МРНТИ 28.23.37
УДК 004.8

https://doi.org/10.51889/2022-2.1728-7901.12

А. Бакыт[1], Е.С. Смагулов[1], С.Б. Маден[1], Д.М. Жексебай[1], Е.Т. Кожагулов[1]*

[1]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан
[*]e-mail: zhexebay92@gmail.com

ГОЛОСОВОЙ ПОМОЩНИК НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ

Аннотация

С развитием интерфейсных технологий в смарт-устройствах голосовые помощники быстро завоевали популярность. Эти помощники предназначены для использования голосовых команд, чтобы обеспечить более удобное взаимодействие с людьми. В связи с этим предлагается один из методов реализации голосового помощника на основе нейронных сетей. Исследованы методы реализации основных этапов создания голосового помощника. В статье представлены результаты тестирования модели на основе сверточной нейронной сети. В качестве речевых команд были выбраны следующие слова: yes, no, up, down, right, left, go, stop. Эта модель классифицирует 8 речевых команд с точностью 86,63%. Модель нейронной сети лучше всего классифицировала команды: yes, up, down, right, left, stop. Для команды go точность 66,67%, а no - 76,4%, это связано со схожим звучанием слов down, go, no.

Ключевые слова: голосовой помощник, распознавания речи, преобразования звукового сигнала, распознавания команд, сверточная нейронная сеть.

Аңдатпа

Ә. Бақытқызы[1], Е.С. Смағұлов[1], С.Б. Мәден[1], Д.М. Жексебай[1], Е.Т. Кожагулов[1]*
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

ТЕРЕҢ ОҚЫТУ НЕГІЗІНДЕГІ ДАУЫС КӨМЕКШІСІ

Смарт құрылғылардағы интерфейстік технологиялардың дамуымен дауыстық көмекшілер тез танымал болды. Бұл көмекшілер адамдармен ыңғайлы әрекеттесу үшін дауыстық пәрмендерді пайдалануға арналған. Осыған байланысты нейрондық желілерге негізделген дауыстық көмекшіні енгізу әдістерінің бірі ұсынылды. Дауыстық көмекші құрудың негізгі кезеңдерін жүзеге асыру әдістері зерттелді. Мақалада үйірткілі нейрондық желіге негізделген модельді тестілеу нәтижелері берілген. Сөйлеу пәрмені ретінде мына сөздер таңдалды: yes, no, up, down, right, left, go, stop. Бұл модель 86,63% дәлдікпен 8 сөйлеу пәрменін жіктейді. Нейрондық желі моделі ең жақсы жіктелген командалар: yes, up, down, right, left, stop. Go командасының дәлдігі 66,67%, ал no - 76,4%, бұл down, go, no сөздерінің ұқсас дыбысталуына байланысты.

Түйін сөздер: дауыстық көмекші, сөйлеуді тану, аудио сигналды түрлендіру, командаларды тану, үйірткілі нейрондық желі.

Abstract

DEEP LEARNING VOICE ASSISTANT

Bakytkyzy A.[1], Smagulov Y.S.[1], Maden S.B.[1], Zhexebay D.M.[1], Kozhagulov Y.T.[1]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

With the development of interface technologies in smart devices, voice assistants quickly gained popularity. These assistants are designed to use voice commands to provide a more convenient interaction with people. In this regard, one of the methods for implementing a voice assistant based on neural networks is proposed. Methods for implementing the main stages of creating a voice assistant have been studied. The article presents the results of testing a model based on a convolutional neural network. The following words were chosen as speech commands: yes, no, up, down, right, left, go, stop. This model classifies 8 speech commands with an accuracy of 86.63%. The neural network model best classified commands: yes, up, down, right, left, stop. The command to go is 66.67% accurate, and no is 76.4%, this is due to the similar sounding of the words down, go, no.

Keywords: voice assistant, speech recognition, audio signal conversion, command recognition, convolutional neural network.

Введение

Голосовой пользовательский интерфейс и технологии распознавания речи вошли в жизнь людей относительно недавно и с тех пор стали незаменимыми. Разработка этих систем – сложная задача, так как требует знания правил определенного языка программирования, технологий преобразования речи, а также вычислительной точности и эффективности при обработке голосовой информации.

Одним из основных принципов интеллектуальных голосовых помощников, особенно с учетом их естественности и простоты, является их способность к обучению и адаптации. Эта распространенность достигается в самых разных сферах, таких как управление умным домом, мобильная помощь или операционные системы. Основными причинами этого являются естественность речи как формы общения в отличие от использования дополнительной внешней периферии и, кроме того, независимость от дополнительного обучения на основе одного приложения. Одна из причин необходимости разработки новых систем, концепций и алгоритмов обработки голоса заключается в том, что голосовое взаимодействие становится более удобным средством управления SMART технологиями. Кроме того, это отличный шанс проявить себя как предприниматель в бизнес сфере, в связи с растущим спросом данной технологии.

А также, это чрезвычайно актуально для людей с ограниченными возможностями, так как данная технология предоставляет возможность облегчить рутинные занятия. Частичная потеря голоса или ее особенности являются основными препятствиями распознавания речи для людей с ограниченными возможностями. Современные системы распознавания речи не могут быть адаптированы к индивидуальным особенностям указанных людей, что мешает им использовать эти системы. Способ преодоления этого препятствия – создание обучаемых зависимых от говорящего систем на основе искусственных нейронных сетей. Основой голосового помощника является распознавания речи, которая включает в себя очень много процессов. На данный момент, существуют очень много систем распознавания речи. Все системы распознавания речи относятся к двум категориям. Первая категория - это системы, зависящие от говорящего, которые в процессе обучения подстраиваются под его речь. Эти системы требуют полной перенастройки для работы с другим говорящим. Данная категория систем зависит от голосовых данных говорящего. Системы, зависящие от одного говорящего, обеспечивают эффективное распознавание голоса, хотя настройка системы для каждого отдельного пользователя является довольно трудоемкой задачей. Индивидуальные особенности произношения (такие как темп речи, тембр и т. д.) усложняют разработку этих систем [1].

В настоящее время разработчики больше сосредоточены на создании систем, не зависящих от говорящего, которые используют огромные объемы данных и не принимают во внимание то, как люди с нарушениями речи произносят команды. Как пример отличного голосового помощника можно отметить, Siri от Apple, Google Assistant, Amazon Alexa и Алиса. Данные мировые бренды помогают разработчикам получить бесплатное программное обеспечение с открытым исходным кодом, что дает возможность дорабатывать и адаптировать их для различных ситуаций. Самыми популярными системами с открытым исходным кодом являются CMU Sphinx, Julius, Kaldi, RWTH ASR, iATROS.

В большинстве случаев (например, как Google и Яндекс) распознавание голоса осуществляется алгоритмами, основанными на наборе данных говорящих, которые по-разному произносят слова правильно. Создание инструмента, который распознает относительно ограниченный набор команд, при этом неважно, может ли пользователь произносить слова правильно, поскольку его речь может быть непонятной и нечеткой. Чтобы распознать ограниченный набор команд, можно использовать классический многослойный персептрон или сверточную нейронную сеть с дополнительными сверточными и объединяющими слоями. Такие сети доказали свою эффективность в распознавании тестовых данных, но для них требуется сравнительно большое количество обучающих данных [2-4].

Другой способ распознавания речи с помощью искусственных нейронных сетей – это рекуррентные нейронные сети. В этих структурах имеется обратная связь, которой выход одного нейрона влияет на другие. В рамках данного исследования большой интерес представляет ассоциативная память, предназначенная для распознавания образов. На вход сети поступает ряд векторов признаков, которые формируют определенные состояния сети. После получения нового вектора признаков, сеть перейдет в состояние, напоминающее один из запомненных векторов, наиболее похожее на новый [5]. Самым простым в реализации ассоциативной памятью, которая сравнивается с обучающей выборкой, является сеть Хопфилда, которая в исходном состоянии имеет довольно ограниченный объем памяти [6]. Ассоциативная память получила дальнейшее развитие в трехслойной сети Хэмминга и двунаправленной автоассоциативной сети Коско.

Некоторые исследования предполагают динамическую ассоциативную память, которая повторно измеряет веса не только во время обучения, но и в качестве реакции на ввод. Эта сеть используется для распознавания звукового сигнала голосовых команд [7].

Комбинированное использование многослойных сетей с прямой связью и рекуррентных сетей имеет большое значение во время обучения. Согласно исследованием, для использования ассоциативной памяти на основе многослойного персептрона и связанной рекуррентной сети при распознавании шумной речи предлагается многослойная сеть с полиномиальными функциями активации для распознавания образцов [8]. Первоначально, сети с ассоциативной памятью использовались для хранения рядов нулей и единиц. Однако аудиоинформация более сложна и необходимо хранить значения с плавающей запятой в диапазоне от 0 до 1, поэтому предлагается использовать сеть Хопфилда с дополнительным слоем нейронов [9]. В самоорганизующихся сетях, согласно правилам конкуренции, активация набора параметров определяет нейрон-победитель, который соответствует этому набору. Это позволяет разбить входные данные на кластеры и выделить каждый новый вектор в свою категорию. Ученые предлагают самоорганизующуюся инкрементную нейронную сеть на основе искусственной нейронной сети для распознавания как изображений, так и голосовых команд [10]. Так же стоить отметить, обучение нейронной сети непосредственно на звуковом сигнале неэффективно. Поэтому, исследователями данной технологии предлагаются разные методы для преобразования звукового сигнала на другой формат. Преобразования звукового сигнала является одним из важных этапов для дальнейшей обработки данных и для обучения модели. Наиболее популярными методами считаются:

1. Преобразования звукового сигнала с помощью MFCC;
2. Преобразования звукового сигнала с помощью системы LinTo.

Данные методы применяются не только для преобразования, а также, для дальнейшего обучения. Из-за особенностей восприятия звука наиболее подходящим набором функций для обработки является MFCC. Для этого весь аудиовход разбивается на кадры, каждый из которых проходит преобразования Фурье, спектры сигнала, полученные коэффициенты применяются для дискретного косинусного преобразования. Сначала необходимо записать обучающую выборку, состоящую из аудиозаписей. Записанные слова преобразуются в MFCC, и каждому набору MFCC назначается транскрипция записанного слова. После этого вектор признаков MFCC используется для обучения нейронной сети. По окончании обучения программа может работать в режиме распознавания команд [11]. Согласно исследованиям, для распознавания команд можно использовать следующие виды искусственных нейронных сетей [11]:

1. Многослойный персептрон с двумя скрытыми слоями, содержащий многочисленное количество нейронов. Функции ReLu, tanh, Sigmoid с выходным слоем и функция активация SoftMax;
2. Однослойная сверточная нейронная сеть;
3. Многослойная сверточная нейронная сеть, состоящая из различных слоев: сверточной, слои подвыборки и слои полностью подключенной нейронной сети;
4. Двунаправленная авто ассоциативная сеть Коско;
5. Самоорганизующаяся карта (карта Кохонена, SOM).

Вторым способом для создания модели, а также для преобразования звукового сигнала на другой формат является платформа на основе системы LinTO. LinTO – это клиент-серверная система с открытым исходным кодом, которая позволяет клиентов обеспечивать с голосовым пользовательским интерфейсом, которое дает возможность администраторам сопровождать программным обеспечениям и управлять аппаратом пользователей. Система повышает простоту использования и производительность как в административном управленим, так и в контексте клиентских приложений.

Кроме того, предлагаются процессы управления голосом без помощи рук и нескольких динамиков для распознавания речи и голосовой доступ для приложения клиента. Помощник состоит из моделей и ресурсов, необходимые для выполнения желаемых функций, и может включать любое количество навыков или действия, определенные для конкретного процесса, например, устный прогноз погоды. LinTO – наиболее подходящая платформа для администрации и для пользователя [12]. Цель данной работы реализовать модель нейронной сети для классификации речевых команд: yes, no, up, down, right, left, go, stop.

Методы

Как ранее было упомянуто, для создания голосового помощника необходимо соблюдать последовательность процессов, такие как: обработка голосовых данных потребителя, распознавание речи, преобразования речи для дальнейшего обучения, обучение модели, определения и понимания языка потребителя, отклик и обратная связь потребителю. Данная статья включает в себе

эксперимент моделирования всех процессов с исключением отклика. На экспериментальном этапе данной работы было достигнуто правильное понимание различных команд, например, «go», «down» и т.д. Для создания такой модели выбрали рабочую платформу Python, как основу для создания и для визуализации использовали Seaborn и Tensorflow. Seaborn – это библиотека визуализации данных Python, основанная на matplotlib. Оно предоставляет высокоуровневый интерфейс для рисования привлекательных и информативных статистических графиков.

Звуковые данные были преобразованы в спектрограммы, потому что работа с аудиосигналами приводит к большим потерям. Соответственно, спектрограммы представляют собой изображению которая несет аналогичную информацию как в аудио сигнале. В дальнейшем использовали изображения, для обучения модели выбрали сверточную нейронную сеть. Совместная работа сверточной нейронной сети и Tensorflow обеспечивает высокие результаты в работе с изображениями. В глубоком обучении сверточная нейронная сеть представляет собой класс глубоких нейронных сетей, наиболее часто применяемых для анализа визуальных образов, использующую особую технику, называемой сверткой.

Функция активации ReLU была использована для получения выходных данных. Функция активации ReLU (Rectified Linear Activation) является наиболее распространенным выбором функции активации в мире глубокого обучения. ReLU обеспечивает самые современные результаты и в то же время очень эффективен с точки зрения вычислений. Основная концепция функции активации ReLU заключается в следующем: «Верните 0, если ввод отрицательный, в противном случае оставить ввод как есть». Для визуализации полученных результатов тестирования использовали матрицу неточности. На практике значения точности гораздо более удобней рассчитывать с использованием матрицы неточностей (confusion matrix). В случае если количество классов команд относительно невелико (не более 100-150 классов), этот подход позволяет довольно наглядно представить результаты работы классификатора. Матрица неточностей – это матрица с размером N*N, где N – это количество классов команд. Столбцы этой матрицы для экспертного решения, а строки для решения классификатора. Когда мы классифицируем команды из тестовой выборки мы инкрементируем число, стоящее на пересечении строки класса, который вернул классификатор и столбца класса к которому действительно относится.

Результаты и обсуждение

Для обучения использовались 16000 аудиозаписей, для тестирования 800 аудиозаписей, разбитых на восемь класса, каждому назначена определенная команда, например, «go», «down» и т.д. Применялась библиотека Tensorflow для создания обучающего набора чтобы извлечь пар аудио-меток и проверки результатов. Позже создались наборы проверки и тестирования, используя аналогичную процедуру.

На рисунке 1 показаны реализации звуковых сигналов с соответствующими метками. Анализируя звуковые сигналы, можно отметить, что при одинаковых метках (например, как «yes») звуковые сигналы могут иметь разные характеристики.

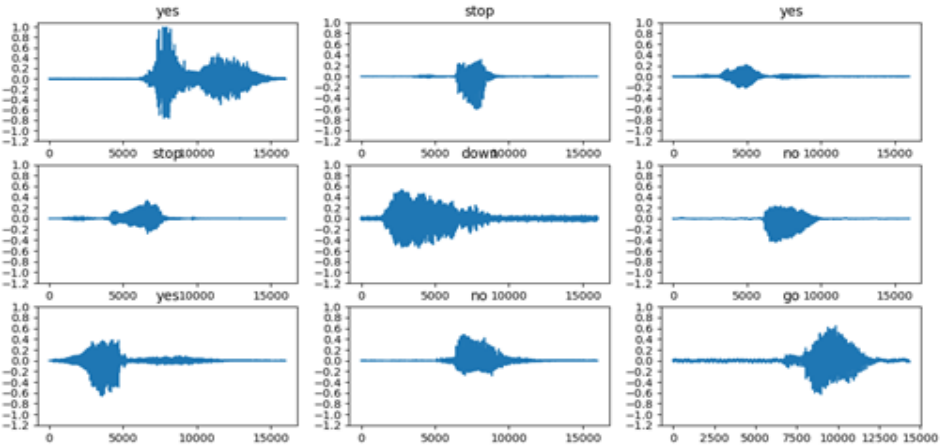

Рисунок 1. Звуковые сигналы с соответствующими метками

Звуковые сигналы были преобразованы в частоты для логарифмической шкалы и транспонировались так, чтобы время было представлено на оси x (столбцы). На рисунке 2 показаны реализация звукового сигнала и спектрограмма данной реализации.

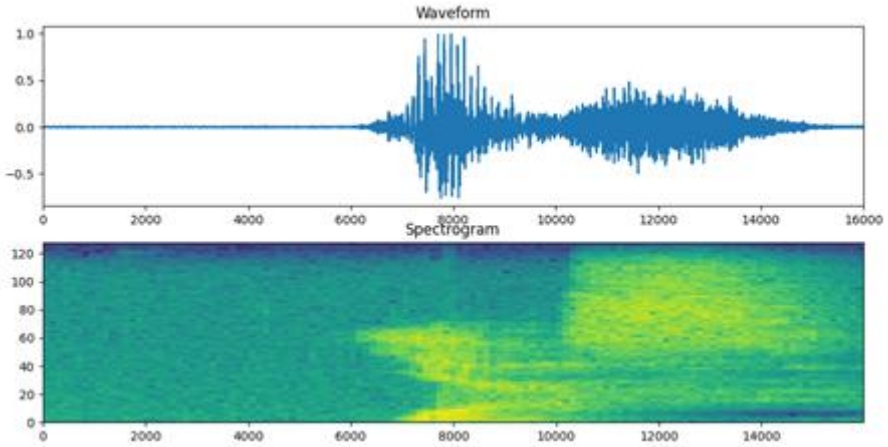

Рисунок 2. Звуковой сигнал и соответствующий спектр

В качестве модели использована простая сверточная нейронная сеть (CNN), поскольку все аудиофайлы преобразовались в изображения спектрограмм (рисунок 3). Анализируя спектрограммы, можно отметить зеленые и оранжево-желтые зоны. В спектрограммах информационную часть описывают оранжево-желтые части. Размеры исходных изображения спектрограмм были сформированы на размер 32x32 для дальнейшего обучения, а также, для активации была использована функция ReLU.

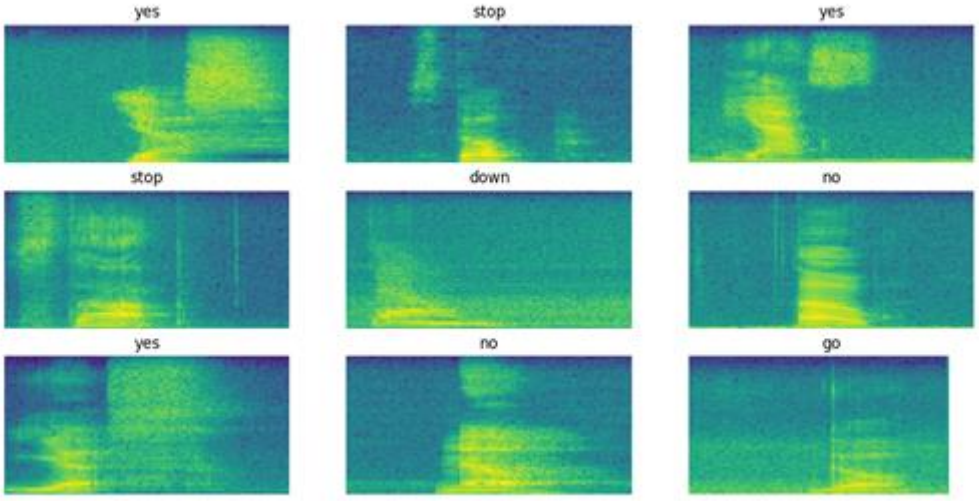

Рисунок 3. Спектры с соответствующими метками

После обучения модели были проверены кривые потерь при обучении, для проверки эффективности модели во время обучения. Количество эпох составляет 10. Ниже показаны (Рисунок 4) кривые потерь при обучении и проверке. Точка пересечения двух кривых должны стремиться к нулю. Чем ближе точка пересечения к точке 0, тем меньше потерь и больше эффективности в обучения модели.

Следующий этап был предназначен для получения результата тестирования. На рисунке 5 показана матрица неточности. Анализируя матрицу, можно заметить световую гамму от темной до светлой, соответственно, с нуля до высокого значения. Общее количество звуковых сигналов для тестирования составило 800 аудиозаписей. Количество правильных классификации составляет 693, то есть точность данной модели 86,63%.

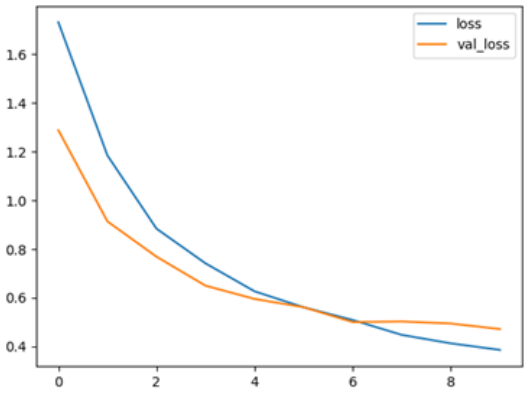

Рисунок 4. Кривые потерь при обучении и проверке

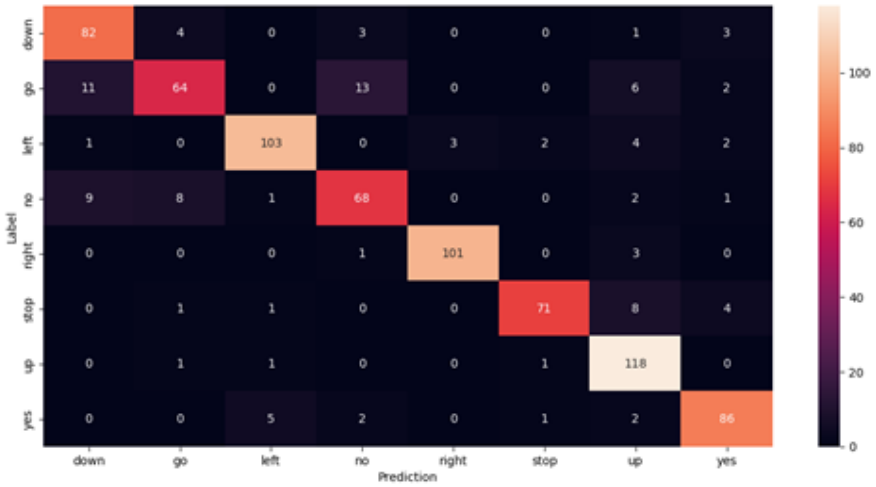

Рисунок 5. Матрица неточности. Результаты тестирования

Заключение

В данной статье были предоставлены методы реализации основных этапов создания голосового помощника. В рабочей платформе Python была тестирована обучающая модель на основе сверточной нейронной сети для распознавания команд. Были задействованы все основные процессы, такие как: обработка голосовых данных, распознавание речи, преобразования речи для дальнейшего обучения, обучение модели, определение и понимание языка.

Для дальнейшего обучения нейронной сети звуковые сигналы были преобразованы в спектрограммы. Обученная модель классифицирует звуковые сигналы с точностью 86,63% для тестовых данных.

Благодарность

Данная работа поддержана грантом Министерства образования и науки Республики Казахстан в рамках проекта №AP09058525 «Разработка цифровых радиомодулей 5G и приемных станции СВЧ сигналов на основе SoC».

Список использованной литературы:

1 Rybka, Jan, and Artur Janicki. "Comparison of speaker dependent and speaker independent emotion recognition." International Journal of Applied Mathematics and Computer Science 23, no. 4 (2013): 797-808. https://doi.org/10.2478/amcs-2013-0060

2 Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review." IEEE Signal processing magazine 32, no. 6 (2015): 74-99. https://doi.org/10.1109/MSP.2015.2462851

3 Kozhagulov, Y. T., Zhexebay D. M., Sarmanbetov S. A., Sagatbayeva A. A., and Zholdas D. "Comparative analysis of object detection processing speed on the basis of neuroprocessors and neuroaccelerators". Известия НАН РК. Серия физико-математических наук 4 (2020): 61-67.

4 Сарманбетов, С., Максутова А.А., Кожагулов Е.Т. "Классификатор изображений микросхем при помощи сверточной нейронной сети." Известия НАН РК. Серия физико-математических наук 6 (2021): 59-65.

5 Stöckel, Andreas. "Design space exploration of associative memories using spiking neurons with respect to neuromorphic hardware implementations." (2016).

6 Tampel, I. B. "Automatic speech recognition–the main stages over last 50 years." Nauchno-Tekhnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki 15, no. 6 (2015): 957. https://doi.org/10.17586/2226-1494-2015-15-6-957-968

7 Vaishnavi, Y., R. Shreyas, S. Suhas, U. N. Surya, Vandana M. Ladwani, and V. Ramasubramanian. "Associative memory framework for speech recognition: adaptation of hopfield network." In 2016 IEEE Annual India Conference (INDICON), pp. 1-6. IEEE, 2016. https://doi.org/10.1109/INDICON.2016.7839105

8 Krotov, Dmitry, and John J. Hopfield. "Dense associative memory for pattern recognition". Advances in neural information processing systems 29 (2016).

9 Sussner, Peter, Estevão L. Esmi, Ivan Villaverde, and Manuel Graña. "The Kosko subsethood fuzzy associative memory (KS-FAM): Mathematical background and applications in computer vision." Journal of Mathematical Imaging and Vision 42, no. 2 (2012): 134-149. https://doi.org/10.1007/s10851-011-0292-0

10 Shen, Furao, Qiubao Ouyang, Wataru Kasai, and Osamu Hasegawa. "A general associative memory based on self-organizing incremental neural network." Neurocomputing 104 (2013): 57-71. https://doi.org/10.1016/j.neucom.2012.10.003

11 Khorosheva, T., Novoseltseva M., Geidarov N., Krivosheev N., and Chernenko S. "Neural network control interface of the speaker dependent computer system «Deep Interactive Voice Assistant DIVA» to help people with speech impairments." In International Conference on Intelligent Information Technologies for Industry, pp. 444-452. Springer, Cham, 2018. https://doi.org/10.1007/978-3-030-01818-4_44

12 Rebai, Ilyes, Sami Benhamiche, Kate Thompson, Zied Sellami, Damien Laine, and Jean-Pierre Lorre. "LinTO Platform: A Smart Open Voice Assistant for Business Environments." In Proceedings of the 1st International Workshop on Language Technology Platforms, pp. 89-95. 2020.

References:

1 Rybka, Jan, and Artur Janicki. "Comparison of speaker dependent and speaker independent emotion recognition." International Journal of Applied Mathematics and Computer Science 23, no. 4 (2013): 797-808. https://doi.org/10.2478/amcs-2013-0060

2 Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review." IEEE Signal processing magazine 32, no. 6 (2015): 74-99. https://doi.org/10.1109/MSP.2015.2462851

3 Kozhagulov, Y. T., Zhexebay D. M., Sarmanbetov S. A., Sagatbayeva A. A., and Zholdas D. "Comparative analysis of object detection processing speed on the basis of neuroprocessors and neuroaccelerators." Izvestija NAN RK. Serija fiziko-matematicheskih nauk 4 (2020): 61-67.

4 Sarmanbetov, Sanzhar, A. A. Maksutova, and E. T. Kozhagulov. "Klassifikator izobrazhenij mikroshem pri pomoshhi svertochnoj nejronnoj seti." Izvestija NAN RK. Serija fiziko-matematicheskih nauk 6 (2021): 59-65.

5 Stöckel, Andreas. "Design space exploration of associative memories using spiking neurons with respect to neuromorphic hardware implementations." (2016).

6 Tampel, I. B. "Automatic speech recognition–the main stages over last 50 years". Nauchno-Tekhnicheskii Vestnik Informatsionnykh Tekhnologii, Mekhaniki i Optiki 15, no. 6 (2015): 957. https://doi.org/10.17586/2226-1494-2015-15-6-957-968

7 Vaishnavi, Y., R. Shreyas, S. Suhas, U. N. Surya, Vandana M. Ladwani, and V. Ramasubramanian. "Associative memory framework for speech recognition: adaptation of hopfield network." In 2016 IEEE Annual India Conference (INDICON), pp. 1-6. IEEE, 2016. https://doi.org/10.1109/INDICON.2016.7839105

8 Krotov, Dmitry, and John J. Hopfield. "Dense associative memory for pattern recognition." Advances in neural information processing systems 29 (2016).

9 Sussner, Peter, Estevão L. Esmi, Ivan Villaverde, and Manuel Graña. "The Kosko subsethood fuzzy associative memory (KS-FAM): Mathematical background and applications in computer vision." Journal of Mathematical Imaging and Vision 42, no. 2 (2012): 134-149. https://doi.org/10.1007/s10851-011-0292-0

10 Shen, Furao, Qiubao Ouyang, Wataru Kasai, and Osamu Hasegawa. "A general associative memory based on self-organizing incremental neural network." Neurocomputing 104 (2013): 57-71. https://doi.org/10.1016/j.neucom.2012.10.003

11 Khorosheva, T., Novoseltseva M., Geidarov N., Krivosheev N., and Chernenko S. "Neural network control interface of the speaker dependent computer system «Deep Interactive Voice Assistant DIVA» to help people with speech impairments." In International Conference on Intelligent Information Technologies for Industry, pp. 444-452. Springer, Cham, 2018. https://doi.org/10.1007/978-3-030-01818-4_44

12 Rebai, Ilyes, Sami Benhamiche, Kate Thompson, Zied Sellami, Damien Laine, and Jean-Pierre Lorre. "LinTO Platform: A Smart Open Voice Assistant for Business Environments." In Proceedings of the 1st International Workshop on Language Technology Platforms, pp. 89-95. 2020.