СИСТЕМА РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО ГОЛОСУ НА ОСНОВЕ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Марчуков В.Ю.
Научный руководитель – Усик В.В.
Харьковский национальный университет радиоэлектроники
(61166, Харьков, просп. Науки,14, каф. МИРЭС тел. 70-21-587)
e-mail: vitalii.marchukov@nure.ua

Emotion recognition by voice finds application in many areas: the development of social assistance robots, autonomous vehicles, equipment for neuro-feedback, etc. However, at the moment there is no sufficiently effective solution to this problem. According to recent studies, high efficiency in various tasks of recognition are convolutional neural networks.

Распознавание эмоций по голосу находит применение во множестве областей: разработке социальных вспомогательных роботов, автономных транспортных средств, оборудовании для нейро-обратной связи и т. д. Однако в настоящий момент не существует достаточно эффективного решения данной задачи. Согласно последним исследованиям высокую эффективность в различных задачах распознавания имеют сверточные нейронные сети.

Первый шаг в распознавании - необходимо преобразовать звуковую волну в цифровой вид. Для этого выполняется процедура дискретизации звуковой волны. Частотный диапазон человеческой речи укладывается в полосу 4кГц, тогда, согласно теореме Котельникова, для восстановления сигнала достаточно частоты дискретизации 8кГц. После выполнения этой операции на выходе будет получен массив чисел, каждое из которых представляет амплитуду звуковой волны через интервалы 1/8000 секунды.

В качестве характерных особенностей исходного сигнала используются MFCC (Mel-Frequency Cepstral Coefficients) . Для извлечения MFCC необходимо разложить звуковую волну на отдельные составляющие. Делается это при помощи дискретного преобразования Фурье.Полученные звуковые волны необходимо разложить на мел шкале, используя треугольную оконную функцию, в данной работе мел-частотное пространство разбивалось на 12 отрезков.Полученные мел-частотные кепстральные коэффициенты, можно в дальнейшем использовать в качестве уникальной характеристики входной звуковой волны.

Загрузить аудиоданные и преобразовать их в формат MFCC можно легко с помощью пакета Python librosa.

Построение нейронной сети и ее обучение

В данной работе используется сверточная нейронная сеть, которая имеет следующую архитектуру:

1. Первый уровень обработки представляет собой два сверхточных слоя и max pooling с 128 фильтрами размера 5, имея на выходе 216x128x2

нейронов.

2. Второй уровень обработки состоит из трех слоев свертки также с 128 фильтрами. На выходе второго слоя модель имеет 216x128x3 нейронов.

3. Последний уровень обработки представляет собой полносвязный слой, который производит классификацию по 10 эмоциям.

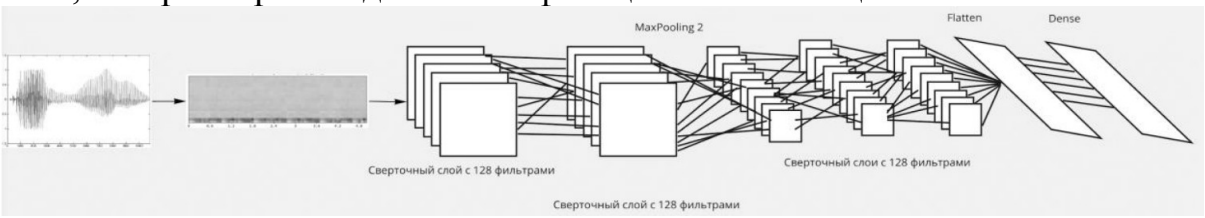

Рис. 1. Алгоритм определения эмоций
при помощи сверточной нейронной сети

Сверточная нейронная сеть была реализована на языке Python с использованием библиотек TensorFlow и Keras. На вход сеть принимает 3-х мерную матрицу (Количество входных аудио файлов, количество строк, количество каналов входных аудио файлов).

В проведенной работе была разработана система распознавания эмоций по голосу на основе сверточной нейронной сети. По результатам экспериментов точность данной системы составила 73%, что сопоставимо с другими алгоритмами, но не является пределом. В дальнейшем планируется расширить класс акустических признаков, а также использовать двухмерную сверточную нейронную сеть, что в конечном итоге должно повлиять на точность распознавания эмоций диктора.

Использованые источники:

1. H. Kun, Yu. Dong, and I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, proceedings of INTERSPEECH, ISCA, Singapore, pp. 223-227, 2014.

2. Белов Ю.С., Гришунов С.С. Основные математические методы выделения речевых особенностей в системах распознавания диктора 2015. С. 57-62..

3. Галушкин, А.И. Нейронные сети: основы теории.