Моделирование и анализ данных
2023. Том 13. № 3. С. 39–51
DOI: https://doi.org/10.17759/mda.2023130303
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 3, pp. 39–51
DOI: https://doi.org/10.17759/mda.2023130303
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.8

Применение вероятностных сетей к проблеме распознавания речи

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

Козырев А.Д.**
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ)
Государственный научно-исследовательский институт авиационных систем (ФАУ «ГосНИИАС»), г. Москва, Российская Федерация
ORCID: https://orcid.org/0009-0008-1769-4121
e-mail: adkozyrev@2100.gosniias.ru

В статье рассмотрен подход к решению задачи распознавания речи, основанный на использовании вероятностных нейронных сетей. 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 разработан алгоритм распознавания, ядром которого является вероятностная сеть, распознающая модифицированные оценки спектральных плотностей. Представленный алгоритм позволяет получить высокую точность распознавания, достаточную для создания голосовых интерфейсов.

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

Для цитаты:
Левонович Н.И., Козырев А.Д. Применение вероятностных сетей к проблеме распознавания речи // Моделирование и анализ данных. 2023. Том 13. № 3. С. 39–51. DOI: https://doi.org/10.17759/mda.2023130303

***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

****Козырев Алексей Денисович,** аспирант, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), инженер, Государственный научно-исследовательский институт авиационных систем (ФАУ «ГосНИИАС»), г. Москва, Российская Федерация, ORCID: https://orcid.org/0009-0008-1769-4121, e-mail: adkozyrev@2100.gosniias.ru

1. ВВЕДЕНИЕ

Распознавание речи – важная часть современных интерфейсов взаимодействия человека и машины. Оно необходимо для реализации голосового управления – наиболее естественного для человека способа управления. Система голосового управления может иметь в своем составе следующие модули: модуль распознавания речи, модуль синтеза речи, модуль обработки естественного языка и интеллектуальной интерпретации речи.

Модуль распознавания речи в свою очередь состоит из подмодулей – акустическая модель, языковая модель и декодер. В настоящее время проблема распознавании речи решается с помощью комбинации различных методов, таких как дискриминантный анализ, основанный на теореме Байеса, скрытые марковские модели, нейронные сети. Среди нейронных сетей при распознавании речи наибольшую популярность приобрели свёрточные сети LSTM-сети.

Акустическая модель позволяет по признакам входного сигнала на фрейме получить распределение вероятностей нахождения акустических единиц, например фонем, на данном участке сигнала.

Языковые модели предназначены для учета контекста, они позволяют выявить наиболее вероятные последовательности фонем и слов с точки зрения структуры языка и текущего контекста

Декодер на базе вероятностей, которые являются результатом работы акустической модели с учетом языковой модели выбирает конкретную речевую единицу.

В данной статье рассмотрен 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.

Граф устройства голосовой команды изображен на рисунке 1.

2. ПОДХОД К РАСПОЗНАВАНИЮ РЕЧИ

Основой предложенного подхода служат вероятностные сети (сети на радиальных базисных элементах). Для реализации распознавания речи с их помощью, необходимо извлечь из звукового сигнала статичные образцы (наборы признаков), на которых будет обучаться сеть.

Для извлечения признаков из сигнала используется спектральный анализ в форме оценки спектральных плотностей. Оценки спектральных плотностей представляются в виде спектрограмм и преобразуются с помощью специального преобразования, вид которого приведен в статье «О методе распознавания голосовых команд с применением особого преобразования спектральных плотностей» [1]. Для понижения размерности спектрограмм (с целью использования вероятностных сетей) используются субдискретизирующие слои [2], подобные тем, что используются внутри сверточных нейронных сетей.

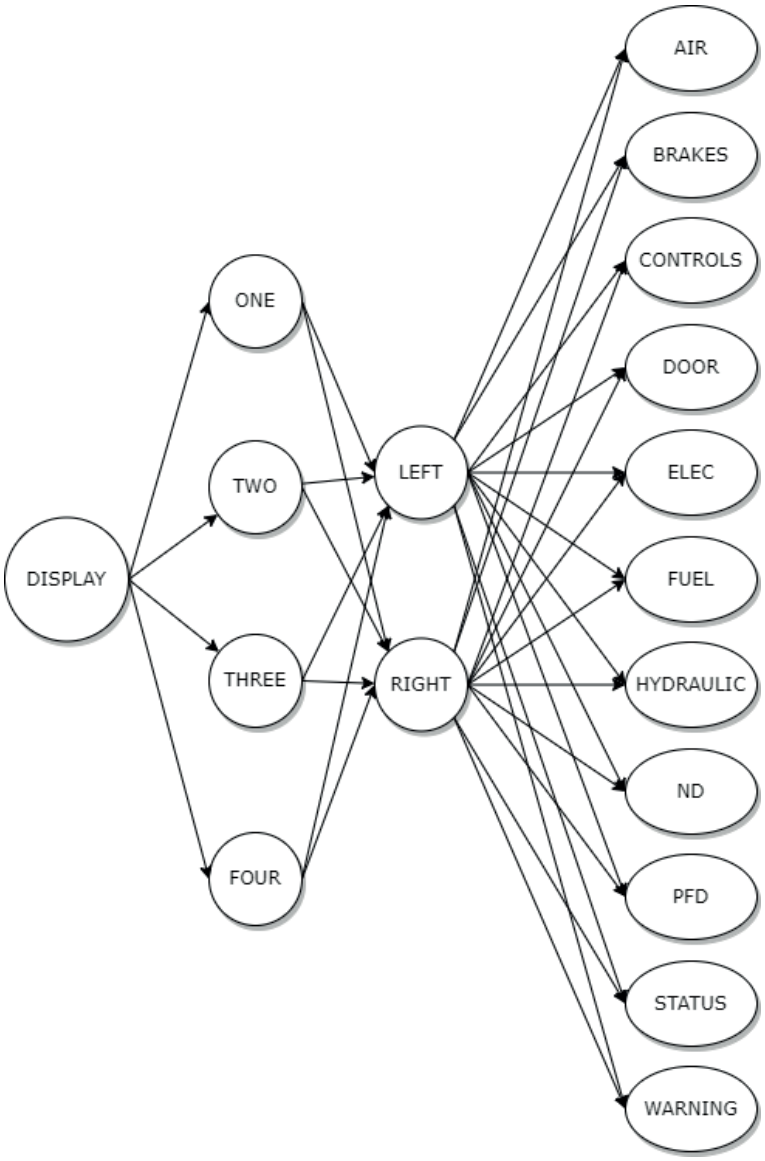

Рис. 1. Граф устройства голосовой команды

Признаки сигнала извлекаются на коротких пересекающихся интервалах с помощью оконной функции, шаг перемещения оконной функции меньше её длинны. Пример перемещения окна изображен на рисунке 2. На рисунке 3 изображен отрезок сигнала входящий в одно окно. В ходе экспериментов было установлено, что оптимальным является окно в 500 мс, а оптимальным шагом 250 мс.

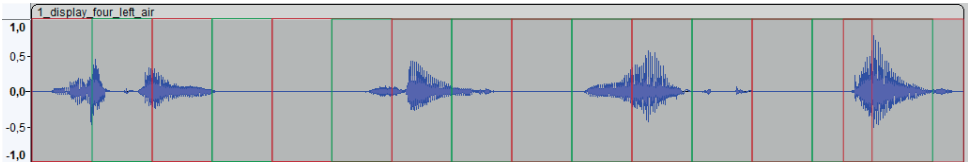

Рис. 2. Пример перемещения окна по сигналу

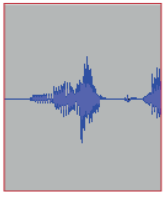

Рис. 3. Пример отрезка сигнала в окне

Примеры исходной спектрограммы и спектрограммы, преобразованной субдискретизирущим слоем с функцией максимума изображен на рисунке 4 (на нем обработанная спектрограмма изображена в том же размере, тогда как на самом деле она меньше в 3 раза по каждой из осей). Наиболее качественно на имеющихся данных показал размер ядра 3x3.

После перехода от исходной спектрограммы к её сжатой версии, обучается вероятностная сеть на отрезках сигнала, которые получаются в результате прохода окна. Для обучения сети образцы предварительно размечаются вручную. Фрагменты аудиосигнала размечаются по следующему принципу, если во фрагменте слово или его часть занимает больше половины фрагмента, и при этом во фрагменте не содержатся части другого слова, то фрагмент помечается меткой этого слова, в противном случае файл помечается специальной меткой, которая обозначает отсутствие полезной информации во фрагменте. В данной работе в качестве специальной пометки используется слово «trash».

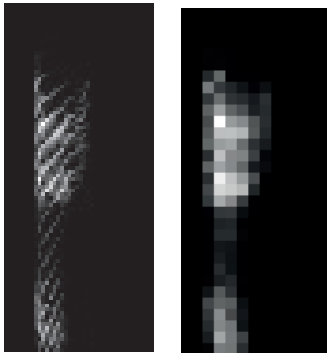

Рис. 4. Пример преобразования спектрограммы субдискретизирующим слоем

Основу вероятностной сети [3, с. 118] составляют радиальные базисные элементы, которые реализуют функцию, вычисляющую значение некоторой радиальной базисной функции с центром в точке $A(\mathbf{c})$, которая в общем виде задается формулой 1.

$$\phi_A(\mathbf{x}) = \phi(\| \mathbf{x} - \mathbf{c} \|) \tag{1}$$

В рамках данной статьи в качестве радиальной базисной функции в вероятностных сетях Гаусса (формула 2). Она имеет один настраиваемый параметр ϵ.

$$\phi(r) = e^{-(er)^2} \tag{2}$$

Топология сети на радиальных базисных элементах изображена на рисунке 5.

Для обучения вероятностных сетей расчет исходных спектрограмм слов осуществлялся на частотах 0–3008 Гц с шагом 32 Гц. По времени спектрограмма отрезка была посчитана с 32-ти миллисекундным шагом. Размер исходных спектрограмм составил 36x94 (3348 точек). Размер спектрограмм, прошедших субдискретизирующий слой – 12x31 (372 точки).

Обучение вероятностной сети производится с помощью задания правильной топологи, центры радиальных базисных элементов устанавливаются в координаты образцов обучающей выборки, радиальные базисные элементы одного класса, связываются с одним нейроном-сумматором, который отвечает за вычисление веса данного класса. Далее веса нормируются с целью получения вероятности.

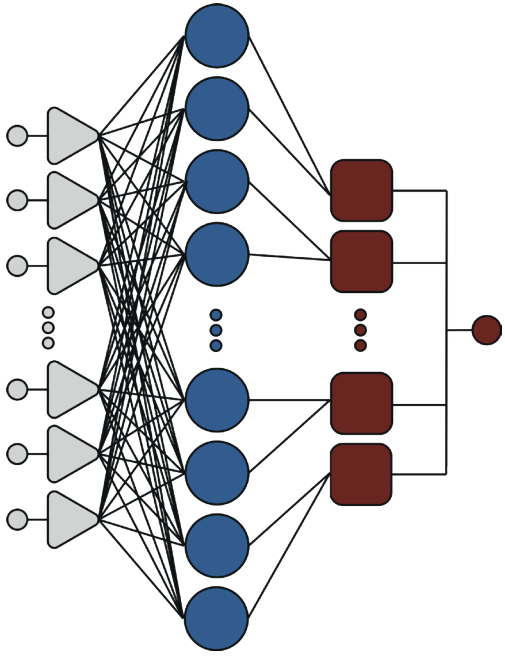

Рис. 5. Топология вероятностной сети

Доля верных ответов при распознавании единичного окна, как правило, не очень высока и колеблется на уровне 60–70 %, однако данный классификатор может служить основой для алгоритма распознавания команды. Рассмотрим пример такого алгоритма, в данном примере окно движется с половинным шагом окна (125мс).

Если пройтись по аудио записи окном и в каждом окне применить вероятностную сеть, то результат можно представить в виде таблицы таблица 1. В качестве исходной аудиодорожки взята запись команды «display four left air».

Таблица 1

Результат распознавания фрагментов команды «display four left air» вероятностной сетью

слово	вероятность нахождения в окне
display	0,999735
display	0,710781
display	0,994761
two	0,217422
display	0,688826
three	0,517768
trash	0,941634
trash	0,999633
trash	0,999974
four	0,991796
two	0,371546
four	0,526390
four	0,737630
warning	0,293155
trash	0,980138
trash	0,993597
one	0,884373
left	0,381266
left	0,987561
left	0,604423
elec	0,343962
status	0,332910
trash	0,999405
trash	0,790388
fuel	0,876582
air	0,664482
air	0,697840
air	0,975710

В результате распознавания появляется последовательность слов с вероятностями их нахождения в конкретном окне. Из анализа таблицы следует, что в некоторых окнах были распознаны слова, которых нет в команде, причем некоторые ошибочно распознанные слова имеют достаточно высокую оценку вероятности.

Второй шаг алгоритма – отбрасывание мусорных спецслов и определение веса слов, путем суммирования вероятностей подряд идущих слов. Дополнительно данный шаг нумерует результаты. Результаты представлены в таблице 2.

Таблица 2

Результат второго шага алгоритма для команды «display four left controls»

№	слово	Вес
1	display	2,705277
2	two	0,217422
3	display	0,688826
4	three	0,517768
5	four	0,991796
6	two	0,371546
7	four	1,264020
8	warning	0,293155
9	one	0,884373
10	left	1,973250
11	elec	0,343962
12	status	0,332910
13	fuel	0,876582
14	air	2,338032

Третий шаг алгоритма строит графовое представление результата. Для команды из примера графовое представление изображено на рисунке 6.

Графовое представление результата строится следующим образом: каждому слову из таблицы 2 ставится в соответствие вершина, которая имеет две пометки, словесную и числовую, а также добавляются 2 специальные вершины, «BEGIN» помечается числом 0 и «END» помечается числом $N+1$, где N количество слов. Числовая и словесная пометка остальных вершин берется из таблицы.

Далее вершины разбиваются на 6 классов эквивалентности. Эти классы эквивалентности имеют порядковые номера от 1 до 6. Классы 2–5 определяются на основании графа устройства команды, а еще 2, первый и шестой, содержат по одной вершине с пометками «BEGIN» и «END» соответственно.

Ребра графа являются ориентированными и взвешенными или, применяя другой термин, взвешенными дугами.

Дуга графа соединяет вершины тогда и только тогда, когда выполняется набор условий:

- конечная вершина лежит в классе с порядковым номером на 1 больше номера класса начальной вершины;
- конечная вершина имеет числовую пометку, которая больше числовой пометки начальной вершины.

Все входящие в вершину дуги помечаются весом слова-пометки данной вершины, вес вершины с пометкой «END» полагается равным 1.

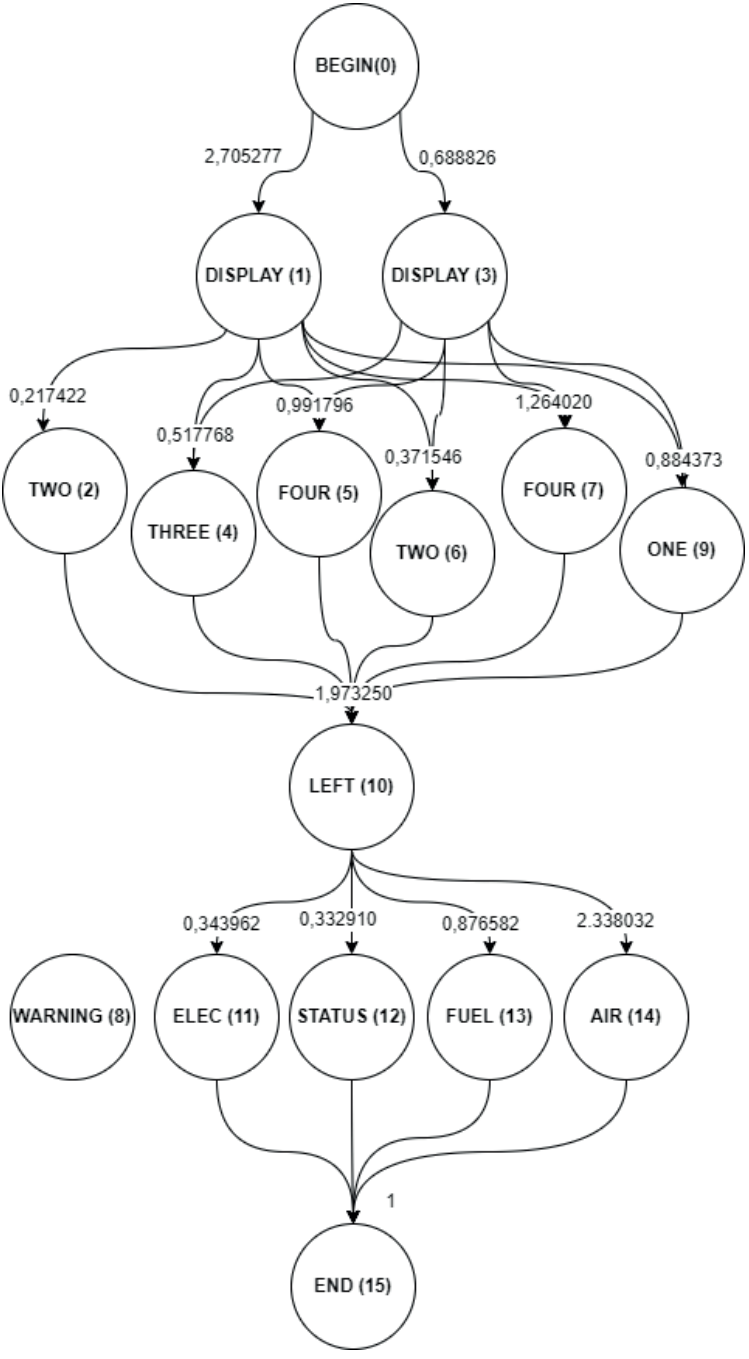

Рис. 6. Графовое представление результата

Далее, на четвертом шаге анализируются все пути из вершины пометкой «BEGIN» в вершину с пометкой «END», для каждого такого пути вычисляется метрика равная произведению весов дуг, по которым проходит данный путь – вес пути. Веса путей для данного графа приведены в таблице 3.

Таблица 3

Пути в графе из вершины «BEGIN» в вершины «END»

Команда	Путь	Вес	Команда	Путь	Вес
display two left elec	1 2 10 11	0,399216	display one left fuel	1 9 10 13	4,138299
display two left status	1 2 10 12	0,386388	display one left air	1 9 10 14	11,03773
display two left fuel	1 2 10 13	1,017396	display three left elec	3 4 10 11	0,242068
display two left air	1 2 10 14	2,713612	display three left status	3 4 10 12	0,23429
display three left elec	1 4 10 11	0,950691	display three left fuel	3 4 10 13	0,616907
display three left status	1 4 10 12	0,920144	display three left air	3 4 10 14	1,645422
display three left fuel	1 4 10 13	2,422823	display four left elec	3 5 10 11	0,463687
display three left air	1 4 10 14	6,462187	display four left status	3 5 10 12	0,448788
display four left elec	1 5 10 11	1,82107	display four left fuel	3 5 10 13	1,181698
display four left status	1 5 10 12	1,762556	display four left air	3 5 10 14	3,151842
display four left fuel	1 5 10 13	4,64097	display two left elec	3 6 10 11	0,173706
display four left air	1 5 10 14	12,37846	display two left status	3 6 10 12	0,168125
display two left elec	1 6 10 11	0,682208	display two left fuel	3 6 10 13	0,442687
display two left status	1 6 10 12	0,660288	display two left air	3 6 10 14	1,180741
display two left fuel	1 6 10 13	1,738597	display four left elec	3 7 10 11	0,589358
display two left air	1 6 10 14	4,637211	display four left status	3 7 10 12	0,570421
display four left elec	1 7 10 11	2,31463	display four left fuel	3 7 10 13	1,501971
display four left status	1 7 10 12	2,240258	display four left air	3 7 10 14	4,006078
display four left fuel	1 7 10 13	5,8988	display one left elec	3 9 10 11	0,413464
display four left air	1 7 10 14	15,73336	display one left status	3 9 10 12	0,400179
display one left elec	1 9 10 11	1,623827	display one left fuel	3 9 10 13	1,053707
display one left status	1 9 10 12	1,571651	display one left air	3 9 10 14	2,810461

На пятом шаге алгоритма (который является опциональным) происходит свертка путей – веса путей, имеющих одинаковые команды, складываются (таблица 4).

На шестом шаге алгоритма выбирается команда, имеющая максимальный вес, в данном случае это команда «display four left air», она имеет вес 35,26975.

Применение данного алгоритма позволяет верно распознать 84 из 88 команд (95 %) имеющейся выборки.

Блок-схема алгоритма изображена на рисунке 7.

Таблица 4

Веса кандидатов команд

display two left elec	4,233266
display two left status	1,220382
display two left fuel	3,19868
display two left air	8,531565
display three left elec	1,192759
display three left status	1,154434
display three left fuel	3,039729
display three left air	8,107609
display four left elec	5,188745
display four left status	5,022023
display four left fuel	13,22344
display four left air	35,26975
display one left elec	2,037291
display one left status	1,97183
display one left fuel	5,192006
display one left air	13,84819

Для демонстрации работы алгоритма было разработано программное обеспечение, имеющее графический интерфейс. на языке Python с использованием графического интерфейса Kivy [4]. Демонстрационное программное обеспечение представляет собой, однооконное приложение, которое позволяет выбрать и распознать файл. В результате своей работы программа выводит исходный сигнал аудиофайла, сигнал, обработанный модулем шумоподавления, отрезки сигнала, вырезанные скользящим окном, граф и результат распознавания. При этом граф распознавания находится в прокручиваемой области. Результат работы программы изображен на рисунке 8. На рисунке 9 отдельно изображен автоматически построенный программой граф распознавания команды.

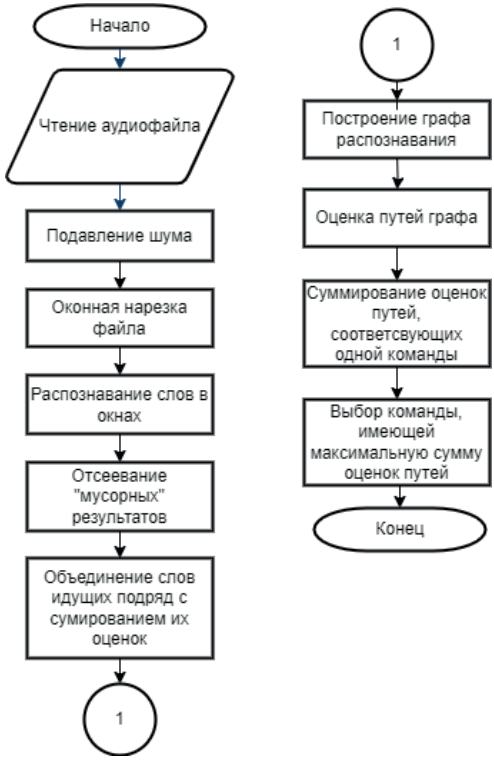

Рис. 7. Блок-схема алгоритма распознавания речи

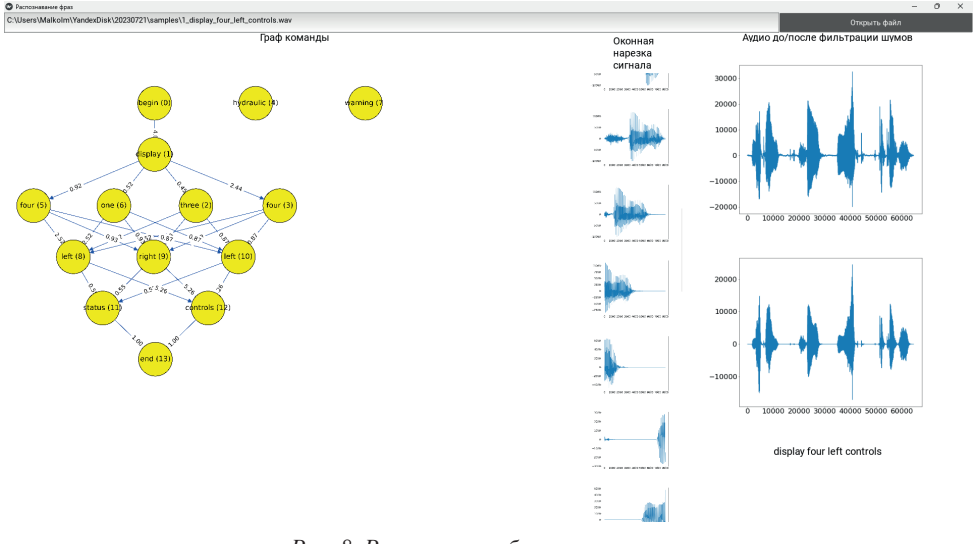

Рис. 8. Результат работы программы

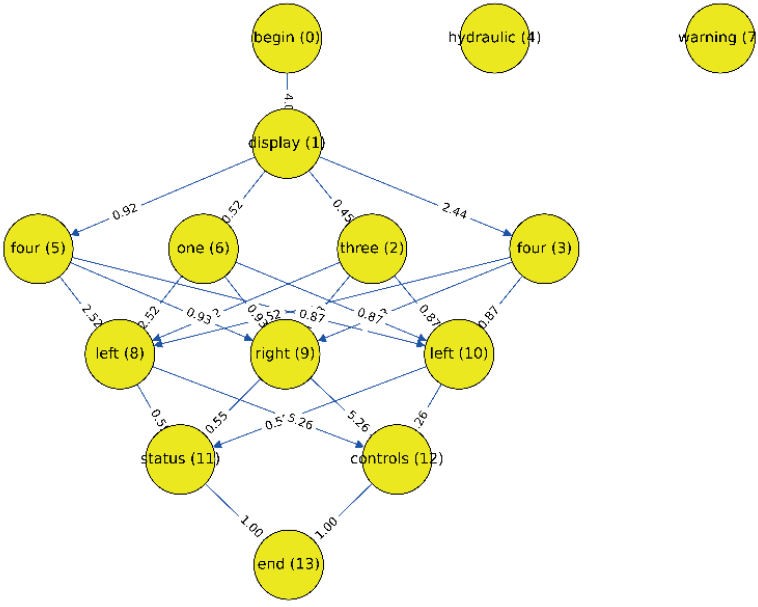

Рис. 9. Результирующий граф команды

3. ЗАКЛЮЧЕНИЕ

В данной работе предложен метод распознавания речи, основанный на применении вероятностных сетей. Данный метод применим для распознавания команд определенной структуры, и дает существенный прирост точности распознавания в сравнении с широко распространенными методами, основанными на сверточных нейронных сетях (95 % против 87 %). Данный результат является достаточно высоким для практического применения в сфере голосового управления.

Литература

1. *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: 10.17759/mda.2022120304
2. *Dumoulin, Vincent, and Francesco Visin.* “A guide to convolution arithmetic for deep learning.” arXiv preprint arXiv:1603.07285 (2016).
3. *Куравский Л.С., Баранов С.Н.* Компьютерное моделирование и анализ данных. Конспекты лекций и упражнения: Учеб. пособие. – М.: РУСАВИА, 2012. – 218 с.:
4. *Ulloa, Roberto.* Kivy–Interactive Applications and Games in Python. Packt Publishing Ltd, 2015.

Probabilistic Neural Network Application to Speech Recognition Problem

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

Alexey D. Kozyrev**
Moscow State University of Psychology and Education (MSUPE)
State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia
ORCID: https://orcid.org/0009-0008-1769-4121
e-mail: adkozyrev@2100.gosniias.ru

This article discusses an approach to solving the problem of speech recognition based on probabilistic neural networks. The problem is formulated as a problem of command recognition. Commands have equal lengths (in words). Each word position has its own set of candidates. The recognition algorithm for solving this problem was developed. The core of the algorithm is a probabilistic network that recognizes modified estimates of spectral densities. The algorithm allows for high precision of recognition, which is sufficient for the creation of a voice user interface.

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

For citation:
Levonovich N.I., Kozyrev A.D. Probabilistic Neural Network Application to Speech Recogni tion Problem. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2023. Vol. 13, no. 3, pp. 39–51. DOI: https://doi.org/10.17759/mda.2023130303 (In Russ.,abstr. in Engl.).

References

1. Levonovich N.I. Voice Commands Recognition Method that Uses Special Spectral Density Transform. Modelirovanie i analiz dannikh = Modelling and Data Analysis, 2022. Vol. 12, no. 3, pp. 49–57. DOI: 10.17759/mda.2022120304. (In Russ., abstr. in Engl.)
2. Dumoulin, Vincent, and Francesco Visin. “A guide to convolution arithmetic for deep learning.” arXiv preprint arXiv:1603.07285 (2016).
3. L.S. Kuravskii, S.N. Baranov Komp’yuternoe modelirovanie i analiz dannykh. Konspekty lektsii i uprazhneniya: Ucheb. posobie. – M.: RUSAVIA, 2012. – 218 s.:
4. Ulloa, Roberto. Kivy–Interactive Applications and Games in Python. Packt Publishing Ltd, 2015.

***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

****Alexey D. Kozyrev,** Post-Graduate Student, Moscow State University of Psychology and Education (MSUPE); Engineer, State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia, ORCID: https://orcid.org/0009-0008-1769-4121, e-mail: adkozyrev@2100.gosniias.ru

Получена 25.08.2023 Received 25.08.2023
Принята в печать 09.09.2023 Accepted 09.09.2023