АЛГОРИТМ ОБНАРУЖЕНИЯ РЕЧЕВОЙ АКТИВНОСТИ В АКУСТИЧЕСКОМ СИГНАЛЕ С ПРИМЕНЕНИЕМ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

А.Б. Тепляков, В.Г. Спицын
Томский политехнический университет
E-mail: abt4@tpu.ru

Введение

В настоящее время голосовой пользовательский интерфейс приобрел широкую популярностью. Такие голосовые помощники как Cortana, Siri, Ok Google, Алиса ежедневно обрабатывают значительное количество запросов, увеличивая удобство выполнения рутинных операций [1].

Также усовершенствованная обработка речевого сигнала может помочь людям с нарушениями слуха. Современные слуховые аппараты усиливают желаемый речевой сигнал и подавляют шумовые компоненты [2].

Хотя существуют различные варианты применения технологий обработки речевого сигнала, разработанные алгоритмы сталкиваются с общей проблемой: необходимо обнаружить присутствие речи в акустическом сигнале, который зачастую искажен шумом.

В данной работе рассматривается алгоритм обнаружения голосовой активности (Voice Activity Detection) во входном акустическом сигнале для отделения активной речи от фонового шума или тишины с помощью свёрточной нейронной сети.

Характеристики речевой активности

Как известно, в вычислительной технике звуковой сигнал представлен в виде последовательности двоичных чисел, полученных аналогово-цифровым преобразователем через равные промежутки времени (период дискретизации) [3]. Такие последовательности могут быть достаточно длинными, поэтому для изучения акустического сигнала принято рассматривать его порции методом скользящего окна.

Часто используемой характеристикой для детектирования речевой активности во временной области сигнала, является краткосрочная энергия (Short Term Energy).

Предполагая, что речевые компоненты имеют более высокие значения мощности по сравнению с фоновым шумом, для обнаружения речи может применяться пороговое значение. Однако фиксированный порог требует априорных знаний об уровнях шума и речи. Нормализация мощности увеличивает различимость между компонентами речи и шума, однако нестационарные помехи, такие как ударные шумы, вызывают ложные срабатывания детектора речевой активности на основе краткосрочной энергии.

Бóльшая устойчивость к шумам достигается при отображении сигнала из временной области в частотную, что можно осуществить дискретным преобразованием Фурье [4].

Отсчеты для преобразования выбираются тем же методом скользящего окна, а затем отображаются на графике, который называется спектрограммой.

При применении спектрограмм в качестве входных данных можно, используя методы глубокого обучения, натренировать модель автоматически выделять признаки, отличающие речевую активность от всего остального.

В силу того, что спектрограмма является изображением, следует использовать свёрточную нейронную сеть, так как данная модель успешно решает задачу классификации [5].

Подготовка обучающей выборки

Для получения высокой обобщающей способности у модели глубокого обучения необходимо иметь большой и качественно размеченный набор данных. В данном случае он должен включать в себя как можно более разнообразную голосовую активность индивидов. Для этих целей были использованы открытые наборы данных Common Voice [6] и Voxforge [7].

Также для обучения модели необходимы примеры, не являющиеся речевой активностью. Для этой цели используются наборы данных ESC-50 [8] и Urban Noises [9].

Все звуковые файлы проходят через децимацию на частоту дискретизации 4000 Гц. После этого, из файлов с речевой активностью вручную были удалены участки тишины. Затем к аудиозаписям, содержащим только речевую активность, были добавлены шумы из набора Urban Noises. Также, использовались такие аугментации звука, как изменение громкости и скорости воспроизведения.

После этого из каждого файла была получена спектрограмма. Входными изображениями для модели становятся изображения в оттенках серого 128 на 128 пикселей, то есть участки спектрограммы сигнала продолжительностью 128 миллисекунд и с мощностями гармоник сигнала на частотах от 0 до 1270 Гц с шагом 10 Гц. Примеры спектрограмм можно увидеть на рисунке 1.

Рис. 1. Спектрограммы для обучения модели (слева направо: речь, дрель, пианино)

Архитектура модели и её обучение

Для реализации алгоритма используется язык программирования Python с библиотекой глубокого обучения PyTorch.

В ходе работы были исследованы несколько архитектур сверточных нейронных сетей, однако при поиске компромисса между точностью классификации и вычислительной сложностью алгоритма наиболее успешной оказалась модель, представленная на рисунке 2.

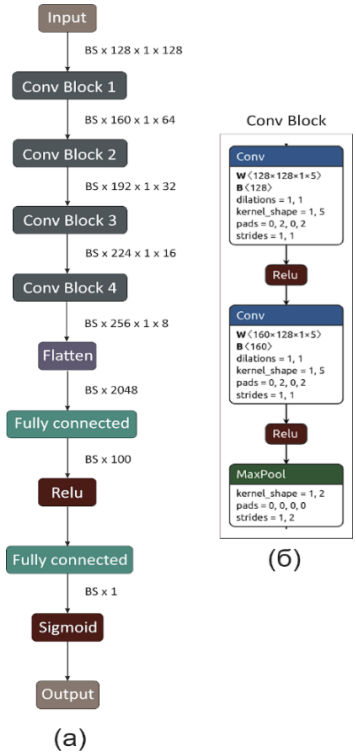

Рис. 2. а) архитектура сети; б) содержание свёрточного блока

Набор данных содержит порядка 200000 изображений, полученных из 7 часов аудиофайлов, половина из которых относится к классу речевой активности. Разделение на обучающую и тестовую выборки производилось в соотношении 70 на 30 %. В качестве алгоритма оптимизации использовался Adam, скорость обучения равна 0,0001, размер пакета 64, коэффициент регуляризации 0,001. Обучение длилось 15 эпох. Оценка модели производилась по точности (Precision) и полноте (Recall) [10].

Заключение

Следует отметить, что разработанный алгоритм обнаружения речевой активности показывает точность 0,929 и полноту 0,938 при пороговом значении 0,5 на выборке из порядка 62000 изображений, что является достаточно обнадеживающим показателем. Подход, используемый в данном исследовании может быть использован для других задач классификации акустических событий, например, при решении задач ежегодного соревнования DCASE [11].

В дальнейшем следует провести статистически значимое сравнение разрабатываемой модели с существующими алгоритмами обнаружения речевой активности при различных отношениях сигнал-шум. Также необходимо исследовать другую форму входных данных для модели. Не менее важным является рассмотрение различных топологий нейронных сетей, например, рекуррентных моделей [12].

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-08-00977 А и в рамках Программы повышения конкурентоспособности ТПУ.

Список использованных источников

1. The Best Voice Assistants [Электронный ресурс] / Anne Dennon – URL: https://www.reviews.com/voice-assistant/ (дата обращения: 13.10.2019)
2. Deep Learning Reinvents the Hearing Aid [Электронный ресурс] / DeLiang Wang. – Электрон. журн. – Spectrum IEEE, 2016. – URL: https://spectrum.ieee.org/consumer-electronics/audiovideo/deep-learning-reinvents-the-hearing-aid (дата обращения: 14.08.2019)
3. Sound Representation [Электронный ресурс] / Teach Computer Science – URL: https://teachcomputerscience.com/sound-representation/ (дата обращения: 3.07.2019).
4. Лайонс Р. Цифровая обработка сигналов: Второе издание. Пер. с англ. - М.: ООО “Бином-Пресс”, 2006 г. - 656 с.: ил.
5. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of IEEE 86 (11) (1998) 2278–2324.
6. Common Voice [Электронный ресурс] / Mozilla – URL: http://voice.mozilla.org (дата обращения: 18.07.2019)
7. VoxForge [Электронный ресурс] / URL: http://www.voxforge.org/ (дата обращения: 18.07.2019)
8. K. J. Piczak. ESC: Dataset for environmental sound classification, in Proceedings of the ACM International Conference on Multimedia. ACM, 2015, in press
9. Urban Sound Dataset [Электронный ресурс] / URL: https://urbansounddataset.weebly.com (дата обращения: 18.07.2019)
10. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves. // Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006
11. Detection and Classification of Acoustic Scenes and Events [Электронный ресурс] / URL: http://dcase.community (дата обращения: 20.08.2019)