Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка алгоритма изменения голоса на основе сверточных нейронных сетей

Выполнил: Филиппов Игорь Сергеевич, гр. 7382

Руководитель: Лисс Анна Александровна, к.т.н., доцент

Консультанты: Жангиров Тимур Рафаилович, ассистент каф. МОЭВМ
Зуева Надежда Николаевна, программист-разработчик ООО «В Контакте»

Санкт-Петербург, 2021

Цель ВКР

Цель: разработка алгоритма изменения голоса человека, обладающего свойством обучаться без параллельного корпуса данных и работать с любыми целевыми голосами, включая те, которых не было в тренировочной выборке

Актуальность: современные алгоритмы цифрового изменения голоса человека имеют недостатки:

- требует сбора параллельного корпуса данных,
- могут работать только с голосами из тренировочной выборки.

Задачи ВКР

Задачи:

1. Обзор существующих аналогов
2. Разработка нейросетевых модулей
3. Обучение нейронных сетей
4. Сравнение с аналогом

Обзор существующих аналогов

Название алгоритма / Свойства	Нейросетевой автокодировщик	Генеративно-состязательная сеть	Метод нормализующих потоков
Не требует параллельного корпуса данных для обучения	-	-	+
Работает с любыми целевыми голосами	-	-	-

Таблица сравнения существующих открытых аналогов.

Разработка нейросетевых модулей

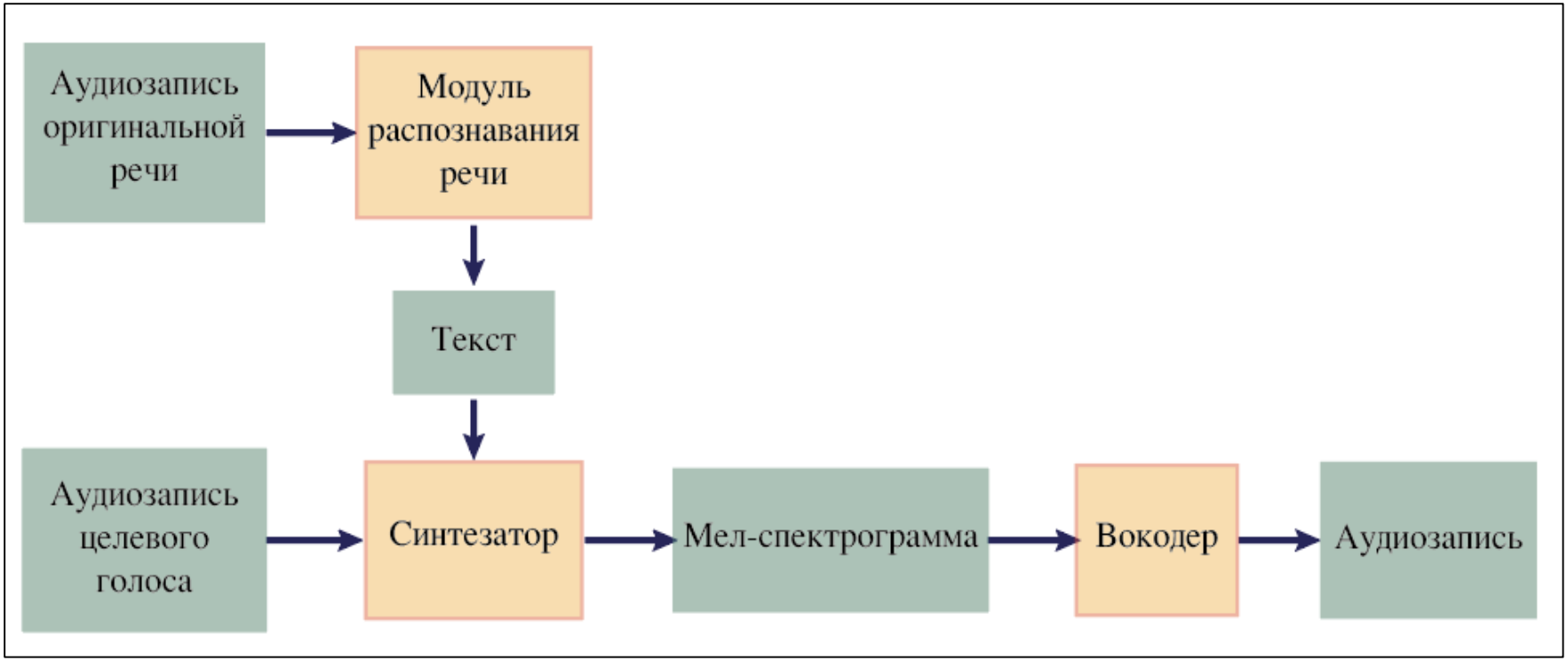

Схема архитектуры предлагаемого алгоритма.

Обучение нейронных сетей

Для обучения предлагаемого в работе алгоритма был использован набор данных VCTK от Университета Эдинбурга, Шотландия. Длина всех доступных записей составляет 46 часов.

Для увеличения размера набора данных, аудиозаписи были аугментированны. В функции аугментации была выбрана комбинация аугментаций рубильник (CutOut) и спектральной аугментации (SpecAugment). В результате размер набора был увеличен в 3 раза.

Для обучения всех нейронных моделей использовался удалённый сервер с характеристиками:

- 8 видеокарт NVIDIA DGX A100
- 1024 Гигабайта ОЗУ
- Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz

Сравнение с аналогом

$$dist(x, \hat{x}) = \frac{dist_{DTW}(x, \hat{x})}{dist_{DTW}\left(x, \hat{x}_{normflow}\right)},$$

где $dist$– расстояние между аудиозаписями речи, $dist_{DTW}$ – расстояние, вычисляемое алгоритмом динамической трансформации временной шкалы, $x, \hat{x}$ – сравнимые аудиозаписи, $\hat{x}_{normflow}$ – соответствующая речь, синтезированная с использованием метода нормализующих потоков.

Алгоритм	Средняя точность верификации, %	Среднее расстояние до эталона
Метод нормализующих потоков	65,01	1,00
Разрабатываемый метод	63,23	1,19

Таблица сравнения алгоритмов.

Заключение

- Был проведен обзор существующих аналогов для решения задачи изменения голоса человека
- Были разработаны необходимые нейросетевые модули на языке программирования Python
- Разработанные нейронные сети были обучены на открытом наборе данных
- Было проведено сравнение с методом нормализующих потоков и сделан вывод, что разработанный алгоритм сравним по качеству синтеза речи, но имеет преимущество в виде работы с любым количеством целевых голосов без переобучения

Апробация работы

Код алгоритма доступ в репозитории GitHub:
https://github.org/my_diploma_sources_repo

Пример изменения голоса доступен по адресу:

или QR-CODE