МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет (национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

Реализация веб-приложения для выделения целевого голоса с использованием нейросетевых технологий

Рецензент: доцент
кафедры ВМиИТ
ЧелГУ, к.ф.-м.н.
А.Ю. Маковецкий

Научные руководители:
доцент кафедры СП, к.ф.-м.н.
С.У. Турлакова,
ст. преподаватель кафедры СП
Н.С. Силкина

Автор:
студент группы КЭ-228
В.И. Капичай

Челябинск, 2024 г.

Цель и задачи

Цель

разработать веб-приложение для выделения целевого голоса из смеси при помощи аудио подсказки

Задачи:

1) выполнить обзор алгоритмов выделения целевого голоса, метрик и тематической литературы
2) сформировать обучающий набор данных, осуществить требуемую предобработку данных
3) реализовать несколько моделей, протестировать их работоспособность
4) провести эксперименты на реальных данных, сравнить эффективность реализованных моделей с двумя и тремя перекрестными дикторами
5) реализовать веб-приложение для выделения целевого голоса

Актуальность

Выделение целевого голоса имеет широкий спектр приложений в таких областях, как:

1) автоматическая транскрипция
2) распознавание речи
3) аудио обработка

Главная проблема, которая решается с помощью выделения целевого голоса – **проблема коктейльной вечеринки**

Определения

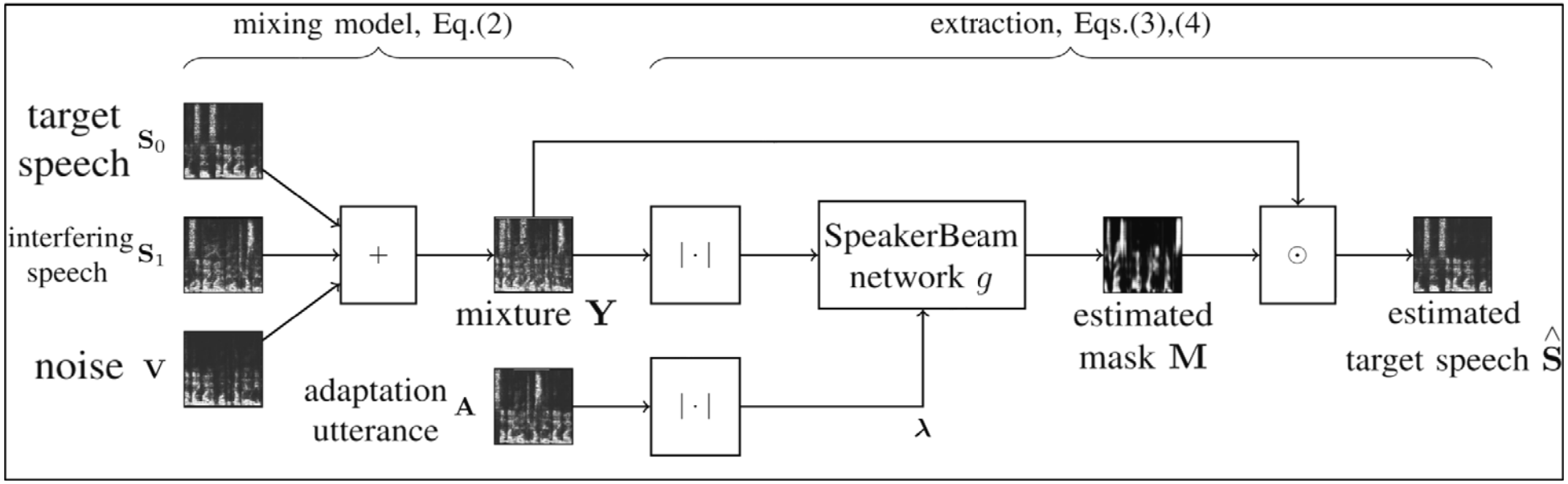
Смесь – аудиофайл с перекрестной речью двух и более дикторов, а также возможными шумами

Подсказка – вспомогательные сигналы в виде записанной речи, видео с речью или данных о местоположении целевого диктора относительно микрофона

Диаризация – процесс разделения аудиопотока на однородные сегменты в соответствии с принадлежностью аудиопотока тому или иному диктору

Модели выделения целевого диктора

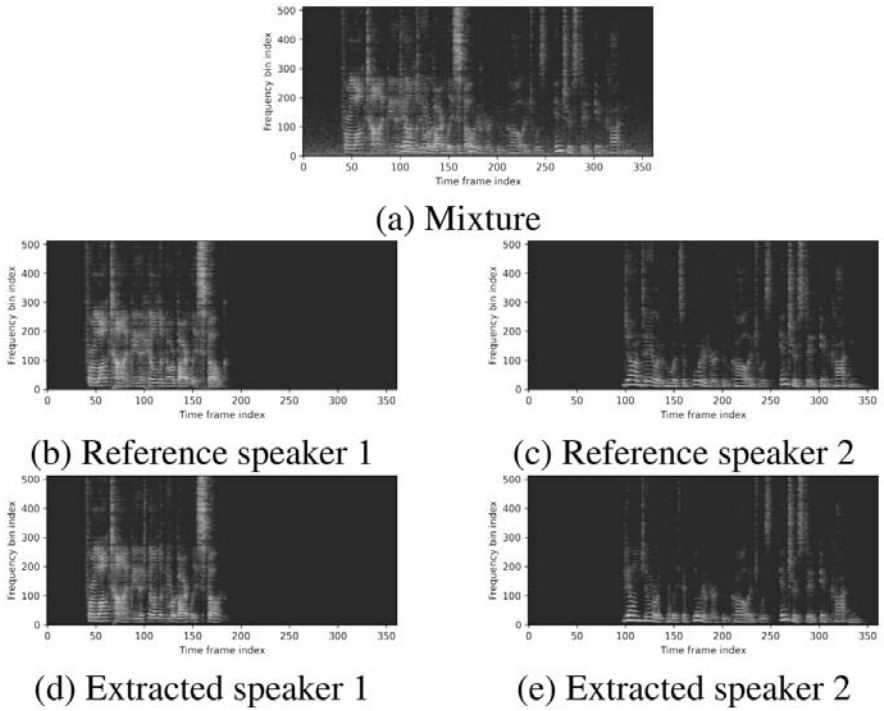
SpeakerBeam[1]



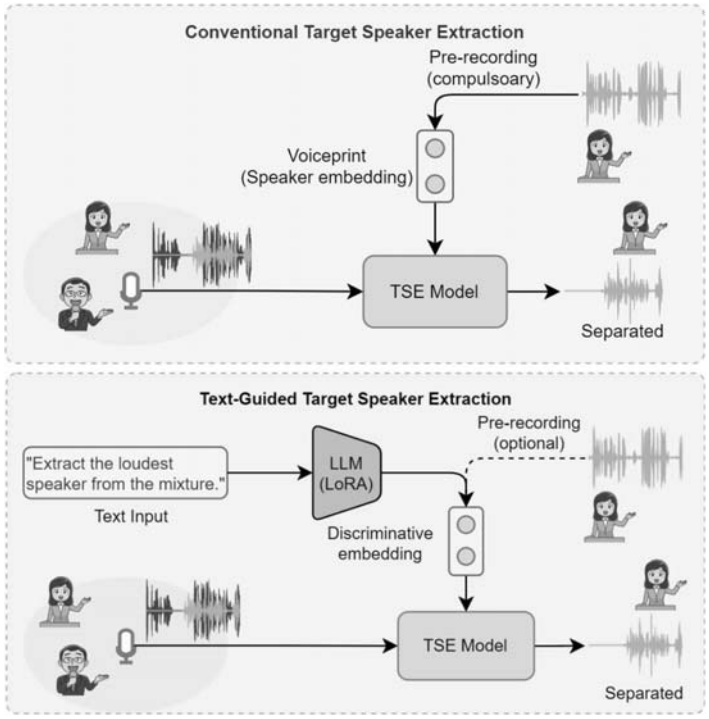
1. Žmolíková K., Delcroix M., Kinoshita K., Ochiai T., Nakatani T., Burget L., Cernocký J. SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures. // IEEE Journal of Selected Topics in Signal Processing, Volume: 13, 2019. – С. 800–814.

Модели выделения целевого диктора (2)

ADEnet[2]



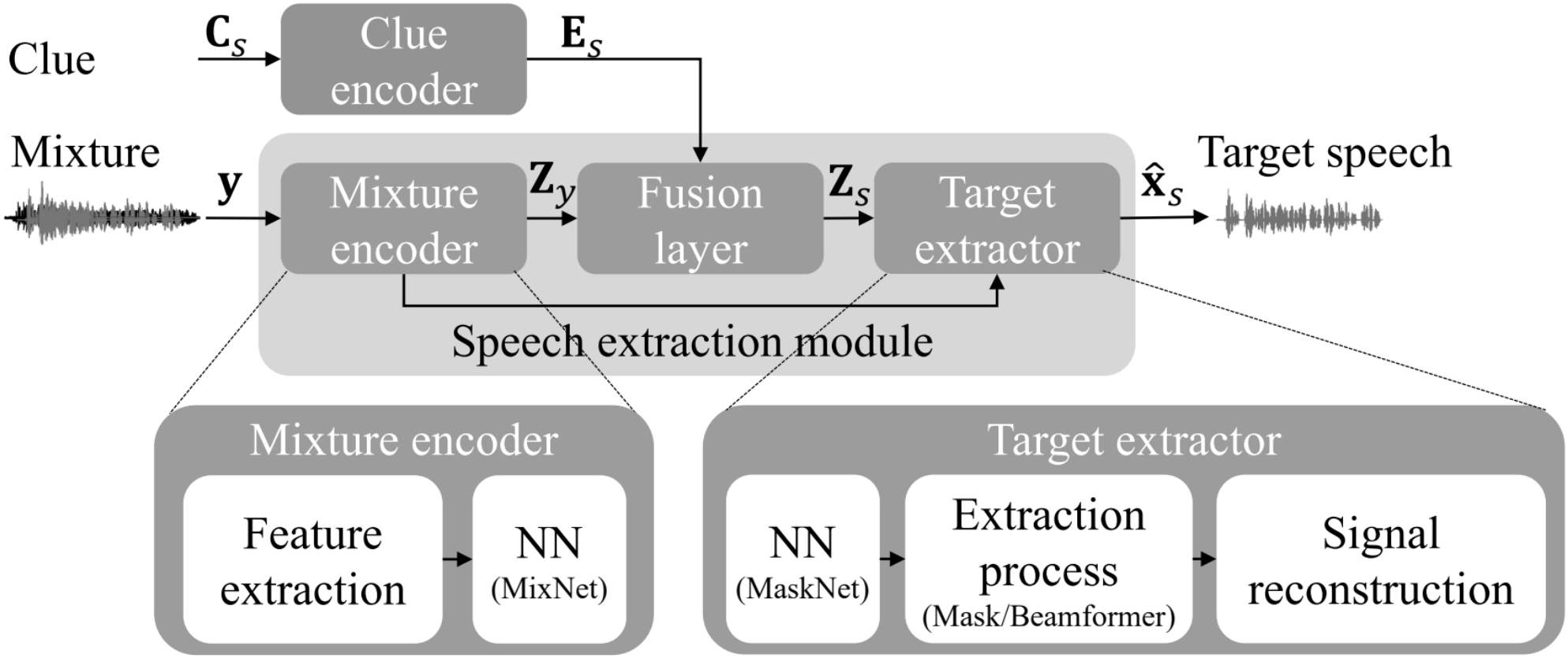
LLM–TSE[3]



2. Delcroix M., Žmolíková K., Ochiai T., Kinoshita K., Nakatani T. Speaker activity driven neural speech extraction.

3. Hao X., Wu J., Yu J., Xu C., Chen Tan K. Typing to Listen at the Cocktail Party: Textguided Target Speaker Extraction.

Общая структура модели TSE*



***TSE** (Target Speech Extraction) – выделение целевой речи

Референсная модель: SpeakerBeam

Преимущества:

1) простота запуска и обучения
2) открытый и доступный код
3) наличие предобученной модели
4) хорошие результаты выделения целевого голоса

Обучение:

1) GPU: GeForce RTX 3060Ti 8Gb
2) CPU: 12th Gen Intel Core i5-12400F @ 6x2.5GHz
3) общий объем оперативной памяти: 32Gb
4) операционная система: Ubuntu 22.04.4 LTS
5) размер батча: 6 ➡ 2
6) количество эпох: 200 ➡ 20
7) время обучения: 13 часов
8) датасет: LibriMix, 460 Гб

Использованные технологии

Asteroid – выполняет диаризацию смеси

SpeechBrain – определяет целевого диктора

Librosa, Soundfile – отвечают за чтение, сохранение и предобработку файлов смеси и подсказки

Реализованная модель: Предобработка

Предобработка данных:

Смесь и подсказка приводятся

- моно формат
- .wav формат
- частота дискретизации 8 кГц или 16 кГц

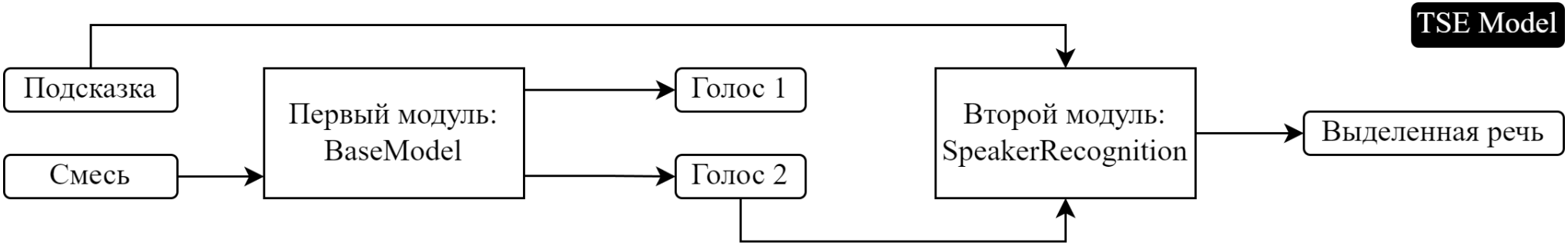
Реализованная модель: Модули

Первый модуль: BaseModel:

- Отвечает за предобработку данных и диаризацию
- Преобразует файлы под стандарт, выбирает предобученную модель и разделяет смесь на отдельные файлы

Второй модуль: SpeakerRecognition:

- Отвечает за определение целевого диктора
- Считывает разделенные голоса, подсказку и проводит сравнение между ними



1. Asteroid. [Электронный ресурс] URL: https://github.com/asteroid-team/asteroid

2. SpeechBrain. [Электронный ресурс] URL: https://github.com/speechbrain/speechbrain

Метрики

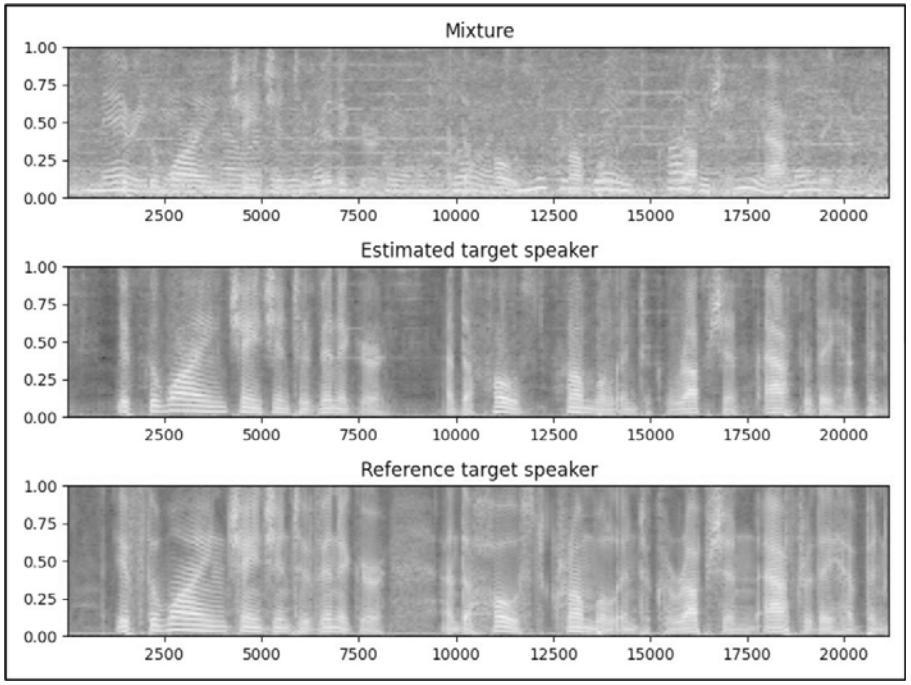
	SI-SDR	SDR	STOI	PESQ
Что измеряет	Отношение между мощностью сигнала и мощностью искажений. Не зависит от громкости сигнала	Отношение между мощностью сигнала и мощностью искажений. Зависит от громкости сигнала	Разборчивость шумной речи	Качество речи
Интервал	Нет границ	Нет границ	От 0 до 100	От -0.5 до 4.5
В чем измеряется	дБ	дБ	%	Баллы
Лучшие показатели	Выше - лучше	Выше - лучше	Выше - лучше	Выше - лучше

Сравнение моделей: Метрики

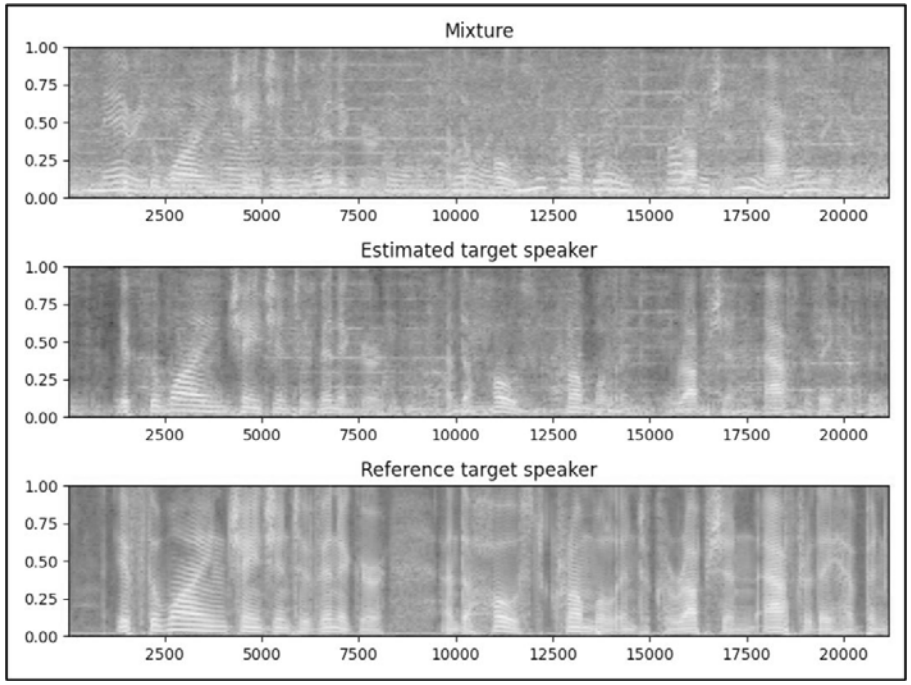
Метрика	Смесь	SpeakerBeam	TSE Model
SI-SDR	-3.809	12.140	6.294
SDR	-3.572	12.507	6.499
STOI	0.658	0.921	0.854
PESQ	1.562	2.706	1.985

Сравнение моделей: Спектрограммы

SpeakerBeam



TSE Model



Сравнение моделей: На слух

SpeakerBeam

выделенный голос четкий, громче чем в смеси, фоновая музыка очень тихо слышна

TSE Model

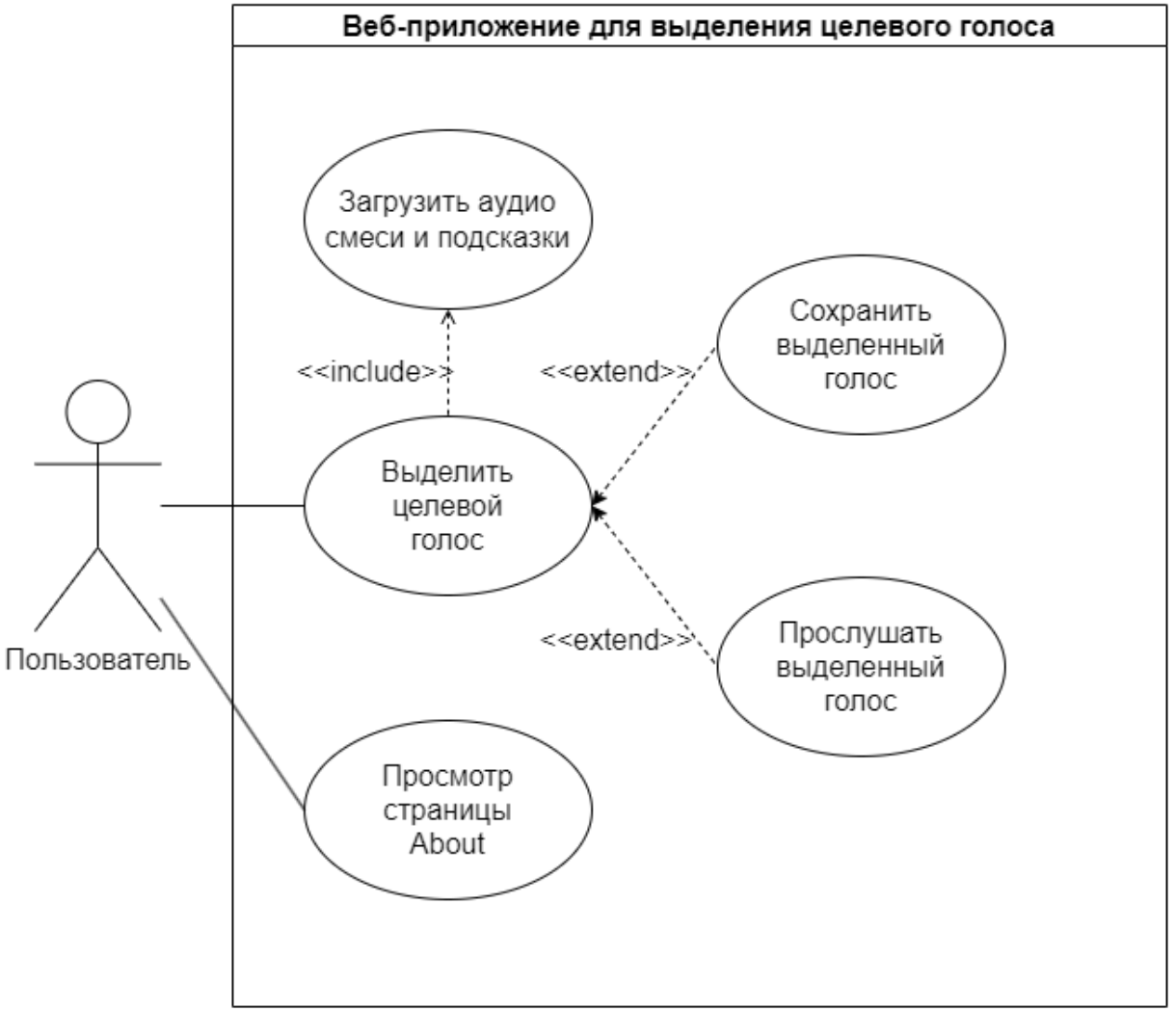
выделенный голос чуть менее четкий, громкость как в смеси, фоновая музыка тише чем в смеси, но громче чем в референсной модели

Смесь

SpeakerBeam

TSE Model

Варианты использования приложения



Средства реализации

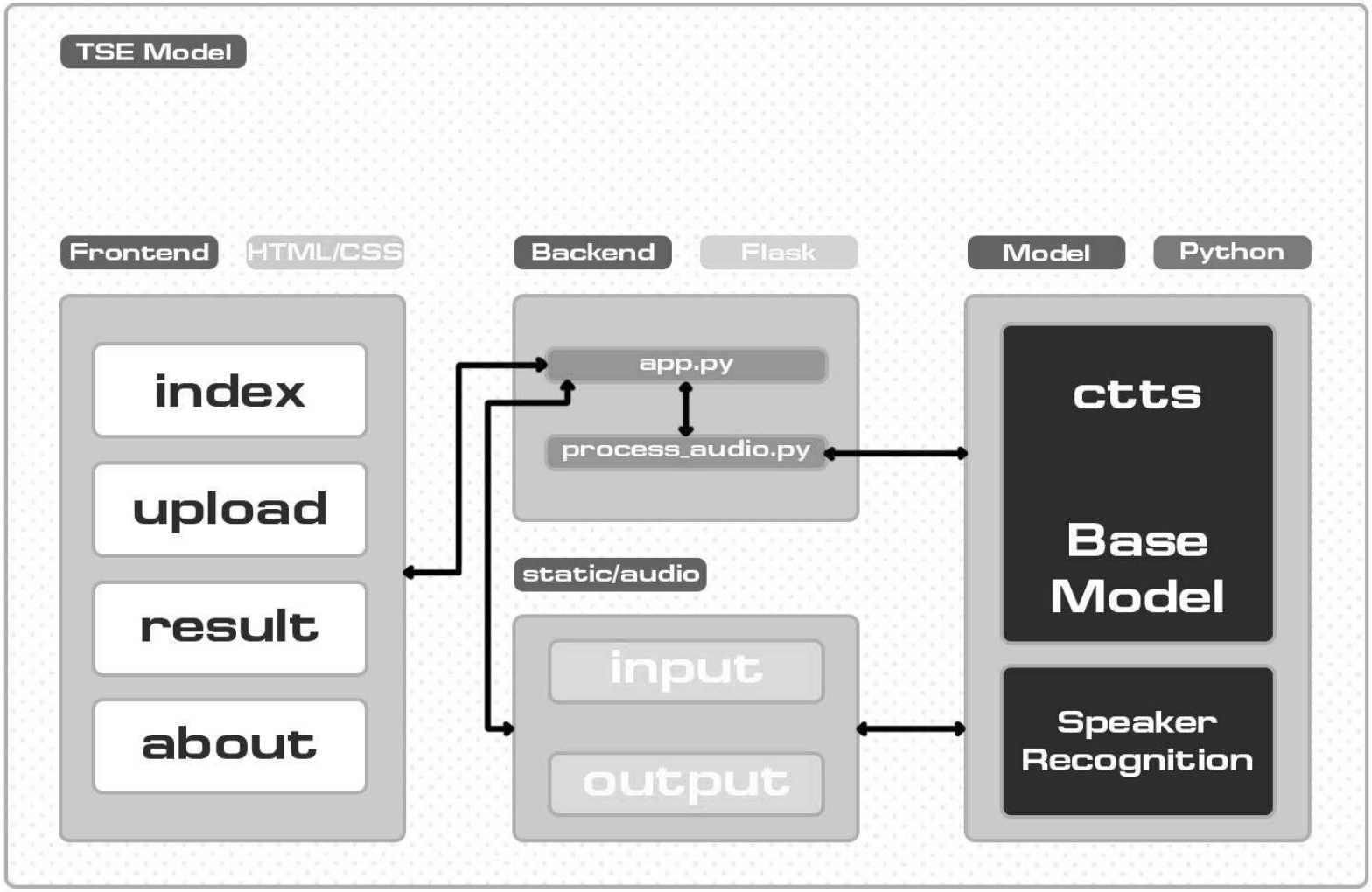
Язык разработки: Python 3.11

Фреймворки: Flask, Bootstrap 5

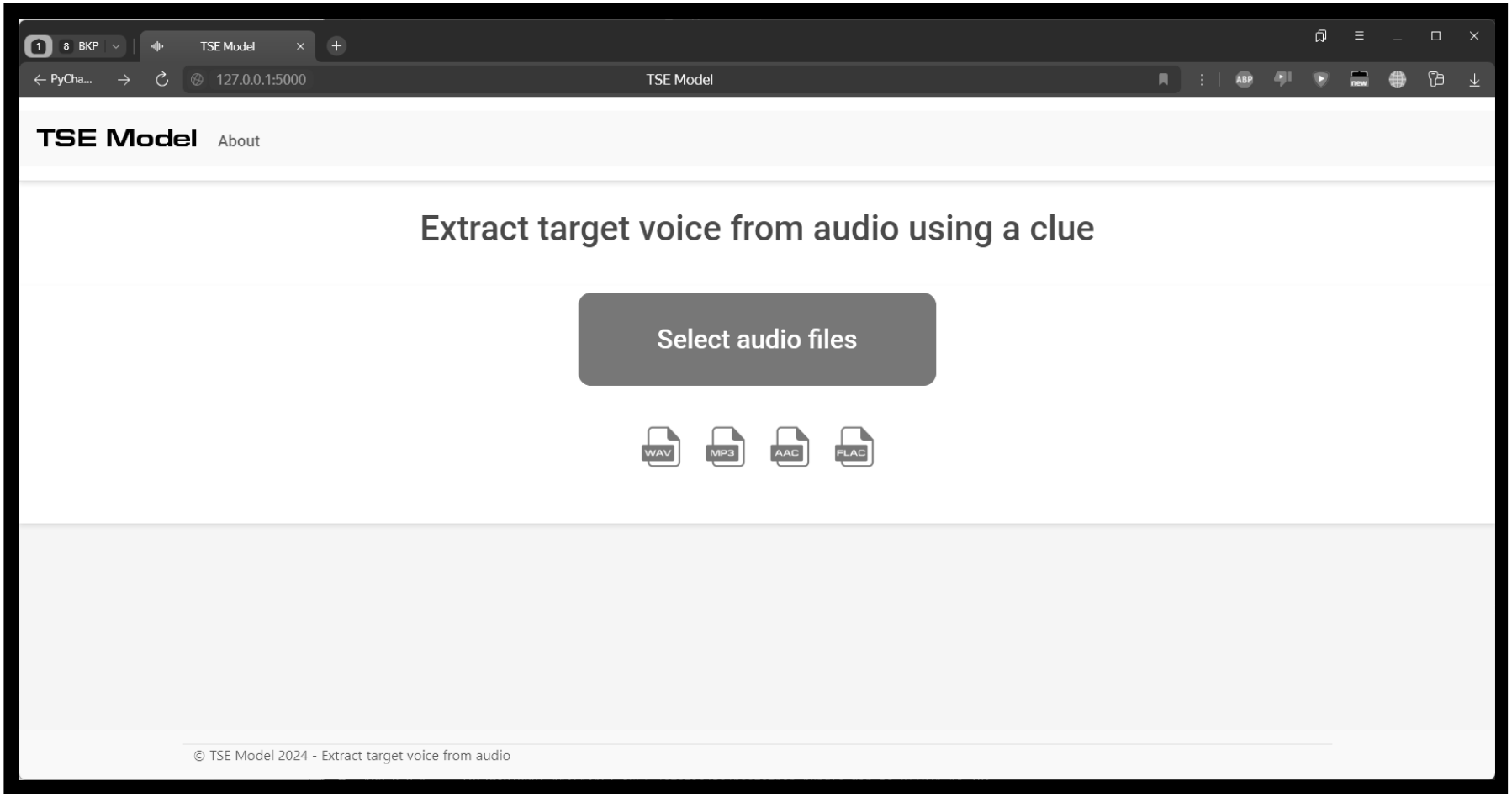
Библиотеки: librosa, soundfile, pydub

IDE: PyCharm Community Edition 2023.3.4

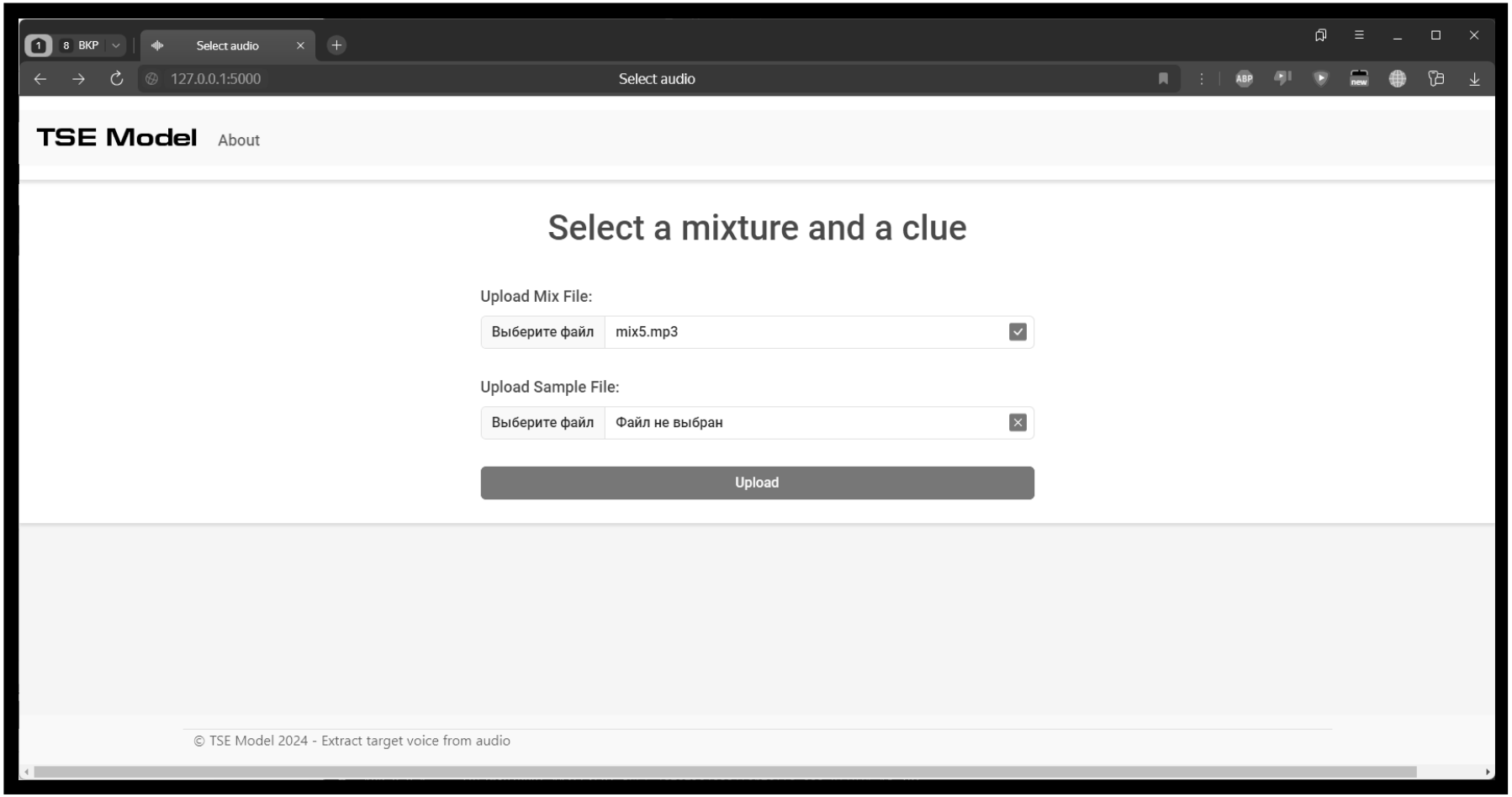
Проектирование



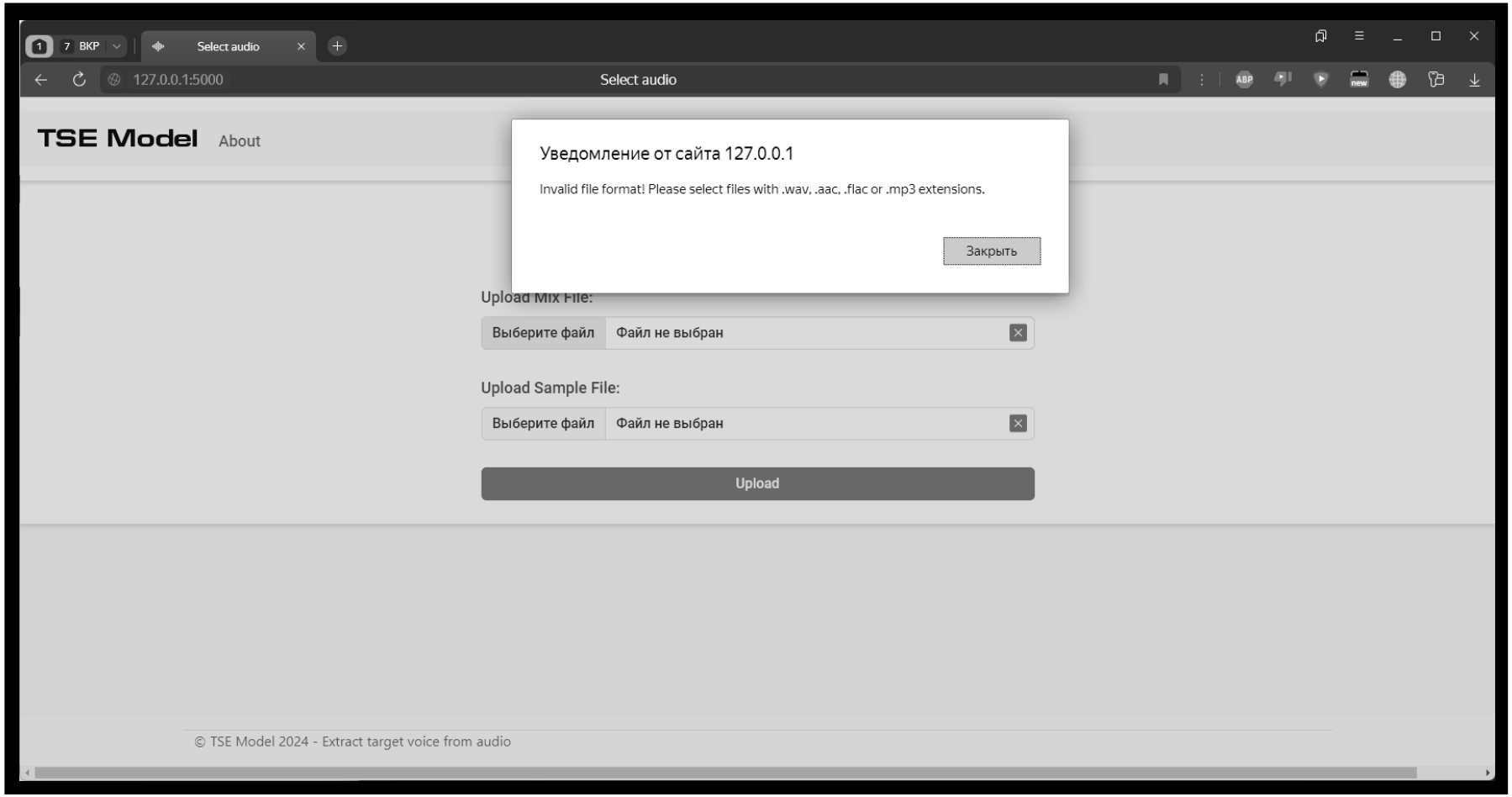
Главная страница



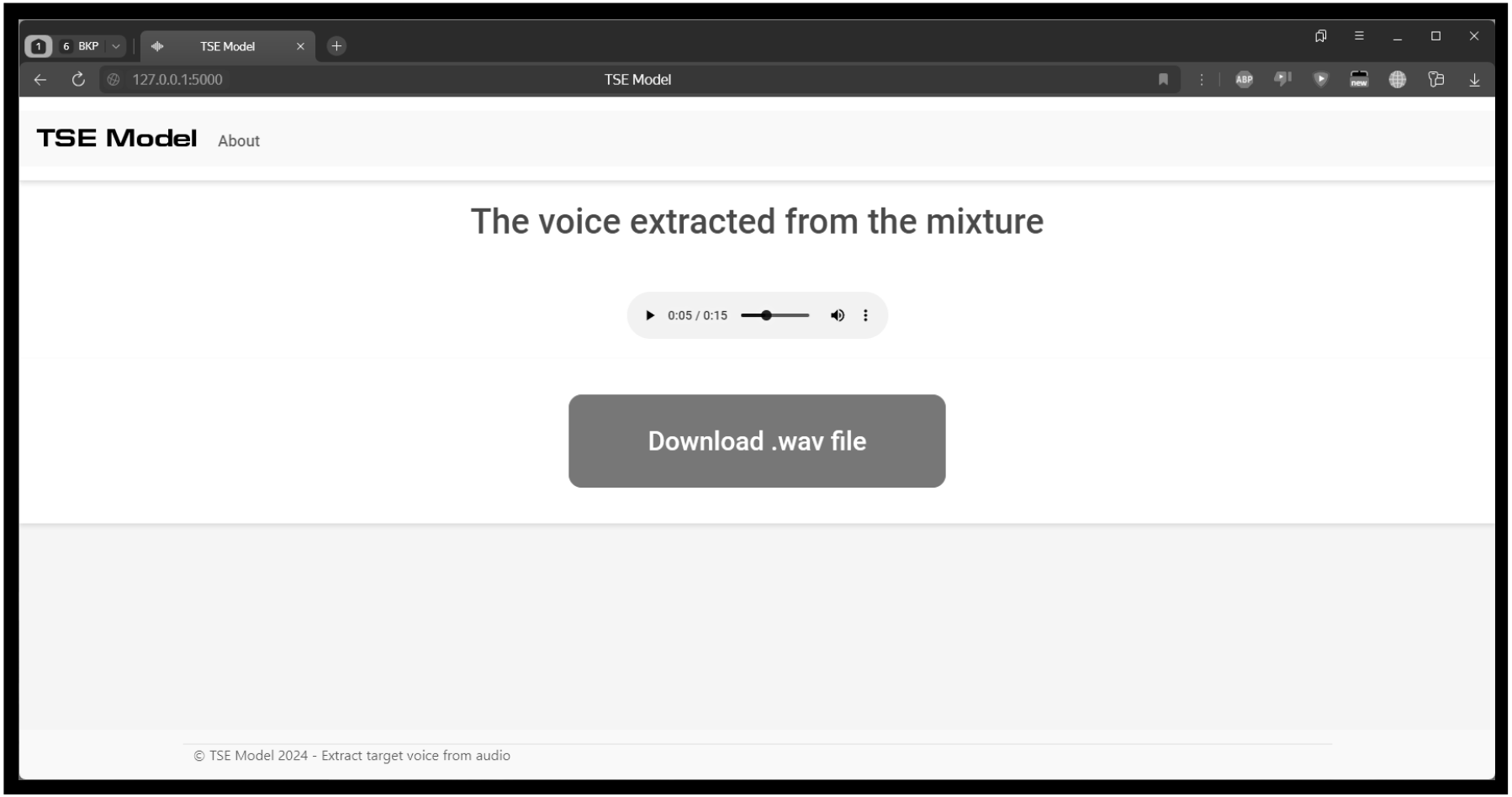
Страница загрузки файлов



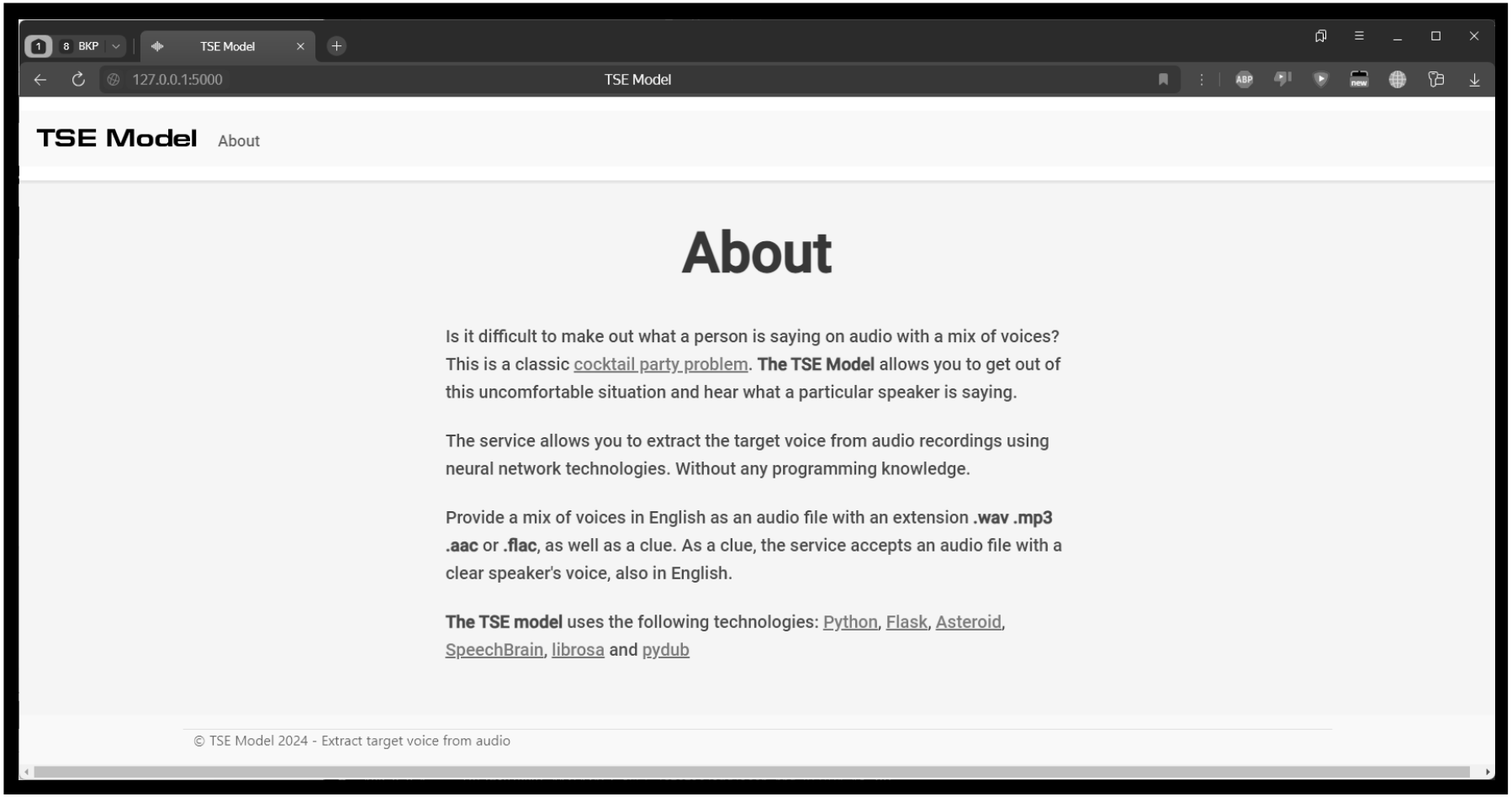
Окно с ошибкой



Страница результата выделения



Страница о сервисе



Функциональное тестирование

- Загрузка пользовательской смеси и подсказки
- Выделение голоса из смеси
- Сохранение файла с выделенным голосом в формате .wav
- Прослушивание файла с выделенным голосом
- Просмотр информации о сервисе

Результат: тестирование прошло успешно, ошибок в работе сервиса не обнаружено

Планы на будущее

1. Реализовать постобработку файлов с шумоподавлением и улучшением качества речи
2. Добавить TSE модель для трех и более перекрестных дикторов
3. Добавить регистрацию и кабинет пользователя с историей и результатами обработки данных

Основные результаты

1. Проведен обзор литературы и аналогов
2. Приведен принцип работы базовой TSE модели
3. Реализована TSE модель, проведена работа с референсной моделью и представлено их сравнение
4. Разработано и протестировано веб-приложение для выделения целевого голоса из смеси с двумя дикторами