УДК 004.42 **Информационные технологии**

Сергунов Дмитрий Игоревич, *студент КФ МГТУ им. Н.Э. Баумана*
Артемова Анна Александровна, *студент КФ МГТУ им. Н.Э. Баумана*
Гришунов Степан Сергеевич, *ассистент кафедры «Программное обеспечение ЭВМ, информационные технологии» КФ МГТУ им. Н.Э. Баумана*

СИСТЕМА РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО ГОЛОСУ НА ОСНОВЕ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Аннотация: В статье приводится описание разработанной системы распознавания эмоций по голосу на основе сверточной нейронной сети: описывается процедура предобработки звуковой информации и архитектура примененной сети, состоящей из трех частей. Приводятся результаты тестирования системы на основе набора данных RAVDESS и сравнение полученных результатов с наиболее популярными алгоритмами.

Ключевые слова: голос, распознавание эмоций, сверточные нейронные сети, RAVDESS, MFCC, Python.

Annotation: The article describes the voice recognition system based on the convolutional neural network. It describes the procedure of pre-processing of sound information and the architecture of the applied network consisting of three parts. The results of testing the system based on the RAVDESS data set and a comparison of the results obtained with the most popular algorithms are presented.

Keywords: voice, emotion recognition, convolutional neural networks, RAVDESS, MFCC, Python.

Введение

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, что неудивительно, ведь, хотя эмоции универсальны, их понимание и интерпретация являются специфическими и частично культурными [1].

Согласно последним исследованиям высокую эффективность в различных задачах распознавания имеют сверточные нейронные сети [4]. Рассмотрим разработанную систему распознавания эмоций по голосу, основанную на нейронной сети такой архитектуры.

Предобработка звука

Первый шаг в распознавании эмоций очевиден - 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 (рис. 1).

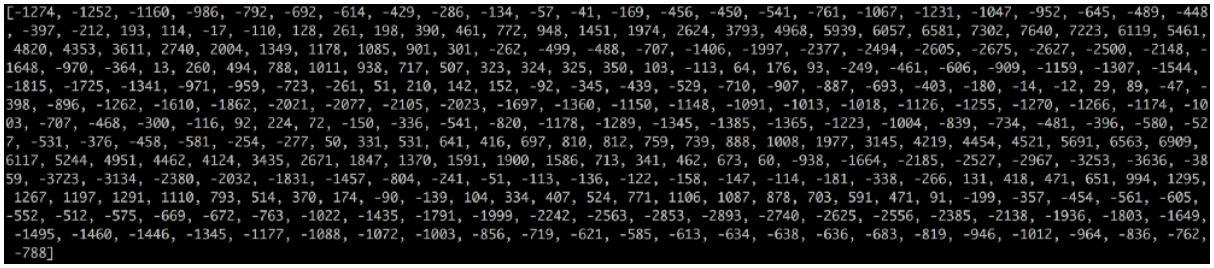

Рис. 1. Дискретизированная звуковая волна.

Можно было бы обучить нейросеть на этих данных, но распознавание речевых моделей путем обработки такого большого количества входных данных затруднительно. Можно облегчить задачу, проведя некоторую предварительную обработку аудио-данных.

В качестве характерных особенностей исходного сигнала используются MFCC (Mel-Frequency Cepstral Coefficients, мел-частотные кепстральные коэффициенты) [3]. 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 (Рис. 2), в данной работе мел-частотное пространство разбивалось на 12 отрезков.

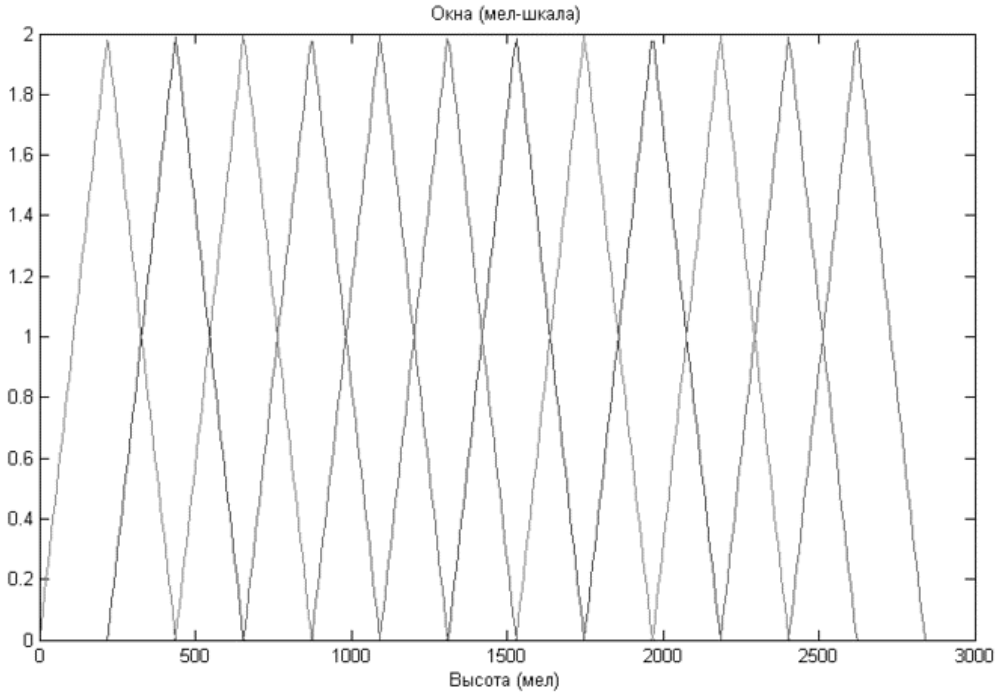

Рис. 2. Оконная треугольная функция.

Далее необходимо вычислить логарифм мощности сигнала в каждом мел-окне и произвести дискретное косинусное преобразование (рис. 3).

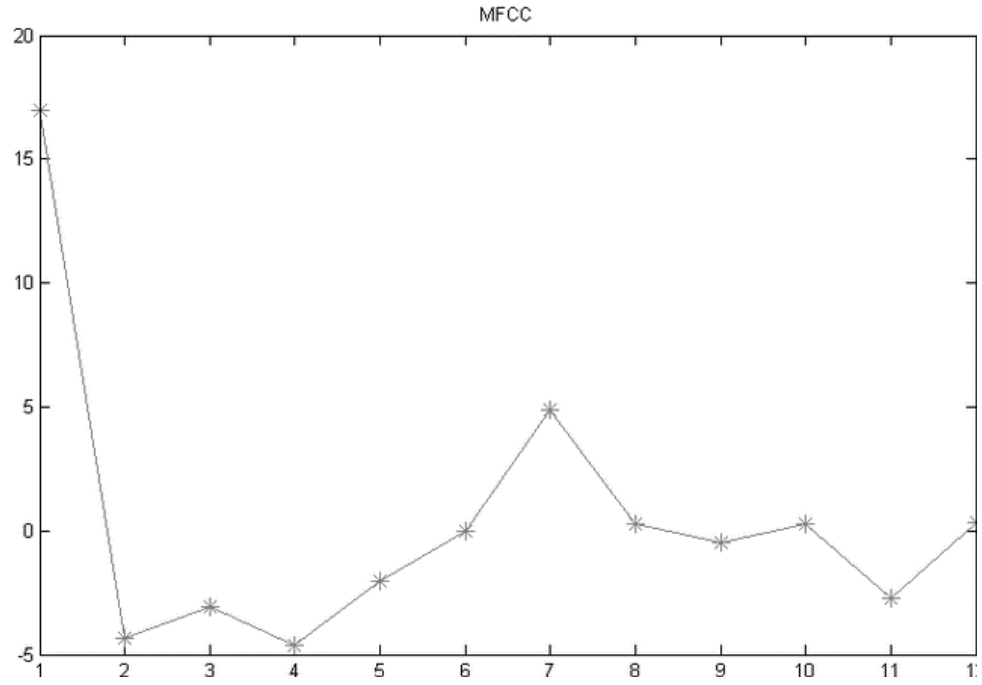

Рис. 3. Линейная функция MFCC.

Полученные мел-частотные кепстральные коэффициенты, можно в дальнейшем использовать в качестве уникальной характеристики входной звуковой волны.

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.

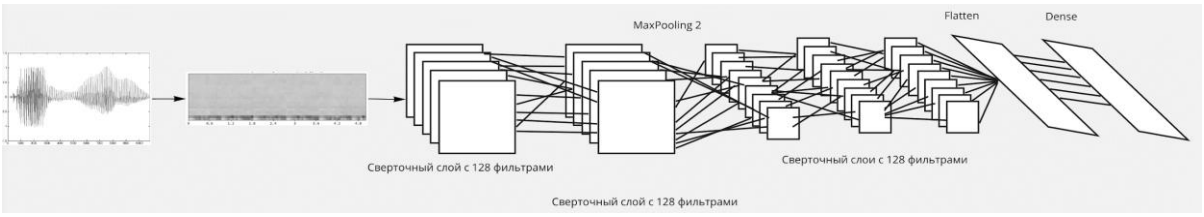

Рис. 4. Алгоритм определения эмоций при помощи сверточной нейронной сети.

Сверточная нейронная сеть была реализована на языке Python [5] 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).

Построение нейронной сети в соответствии с выбранной архитектурой

```
model = Sequential()
model.add(Conv1D(128, 5,padding='same', input_shape=(216,1)))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same'))
model.add(Activation('relu'))
```

```
model.add(Dropout(0.1))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Dropout(0.2))
model.add(Conv1D(128, 5,padding='same',))
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(10))
model.add(Activation('softmax'))
opt = keras.optimizers.rmsprop(lr=0.00001, decay=1e-6)
```

Эксперименты и результаты

В качестве входных данных использовался набор данных от RAVDESS [2], состоящий из аудиозаписей 24 актеров, которые содержат различные эмоции: счастье, злость, грусть, отвращение, спокойствие, удивление. Данный набор данных был разделен на тренировочные и тестовые данные в пропорции 80/20.

Тренировка нейронной сети и подсчет точности на тестовых файлах:

```
model.summary()
model.compile(loss='categorical_crossentropy',
optimizer=opt,metrics=['accuracy'])
cnnhistory=model.fit(x_traincnn, y_train, batch_size=32, epochs=1009,
validation_data=(x_testcnn, y_test))
preds = model.predict(x_testcnn, batch_size=32,verbose=1)
```

Сравнительная точность примененного алгоритма с его аналогами приведена в таблице 1.

Таблица 1. Сравнение с другими алгоритмами

Алгоритм	Достигнутая точность
Linear Discriminant Analysis (LDA)	62
Regularized discriminant Analysis (RDA)	80
Support vector machines (SVM)	75
k nearest neighbor (KNN)	72
Convolutional Neural Networks	**73**

Выводы

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.

Библиографический список:

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. Zenodo [Электронный ресурс]: The Ryerson Audio-Visual Database of Emotional Speech and Song. – 2018. – URL: https://zenodo.org/record/1188976/?f=3#.XCSVOS1eOu4 (дата обращения 29.06.2019).

3. Белов Ю.С., Гришунов С.С. Основные математические методы выделения речевых особенностей в системах распознавания диктора // Электронный журнал: наука, техника и образование. 2015. №3 (3). С. 57-62. URL: http://nto-journal.ru/uploads/articles/a747c0c5bc3f2ec62464e08da44b2878.pdf (дата обращения 30.06.2019).

4. Галушкин, А.И. Нейронные сети: основы теории [Электронный ресурс] / А.И. Галушкин. — Электрон. дан. — Москва : Горячая линия-Телеком, 2017. — 496 с. — Режим доступа: https://e.lanbook.com/book/111043. — Загл. с экрана.

5. Козеева О.О., Чухраев И.В. РАЗРАБОТКА СРЕДСТВАМИ ЯЗЫКА PYTHON МОДУЛЯ АУТЕНТИФИКАЦИИ АВТОМАТИЗИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ // Электронный журнал: наука, техника и образование. 2017. №2 (12). С. 173-180. URL: http://nto-journal.ru/uploads/articles/4f5bef8fbb37cac19978bf31ff8a038e.pdf (дата обращения 30.06.2019).