ДЕТЕКТОР РЕЧЕВОЙ АКТИВНОСТИ НА ОСНОВЕ СВЁРТОЧНОЙ НЕЙРОННОЙ СЕТИ С РЕКУРРЕНТНЫМИ СВЯЗЯМИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Вашкевич Г.С.

Азаров И.С. – д.т.н., доцент

В данной работе исследуется задача обнаружения речевой активности в зашумленном звуковом сигнале. Классические подходы к решению данной задачи используют MFCC в качестве базовых признаков, и вычислительно-сложные модели для их классификации, что плохо сказывается на производительности. Мы представляем модель свёрточной нейронной сети с рекуррентными связями использующую информацию о гармонической природе речевого сигнала в качестве базовых признаков. В сравнении с другими нейросетевыми моделями, предложенная модель имеет на несколько порядков меньше настраиваемых параметров, что позволяет ей быстрее обучаться и обладать низкими вычислительными затратами.

В работе [1] нами была предложена модель детектирования речевой активности на основе свёрточной нейронной сети. Текущая работа развивает идею выделения гармонических компонент речевого сигнала при помощи свёрточной нейронной сети. Мы добавляем еще один свёрточный слой с рекуррентными связями, который должен учитывать небольшие изменения тона речевого сигнала во времени.

Как видно на рисунке 1, частота гармонических компонент речевого сигнала (слева) медленно изменяется во времени, в то же время частотные пики шумового сигнала (справа) почти хаотично разбросаны как вдоль частотной, так и вдоль временной оси. Задача нового слоя заключается в том, чтобы сеть смогла обнаружить именно плавные изменения частотных пиков во времени.

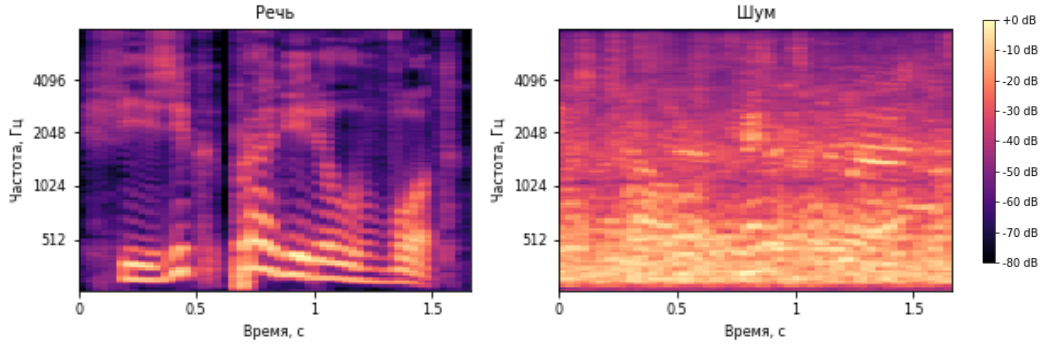

Рис.1 –Спектрограмма речевого и шумового сигналов

На рисунке 2 представлена архитектура усовершенствованной модели нейронной сети. Данная модель отличается лишь тем, что после второго свёрточного слоя базовой модели добавлен еще 1 слой. Данный слой работает следующим образом:

1) каждая из *SEQ_LEN* матриц, размером *N_CANDx1* сворачивается с обучаемым ядром размером *N_CHx1*;
2) полученная матрица умножается на настраиваемый коэффициент обратной связи и результат умножения суммируется с аккумулятором, который хранит сумму всех предыдущих фреймов;
3) к аккумулятору прибавляется настраиваемый параметр смещения, результат пропускается через логистическую функцию активации.

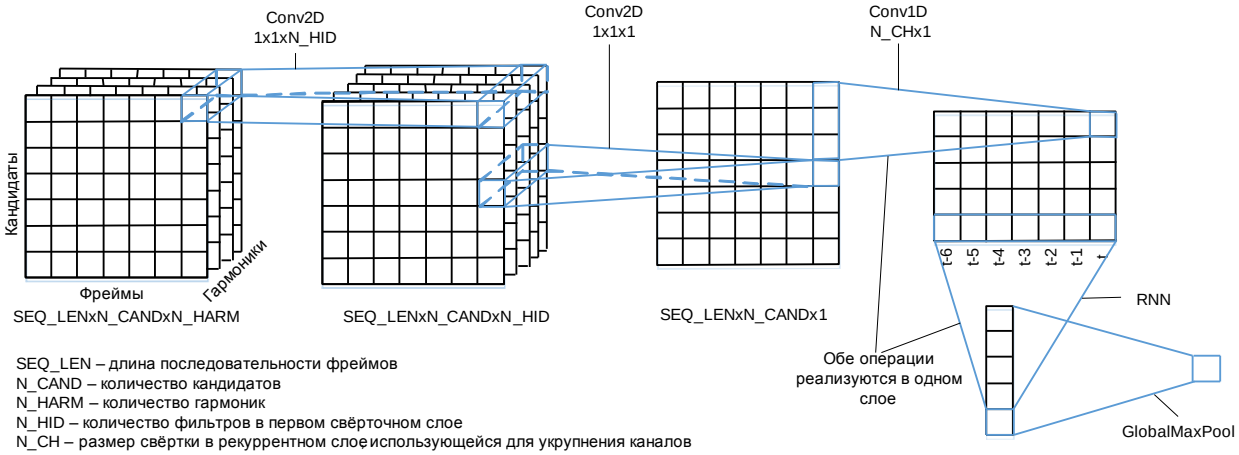

Рис.2 – Архитектура предложенной модели

Результатом работы данного слоя является матрица, размером *N_CHx1*, максимальное значение которой и будет выходом всей нейронной сети. Т.к. значения выхода сети представляют собой активации логистической функции – их можно интерпретировать как вероятность наличия шума в заданном входном примере. Высокое значение активации свидетельствует о высокой вероятности наличия во входном примере голосовой активности, низкое значение активации, в свою очередь, свидетельствует о наличии во входном примере шума.

Обучение модели проходило на открытом наборе данных Musan [2]. Данный набор данных содержит 3 типа звуковых записей – речь, шум, и музыка. В своих экспериментах мы использовали лишь записи речи и шума. Шумовые записи подмешивались в записи с речью с определенным коэффициентом, который определялся требуемым соотношением сигнал/шум (SNR). Значения SNR (в дБ) выбирались случайным образом из ряда [-10, -5, 0, 5, 10, 15, 20]. Коэффициент *k*, на который необходимо было умножить шумовой сигнал рассчитывается следующим образом:

$$SNR = 20 * log_{10}\left(\frac{A}{k * A_n}\right) => k = 10^{log_{10}\left(\frac{A}{A_n}\right) - \frac{SNR}{20}},$$

где *A* – мощность речевого сигнала;
A_n – мощность сигнала с шумом.

Т.к. мощность сигнала – это квадратичная величина, то окончательное выражение для расчёта коэффициента *k* имеет вид:

$$k = 10^{\frac{log_{10}\left(\frac{A}{A_n}\right) - \frac{SNR}{20}}{2}}.$$

Значение параметра *N_CH* было принято равным 32. Все остальные значение гипперпараметров по сравнению с базовой работой остались неизменными.

Обучение модели производится в 2 этапа. Необходимо отдельно обучать свёрточную и рекуррентную части всей сети. Свёрточная часть обучается аналогичным образом, как это происходило в базовой работе. При обучении рекуррентной части обучаются только веса рекуррентного слоя, который включает в себя параметры свёртки, укрупняющей каналы, коэффициент обратной связи, и коэффициент смещения.

Общее количество настраиваемых параметров предложенной сети составляет 2083 коэффициента, что является очень малым значением по сравнению с современными архитектурами нейронных сетей.

Результаты тестирования предложенной и базовой модели на одном и том же наборе данных представлен в виде матриц запутанности на рисунке 3.

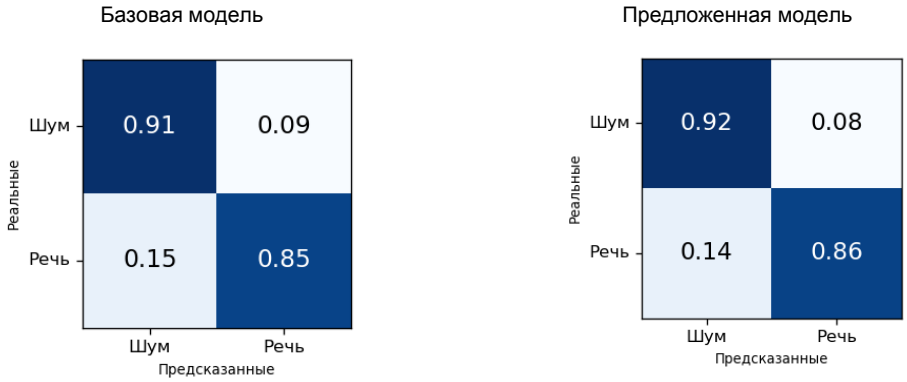

Рис.3 – Результат тестирования моделей

Из матриц запутанности на рисунке 3 видно, что предложенная в данной работе модель обладает немного лучшей точностью детектирования речевого сигнала. При этом по сравнению с базовой, новая модель стала сложнее лишь на 34 настраиваемых параметра, что почти не сказывается на вычислительной сложности.

Список использованных источников:

1. Вашкевич, Г. Свёрточная нейронная сеть с семантически-значимыми активациями для анализа речи / Г. Вашкевич, И. Азаров // Открытые семантические технологии проектирования интеллектуальных систем: материалы международной научно-технической конференции, Минск, 15-17 февраля 2018 г. / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: В. В. Голенков (гл. ред.) [идр.]. – Минск, 2018. – С.227.

2. Snyder, D. MUSAN: A Music, Speech, and Noise Corpus // arxiv preprint, arXiv:1510.08484v1. – 2015.