
- **Validation:** 200 + *2000 unknown*
- **Test:** 200 + *2000 unknown*

Размер пакета: 440

Количество итераций: 6000

Переменный темп обучения



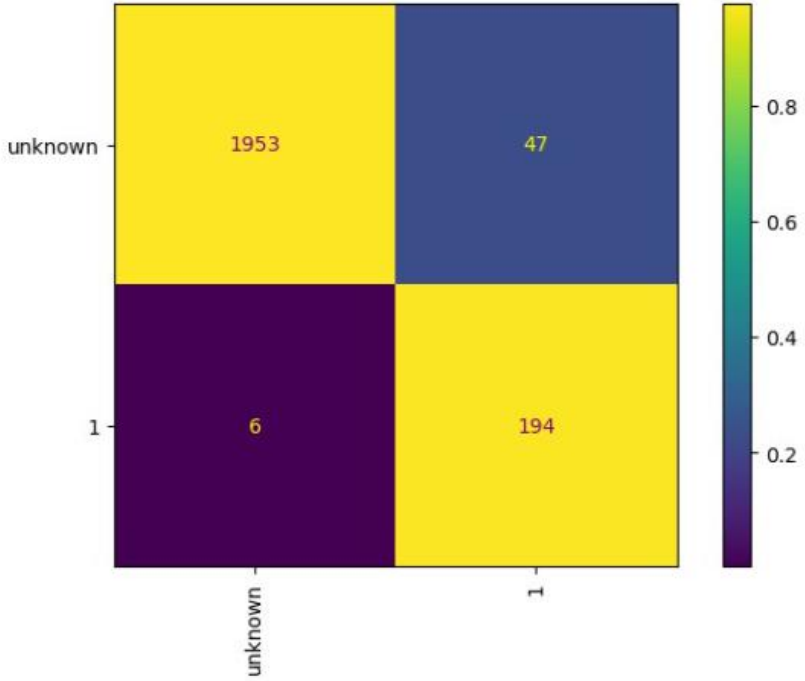
МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Результаты

Accuracy 0,9759

F1-мера 0,8798

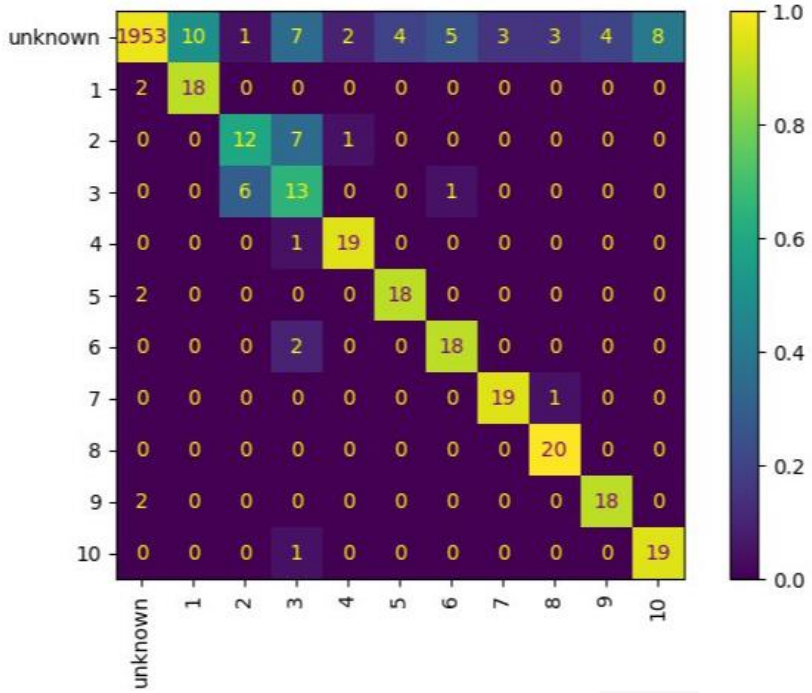
ROC-AUC 0,9732



Accuracy 0,9677

F1-мера (взвешенная) 0,9690

ROC-AUC (взвешенная) 0,9696

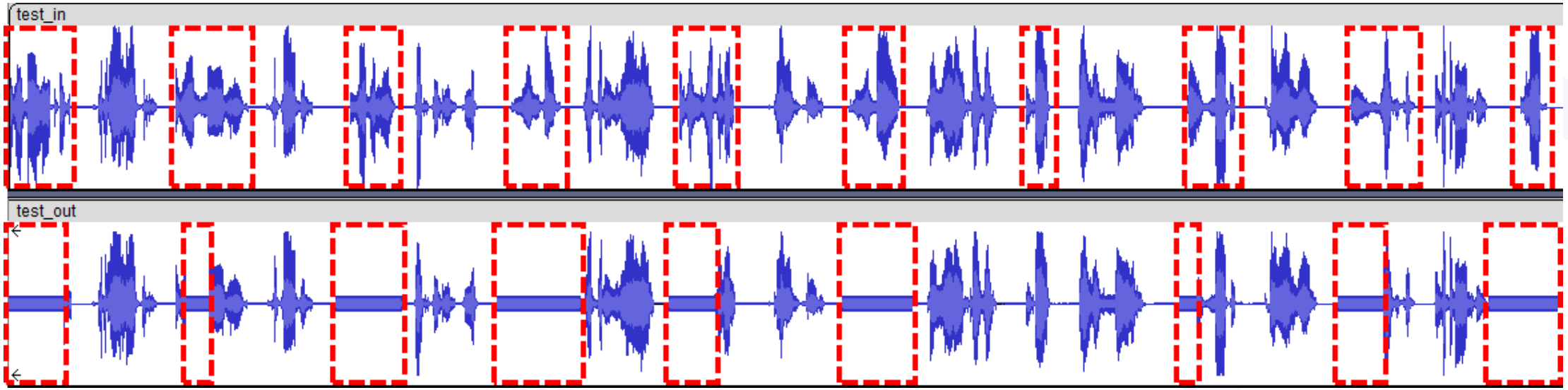


МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Разработка и тестирование приложения

Входные данные: аудиопоток получается с устройства ввода

Выходные данные: аудиопоток с отфильтрованными нецензурными словами отправляется на устройство вывода



МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Выводы

В результате работы:

- Собран датасет, содержащий нецензурные слова на русском языке
- Обучена модель, позволяющий находить нецензурную лексику в аудиопотоке
- Модель имеет низкую задержку и достаточно высокую точность благодаря использованию современной архитектуры
- Разработано приложение, позволяющее применять полученную модель для обнаружения нецензурной лексики и замены её на звуковой сигнал

МТУСИ
МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ

Спасибо за внимание!