УДК 004.89:004.934
doi: 10.26583/bit.2024.2.09

Владислав В. Жолондковский[1], Юрий И. Бочаров[2], Владимир А. Бутузов[3]
[1-3]Национальный исследовательский ядерный университет «МИФИ»,
Каширское ш., 31, Москва, 115409, Россия
[1]АО НТЦ «Модуль»,
4-я улица 8 Марта, 3, Москва, 125319, Россия
[1]e-mail: ofisse884@mail.ru, https://orcid.org/0009-0007-4150-4589
[2]e-mail: yibocharov@mephi.ru, https://orcid.org/0000-0002-0092-0590
[3]e-mail: butuzov@inbox.ru, https://orcid.org/0000-0002-6071-3905

НЕЙРОСЕТЕВОЕ УСТРОЙСТВО ДЛЯ РАСПОЗНАВАНИЯ РЕЧЕВЫХ КОМАНД НА БАЗЕ АППАРАТНЫХ УСКОРИТЕЛЕЙ СЕМЕЙСТВА NEUROMATRIX

Аннотация. Рассматриваются различные подходы к решению задачи распознавания речевой команды в реальном времени с помощью искусственных нейронных сетей (ИНС) нескольких типов. Приведен краткий обзор глубоких полносвязных, сверточных и рекуррентных нейронных сетей. Показано, что ИНС сверточного и рекуррентного типа способны обеспечить необходимый уровень производительности и точности при формировании логического вывода (inference) в масштабе реального времени. Дается сравнение производительности сверточного и рекуррентного алгоритмов при их портировании на аппаратную платформу К1879ВМ6Я из семейства высокопроизводительных цифровых сигнальных процессоров NeuroMatrix. При портировании учтены особенности архитектуры микропроцессора. Все основные операции исполняются на векторно-матричном сопроцессоре. Продемонстрировано преимущество сверточной ИНС. С учетом результатов исследований нейросети этого типа на платформе К1879ВМ6Я осуществлен ее перенос на более производительный аппаратный ускоритель – систему на кристалле К1879ВМ8Я. Рассматриваются перспективы создания системы распознавания речи на платформе К1879ВМ8Я.
Ключевые слова: распознавание речи, искусственная нейронная сеть, сверточная нейронная сеть, микропроцессор, К1879ВМ8Я, К1879ВМ6Я.
Для цитирования: ЖОЛОНДКОВСКИЙ, Владислав В.; БОЧАРОВ, Юрий И.; БУТУЗОВ, Владимир А. НЕЙРОСЕТЕВОЕ УСТРОЙСТВО ДЛЯ РАСПОЗНАВАНИЯ РЕЧЕВЫХ КОМАНД НА БАЗЕ АППАРАТНЫХ УСКОРИТЕЛЕЙ СЕМЕЙСТВА NEUROMATRIX. Безопасность информационных технологий, [S.l.], т. 31, № 2, с. 134–141, 2024. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1636. DOI: http://dx.doi.org/10.26583/bit.2024.2.09.

Vladislav V. Zholondkovskiy[1], Yuri I. Bocharov[2], Vladimir A. Butuzov[3]
[1-3]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashirskoe sh., 31, Moscow, 115409, Russia
[1]STC Module,
4th Str., March 8, 3, Moscow, 125319, Russia
[1]e-mail: ofisse884@mail.ru, https://orcid.org/0009-0007-4150-4589
[2]e-mail: yibocharov@mephi.ru, https://orcid.org/0000-0002-0092-0590
[3]e-mail: butuzov@inbox.ru, https://orcid.org/0000-0002-6071-3905

Neural network device for recognizing speech commands based on hardware accelerators of the NEUROMATRIX family

Abstract. Various approaches to solving the problem of speech command recognition in real time using artificial neural networks (ANN) of several types are considered. A brief overview of deep fully-connected, convolutional and recurrent neural networks is given. It is shown that ANNs of convolutional and recurrent types are capable of providing the required level of performance and accuracy when performing inference

in real time. A comparison is given of the performance of convolutional and recurrent algorithms when ported to the K1879VM6Ya hardware platform from the NeuroMatrix family of high-performance digital signal processors. When porting, the features of the microprocessor architecture were considered. All basic operations are performed on a vector matrix coprocessor. The advantage of convolutional ANN has been demonstrated. Considering the results of research into a neural network of this type on the K1879VM6Ya platform, it was transferred to a more productive hardware accelerator - the K1879VM8Ya system-on-chip. The prospects for creating a speech recognition system on the K1879VM8Ya platform are considered.
Keywords: speech recognition, artificial neural network, convolutional neural network, microprocessor, K1879VM8Ya, K1879VM6Ya.
For citation: ZHOLONDKOVSKIY, Vladislav V.; BOCHAROV, Yuri I.; BUTUZOV, Vladimir A. Neural network device for recognizing speech commands based on hardware accelerators of the NEUROMATRIX family. IT Security (Russia), [S.l.], v. 31, no. 2, p. 134–141, 2024. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1636. DOI: http://dx.doi.org/10.26583/bit.2024.2.09.

Введение

Одним из ключевых направлений в разработке человеко-машинного интерфейса является создание устройств распознавания речи. Современный подход к решению задач, связанных с распознаванием речи, заключается в преобразовании голосового аудиоконтента в текстовый при помощи искусственной нейронной сети (ИНС). В зарубежных источниках это направление обычно обозначают как LVCSR (Large Vocabulary Continuous Speech Recognition) или STT (Speech-To-Text). В его основе – система, которая прямо отображает последовательность акустических признаков в последовательность графем/фонем или слов. Такие системы реализованы на основе ИНС различных топологий, которые обучены на большом массиве слов и требуют большой вычислительной мощности для своей работы [1].

Среди широкого спектра различных задач распознавания речи можно выделить некоторые частные задачи, решение которых необходимо для практической реализации устройств распознавания голосового ввода. Одной из таких задач является обнаружение и распознавание речевой команды в голосовом сигнале. Устройства, создаваемые в рамках этого направления, находят применение в качестве компонентов систем голосовых помощников [2].

При создании устройств для распознавания речи используют двухэтапный подход. На первом этапе устройство «пробуждается» речевой командой, а на втором этапе происходит распознавание речи. Это делается для снижения энергетических затрат и требований к производительности аппаратных средств для работы в реальном времени, поскольку ИНС, работающие с речевым аудиопотоком, достаточно сложны и требуют больших вычислительных ресурсов для инференса. Для первого этапа требуется «легкая» нейронная сеть, которая могла бы непрерывно, в режиме в реальном времени обрабатывать входную информацию для того, чтобы отозваться на команду и «разбудить» устройство. Центральный процессор персонального компьютера (ПК), мобильного устройства или многоядерный графический процессор имеют для реализации такой сети необходимую вычислительную мощность. Для автономных устройств с ограниченными вычислительными и энергетическими ресурсами, как правило, требуется использование аппаратных ускорителей. К ним предъявляются такие требования, как возможность длительной работы с автономным питанием, возможность потоковой обработки сигналов в реальном времени, а также эффективное выполнение реализуемого алгоритма средствами конкретной аппаратной платформы с учетом ее особенностей.

В качестве аппаратной платформы устройств распознавания речевых команд перспективно применение СБИС семейства NeuroMatrix, которые содержат специальные средства векторно-матричных вычислений и демонстрируют высокую производительность

при выполнении базовых математических операций, лежащих в основе нейросетевых алгоритмов. При этом, они обладают еще целым рядом достоинств, таких как набор удобных программных инструментов для взаимодействия с различными платформами (в первую очередь с ПК) и наличие аппаратных модулей, которые содержат эти СБИС (для К1879ВМ6Я это NMStick, для К1879ВМ8Я – NMCard), реализованных на отечественной компонентной базе [3].

С целью выбора наиболее подходящей нейросети для распознавания голосовой команды с точки зрения удобства ее аппаратной реализации и портирования на вычислительные платформы семейства NeuroMatrix, проведен сравнительный анализ свойств различных ИНС, реализуемых на процессорах этого семейства. Этому посвящена настоящая работа.

В первой части работы анализируются наиболее широко применяемые сейчас ИНС: глубокие (Deep Neural Networks, DNN), сверточные (Convolutional Neural Networks, CNN) и рекуррентные (Recurrent Neural Networks, RNN). Показано их портирование на СБИС К1879ВМ6Я. Даны результаты сравнения производительности этих ИНС при выполнении инференса.

Во второй части работы показано портирование выбранной ИНС на СБИС К1879ВМ8Я с применением вычислительного модуля, содержащего этот процессор. Показаны возможности применения этого модуля наряду с модулем на базе процессора К1879ВМ6Я для решения задачи распознавания голосовой команды и создания перспективной системы распознавания речи.

1. Процессоры семейства NeuroMatrix

СБИС семейства NeuroMatrix – это высокопроизводительные RISC-процессоры с архитектурой SIMD (Single Instruction, Multiple Data) и VLIW (Very Long Instruction Word), элементами суперскалярности и аппаратной поддержкой матричных и векторных операций. Программная поддержка микропроцессоров семейства NeuroMatrix включает комплект высокоуровневых инструментальных средств. Пакет nmDL содержит набор специализированного программного обеспечения. Пакет DLCL (Deep Learning Compute Library) содержит оптимизированные функции базовых математических операций над векторными и матричными примитивами.

СБИС К1879ВМ6Я

СБИС К1879ВМ6Я – это высокопроизводительный многоядерный микропроцессор, предназначенный в первую очередь для цифровой обработки сигналов. Он содержит два вычислительных ядра – NMPU0 и NMPU1 с архитектурой NeuroMatrix NMC4, каждое из которых включает в себя 32/64-разрядный RISC процессор и векторно-матричный блок (рис. 1).

Первое ядро выполняет операции над данными в формате с плавающей точкой и содержит четыре вычислительных ячейки. Второе ядро предназначено для выполнения операций над целочисленными данными переменной длины от 2 до 64 бит. Базовая целочисленная операция – матричное умножение. Тактовая частота ядер процессора К1879ВМ6Я – 500 МГц. Использование СБИС К1879ВМ6Я в качестве многоцелевого аппаратного ускорителя базируется на наличии в его составе средств цифровой обработки сигналов, и аппаратной поддержки функций активации, дискретного преобразования Фурье и дискретного косинусного преобразования.

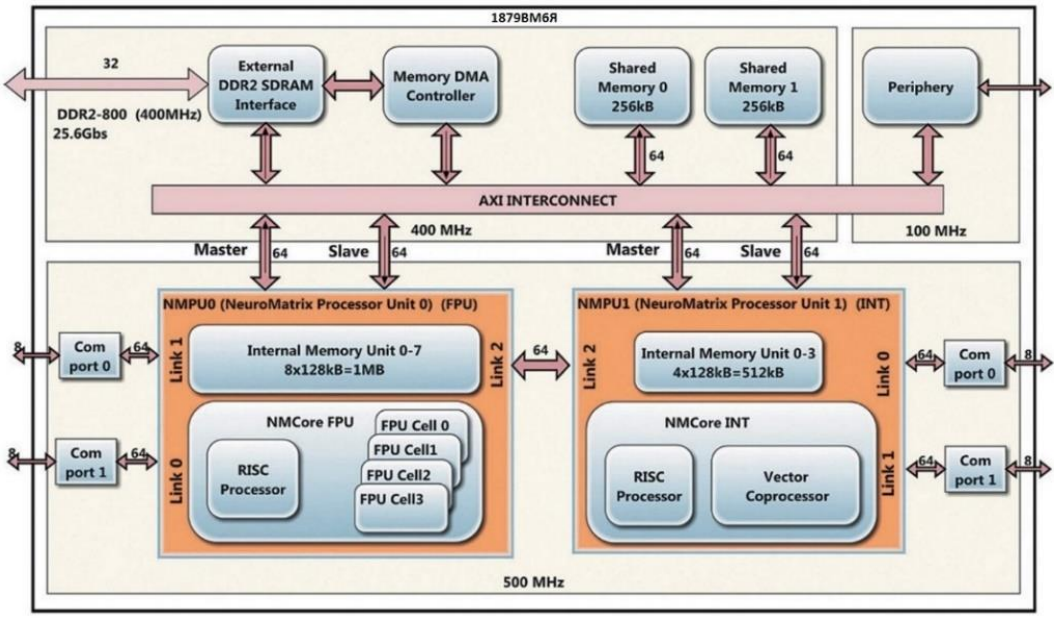

Рис. 1. Архитектура микропроцессора К1879ВМ6Я

СБИС К1879ВМ8Я

СБИС К1879ВМ8Я – это гетерогенная многопроцессорная система на кристалле (СнК), в состав которой входят 16 процессорных ядер NeuroMatrix Core 4 (NMC4) и 5 ядер ARM Cortex-A5, которые объединены в кластерную структуру (рис. 2).

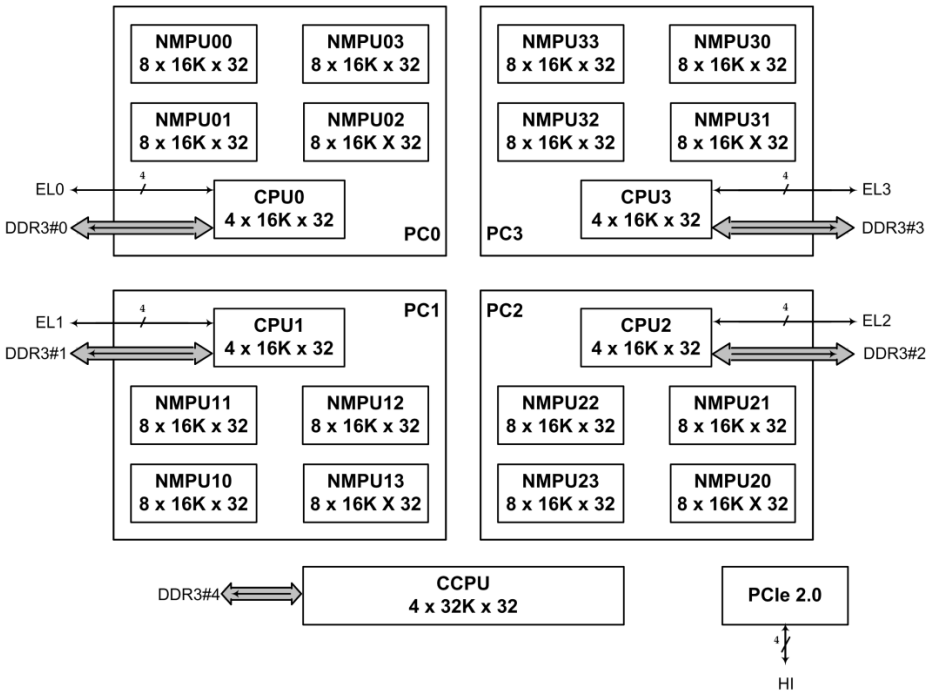

Рис. 2. Кластерная структура СнК 1879ВМ8Я

Микросхема содержит четыре кластера, каждый из которых содержит одно ядро ARM Cortex-A5 (на рис. 2 – CPU) и четыре процессорных ядра NMC4 (на рис. 2 – NMPU). Кроме этого, присутствует управляющее системой процессорное ядро ARM – CCPU. СБИС имеет пять интерфейсов с внешней памятью типа DDR3 (по одному интерфейсу в каждом

кластере и один для выделенного процессорного ядра). Тактовая частота ядер процессора К1879ВМ8Я – 800 МГц. Эта СнК может быть использована и для создания устройства распознавания речевой команды, и для создания перспективной системы распознавания речи в силу наличия сравнительного большого вычислительного ресурса.

2. Сравнение нейронных сетей и реализация ИНС на базе СБИС К1879ВМ6Я

Проведено сравнение трех типов ИНС для распознавания речевой команды – DNN, CNN и RNN [4–6] с точки зрения возможности их реализации на базе СБИС К1879ВМ6Я. Анализ показал, что все три типа нейронных сетей могут быть использованы для решения такой задачи. Точность распознавания при использовании сверточных и рекуррентных ИНС выше, чем при использовании глубоких ИНС. Основные методы подготовки звуковых данных к их вводу в нейросеть [7] эффективно сочетаются со сверточным и рекуррентным подходами. Полученная в результате предобработки спектрограмма может быть рассмотрена как изображение и классифицирована сверточной ИНС по стандартной методике распознавания образов. Поэтому для практической реализации на СБИС К1879ВМ6Я были отобраны CNN и RNN. Они были портированы на СБИС и протестированы.

Основой устройства для решения задачи распознавание речевой команды является программно-аппаратный комплекс (хост-процессор), содержащий ПК и специализированный вычислительный модуль NMStick. Возможен перенос встроенной версии устройства и на другие программно-аппаратные платформы. СБИС К1879ВМ6Я имеет только одно векторно-матричное ядро и мало подходит для выполнения предобработки звуковых данных. Поэтому она осуществляется средствами хост–процессора. Для преобразования входного звукового сигнала в спектрограмму применен алгоритм, аналогичный рассмотренному в [5]. Сформированная спектрограмма, может быть подана на вход нейросети.

Первой была создана и портирована сверточная нейронная сеть для распознавания речевой команды, основанная на архитектуре resnet. Обученный на 8 различных командах сверточный классификатор, основываясь на базе данных, лежащей в открытом доступе [8], показал точность, которая сравнима с точностью подобных нейросетевых решений [9]. Тестовая выборка составляла 1600 образцов (10% от общей выборки в 16000 слов, разделенной на 8 одинаковых классов). Далее, при помощи пакета NMDL нейронная сеть была перенесена на СБИС К1879ВМ6Я.

Основные операции в ИНС с архитектурой resnet – это вычисление функций свертки, активации, субдискретизации, а также поэлементные матричные умножения – произведения Адамара [10]. Средствами этой СБИС вычисление многоканальной свертки с ядром типа квадратной матрицы, а также выполнение матричных умножений выполняется эффективно [3]. Дополнительно преобразуется и выходной слой линейного преобразования. Эксперимент, основанный на методике распознавания образов, показал, что аппаратная платформа на базе СБИС К1879ВМ6Я обладает достаточной производительностью, чтобы удовлетворить условию обработки данных в реальном времени. Разработанная программная часть комплекса также позволяет обрабатывать входящий непрерывный звуковой поток в режиме реального времени.

Одной из задач, которая решалась при проектировании системы, была регулировка чувствительности. Обычно она регулируется либо внесением дополнительных классов (шум, другие слова), либо выставлением высокого порога отсечения по уверенности сети. Был избран второй способ, в связи с высокой точностью предсказаний на тестовом множестве. После того, как сеть классифицировала входящее слово как команду, в течение некоторого временного интервала эта команда все еще будет определяться по спектрограмме. Во избежание возникновения ложных срабатываний в поточном режиме, после предсказания

добавлена задержка порядка длины одной команды, во время которой входные данные не обрабатываются.

В СБИС К1879ВМ6Я была также портирована рекуррентная нейронная сеть типа LSTM (Long-Short Term Memory) [11]. Она обучена на том же самом датасете и с точностью того же уровня, как и у сверточной сети. Несмотря на поддержку этой вычислительной платформой базовых операций – умножения с накоплением и вычисления функции нелинейной сигмоидальной активации, вычисление функции гиперболического тангенса в составе операций, поддерживаемых компилятором NMDL, отсутствует, поэтому портирование ИНС выполнено с использованием пакета DLCL (набора базовых примитивов, написанных на языке ассемблера процессора).

При этом даже с учетом неоптимальной работы с памятью (весовые коэффициенты матриц хранились во внешнем запоминающем устройстве), была обеспечена работа в реальном времени благодаря тому, что рекуррентная сеть работает не со всей спектрограммой, а с каждым фреймом в отдельности, сохраняя при этом предыдущую информацию в своих скрытых слоях.

При проведении эксперимента в режиме реального времени ИНС типа LSTM показывала частоту ложных срабатываний при распознавании одиночных команд порядка 10 в час, а ИНС сверточного типа примерно вдвое меньше.

При оценке результата следует принять во внимание, что поскольку исследовалась принципиальная возможность реализации системы распознавания речевых команд на базе СБИС семейства NeuroMatrix, то при обучении не были приняты возможные меры для повышения шумоустойчивости (робастности) системы, а также не применялись даже простейшие фильтры для обработки сигналов.

Для того чтобы уменьшить вероятность ложных срабатываний, был применен подход с парной командой: если голосовая последовательность содержит пару определенных команд, то в силу их статистической независимости, вероятность поступления ложной пары снижается на порядок. В этом случае частоты ложных срабатываний следующие: для рекуррентной ИНС типа LSTM – примерно 2 в час, для ИНС сверточного типа – примерно 0,5 в час. Основываясь на данных результатах, можно сделать вывод, что, несмотря на более простой механизм инференса, сверточная сеть демонстрирует преимущество перед ИНС рекуррентного типа, поскольку имеет более низкую вероятность ложного срабатывания и более чувствительный отклик на подачу голосовой команды.

При этом, необходимо отметить, что как уже неоднократно упоминалось, сверточная сеть работает со спектрограммой в целом и решает задачу определения команды, которая сведена к распознаванию образов. Таким образом, неважно, какого рода звуковая последовательность будет рассматриваться на этой спектрограмме – это необязательно голосовая команда. Данный подход можно использовать и для более широкого круга задач выявления специфических фрагментов в звуковых информационных потоках.

3. Портирование ИНС сверточного типа на СБИС К1879ВМ8Я

Выбранная по результатам проведенного исследования ИНС сверточного типа портирована в СБИС К1879ВМ8Я в составе модуля NMCard при помощи пакета NMDL. Задача предобработки данных решалась с использованием ПК. Инференс нейронной сети выполнялся с помощью модуля NMCard. Обмен данными осуществлялся с применением высокоскоростного интерфейса PCIe 4.

СБИС К1879ВМ8Я использовалась в конфигурации с одним ядром на одном кластере. Данная СБИС предоставляет возможность проводить параллельные вычисления с разбиением подаваемого на ИНС изображения (разделение на 4 квадранта – тайлинг) и с использованием нескольких ядер одновременно. В данном случае в силу малого размера

картинки, тайлинг не имел смысла, поскольку выигрыш в производительности нивелировался бы временем на пересылки фрагментов на разные ядра, а параллельное вычисление не использовалось в силу характера обрабатываемых данных, разнесенных во времени. Несмотря на это выигрыш в производительности при выполнении инференса был достигнут, поскольку тактовая частота данной СБИС выше, чем у К1879ВМ6Я.

4. Перспективы исследований

На основе полученных результатов в качестве следующего этапа работы рассматриваются два пути.

Первый путь – это создание системы распознавания речи с использованием всех вычислительных ресурсов К1879ВМ8Я для портирования полновесных ИНС, хоть и облегченных по сравнению с DeepSpeech2 [12] и QuartzNet [13], и использование одного ядра СБИС для распознавания команд для них.

Второй путь – использование ИНС для валидации данных, полученных с облегченного устройства распознавания команд, то есть создание конвейера данных, куда при любом срабатывании «легкой сети», загружались бы временные отрезки с возможными командами. После этого малая ИНС валидировала бы данные (не обязательно в реальном времени). Такой подход называется мультимоделинг [14] и применяется в современных коммерческих устройствах.

Первый путь дает возможность создания прототипа полноценного устройства, однако его потенциальная точность может быть недостаточно высокой в условиях зашумленных аудиоданных. Второй путь позволяет значительно повысить точность и уменьшить количество ложных срабатываний в случае зашумленного аудиоконтента, а также возможного ложного срабатывания при подаче похожего по звуковому паттерну слова.

Заключение

Для решения задачи распознавания голосовой команды с использованием аппаратной платформы на базе СБИС семейства NeutoMatrix на основе проведенного анализа выбраны ИНС сверточного и рекуррентного типа, которые были портированы в СБИС К1879ВМ6Я и К1879ВМ8Я с учетом особенностей их архитектуры. Проведенные эксперименты показали преимущество ИНС сверточного типа в точности определения голосовой команды. Поскольку такая ИНС обрабатывает данные, содержащиеся в спектрограмме звукового сигнала, это позволяет использовать ее не только для распознавания голосовых команд, но и для более широкого круга задач выявления специфических фрагментов в звуковых информационных потоках.

СПИСОК ЛИТЕРАТУРЫ:

1. Bahdanau D., Chorowski J., Serdyuk D., et al. End-to-end attention-based large vocabulary speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016, p. 4945–4949. DOI: 10.1109/ICASSP.2016.7472618.
2. Bonet D., Cámbara G., et al. Speech Enhancement for Wake-Up-Word detection in Voice Assistants. IberSPEECH 2021. 2021, p. 41–45. DOI:10.21437/IberSPEECH.2021-9.
3. Бирюков А.А., Таранин М.В., Таранин С.В. Процессор 1879ВМ6Я. Реализация глубоких сверточных нейронных сетей. DSPA: Вопросы применения цифровой обработки сигналов. 2018, № 4, т. 8, с. 191–195. URL: https://www.module.ru/uploads/media/1534165451-18796-68958c488f.pdf (дата обращения: 09.04.2024)
4. Wang Z., Li X., Zhou J. Small-footprint Keyword Spotting Using Deep Neural Network and Connectionist Temporal Classifier, 2017. URL: https://arxiv.org/abs/1709.03665 (дата обращения: 02.04.2024).
5. Jose C., Mishchenko Yu., Senechal T. et al. Accurate Detection of Wake Word Start and End Using a CNN. Interspeech. 2020. DOI: 10.21437/Interspeech.2020-1491.
6. Bluche T., Primet M., Gisselbrecht T. Small-Footprint Open-Vocabulary Keyword Spotting with Quantized LSTM Networks, 2020. URL: https://arxiv.org/abs/2002.10851 (дата обращения: 02.04.2024).

7. Muda L., Begam M., Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, 2010. URL: https://arxiv.org/abs/1003.4083 (дата обращения: 02.04.2024).
8. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, 2018. URL: https://arxiv.org/abs/1804.03209 (дата обращения: 02.04.2024).
9. Yundong Z., Suda N., Lai L., Chandra V. Hello Edge: Keyword Spotting on Microcontrollers, 2017. URL: https://arxiv.org/abs/1711.07128 (дата обращения: 02.04.2024).
10. He K., Zhang X., et al. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016, p. 770–778. DOI: 10.1109/CVPR.2016.90.
11. Hochreiter S., Schmidhuber J. Long short-term memory. Neural computation. 1997, v. 9(8), p. 1735–1780. DOI:10.1162/neco.1997.9.8.1735.
12. Amodei D., Anubha R., et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, 2015. URL: https://arxiv.org/abs/1512.02595 (дата обращения: 02.04.2024).
13. Kriman S., Beliaev S., Ginsburg B., et al. QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, 2019. URL: https://arxiv.org/abs/1910.10261 (дата обращения: 02.04.2024).
14. López F., Luque J., et al. Robust Wake-Up Word Detection by Two-stage Multi-resolution Ensembles, 2013. URL https://arxiv.org/abs/2310.11379 (дата обращения: 02.04.2024).

REFERENCES:

[1] Bahdanau D., Chorowski J., Serdyuk D., et al. End-to-end attention-based large vocabulary speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016, p. 4945–4949. DOI: 10.1109/ICASSP.2016.7472618.
[2] Bonet D., Cámbara G., et al. Speech Enhancement for Wake-Up-Word detection in Voice Assistants. IberSPEECH 2021. 2021, p. 41–45. DOI:10.21437/IberSPEECH.2021-9.
[3] Biryukov A.A., Taranin M.V., Taranin S.V. Processor 1879VM6Ya. Implementation of deep convolutional neural networks. DSPA: Issues in the application of digital signal processing. 2018, no. 4, v. 8, p. 191–195. URL: https://www.module.ru/uploads/media/1534165451-18796-68958c488f.pdf (accessed: 04.09.2024) (in Russian).
[4] Wang Z., Li X., Zhou J. Small-footprint Keyword Spotting Using Deep Neural Network and Connectionist Temporal Classifier, 2017. URL: https://arxiv.org/abs/1709.03665 (accessed: 02.04.2024).
[5] Jose C., Mishchenko Yu., Senechal T. et al. Accurate Detection of Wake Word Start and End Using a CNN. Interspeech. 2020. DOI: 10.21437/Interspeech.2020-1491.
[6] Bluche T., Primet M., Gisselbrecht T. Small-Footprint Open-Vocabulary Keyword Spotting with Quantized LSTM Networks, 2020. URL: https://arxiv.org/abs/2002.10851 (accessed: 02.04.2024).
[7] Muda L., Begam M., Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, 2010. URL: https://arxiv.org/abs/1003.4083 (accessed: 02.04.2024).
[8] Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, 2018. URL: https://arxiv.org/abs/1804.03209 (accessed: 02.04.2024).
[9] Yundong Z., Suda N., Lai L., Chandra V. Hello Edge: Keyword Spotting on Microcontrollers, 2017. URL: https://arxiv.org/abs/1711.07128 (accessed: 02.04.2024).
[10] He K., Zhang X., et al. Deep Residual Learning for Image Recognition.2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016, p. 770–778. DOI: 10.1109/CVPR.2016.90.
[11] Hochreiter S., Schmidhuber J. Long short-term memory. Neural computation. 1997, v. 9(8), p. 1735–1780. DOI:10.1162/neco.1997.9.8.1735.
[12] Amodei D., Anubha R., et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, 2015. URL: https://arxiv.org/abs/1512.02595 (accessed: 02.04.2024).
[13] Kriman S., Beliaev S., Ginsburg B., et al. QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, 2019. URL https://arxiv.org/abs/1910.10261 (accessed: 02.04.2024).
[14] López F., Luque J., et al. Robust Wake-Up Word Detection by Two-stage Multi-resolution Ensembles, 2013. URL: https://arxiv.org/abs/2310.11379 (accessed: 02.04.2024).

Поступила в редакцию – 02 апреля 2024 г. Окончательный вариант – 24 мая 2024 г.
Received – April 02, 2024. The final version – May 24, 2024.