Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 534.773

Лайша
Анастасия Игоревна

Детектор голоса в слуховом аппарате на основе нейронной сети

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности «1-40 80 01 Элементы и устройства вычислительной техники и систем управления»

Научный руководитель:
Вашкевич Максим Иосифович
Доцент, кандидат технических наук

Минск 2020

КРАТКОЕ ВВЕДЕНИЕ

В настоящее время множество людей по всему миру страдает из-за проблем со слухом. Примерно 25% больных пользуется слуховыми аппаратами.

Слуховые аппараты — это электронные устройства, которые усиливают звуки выше порога слышимости пользователя с нарушениями слуха. Многие из пользователей слуховых аппаратов жалуются на дискомфорт восприятия их собственного голоса. При закрытии слухового канала слуховым аппаратом усиливаются низкочастотные компоненты голоса, а высокочастотные компоненты ослабляются ввиду звуковой проводимости костей. Это явление называется «эффект окклюзии» и является одной из критических проблем при ношении слуховых аппаратов.

Это приводит к снижению пользы слухового аппарата для пользователя и часто удерживает людей от ношения слуховых аппаратов. Этим фактом объясняется актуальность разработки системы детектирования собственного голоса в слуховом аппарате.

Для достижения данной цели необходимо решить следующие задачи:

1) разработать метод параметризации речевого сигнала, пригодного для последующего обучения нейронной сети;
2) разработать архитектуру нейронной сети для детектирования голоса пользовательского аппарата;
3) экспериментально проверить эффективность разработанного детектора.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

В настоящее время множество людей по всему миру страдает из-за проблем со слухом. Примерно 25% больных пользуется слуховыми аппаратами, однако из-за эффекта окклюзии - когда при закрытии слухового канала слуховым аппаратом усиливаются низкочастотные компоненты голоса, а высокочастотные компоненты ослабляются ввиду звуковой проводимости костей, многие из пользователей слуховых аппаратов жалуются на дискомфорт восприятия их собственного голоса.

Это приводит к снижению пользы слухового аппарата для пользователя и часто удерживает людей от ношения слуховых аппаратов. Этим фактом

объясняется актуальность разработки системы детектирования собственного голоса в слуховом аппарате.

Цель и задачи исследования

Целью данного исследования является разработка детектора голоса в слуховом аппарате на основе нейронной сети. В соответствии с поставленной целю в работе сформулированы и решены следующие **задачи**:

1) разработать метод параметризации речевого сигнала, пригодного для последующего обучения нейронной сети;

2) разработать архитектуру нейронной сети для детектирования голоса пользовательского аппарата;

3) экспериментально проверить эффективность разработанного детектора.

Объектом исследования выступает слуховой аппарат.

Предметом исследования является модель на основе нейронной сети для распознания дикторов.

Область исследования и содержание диссертационной работы соответствуют образовательному стандарту высшего образования второй ступени (магистратуры) специальности 1-40 80 01 «Элементы и устройства вычислительной техники и систем управления».

Научная новизна диссертационной работы заключается в построении новой модели на базе нейронной сети для решения задачи распознания дикторов.

Положения, выносимые на защиту:

1) разработанный метод параметризации речевого сигнала, пригодного для последующего обучения нейронной сети;

2) разработанная и обученная нейронная сеть для детектирования голоса пользовательского аппарата;

3) экспериментальные исследования эффективности работы разработанного детектора.

Апробация результатов диссертации

Основные положения и результаты диссертационной работы докладывались и обсуждались на 55-я научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, 2019).

Опубликованность результатов исследования

По результатам исследований, представленных в диссертации, опубликован тезис в сборнике и материале научной конференции.

Структура и объем диссертации

Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, четырёх глав, заключения,

библиографического списка и приложений. Общий объем диссертации – 63 страницы. Работа содержит 40 рисунков, 2 таблицы. Библиографический список включает 37 наименований, графический материал включает 15 слайдов презентации (Приложение Б).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрена структура слухового аппарата, а также описано обоснование актуальности темы и задачи магистерской диссертации.

В **общей характеристике работы** показана актуальность исследования, описанного в работе, сформулированы цель и задачи диссертации, обозначены предмет, объект и область исследования, научная (теоретическая и практическая) значимость исследования, а также апробация работы.

В первой главе приведен обзор существующих схем подавления окклюзии.

Во второй главе приведен обзор существующих систем диаризации речи, описание схемы реализации слухового аппарата со встроенным детектором голоса, а также описана структура типовой системы диаризации.

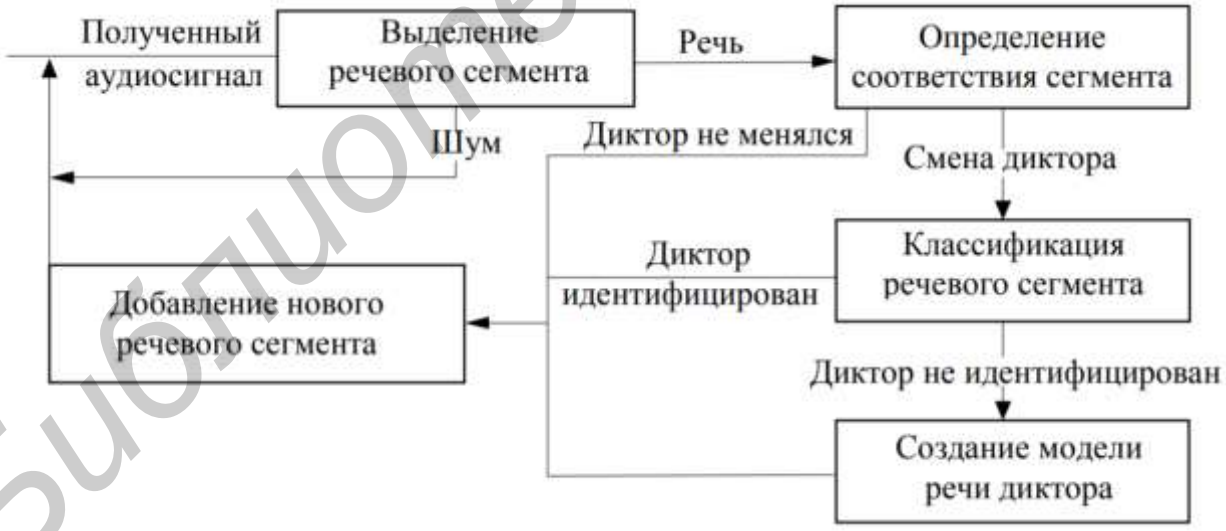

Рисунок 1– Структура типовой системы диаризации

Процесс диаризации речи сам по себе является достаточно сложной задачей, и обзор существующих систем диаризации речи на базе нейронных сетей позволяет понять, что использование большинства существующих систем требует большой вычислительной сложности, преимуществом этой системы

распознания речи являются умеренные вычислительные затраты, позволяющие использовать ее в слуховом аппарате.

Хотелось получить как можно более высокую точность. Но для обучения более сложной модели (которая должна обеспечить большую точность) потребуется больше оперативной памяти (памяти видеоплаты в случае использования графического процессора).

В третьей главе приведен выбор инструментов и модели для системы диаризации речи.

Помимо этого глава содержит подробное описание процесса обработки аудио для последующего использования в нейронной сети, а также описание архитектуры модели на основе нейронной сети с последующим описанием составных компонентов модели.

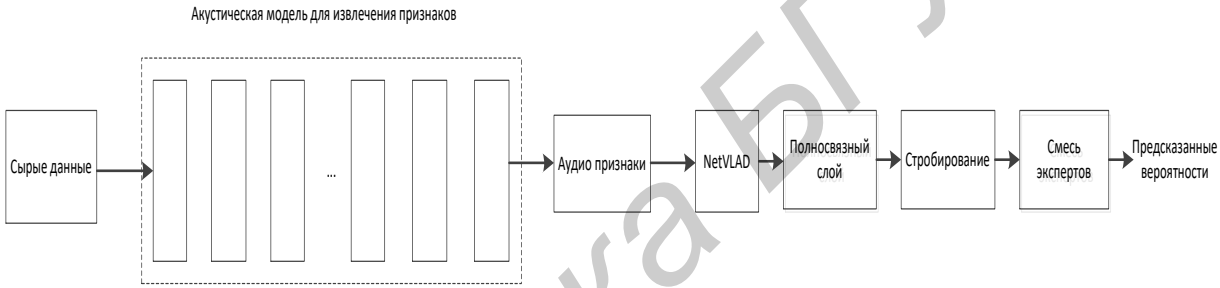

Рисунок 2– Архитектура модели

Модель классификатора состоит из слоя NetVLAD, полносвязного слоя, слоя стробирования и финального слоя Mixture of experts (Смесь экспертов).

Также в главе приведено описание метрик, используемых в оценке точности модели.

В четвёртой главе приведено описание датасета, использованного для обучения нейронной сети, а также результаты обучения модели.

Модель построена с использованием Tensorflow. Лучшую производительность для модели дает скорость обучения 0,0002 с затуханием скорости обучения 0,8 каждые 1000000 шагов.

Чтобы не возникло переобучения (проблем работы с новыми данными из-за высокой скорости), необходимо использовать регуляризацию – понижение сложности модели с сохранением параметров. В данной работе использовалось несколько методов регулязации: пакетная нормализация, а также метод адаптивной оценки моментов с L2-регулязацией.

Команда для запуска скрипта выглядит следующим образом:

```
python train.py --
train_data_pattern=/path_to_data/audioset_v1_embeddings/bal_train/*.tfrecord --
num_epochs=100 --learning_rate_decay_examples=400000 --
feature_names=audio_embedding --feature_sizes=128 --frame_features --
batch_size=512 --num_classes=527 --train_dir=/path_to_logs --model=ModelName
```

Рисунок 3 – Команда для запуска скрипта

Процесс обучения и оценки модели показан на рисунках ниже:

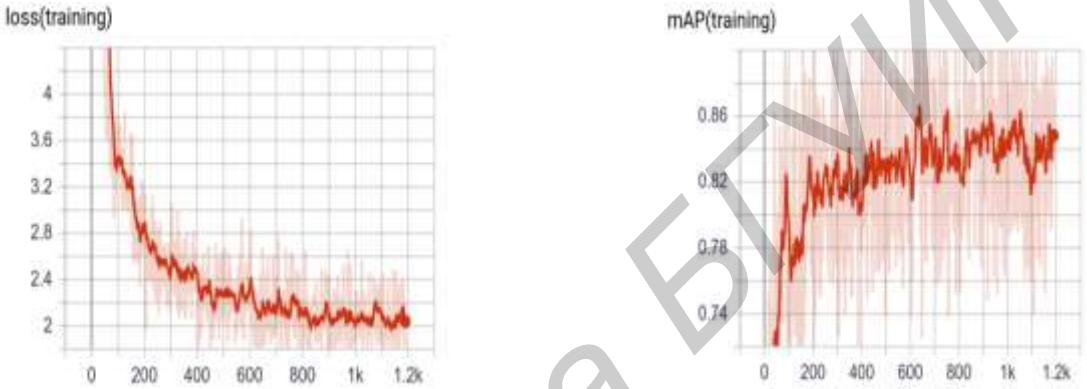

Рисунок 4 – Процесс обучения модели классификатора

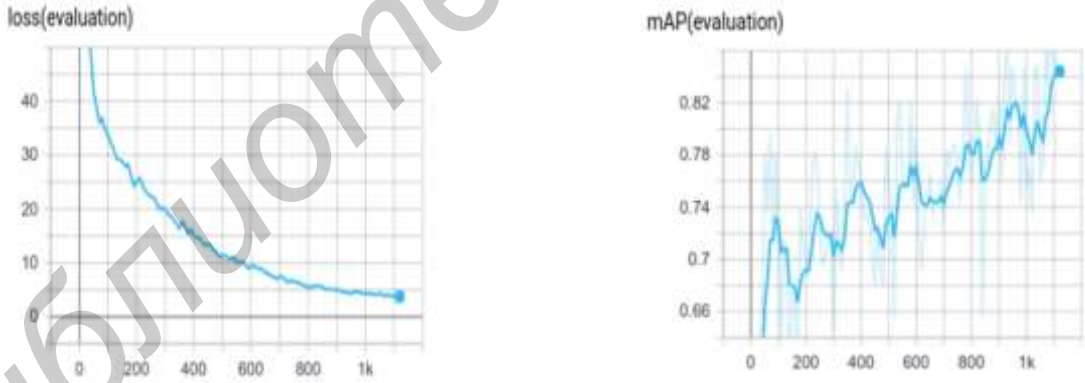

Рисунок 5 – Процесс оценки модели классификатора

С целью сопоставления предсказаний и реальности в работе используется матрица ошибок (confusion matrix). Матрица ошибок была построена с помощью функции из документации sklearn.

Матрица ошибок — это способ разбить объекты на четыре категории в зависимости от комбинации истинного ответа и ответа алгоритма.

Основные термины:

1. TP — истино-положительное решение;

2. TN — истино-отрицательное решение;
3. FP — ложно-положительное решение (ошибка первого рода);
4. FN — ложно-отрицательное решение (ошибка второго рода).

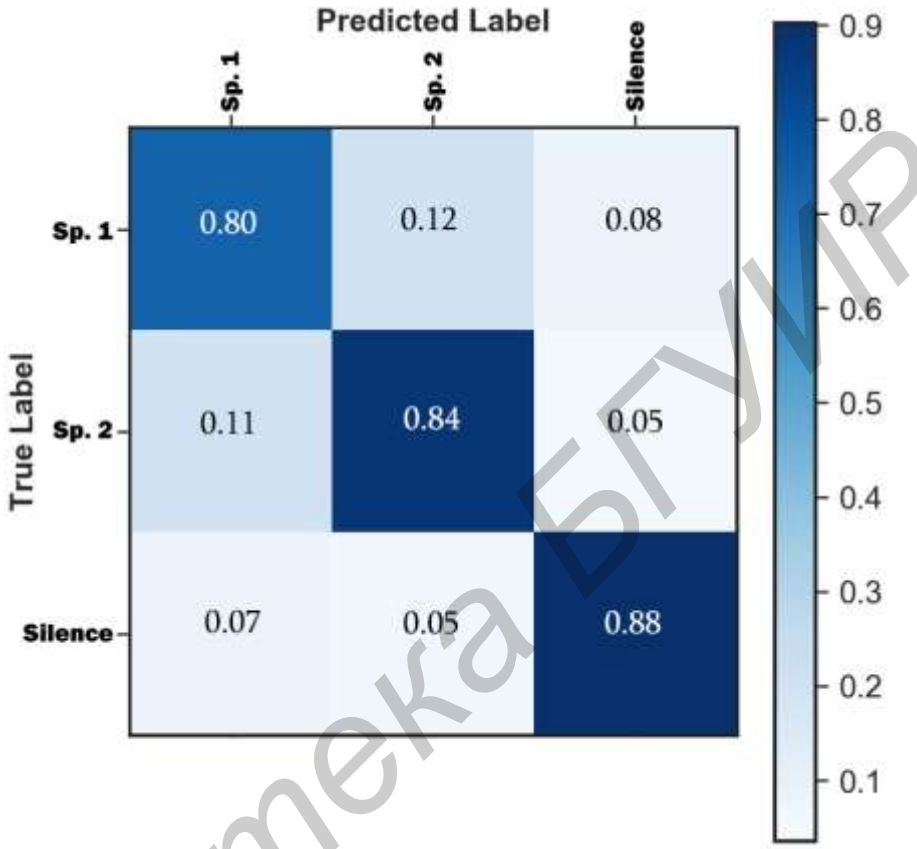

Рисунок 6 – Матрица ошибок

Чтобы получить значение точности их матрицы спутывания, мы делим общее количество правильно классифицированных положительных точек на общее количество предсказанных положительных точек. Высокая точность указывает, что точка, помеченная как положительная, действительно является положительной (небольшое количество FP).

$$Precision = \frac{TP}{TP + FP} \tag{4.8}$$

где $precision$ показывает, какая доля объектов, выделенных классификатором как положительные, действительно является положительными:

Для рисунка 6 , mAP (итоговая метрика) будет рассчитываться, как:

$$Mean\ average\ precision = \frac{P(Speaker\ 1) + P(Speaker\ 2) + P(Silence)}{3} \quad (1)$$

где

$$P(Speaker\ 1) = \frac{0.8}{0.8 + 0.12 + 0.08} = 0.8 \quad (2)$$

$$P(Speaker\ 2) = \frac{0.84}{0.84 + 0.11 + 0.05} = 0.84 \quad (3)$$

$$P(Silence) = \frac{0.88}{0.88 + 0.07 + 0.05} = 0.88 \quad (4)$$

Итоговая точность модели составляет 84%.

Результаты сравнения модели с имеющейся моделью для идентификации и кластеризации дикторов с использованием сверточной нейронной сети:

Таблица 1 – Результат сравнения моделей

название модели	итоговая точность	время обучения
Модель VLAD	84%	1h 45m 53s
Модель на основе CNN	96.5%	9h 42m 52s

Несмотря на большую итоговую точность второй модели, время обучения модели существенно больше первой в виду большего количества слоев.

ЗАКЛЮЧЕНИЕ

В работе предложена разработка и экспериментальное исследование разработка детектора голоса в слуховом аппарате на основе нейронной сети.

Сперва приведены способы уменьшения окклюзии в слуховом аппарате. Затем был произведен обзор существующих систем диаризации речи. По результатам обзора была выбрана модель на основе нейронной сети LSTM, так как данная сеть позволяет решить поставленную задачу с учетом выбранного набора данных.

Для обучения модели необходимы данные, на основе которых обучается модель. Вариантом решения для данной работы стал набор данных, который основан на размеченных видео фрагментах YouTube и доступен для загрузки в двух форматах:

1. CSV-файлы, в которых содержится следующая информация о каждом фрагменте: ID размещенного видео, время начала и окончания фрагмента, одна или несколько присвоенных отрывку меток.

2. Извлеченные аудиопризнаки, которые сохраняются в виде файлов TensorFlow – это 128-мерные аудиопризнаки, извлеченные с частотой 1 Гц.

Для использования аудиоданных в обучении модели была произведена предварительная обработка данных с помощью VGG-алгоритма.

В ходе работы представлена архитектура модели классификатора. Сырые аудиоданные проходят через акустическую модель, Извлеченные признаки представляют собой последовательность 128-мерных векторов 8-битных целых чисел без знака. Затем извлеченные признаки подаются на NetVLAD-слой, и каждый признак преобразуется в единое представление. Эти индивидуальные представления затем объединяются и подаются на полносвязный слой для уменьшения их размерности. Выходом этого слоя является компактный 1024-мерный вектор. Затем следует слой стробирования - обучаемая нелинейная единица, целью которой является моделирование взаимозависимостей между активациями сети с помощью стробирования. Цель слоя - перераспределить веса признаков в векторе, выявляя и фиксируя зависимости между признаками. После этого следует классификатор “Смесь экспертов”, на вход он принимает итоговое аудио представление и выводит набор меток для аудио вместе с их баллами.

Для проверки качества обучения модели классификатора была использована mAP (Mean average precision) в качестве метрики качества.

В конечном итоге создана обученная нейронная сеть для распознания дикторов. Точность модели составляет 84%. Итоговая модель была сравнена с имеющейся моделью для идентификации и кластеризации дикторов с использованием сверточной нейронной сети. Сравнение моделей показало меньшую точность разработанной модели, однако разработанная модель затрачивает меньше времени на обучение. Учитывая специфику поставленной задачи - разработка детектора голоса в слуховом аппарате, более оптимальным критерием является время обучения, а не сложность модели, так как для обучения более сложной модели (которая должна обеспечить большую точность) потребуется больше оперативной памяти (памяти видеоплаты в случае использования графического процессора).

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1.А. Лайша, А. И. Применение детектора голоса в слуховом аппарате / А. И. Лайша // Компьютерные системы и сети: 55-я юбилейная научная

конференция аспирантов, магистрантов и студентов, Минск, 22-26 апреля 2019 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2019. – С. 271 – 273. (https://libeldoc.bsuir.by/handle/123456789/35269)