# ИСПОЛЬЗОВАНИЕ ГЛУБОКОГО ОБУЧЕНИЯ НЕЙРОСЕТИ ДЛЯ РАСПОЗНАВАНИЯ ГОЛОСОВЫХ КОМАНД ПОЛЬЗОВАТЕЛЯ

**А. Г. Романюк[1] , А. Н. Смирнов[1], В. М. Антонова[1,2]**

**[1] Московский государственный технический университет им. Н.Э. Баумана, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1**
**[2] Институт радиотехники и электроники им. В.А. Котельникова РАН, 125009, Москва, ул. Моховая, 11-7**

Статья поступила в редакцию 19 ноября 2019 г.

**Аннотация.** Данная работа посвящена использованию сверточной нейронной сети для распознавания речи. Исследован способ обучения нейросети, произведенный на архиве из 7100 звуковых дорожек с проиндексированными метками, речевые сигналы в которых были преобразованы в log-mel спектрограммы. Обучение нейронной сети происходило на входящем сигнале, имеющем плавное распределение и нормализацию. В статье описана способность созданной сети распознавать разные произнесенные слова и определять, является ли входящий сигнал тишиной или фоновым шумом, что было достигнуто путем проработки 4000 образцов клипов шума. Рассматривается способность сети одновременно классифицировать несколько преобразованных входящих сигналов, независимо от точного положения речи во времени. Описан процесс создания виртуального устройства, способного считывать сигнал с микрофона с определенной частотой дискретизацией звука. В настоящей работе была получена нейросеть, которая может быть усовершенствована для понимания большего числа голосовых команд и использована в нескольких сферах жизнедеятельности человека.

**Ключевые слова:** нейронные сети, глубокое обучение, распознавание речи.

**Abstract.** This work is devoted to the use and development of speech recognition of neural networks. The process of neural network learning has been explored with the archive containing 7100 tracks with indexed tags. Speech signals in those tracks were converted into log-mel spectrograms. Neural network training has occurred onto an entering signal which possessed smooth distribution and normalization. The article describes the ability of the created network to recognize different spoken words and

to determine whether the incoming signal is silence or a background noise which was achieved by working out 4000 samples of noise clips. The ability of the network to classify several converted incoming signals simultaneously regardless of the exact position of speech in time is investigated. The process of creating a virtual device that capable of reading the signal from a microphone with a certain sampling frequency of sound is described. The neural network has been obtained in this very project. It may be perfected for the comprehension of a bigger number of voice commands and use in various human activity spheres.

**Keywords:** neural networks, deep learning, speech recognition.

## Введение

В современном мире человек старается автоматизировать все процессы для облегчения собственной жизни, что способствует развитию технологий. Один из инструментов для оптимизации - искусственные нейронные сети. Они появились совсем недавно, но уже применяются для решения множества задач: систем распознавания и классификации объектов на изображениях, создания контекстной рекламы, распознавания рукописных текстов и т.д. Данная статья посвящена разработке простой модели глубокого обучения, которая обнаруживает речевые команды в аудио, с разбором принципов обработки информации. Результатом работы является сверточная нейронная сеть.

## 1. Теоретические сведения

Сверточная нейронная сеть основана на математической операции свертки - упорядоченной процедуры смешивания двух источников информации. Например, есть два мешка с зерном, зерна - это информация, зерна ссыпаются в один большой контейнер и затем перемешиваются определённым способом. Каждый мешок с информацией имеет своё собственное правило, которое описывает, как зерно из одного мешка смешивается с другим.

Для наглядного объяснения представим матрицу входных данных размером 5x5 (карту признаков), сканирующее ядро этой матрицы будет

представлять собой матрицу 3x3, которая получается поэлементным умножением части исходной матрицы и суммированием всех полученных значений в один выходной пиксель (см. Рисунок 1).

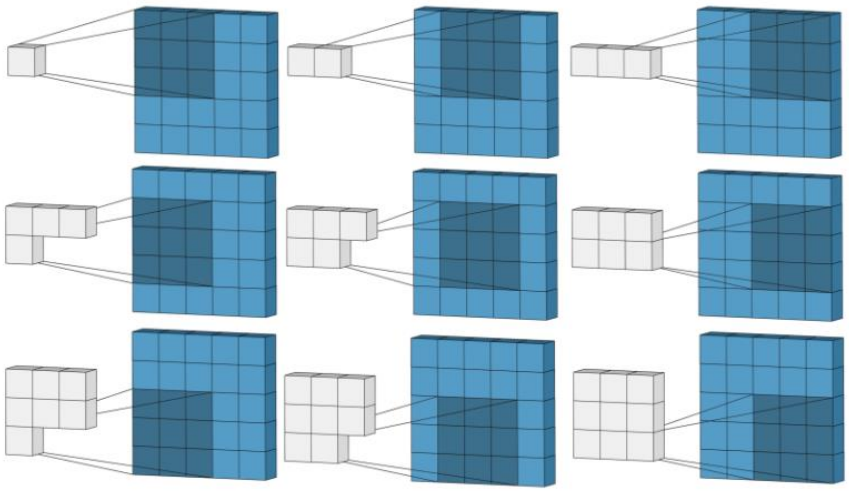

Рис. 1. Принцип свертки.

Фильтр представляет собой коллекцию ядер, причем для каждого отдельного входного канала этого слоя есть одно ядро, и каждое ядро уникально.

Результат для любого количества фильтров идентичен: каждый фильтр обрабатывает вход со своим отличающимся от других набором ядер и скалярным смещением по описанному выше процессу, создавая один выходной канал (см. Рисунок 2).

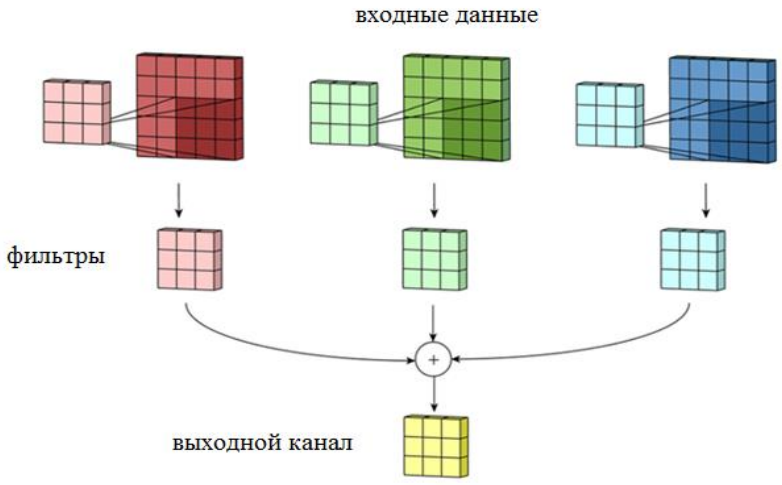


Рис. 2. Принцип создания выходного канала.

## 2. Работа с выходными данными

### *2.1 Подготовка*

На основе базы данных речевых команд из Google [1] было создано хранилище, которое содержит около 71000 звуковых дорожек с проиндексированными метками, распределенные по папкам. Вначале мы не будем включать в тренировочный набор длинные файлы с фоновым шумом, а добавим на отдельном этапе. После указания ограниченного множества слов, которые модель должна распознавать как команды и подмножество неизвестных слов, нейросеть проработала аудиофайлы:

- "yes" в количестве 2377 штук;
- "no" - 2375;
- "up"- 2375;
- "down" - 2359;
- "left"- 2353;
- "right" - 2367;
- "on"- 2367;
- "off"- 2357;
- "stop"- 2380;
- "go"- 2372

и 8193 неизвестных ("unknown") шумов

### *2.2 Вычисление речевых спектрограмм и визуализация данных*

Чтобы подготовить данные для эффективного обучения сверточной нейронной сети, речевые сигналы были преобразованы в log-mel спектрограммы с параметрами расчета: продолжительность каждого речевого клипа (в секундах), продолжительность каждого кадра, шаг по времени между каждым столбцом, количество log-mel-фильтров (высота каждой спектрограммы).

Проводя проверку на файлах из нашего хранилища, мы убеждаемся в том, что все действительно работает (см. Рисунок 3).

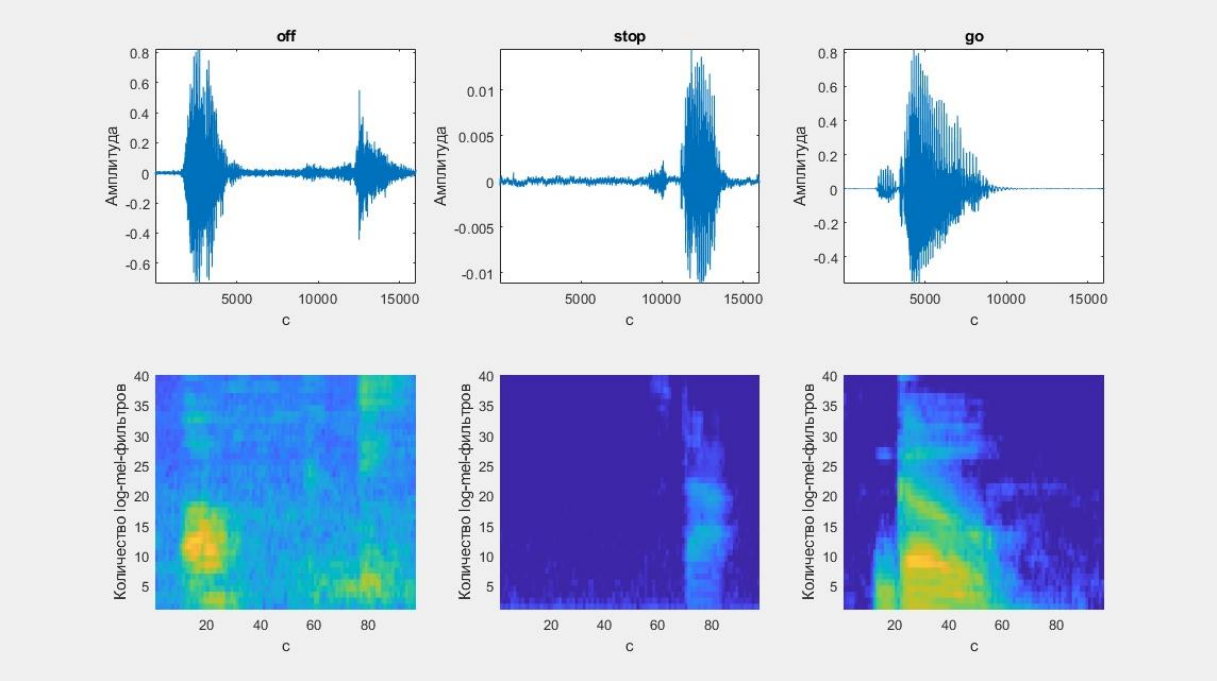


Рис. 3. Формы волны и спектрограммы нескольких обучающих примеров.

Обучение нейронных сетей происходит проще всего, когда входы в сеть имеют плавное распределение и нормализованы. Чтобы проверить, что распределение данных является плавным, представим обучающие данные в виде гистограммы значений пикселей (см. Рисунок 4).

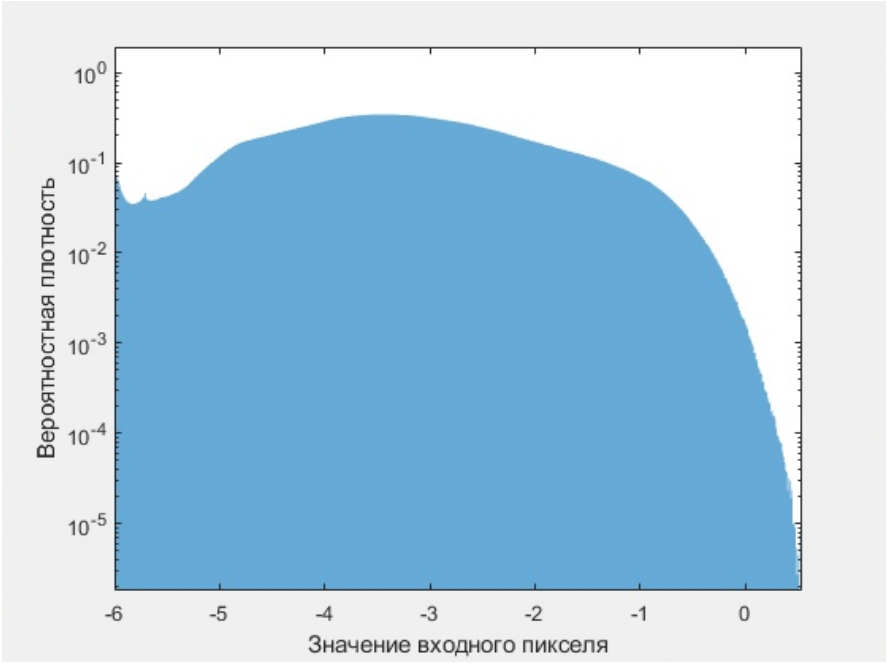


Рис. 4. Гистограмма значений пикселей обучающих данных.

### *2.3 Фоновый шум*

Созданная сеть способна не только распознавать разные произнесенные слова, но и определять, содержит ли вход тишину или фоновый шум. Создав из аудиозаписей образцы 4000 клипов фонового шума, длительностью в одну секунду и отмасштабировав каждый на число от $10^{-4}$ до 1, получим спектрограммы от тишины до громкого шума.

Разделим спектрограммы фонового шума между обучающими, проверочными и тестовыми наборами, фоновые выборки в разных наборах данных сильно коррелированы (см. Рисунок 5).

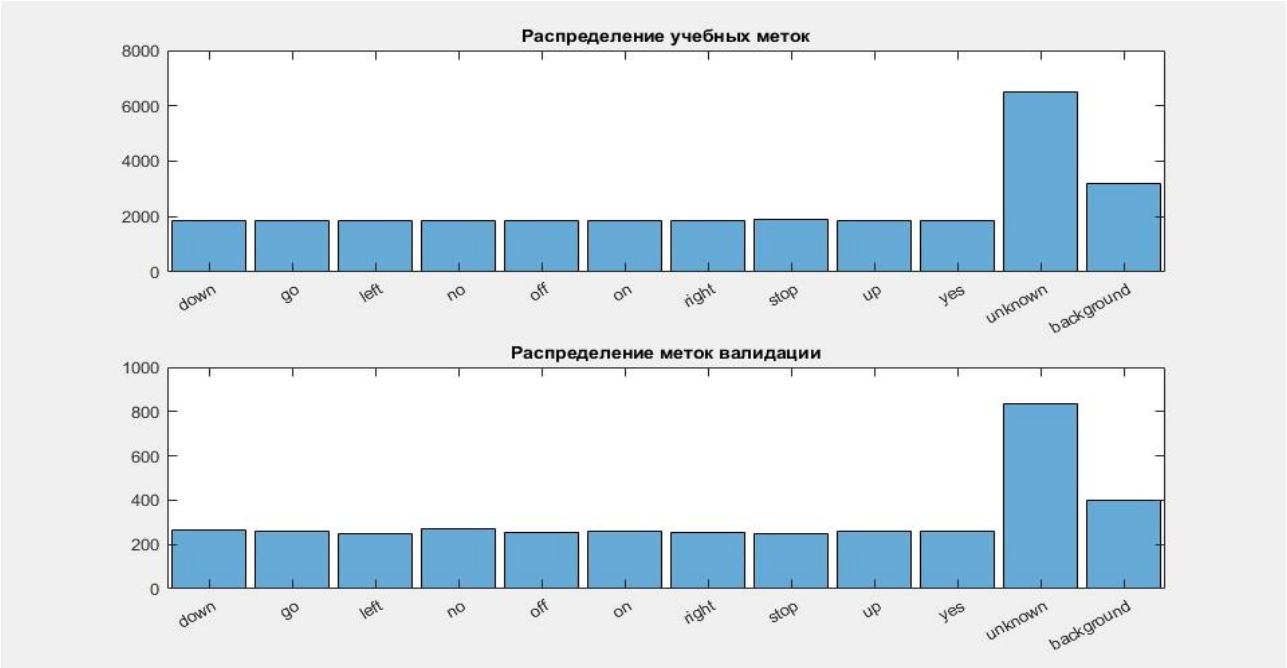


Рис. 5. Распределение меток.

Также на данном этапе был увеличен эффективный размер обучающих данных для предотвращения перегрузки сети.

## 3. РАБОТА С НЕЙРОСЕТЬЮ

### *3.1 Определение архитектуры нейронной сети*

В ходе работы была создана простая сетевая архитектура в виде массива слоев нейросети. Использованы сверточные и пакетные слои нормализации и уменьшена выборка карт объектов (по времени и частоте), использованы максимальные уровни объединения. Окончательный максимальный слой пула со временем объединяет карту входных объектов, что обеспечивает

инвариантность преобразования времени во входных спектрограммах, позволяя сети выполнять одинаковую классификацию независимо от точного положения речи во времени.

Размерность сети была взята небольшая: в ней всего пять сверточных слоев с несколькими фильтрами и пользовательским слоем, который вычисляет перекрестную энтропийную потерю с взвешенными наблюдениями.

### *3.2 Обучение и тренировка нейросети*

В качестве варианта обучения был использован оптимизатор Adam (adaptive moment estimation - оптимизационный алгоритм) с размером мини-пакета 128. Adam - это метод адаптивной скорости обучения, то есть он рассчитывает индивидуальные скорости обучения для различных параметров. Его название происходит от адаптивной оценки моментов, и причина, по которой его так называют, заключается в том, что Адам использует оценки первого и второго моментов градиента, чтобы адаптировать скорость обучения для каждого веса нейронной сети. Тренировочный цикл длится в течение 25 полных проходов через весь набор тренировок и скорость обучения уменьшится в 10 раз после 20 прохода (см. Рисунок 6-7).

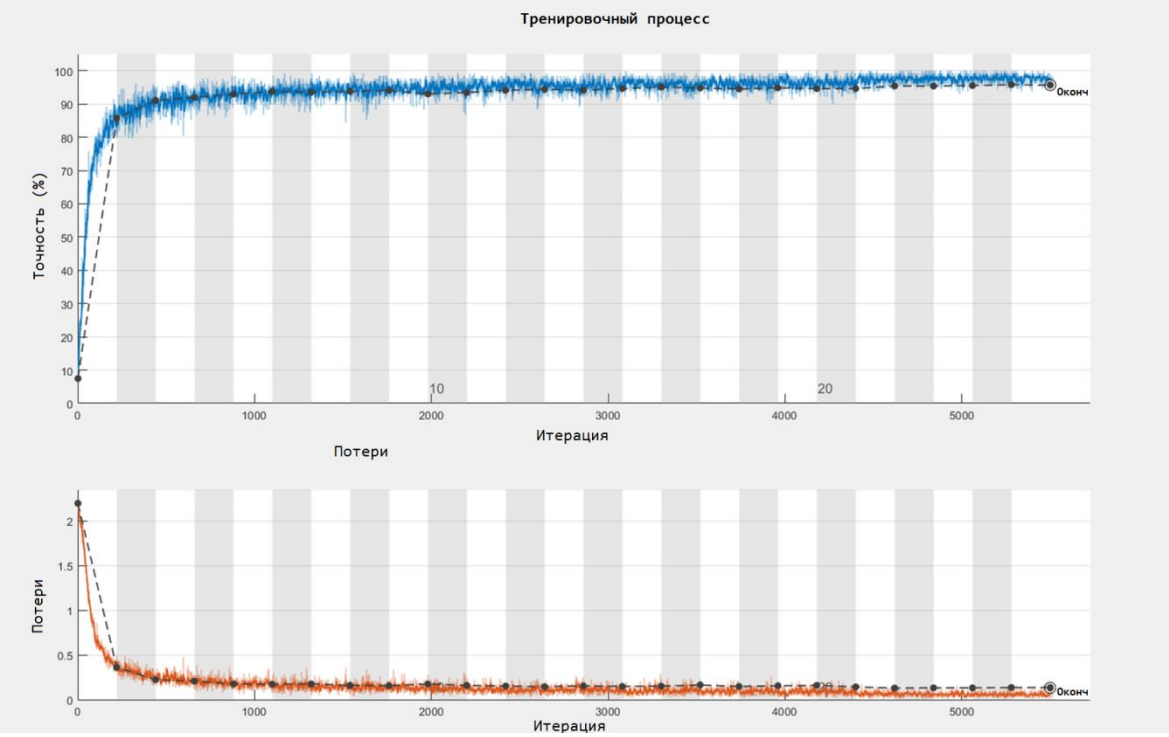


Рис. 6. Процесс обучения.


Рис. 7. Обозначения и параметры обучения.

После обучения слои нейросети выглядят следующим образом:

21x1 Массив слоев со слоями:

| | | | |
|---|---|---|---|
| 1 | 'imageinput' | Image Input | 40x98x1 изображения с нормализацией |
| 2 | 'conv_1' | Convolution | 12 3x3x1 свертки с шагом [1 1] и дополнением "то же самое" |
| 3 | 'batchnorm_1' | Batch Normalization | Пакетная нормализация с 12 каналами |
| 4 | 'relu_1' | ReLU | ReLU |
| 5 | 'maxpool_1' | Max Pooling | 2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое' |
| 6 | 'conv_2' | Convolution | 24 3x3x12 свертки с шагом [1 1] и дополнением "то же самое" |
| 7 | 'batchnorm_2' | Batch Normalization | Пакетная нормализация с 24 каналами |
| 8 | 'relu_2' | ReLU | ReLU |
| 9 | 'maxpool_2' | Max Pooling | 2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое' |
| 10 | 'conv_3' | Convolution | 48 3x3x24 свертки с шагом [1 1] и дополнением "то же самое" |
| 11 | 'batchnorm_3' | Batch Normalization | Пакетная нормализация с 48 каналами |
| 12 | 'relu_3' | ReLU | ReLU |
| 13 | 'maxpool_3' | Max Pooling | 2x2 максимальное объединение с шагом [2 2] и дополнением 'то же самое' |
| 14 | 'conv_4' | Convolution | 48 3x3x48 свертки с шагом [1 1] и дополнением "то же самое" |
| 15 | 'batchnorm_4' | Batch Normalization | Пакетная нормализация с 48 каналами |
| 16 | 'relu_4' | ReLU | ReLU |
| 17 | 'maxpool_4' | Max Pooling | 1x13 максимальное объединение с шагом [1 1] и заполнением [0 0 0 0] |
| 18 | 'dropout' | Dropout | исключение 20% |
| 19 | 'fc' | Fully Connected | 12 полностью связанных слоев |
| 20 | 'softmax' | Softmax | softmax |
| 21 | 'classoutput' | Classification Output | Взвешенная перекрестная энтропия |

Рис. 8. Описание слоев нейросети.

### *3.3 Оценка обученной сети*

Ошибка сети на тренировочном наборе (без дополнения данных) составляет 1.3806%и на проверочном наборе 4.2999%. Таким образом, точность нейросети составляет 95.70%. Тем не менее, данные обучения, проверки и тестирования имеют схожие распределения, которые не обязательно отражают реальную среду.

Благодаря матрице путаницы мы можем сделать вывод, что большая путаница между неизвестными словами и командами:"up"и "down", "down"и "no", "go" и "no" (см. Рисунок 9).


Рис. 9. Матрица путаницы для данных проверки.

Общий размер сети составляет 255.709 Кб, а скорость ее прогнозирования при использовании процессора (время для классификации одного входного изображения) 3.944 мс. Сеть имеет возможность классификации нескольких изображений одновременно, что приводит к сокращению времени прогнозирования для каждого изображения.

### *3.4 Работа с микрофоном*

Для тестирования работы сети была взята частота дискретизации звука с нашего микрофона. Затем создано виртуальное устройство, способное считывать звук с нашего микрофона. Инициализирован буфер для аудиосигнала с параметрами для расчетов потоковой спектрограммы,

При проведенных тестах речевые команды были распознаны верно в большинстве случаев (см. Рисунки 10-12).

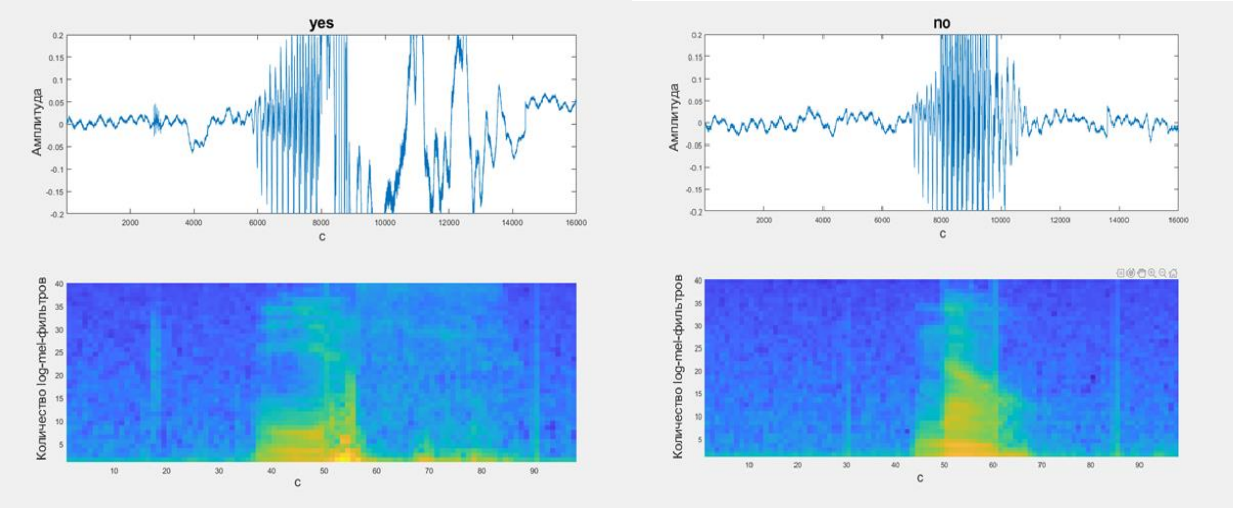


Рис. 10. Распознавание нейросетью произнесенной команды "yes" и команды "no".

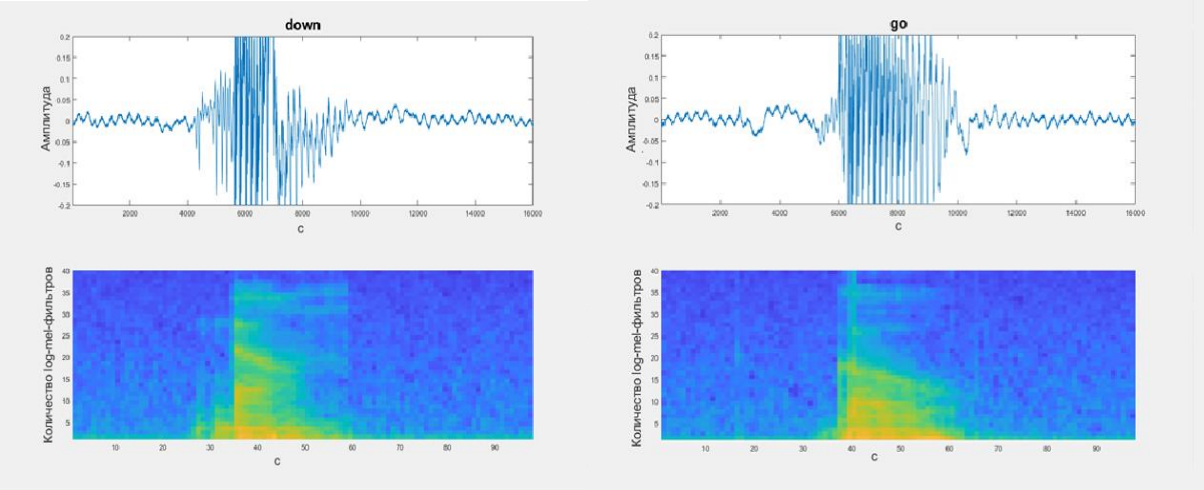


Рис. 11. Распознавание нейросетью произнесенной команды "down" и команды "go".

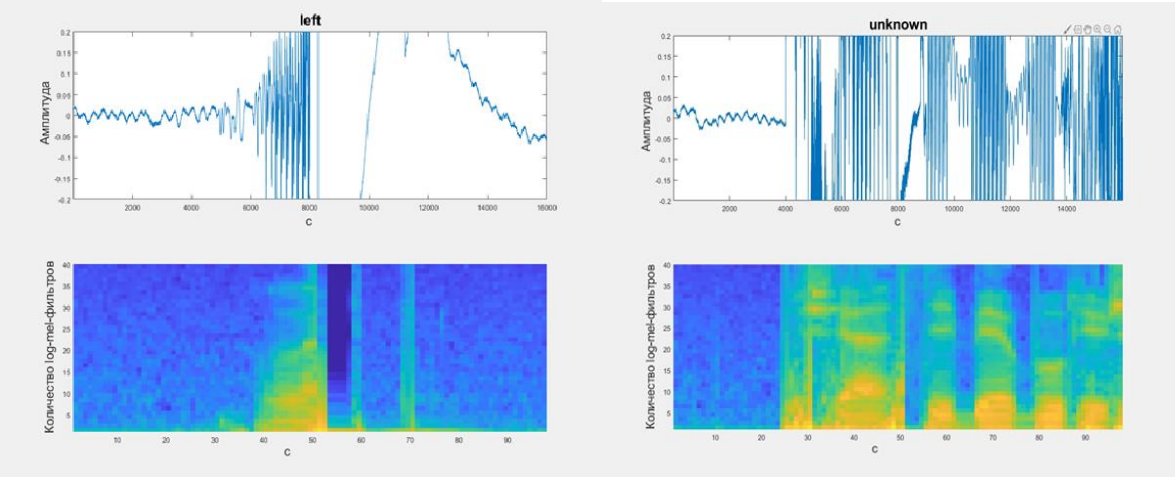


Рис. 12. Распознавание нейросетью произнесенной команды "left" и неизвестного шума.

## Вывод

В результате данной работы была разработана нейросеть, различающая произносимые слова. Оценки, доставляемые алгоритмом, состоятельны при почти произвольных помехах. Также сохраняется работоспособность алгоритма при росте размерности вектора оцениваемых параметров и увеличении количества классов (распознаваемых слов).

Точность распознавания построенной системы достигает 95.7%. Система может быть усовершенствована для случая распознавания большего количества слов. Точность распознавания может быть улучшена за счет увеличения словарного запаса, подразумевающего предоставления большего количества аудиофайлов для тренировки.

Данные результаты могут быть применены при автоматизации первой линии звонков колл-центров в качестве ответа клиента на вопрос.

## Литература

1. База данных <speech_commands> [онлайн ресурс] URL: https://storage.googleapis.com/download.tensorflow.org/data/speech_commands_v0.01.tar.gz (дата обращения 17.11.2019)

2. Справочник по Matlab [онлайн ресурс]. URL: http://radiomaster.ru/cad/matlab/index.php (дата обращения 16.11.2019)

3. Сверточная нейронная сеть, часть 1: структура, топология, функции активации и обучающее множество [онлайн ресурс]. URL: https://habr.com/ru/post/348000/ (дата обращения 16.11.2019)

4. Потемкин В.Г., Медведев В.С. Нейронные сети. MATLAB 6. Москва, «Диалог-МИФИ». 2002. 496 с.

5. Мел-кепстральные коэффициенты (MFCC) и распознавание речи [онлайн ресурс] URL: https://habr.com/ru/post/140828/ (дата обращения 16.11.2019)

**Для цитирования:**
Романюк А.Г, Смирнов А.Н., Антонова В.М. Использование глубокого обучения нейросети для распознавания голосовых команд пользователя. Журнал радиоэлектроники [электронный журнал]. 2019. № 11. Режим доступа: http://jre.cplire.ru/jre/nov19/18/text.pdf. DOI 10.30898/1684-1719.2019.11.18