ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ

УДК 519.7 Обзорная статья
DOI: 10.35330/1991-6639-2022-6-110-127-133

Обзор актуальных открытых решений в области распознавания речи

К. В. Налчаджи

Кабардино-Балкарский научный центр Российской академии наук
360010, Россия, г. Нальчик, ул. Балкарова, 2

Аннотация. Целью данной работы является обзор наиболее успешных открытых решений в области распознавания речи, рассмотрены также процессы распознавания речи и возможности их практического использования. Представлены классические решения, основывающиеся на рекуррентных нейронных сетях, и более современные, которые используют за основу сверточные нейронные сети для удаления шумов и снижения размерности, а также трансформеры, позволяющие запоминать контекст и работать с семантическим смыслом последовательностей вне зависимости от времени.

Ключевые слова: искусственный интеллект, распознавание речи, нейронные сети, обработка естественного языка, сверточные нейронные сети, рекуррентные нейронные сети, трансформеры

Поступила 07.12.2022, одобрена после рецензирования 09.12.2022, принята к публикации 13.12.2022

Для цитирования. Налчаджи К. В. Обзор актуальных открытых решений в области распознавания речи // Известия Кабардино-Балкарского научного центра РАН. 2022. № 6(110). С. 127–133. DOI: 10.35330/1991-6639-2022-6-110-127-133

MSC: 68T50 Review article

Overview of current open solutions in the field of speech recognition

K.V. Nalchadzhi

Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360010, Russia, Nalchik, 2 Balkarov street

Abstract. The purpose of this work is to review the most successful open solutions in the field of speech recognition and also considers the processes of speech recognition and the possibilities of their practical use. The paper presents classical solutions based on recurrent neural networks, as well as more modern ones, which use convolutional neural networks as a basis to remove noise and reduce dimensionality, and transformers that allow to memorize the context and work with the semantic meaning of sequences, regardless of time.

Keywords: artificial intelligence, speech recognition, neural networks, natural language processing, convolutional neural networks, recurrent neural networks, transformers

Submitted 07.12.2022, approved after reviewing 09.12.2022, accepted for publication 13.12.2022

For citation. Nalchadzhi K.V. Overview of current open solutions in the field of speech recognition. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2022. No. 6(110). Pp. 127–133. DOI: 10.35330/1991-6639-2022-6-110-127-133

Введение

В настоящее время все большее количество задач подвергается автоматизации. Не последнюю роль тут играет искусственный интеллект, в частности нейронные сети. Были автоматизированы многие задачи в видеонаблюдении, обработке естественного языка. И несмотря на то, что были достигнуты значимые результаты, в настоящее время исследования продолжаются, в особенности в сфере обработки естественного языка. Проблема обработки человеческой речи составляет важную часть области искусственного интеллекта, и ей придается особое значение.

Цели и задачи исследования. Целью данной работы является обзор наиболее успешных открытых решений в области распознавания речи. Решение данной проблемы позволяет реализовать следующие технологии: «умный дом», голосовое управление для людей с ограниченными возможностями, управление транспортом, генерация субтитров, построение автоответчиков и т.д. Новизна обзора заключается в том, что на данный момент очень мало релевантных статей, в которых рассматриваются современные решения задачи распознавания речи, несмотря на то, что архитектур существует не так много по сравнению с областью компьютерного зрения [1]. Также очень мало существующих обзоров рассматривают возможность адаптации к другим языкам, кроме английского и китайского.

Методы исследования. В процессе достижения цели использован монографический метод, методы анализа и синтеза.

Результаты исследования. В процессе достижения цели исследования были рассмотрены передовые архитектуры нейронных сетей, используемых в процессах распознавания речи. Рассмотрим их более подробно.

Архитектура Deep Speech

На рисунке 1 представлена архитектура искусственной нейронной сети Deep Speech [2].

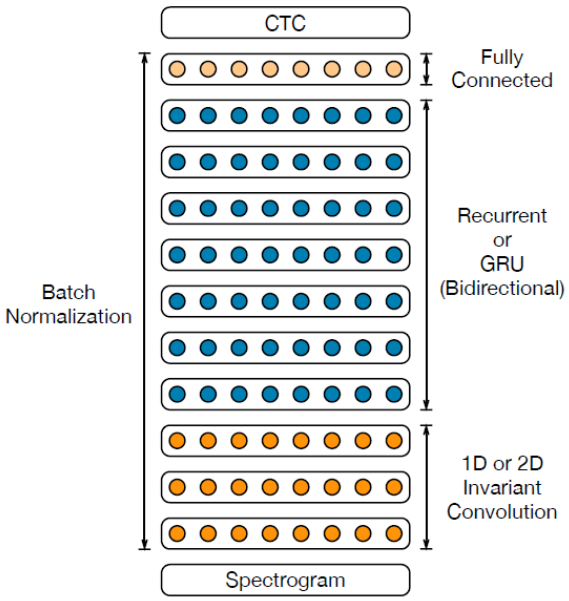

Рис. 1. *Архитектура Deep Speech*
Fig. 1. *Deep Speech Architecture*

На вход нейронная сеть получает спектрограмму звукового сигнала. На первом этапе данные обрабатываются сверточными слоями для кратного понижения длины последовательности [3]. Этот этап позволяет последующим рекуррентным слоям в меньшей мере испытывать проблему «взрывающихся» и «затухающих» градиентов, которая связана с длиной входящей последовательности [4]. На втором этапе преобразованные данные

обрабатываются рекуррентными слоями. Особенность рекуррентных слоев заключается в способности находить закономерности во временных рядах, каковой является человеческая речь. В этой группе слоев наибольшее количество обучаемых параметров, поэтому они являются «узким местом» с точки зрения вычислительной сложности. Далее данные поступают в полносвязный слой, который предназначен для агрегации полученной информации и ее преобразования для получения меньшой размерности [5].

После обработки последовательности на полносвязном слое вычисляется ошибка Connectionist Temporal Classification, которая отражает степень сходства выхода нейронной сети и предсказываемой строки. После подсчета ошибки вычисляются градиенты нейронной сети методом обратного распространения ошибки. Для оптимизации нейронной сети используется алгоритм стохастического градиентного спуска (Stochastic Gradient Descent) [6]. По достижению заданного количества итераций оптимизация нейронной сети завершается. Нейронная сеть с полученными параметрами является искомым алгоритмом распознавания [7]. К преимуществам данной архитектуры относятся легкость обучения и адаптация к другим языкам. К недостаткам – огромное количество данных и GPU часов для обучения, а также слабая обобщающая способность на разных доменах данных.

Архитектура Deep Speech 2

На рисунке 2 представлена архитектура искусственной нейронной сети Deep Speech 2 [8].

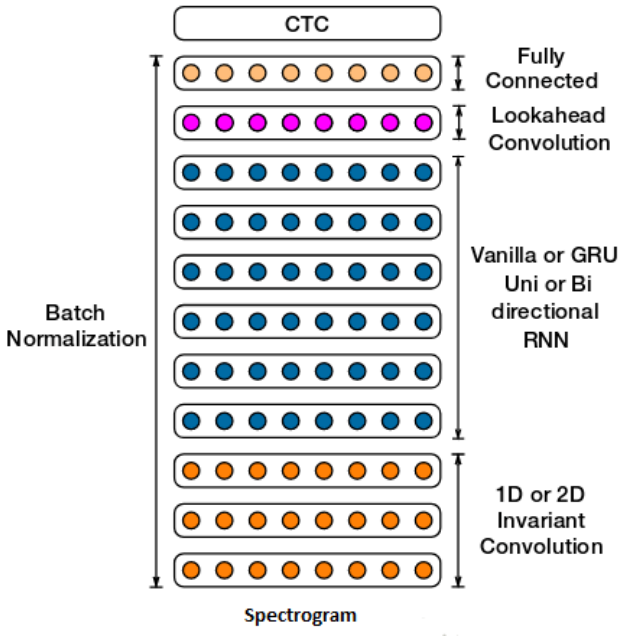

Рис. 2. *Архитектура Deep Speech 2*

Fig. 2. *Deep Speech 2 architecture*

Данная архитектура наследует идеи оригинальной Deep Speech. Первый слой представляет собой 1D или 2D сверточную нейронную сеть, который затем подключается к Bidirectional RNN или Bidirectional GRU. Нововведением является добавление Lookahead свертки, которая обрабатывает два выхода bidirectional RNN. В модели также используется пакетная нормализация для каждого слоя, чтобы уменьшить разрыв в распределении между входом и выходом, повысить способность модели к обобщению и ускорить обучение [9]. На выходе вычисляется функция потери Connectionist Temporal Classification, а на этапе инференса на выходе используется жадный CTC декодер или алгоритм Beam Search. Также архитектура была реализована на фреймворке PaddlePaddle – открытом проекте для исследователей. Deep Speech 2 унаследовала преимущества от Deep Speech, минимизировав

недостатки. Основной проблемой данной архитектуры является лишь использование двусторонних рекуррентных слоев, что многократно снижает скорость работы, особенно в условиях сложного контекста.

Архитектура Wav2Letter

На рисунке 3 представлена архитектура искусственной нейронной сети Wav2Letter [10].

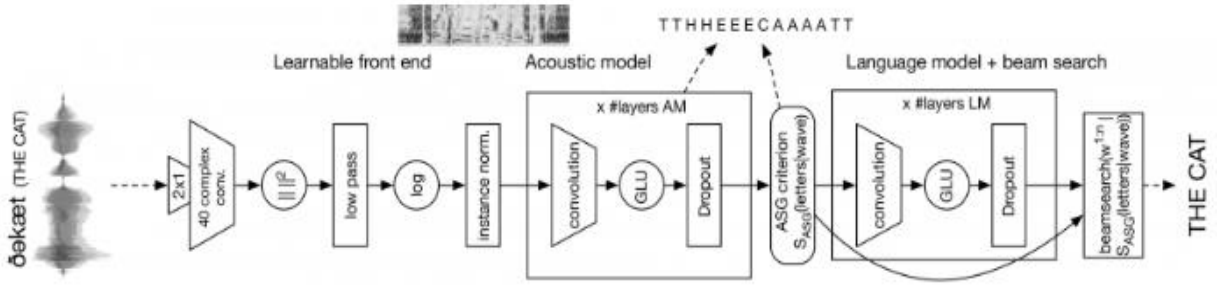

Рис. 3. *Архитектура Wav2Letter*

Fig. 3. *Wav2Letter architecture*

В отличие от предыдущих моделей Wav2Letter использует сверточные слои вместо рекуррентных, поскольку те требуют больших объемов обучающих данных и несоизмеримую вычислительную мощность, что чаще всего доступно только огромным корпорациям. При прямом распространении модель обрабатывает аудиопоток и извлекает его ключевые признаки. Далее следует сверточная акустическая модель, которая пробует прогнозировать буквы. Затем применяется внешняя языковая модель для определения слов и генерации транскрипции. В конце декодирующая сеть генерирует последовательности слов с учетом данных, полученных от акустической модели [11]. Wav2Letter является самой быстрой архитектурой в обзоре, в частности из-за полной реализации на C++ (что означает почти полное отсутствие задержки при обработке данных), однако данное решение тяжело масштабируется и принимает различные изменения [12]. Также для обучения не нужно много данных по сравнению с Deep Speech и Deep Speech 2.

Архитектура FairSeq

На рисунке 4 представлена архитектура искусственной нейронной сети FairSeq [13].

FairSeq является прорывной архитектурой, поскольку использует слои-трансформеры вместо обычных сверток. Эта нейросетевая архитектура избавляется от рекуррентности, то есть от последовательных вычислений. Более нет необходимости ждать, пока закончит работу прежний шаг программы, а проводить подсчеты параллельно, нейронная сеть станет работать быстрее. Данные в трансформере идут по укороченному пути по сравнению с рекуррентной архитектурой. Все благодаря механизму внимания Attention – он фокусируется на отдаленных, но важных словах и отдает их напрямую в обработку. В результате у нейронной сети улучшается долгосрочная память [14]. Изначально трансформеры разрабатывались для обработки текстов, но позже их адаптировали к любым последовательностям и даже к изображениям [15]. К преимуществам этой архитектуры относят повышенную точность распознавания (в среднем Word Error Rate ниже на 5–10 % по сравнению с Wav2Letter или Deep Speech) и небольшое количество данных для обучения. Недостатками являются обязательное наличие большого количества мощностей, медленная сходимость и тяжелая адаптация к другим языкам [16].

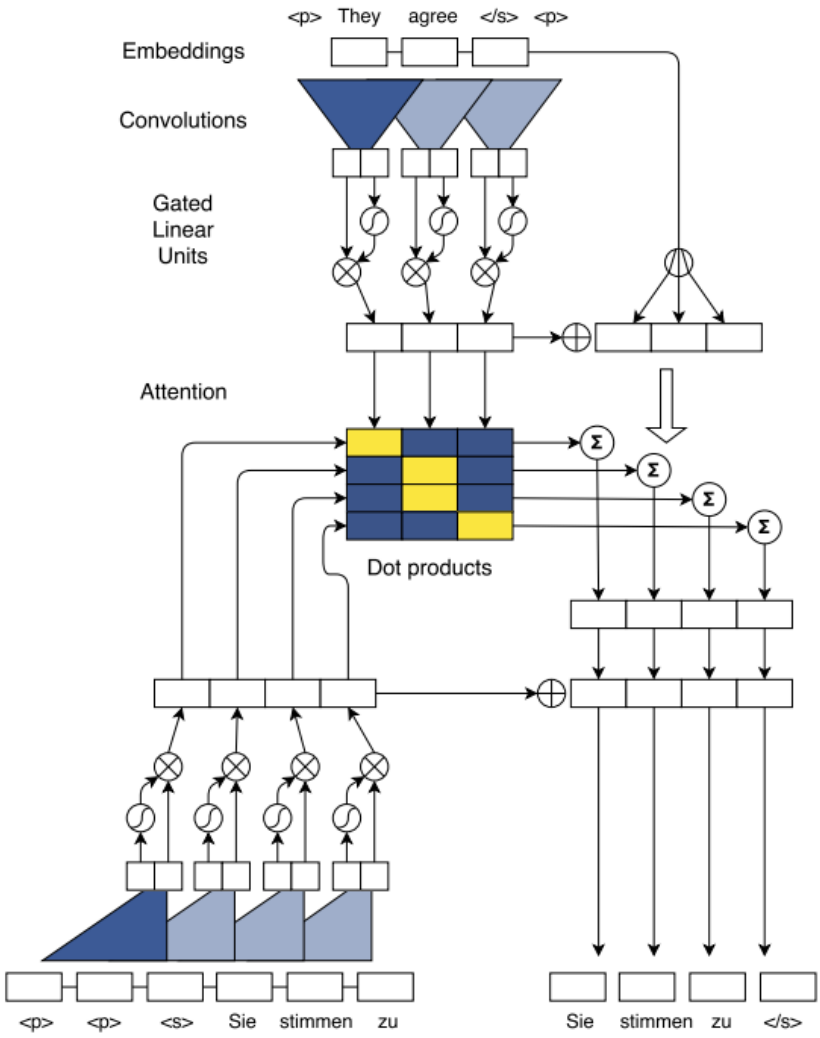

Рис. 4. *Архитектура FairSeq*
Fig. 4. *FairSeq architecture*

Архитектура Wav2Vec

На рисунке 5 представлена архитектура искусственной нейронной сети Wav2Vec [17].

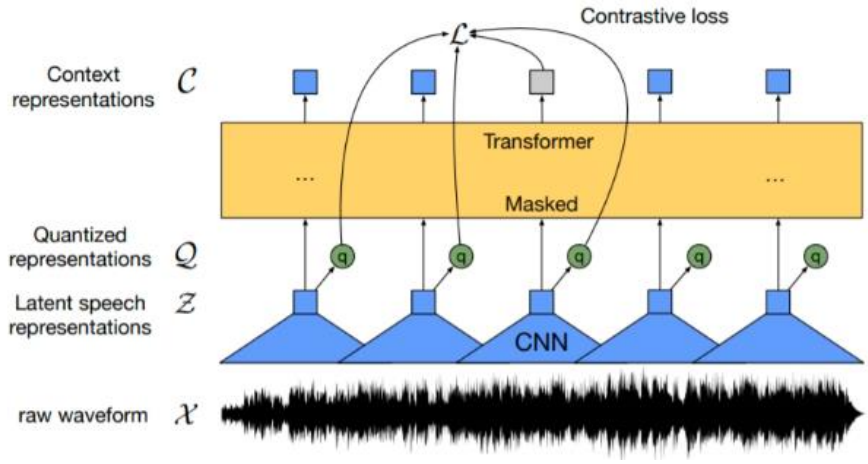

Рис. 5. *Архитектура Wav2Vec*
Fig. 5. *Wav2Vec architecture*

Wav2Vec вышла в то же время, что и FairSeq, и тоже использует трансформеры для работы с контекстом, но благодаря использованию контрастной функции ошибки может работать с неразмеченными данными, что значительно облегчает процесс обучения. Авторы статьи утверждают, что модели не нужны транскрибированные данные и при обучении нужно только загрузить образцы речи и некоторый текст на нужном языке. Система сама распознает слова и фразы и соотнесет их со словарем [18]. Также для обучения оригинальной модели используется состязательная сеть для генерации фонемы, соответствующей звуку на языке. Преимуществом архитектуры является внедрение обучения без учителя, так как неразмеченных данных намного больше и собрать нужный набор с речью становится гораздо легче [19]. К недостаткам относят возросшую сложность обучения из-за наличия дополнительной состязательной сети и сниженную скорость обучения, так как состязательная сеть долго сходится. Без генератора транскрипции будут хуже, и процесс схождения займет еще больше времени.

Заключение

В статье были рассмотрены пять архитектур, которые применяются в задаче распознавания речи. Проведенное исследование позволило выявить перспективные модели, позволяющие решать задачи распознавания речи независимо от языка наиболее эффективно. Обзор и анализ архитектур проводился с целью выявить самый эффективный энкодер-декодер речи по соотношению «скорость/качество». Каждая модель имеет свой ряд преимуществ и недостатков, поэтому решение об применении той или иной модели зависит от требований по точности и скорости распознавания искомого языка, а также наличия возможности работы на разных доменах [20]. Подходящим решением без оглядки на сложность обучения является трансформер Wav2Vec, так как данная модель не требует размеченных данных и легко адаптируется к любым языкам (например, суахили, который не имеет качественного словаря). Самым простым вариантом является Deep Speech, но данная архитектура, как и ее вторая версия, требует большого объема разнообразных размеченных данных, их сложно собрать и разметить без ошибок, которые могут сильно повлиять на итоговый результат. Так как были рассмотрены не все решения, в дальнейшем исследования будут затрагивать и другие архитектуры, например, EspNet, который стремительно набирает популярность в различных подзадачах Speech Recognition [21]. Также необходимо провести исследования для изучения возможности запуска подобных систем в реальном времени на различных устройствах, поскольку только так станут реализуемыми глобальные цели задачи распознавания речи.

REFERENCES

1. Hemant Yadav, Sunayana Sitaram [et al]. A Survey of Multilingual Models For Automatic Speech Recognition. 2022. URL: https://arxiv.org/abs/2202.12576

2. Awni Hannun, Carl Case, Jared Casper [et al]. Deep Deep Speech: Scaling up end-to-end speech recognition. 2007. URL: https://arxiv.org/abs/1412.5567v2

3. Roger Grosse, Helen Kwong, Andrew Y. Ng [et al]. Shift-Invariance Sparse Coding for Audio Classification. 2012. URL: https://arxiv.org/abs/1206.5241

4. Awni Y. Hannun, Daniel Jurafsky, Andrew Y. Ng [et al]. First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs. 2014. URL: https://arxiv.org/abs/1408.2873

5. Anmol Gulati, James Qin, Chung-Cheng Chiu [et al]. Conformer: Convolution-augmented Transformer for Speech Recognition. 2020. URL: https://arxiv.org/abs/2005.08100

6. Andrew L. Maas, Peng Qi, Ziang Xie [et al]. Building DNN Acoustic Models for Large Vocabulary Speech Recognition. 2014. URL: https://arxiv.org/abs/1406.7806

7. Kaitao Song, Xu Tan, Di He, Jianfeng Lu [et al]. Double Path Networks for Sequence to Sequence Learning. *In Proceedings of the 27th International Conference on Computational Linguistics*, 2018. Pp. 3064–3074. URL: https://arxiv.org/abs/1806.04856

8. Dario Amodei, Rishita Anubhai, Eric Battenberg [et al]. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. 2015. URL: https://arxiv.org/abs/1512.02595

9. Tianxiao Shen, Myle Ott, Michael Auli [et al]. Mixture Models for Diverse Machine Translation: Tricks of the Trade. 2019. URL: https://arxiv.org/abs/1902.07816

10. Neil Zeghidour, Qiantong Xu, Vitaliy Liptchinsky [et al] Wav2Letter++: Fully Convolutional Speech Recognition. 2018. URL: https://arxiv.org/abs/1812.06864

11. Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz [et al]. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. 2017. URL: https://arxiv.org/abs/1701.06538

12. Shashi Narayan, Shay B. Cohen, Mirella Lapata [et al]. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. 2018. URL: https://arxiv.org/abs/1808.08745

13. Myle Ott, Sergey Edunov, Alexei Baevski [et al]. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. 2019. URL: https://arxiv.org/abs/1904.01038

14. Stephen Merity, Nitish Shirish Keskar, Richard Socher [et al]. An Analysis of Neural Language Modeling at Multiple Scales. 2018. URL: https://arxiv.org/abs/1803.08240

15. Sebastian Gehrmann, Yuntian Deng, Alexander M. Rush [et al]. Bottom Up Abstractive Summarization. 2018. URL: https://arxiv.org/abs/1808.10792

16. Shamil Chollampatt, Hwee Tou Ng. A Multilayer Convolutional Encoder Decoder Neural Network for Grammatical Error Correction. 2018. URL: https://arxiv.org/abs/1801.08831

17. Steffen Schneider, Alexei Baevski, Ronan Collobert [et al]. wav2vec: Unsupervised Pre-training for Speech Recognition. 2019. URL: https://arxiv.org/abs/1904.05862

18. Gabriel Synnaeve, Qiantong Xu, Jacob Kahn [et al]. End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures. 2019. URL: https://arxiv.org/abs/1911.08460

19. Jayadev Billa. Improving low-resource ASR performance with untranscribed out-of-domain data. 2021. URL: https://arxiv.org/abs/2106.01227

20. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai [et al]. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 2021. URL: https://arxiv.org/abs/2106.07447

21. Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman [et al]. Towards Building ASR Systems for the Next Billion Users. 2021. URL: https://arxiv.org/abs/2111.03945

Информация об авторе

Налчаджи Карен Витальевич, аспирант Научно-образовательного центра, Кабардино-Балкарский научный центр РАН;
360010, Россия, г. Нальчик, ул. Балкарова, 2;
nalkar07@yandex.ru

Information about the author

Nalchadzhi Karen Vitalyevich, Postgraduate student of the Scientific and Educational Center, Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;
360010, Russia, Nalchik, 2 Balkarova street;
nalkar07@yandex.ru