Моделирование и анализ данных
2022. Том 12. № 3. С. 49–57
DOI: https://doi.org/10.17759/mda.2022120304
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 3, pp. 49–57
DOI: https://doi.org/10.17759/mda.2022120304
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.934

О методе распознавания голосовых команд с применением особого преобразования спектральных плотностей

Левонович Н.И.*
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-8580-0490
e-mail: levonikitatech@yandex.ru

В статье рассмотрен новый подход к решению частной задачи распознавания речи. Задача ставиться как задача распознавания голосовых команд заданной длинны в словах, в которой на каждой позиции может стоять определенный набор слов. Наборы слов не пересекаются по позициям. Для решения данной задачи разработаны модели, совмещающие спектральный анализ, свёрточные нейронные сети и наивный байесовский классификатор. Произведен сравнительный анализ нескольких разработанных моделей, выбрана лучшая, разработан графический интерфейс для взаимодействия с моделью. Полученный результат может использоваться в качестве базового при создании методов распознавания речи для голосовых интерфейсов.

Ключевые слова: автоматическое распознавание речи, спектральный анализ, свёрточные нейронные сети.

Для цитаты:
Левонович Н.И. О методе распознавания голосовых команд с применением особого преобразования спектральных плотностей // Моделирование и анализ данных. 2022. Том 12. № 3. С. 49–57. DOI: https://doi.org/10.17759/mda.2022120304

1. ВЕДЕНИЕ

Одной из наиболее актуальных проблем в области человеко-машинных интерфейсов в настоящий момент является создание голосовых интерфейсов. Это направление включает в себя исследования в области распознавания речи, синтеза речи, обработки естественного языка и интеллектуальной интерпретации речи.

В настоящее время существует несколько подходов к распознаванию речи для разных модулей системы распознавания речи. Эти модули – акустическая модель,

***Левонович Никита Ильич**, студент, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-8580-0490, e-mail: levonikitatech@yandex.ru

языковая модель и декодер. Современные средства распознавания речи комбинируют различные методы: алгоритм динамической трансформации временной шкалы, методы дискриминантного анализа, основанные на байесовской дискриминации, скрытые марковские модели, нейронные сети.

Акустическая модель – это функция, принимающая на вход признаки на небольшом участке акустического сигнала (фрейме) и выдающая распределение вероятностей различных фонем на этом фрейме. Самой популярной моделью акустического моделирования являются скрытые марковские модели, однако в свежих работах [1] встречаются модели, использующие рекуррентные нейронные сети, в частности LSTM-сети (сети долгой краткосрочной памяти), и нейросетевую темпоральную классификацию (CTC).

Языковые модели позволяют учитывать контекст и выяснять, какие последовательности слов и фонем являются наиболее вероятными с точки зрения текущего контекста. Современные средства языкового моделирования так же используют рекуррентные нейронные сети.

Декодер на базе вероятностей, предоставленных акустической и языковой моделью, выбирает конкретную речевую единицу.

2. НОВЫЙ ПОДХОД К РАСПОЗНАВАНИЮ РЕЧИ

В данной статье рассмотрен новый подход к решению частной задачи распознавания речи, а именно к распознаванию голосовых команд заданной длинны в словах, в которой на каждой позиции может стоять определенный набор слов и эти наборы не пересекаются по позициям.

Схема команд, представленных в данном наборе (граф переходов между словами), представлена на рисунке 1.

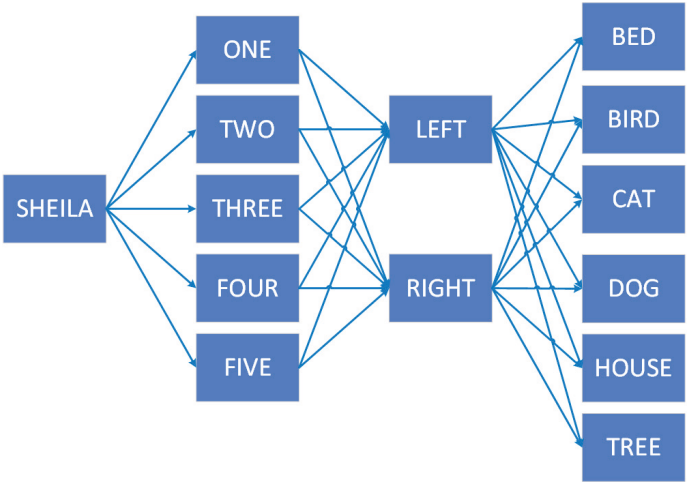

Рис. 1. Граф построения голосовой команды.

С целью создания модели для распознавания был собран набор данных – записей различных голосовых команд, сделанных разными дикторами. Записи набора данных были сделаны шестьюдесятью дикторами. Всего исходный набор данных содержит 1129 голосовых команд. Максимальная продолжительность звукового сигнала команды 4 секунды.

На данном наборе данных можно поставить 2 задачи многоклассовой классификации: классификация произнесенных команд – 60-ти классовая классификация, поиск и классификация произнесенных слов – 14-ти классовая классификация.

В рамках данной статьи будет рассмотрено решение задачи 14-ти классовой классификации произнесенных слов, так как на данном наборе данных она представляется более разрешимой. Базовым методом для решения данной задачи был выбран метод анализа спектральных плотностей [2] с помощью свёрточных нейронных сетей [3].

Разделение сигналов на слова производилось с помощью оконного подсчета энергии сигнала и выборке областей, в которых энергия больше энергии шума.

С целью улучшения разделения слов и дальнейшей дискриминации был предложен метод предобработки спектрограмм. Сущность метода – логарифмирование с предварительным прибавлением перцентиля спектрограммы и минимального значения сигнала (шага амплитуды) $\in$. Одна из реализаций метода выражается формулой 1 (используется 85-тый перцентиль). К полученной модифицированной логарифмированной спектрограмме было применено масштабирование к диапазону [0; 255] и взятие целой части (формула 2).

$$S_{xx_{modified}} = \log(S_{xx} + p_{85} + \epsilon) \quad (1)$$

$$S_{xx_{scaled}} = \left\lfloor \frac{S_{xx_{modified}} - \min\left(S_{xx_{modified}}\right)}{\max\left(S_{xx_{modified}}\right) - \min\left(S_{xx_{modified}}\right)} * 255 \right\rfloor \quad (2)$$

Расчет исходных спектрограмм осуществлялся до разделения на слова, на частотах 8–3064 Гц с шагом 8 Гц. По времени спектрограмма команды была посчитана в 285 точках, что соответствует приблизительно 14-ти миллисекундному шагу.

Модифицированные спектрограммы были разбиты на слова раннее описанным методом. В качестве энергии в этом случае выступала сумма амплитуд. Максимальная длинна слова в выборке при разбиении 105 отсчетов спектрограммы (что примерно соответствует 1,47 секундам). Все полученные спектрограммы слов были преобразованы к размеру 382x105, путем добавления нулевых столбцов в матрицу справа от матрицы спектрограммы.

На подготовленных таким образом данных была обучена свёрточная модель архитектуры, изображенной на рисунках 2–3.

Данная модель была обучена оптимизатором Adam, с шагом обучения 0.002. Перед началом обучения обе выборки были разбиты на обучающую и контрольную подвыборки, в отношении 75 % и 25 %. В подвыборках сохранялось исходное соотношение классов. По результатам обучения данной модели на тестовой выборке было получено 97.09 % (Рисунок 4).

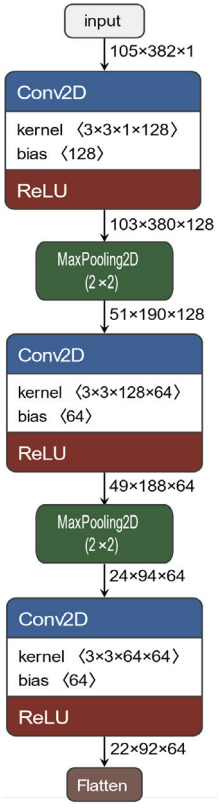

Рис. 2. Архитектура свёрточной модели (свёрточная часть).

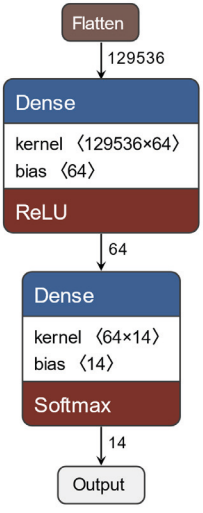

Рис.3. Архитектура свёрточной модели (полносвязная часть).

При использовании данной модели для распознавания команд 873 из 1129 команд были распознаны верно, что соответствует ~77 %.

С целью улучшения качества распознавания была предпринята попытка обучить отдельные полносвязные модели для каждой позиции. Все отдельные модели базировались на общих свёрточных слоях, изображенных на рисунке 2. Архитектура отдельных моделей изображена на рисунке 5 (модели идут слева на право, также как в графе переходов между словами). Модели позволяют различать соответствующие позиции слова между собой и отличать их от других слов.

Данные модели были обучены оптимизатором Adam, с шагом обучения 0.002. Перед началом обучения обе выборки были разбиты на обучающую и контрольную подвыборки, в отношении 75 % и 25 %.

Максимальная достигаемая с применением данного метода доля верных ответов на тестовой выборке – 98 % слов. Однако при распознавании команд данный метод показал себя хуже. 841 из 1129 команд были распознаны верно, что соответствует ~74 %.

Рис. 4. Доля верных ответов на контрольной выборке.

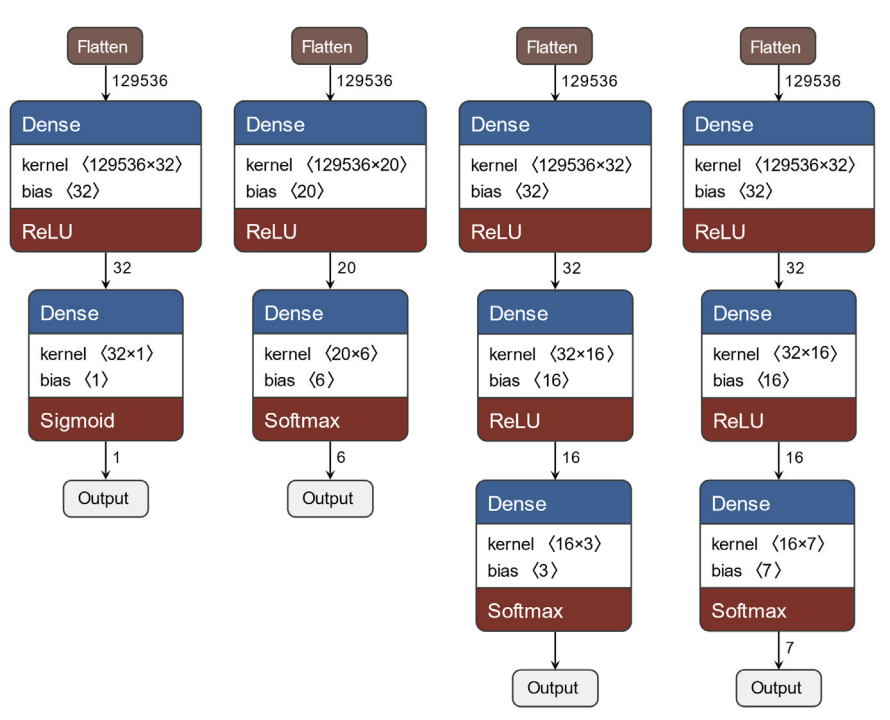

Рис. 5. Архитектуры отдельных полносвязных моделей.

Еще одним подходом, который был применен при решении данной задачи – наивный байесовский классификатор на выходах нейросети, изображенной на рисунках 2 и 3. При обычной интерпретации выходов нейросети берется аргумент максимизации выходного слоя (формула 3).

$$output = \arg\max f\left(x\right) \tag{3}$$

При использовании наивного байесовского классификатора [4] аргумент максимизации берется от произведения входа нейросети на априорную вероятность класса P_{output}, которая задается графом переходов между словами (формула 4).

$$output = \arg\max P_{output} f\left(x\right) \tag{4}$$

Конкретные априорные вероятности слов для каждой позиции приведены в таблице 1.

Таблица 1

Априорные вероятности слов на позиции

	1	**2**	**3**	**4**		**1**	**2**	**3**	**4**
bed	0	0	0	$\frac{1}{6}$	**left**	0	0	$\frac{1}{2}$	0
bird	0	0	0	$\frac{1}{6}$	**one**	0	$\frac{1}{5}$	0	0
cat	0	0	0	$\frac{1}{6}$	**right**	0	0	$\frac{1}{2}$	0
dog	0	0	0	$\frac{1}{6}$	**sheila**	1	0	0	0
five	0	$\frac{1}{5}$	0	0	**three**	0	$\frac{1}{5}$	0	0
four	0	$\frac{1}{5}$	0	0	**tree**	0	0	0	$\frac{1}{6}$
house	0	0	0	$\frac{1}{6}$	**two**	0	$\frac{1}{5}$	0	0

При использовании данного метода верно были классифицированы 977 из 1129 команд, точность классификации составила 87 %.

Для демонстрации, обученной по предложенному методу, нейросети было разработано настольное программное обеспечение на языке Python с использованием графического интерфейса Kivy [5]. Данное приложение использует модель с наивным байесовским классификатором. Разработанное программное обеспечение представляет собой, однооконное приложение, которое позволяет выбрать файл и распознать его. Пример работы приложения изображен на рисунке 6.

Приложение считывает аудиофайл с диска, строит график и спектрограмму исходного сигнала (левая верхняя четверть), фильтрует шумы и строит график и спектрограмму отфильтрованного сигнала (правая верхняя четверть). После чего сигнал разбивается на слова, спектрограммы которых изображаются в левой нижней

четверти. Ответы модели в виде распознанной фразы и исходных (до применения байесовского аргумента максимизации) максимальных вероятностей выводятся в правой нижней четверти.

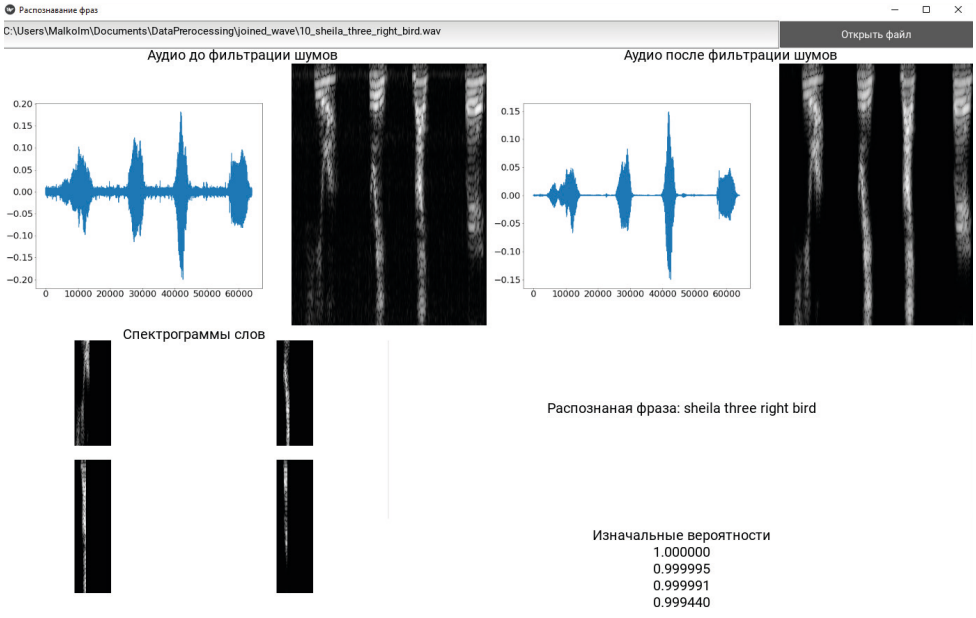

Рис. 6. Пример работы приложения.

3. ЗАКЛЮЧЕНИЕ

В данной работе предложен новый метод обработки спектрограмм, который существенно повышает сходимость обучения свёрточных нейронных сетей для решения задачи классификации произнесенных слов. С помощью него разработан метод распознавания голосовых команд, который на представленной выборке дает точность классификации 87 %. Данный результат является достаточно высоким, однако не достаточным для практического применения, его необходимо усовершенствовать. Одним из путей усовершенствования может быть распознавание фразы целиком, однако такая модель будет более тяжеловесной и, возможно, потребует более объемной выборки.

Литература

1. *Sak, Haşim,* et al. “Fast and accurate recurrent neural network acoustic models for speech recognition.” arXiv preprint arXiv:1507.06947 (2015).
2. *Л.С. Куравский, С.Н. Баранов.* Компьютерное моделирование и анализ данных. Конспекты лекций и упражнения: Учеб. пособие. – М.: РУСАВИА, 2012. – 218 с.:
3. *Dumoulin, Vincent, and Francesco Visin.* “A guide to convolution arithmetic for deep learning.” arXiv preprint arXiv:1603.07285 (2016).

4. *Ch, Read,* and ML MAP. “Bayesian learning.” book: Machine Learning. McGraw-Hill Science/Engineering/Math (1997): 154–200.
5. *Virbel, Mathieu, Thomas Hansen, and Oleksandr Lobunets.* “Kivy–a framework for rapid creation of innovative user interfaces.” Workshop-Proceedings der Tagung Mensch & Computer 2011. überMEDIEN| ÜBERmorgen. Universitätsverlag Chemnitz, 2011.

Voice Commands Recognition Method that Uses Special Spectral Density Transform

Nikita I. Levonovich*
Moscow State University of Psychology and Education (MSUPE), Moscow, Russia
ORCID: https://orcid.org/0000-0002-8580-0490
e-mail: levonikitatech@yandex.ru

This article discusses new approach to solving specific problem of speech recognition. The problem is formulated as problem of command recognition. Commands have equal lengths (in words). Each word position has its own set of word candidates. There were developed the models for solving this problem. The models consist of spectral density estimator, convolutional neural networks and naive Bayes classifier. The author performed a comparative analysis of the models and selected the best. There was developed a graphical user interface for interacting with the best model. The result can be used as a base for creation speech recognition methods for voice user interfaces.

Keywords: automatic speech recognition, spectral density estimation, convolutional neural networks.

For citation:
Levonovich N.I. Voice Commands Recognition Method that Uses Special Spectral Density Transform. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2022. Vol. 12, no. 3, pp. 49–57. DOI: https://doi.org/10.17759/mda.2022120304 (In Russ., abstr. in Engl.).

References

1. Sak, Haşim, et al. "Fast and accurate recurrent neural network acoustic models for speech recognition." arXiv preprint arXiv:1507.06947 (2015).
2. L.S. Kuravskii, S.N. Baranov Komp'yuternoe modelirovanie i analiz dannykh. Konspekty lektsii i uprazhneniya: Ucheb. posobie. – M.: RUSAVIA, 2012. – 218 p.
3. Dumoulin, Vincent, and Francesco Visin. "A guide to convolution arithmetic for deep learning." arXiv preprint arXiv:1603.07285 (2016).
4. Ch, Read, and ML MAP. "Bayesian learning." book: Machine Learning. McGraw-Hill Science/Engineering/Math (1997): 154–200.
5. Virbel, Mathieu, Thomas Hansen, and Oleksandr Lobunets. "Kivy–a framework for rapid creation of innovative user interfaces." Workshop-Proceedings der Tagung Mensch & Computer 2011. überMEDIEN| ÜBERmorgen. Universitätsverlag Chemnitz, 2011.

Nikita I. Levonovich, Student, Moscow State University of Psychology and Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0002-8580-0490, e-mail: levonikitatech@yandex.ru

Получена 21.08.2022
Принята в печать 12.08.2022

Received 21.08.2022
Accepted 12.08.2022