ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

DOI: 10.18137/RNU.V9187.24.04.P.71 УДК 004.93

М.М. Баруздин, М.В. Раскатова, П. Щёголев

РАЗВИТИЕ СОВРЕМЕННЫХ СИСТЕМ ТРАНСКРИБАЦИИ АУДИО- И ВИДЕОКОНТЕНТА

Аннотация. В статье проведен анализ существующих проблем транксрибации. Рассмотрены актуальные технологии, использующиеся в данных системах. Подробно рассмотрены современные open-source-решения и изучены их возможности в решении описанных проблем транскрибации. Описаны четыре наиболее популярные открытые платформы: Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0. В статье проведено сравнение архитектур и особенностей данных моделей, что дает представление об их возможностях и ограничениях. Показано, как модели справляются с проблемами, стоящими перед системами автоматического распознания речи. Выбор модели для автоматического распознания речи зависит от конкретных задач и условий использования.

Ключевые слова: транскрибация, системы распознавания речи, глубокие нейронные сети, Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0.

M.M. Baruzdin, M.V. Raskatova, P. Shchegolev

DEVELOPMENT OF MODERN AUDIO AND VIDEO CONTENT TRANSCRIPTION SYSTEMS

Abstract. The article focuses on the existing problems in transcription. Current technologies used in transcription systems are reviewed. Modern open-source solutions are examined in detail, and their capabilities in addressing the described transcription challenges are explored. The four most popular open-source platforms are described: Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0. Comparing the architecture and features of the models gives an idea of their capabilities and limitations. The article shows how models cope with problems faced by automatic speech recognition systems. The choice of the automatic speech recognition model depends on the specific tasks and conditions.

Keywords: transcription, speech recognition systems, deep neural networks, Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0.

Введение

Современные технологии транскрибации аудио- и видеоконтента в текст становятся неотъемлемой частью цифровой медиаиндустрии. Автоматическая транскрибация значительно ускоряет и упрощает обработку больших объемов данных, позволяя компаниям оперативно создавать текстовые версии аудиоконтента, будь то новостные выпуски, подкасты, интервью или видеоролики.

Однако, несмотря на прогресс, автоматические системы транскрибации все еще сталкиваются с рядом проблем, которые ограничивают их точность и надежность. Одна из ключевых – фоновые шумы и некачественные записи. В реальных условиях аудиозаписи часто содержат разнообразные шумы, как то: звуки транспорта, разговоры других людей, ветер или просто фоновую музыку [1].

Баруздин Михаил Михайлович
магистрант, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: машинное обучение, искусственные нейронные сети, системы распознавания речи. ORCID: 0009-0002-8024-1391.
Электронный адрес: mmb222v@gmail.com

Раскатова Марина Викторовна
кандидат технических наук, доцент кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: разработка программного обеспечения, информационные системы. Автор более 40 опубликованных научных работ. ORCID: 0000-0001-7671-3312, SPIN-код: 8053-5041, Author ID: 609945.
Электронный адрес: marina@raskatova.ru

Щёголев Павел
старший преподаватель кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: языки и методы программирования, web-разработка. Автор семи опубликованных научных работ. ORCID: 0000-0001-9954-8858, SPIN-код: 6914-1637, Author ID: 1246900.
Электронный адрес: Shchegolevsp@mpei.ru

Еще одной проблемой является вариативность акцентов и диалектов. Большинство систем распознавания речи обучается на стандартных версиях языка, но реальная речь может сильно отличаться в зависимости от региона, акцента или даже личных особенностей произношения. Еще одно значительное препятствие – многоголосие. В случаях, когда говорят сразу несколько человек, как в интервью, подкастах или групповых обсуждениях, транскрибационные системы могут не справляться с разделением голосов, воспринимая их как один общий поток.

Недостаточная адаптация к специализированной лексике и терминологии также остается проблемой для большинства современных систем. Сложные термины и профессиональная лексика, которые часто встречаются в юридических, медицинских, научных и технических аудиоматериалах, могут быть некорректно распознаны, что приводит к искажению содержания либо к транскрибированию термина на другой язык. Недостаточное понимание контекста и многозначных слов – еще одна важная проблема. Большинство систем транскрибации пока не способны эффективно обрабатывать слова, которые могут иметь разное значение в зависимости от контекста, интонации или расположения в предложении.

Для преодоления этих трудностей активно развиваются современные алгоритмы и архитектуры, которые позволяют обрабатывать речь с большей точностью и адаптивностью. Благодаря использованию инновационных подходов, таких как глубокое обучение и гибридные модели, становится возможным не только повышать качество распознавания речи, но и оптимизировать системы для специфических задач и условий.

В данной статье рассмотрены актуальные технологии и алгоритмы транскрибации, а также дан сравнительный анализ доступных решений.

Технологии и алгоритмы распознавания речи

Технологии распознавания речи, или Speech-to-Text (далее – STT), эволюционировали на протяжении нескольких десятилетий. На ранних стадиях разработки STT-системы полагались на фонемные и акустические модели, которые были созданы с использованием скрытых марковских моделей (далее – HMM) и имели ограниченные возможности распознавания. Ключевая идея заключалась в том, чтобы представлять речь как последовательность фонем – минимальных звуковых единиц, которые составляют слова. Системы анализа сравнивали фонемы с образцами в своей базе данных и определяли наиболее вероятное сочетание, соответствующее произнесенному слову [2]. Однако этот подход оказался чувствителен к качеству записи и не справлялся с проблемами шумов, акцентов и многоголосия.

Современные системы транскрибации во многом обязаны своим прогрессом развитию архитектур глубоких нейронных сетей (Deep Neural Networks, DNN), которые смогли решить многие проблемы, связанные с традиционными методами. Одной из первых нейросетевых архитектур, применяемых для анализа последовательных данных, стали рекуррентные нейронные сети (далее – RNN). В отличие от обычных нейронных сетей RNN обладают способностью «запоминать» информацию предыдущих шагов, и это делает их подходящими для обработки речи, где текущий звуковой сигнал зависит от предыдущих [3]. Классические RNN, однако, имеют свои ограничения: со временем они «забывают» более ранние элементы последовательности и не могут точно обрабатывать длинные фрагменты речи, что снижает их эффективность при анализе сложных предложений и разговоров. Для решения данной проблемы были разработаны улучшенные версии RNN, такие как Long Short-Term Memory (далее – LSTM) и Gated Recurrent Unit (далее – GRU). Эти архитектуры используют специальные механизмы – так называемые вентили (gates), которые контролируют, какая информация передается на следующую итерацию, а какая удаляется [4]. Это позволяет моделям сохранять долгосрочные зависимости и обрабатывать более длинные последовательности, что необходимо для транскрибации сложных аудиофайлов. Версии LSTM и GRU показали себя достаточно эффективными в ранних системах транскрибации и до сих пор применяются в некоторых решениях, требующих детального анализа временных последовательностей.

Сверточные нейронные сети (Convolutional Neural Networks, далее – CNN) традиционно используются для обработки изображений, а также в системах транскрибации. В контексте обработки звука CNN применяются для выделения ключевых признаков в спектрограммах – визуальных представлениях звука. Сверточные слои позволяют модели выявлять паттерны в звуковых сигналах и эффективно фильтровать шум, что особенно важно для обработки записей с фоновыми звуками [5]. Например, многослойные CNN могут использоваться для предварительной обработки аудиосигналов, позволяя извлекать из них спектральные признаки, которые затем передаются в другие архитектуры, такие как RNN или трансформеры, для более детального анализа и расшифровки текста. Этот подход позволяет значительно повысить точность транскрибации в условиях шумного окружения.

Наиболее революционным достижением в области распознавания речи стали *трансформеры* – архитектуры, которые кардинально изменили подход к обработке последовательных данных. В отличие от RNN, которые обрабатывают информацию последовательно, трансформеры могут анализировать все элементы последовательности одновремен-

но, используя механизмы внимания (attention) [6]. Это позволяет моделям трансформеров учитывать как ближайший, так и более отдаленный контекст, что критически важно для корректного распознавания речи в длинных аудиофайлах.

С развитием новых технологий в области распознавания речи и машинного обучения, а также с ростом интереса к автоматическому транскрибированию появляются новые open-source-платформы, предлагающие разработчикам эффективные инструменты для работы с аудиоданными. Эти решения не только облегчают процесс транскрибирования, но и позволяют адаптировать алгоритмы под специфические нужды пользователей и различные языковые контексты. В этом плане рассмотрим четыре наиболее популярные открытые платформы: Kaldi, Mozilla Deep Speech, Whisper, Wav2Vec 2.0.

Kaldi – это фреймворк для автоматического распознавания речи (ASR), разработанный с акцентом на гибкость и эффективность. Основная архитектура Kaldi построена вокруг гибридной модели HMM-DNN (Hidden Markov Model – Deep Neural Network). Модель HMM используется для моделирования последовательностей и временных зависимостей речи, DNN – для обработки ее акустических особенностей [7]. Современные реализации Kaldi также поддерживают TDNN (Time-Delay Neural Networks) и архитектуры на основе RNN, что позволяет лучше обрабатывать зашумленные аудиофайлы и улучшает качество транскрибации записей с акцентом, распознавать слова с растянутыми или смещенными фонемами. Эти архитектуры усиливают модель, позволяя эффективно анализировать длинные временные зависимости в аудиопотоке.

Kaldi широко использует такие алгоритмы, как GMM-HMM (Gaussian Mixture Model – Hidden Markov Model) для начальной сегментации данных[1], извлечения акустических признаков MFCC (Mel-Frequency Cepstral Coefficients) или PLP (Perceptual Linear Prediction). Для работы с языковыми моделями Kaldi интегрируется с инструментами для построения n-грамм-моделей, а также с современными методами на основе трансформеров. Гибкая обработка данных, высокая точность и богатый инструментарий делают Kaldi мощной базой для решения задач ASR.

Следующий инструмент – *Mozilla Deep Speech* – это ASR-система с открытым исходным кодом, вдохновленная архитектурой Deep Speech 2, предложенной Baidu. Deep Speech построен на базе архитектуры RNN с компонентами LSTM или GRU, что делает его способным к обучению на длинных последовательностях данных. Для входного аудиосигнала сначала извлекаются акустические признаки MFCC, которые затем передаются в сеть для обработки. Сеть состоит из 5 скрытых слоев, из которых только 4-й слой является рекуррентным[2]. Полная модель представлена на Рисунке 1.

Важной особенностью Deep Speech является использование технологии CTC (Connectionist Temporal Classification) для обработки последовательностей с неопределенными выравниваниями между входными аудиоданными и текстовыми метками. Это упрощает процесс обучения модели, исключая необходимость в предварительном сегментировании данных. Для языковых моделей Deep Speech поддерживает интеграцию с моделями на основе n-грамм или нейронных языковых моделей, что позволяет улучшать распознавание речи в зависимости от контекста.

---

[1] Kaldi toolkit // Kaldi. URL: https://kaldi-asr.org/doc/index.html (дата обращения: 03.10.2024)

[2] Welcome to DeepSpeech's documentation! // Mozilla Deep Speech. URL: https://deepspeech.readthedocs.io/en/r0.9/ (дата обращения: 30.09.2024).

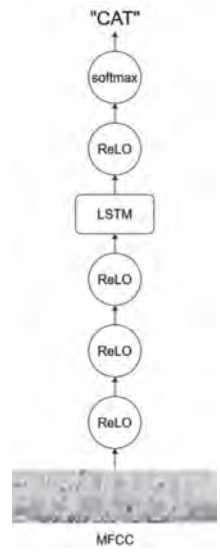

Рисунок 1. Архитектура Mozilla Deep Speech
Источник: Mozilla Deep Speech. URL: https://deepspeech.readthedocs.io/en/r0.9

Whisper – это мощная система автоматического распознавания речи (ASR), разработанная Open AI, которая основана на сквозной архитектуре кодер-декодер с использованием трансформеров. Обученная на более чем 680 000 часов многоязычных данных, система показывает высокую устойчивость к акцентам, шуму и сложным техническим текстам. Whisper является универсальной моделью, объединяющей функции распознавания речи, перевода, идентификации языка и маркировки временных меток. Ключевой компонент архитектуры Whisper – трансформер, используемый как в кодировщике, так и декодировщике. На вход подаются аудиосигналы, разбитые на фрагменты длиной 30 секунд, которые конвертируются в спектрограммы log-Mel. Кодировщик преобразует эти спектрограммы в представления высокой размерности, которые затем обрабатываются декодировщиком для генерации текстовых токенов. Эти токены могут содержать как распознанную речь, так и метаданные, такие как языковая идентификация и временные метки[3].

Модель обучена в формате многозадачного обучения (MTL), что позволяет ей решать несколько задач одновременно. Специальные токены направляют модель к выполнению конкретной задачи, например, транскрипции речи, перевода на английский или генерации временных меток. Такой подход заменяет сложные конвейеры, традиционно используемые в обработке речи, одной универсальной моделью. Архитектура модели представлена на Рисунке 2.

[3] Introducing Whisper // Open AI. 2022. September 21. URL: https://openai.com/index/whisper/ (дата обращения: 05.10.2024).

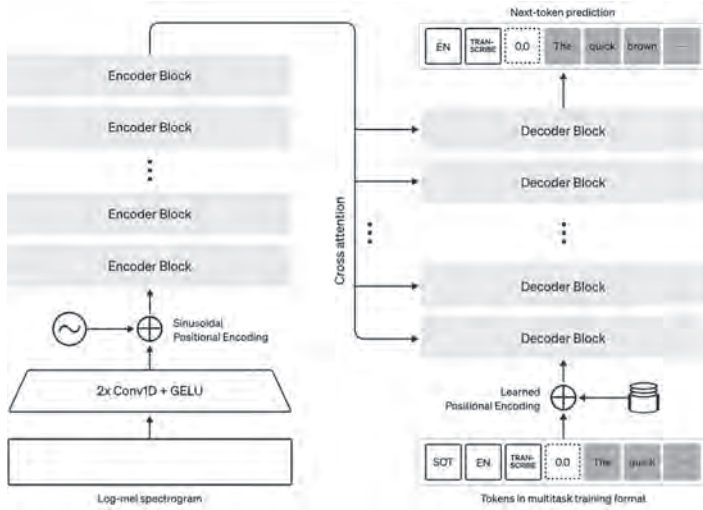

Рисунок 2. Архитектура модели Whisper
Источник: [8].

Архитектура *Wav2vec 2.0* – передовая система ASR, разработанная Facebook AI. Она позволяет моделям учиться на «сырых» аудиоданных, минимизируя зависимость от вручную размеченных наборов данных. Система представляет собой одно из первых успешных внедрений метода обучения с самонаблюдением (Self-Supervised Learning) в задачи распознавания речи. Архитектура основана на трансформерах, которые применяются для извлечения скрытых представлений из непрерывного аудиопотока. На первом этапе модель обучается восстанавливать «замаскированные» фрагменты аудиосигнала, что позволяет ей понимать структуру данных и извлекать полезные признаки. На втором этапе уже обученная модель дообучается с помощью небольшой размеченной выборки данных для выполнения задач ASR [9]. Wav2vec 2.0 также отличается использованием кластеризации признаков для представления аудиосигналов в виде дискретных токенов, аналогичных текстовым, что делает процесс обработки речи более эффективным и позволяет объединять преимущества обработки текстовых и звуковых данных.

Сравнение архитектур и особенностей моделей дает представление об их возможностях и ограничениях. Благодаря этому можно рассмотреть, как модели справляются с проблемами, стоящими перед системами автоматического распознания речи.

Kaldi за счет гибридной архитектуры и гибкой настройки хорошо работает с зашумленными записями, добавляет модули обработки многоголосия и настраивает языковую модель на конкретную задачу, что улучшает понимание контекста. Для эффективной работы с акцентами, диалектами и специализированной лексикой требуется тщательное обучение на узкоспециализированных наборах размеченных данных. Mozilla Deep Speech, использующая RNN с CTC, относительно устойчива к фоновому шуму, но ограничена в работе с многоголосием, так как ее архитектура плохо разделяет параллельные голосовые потоки. Кроме того, слабая языковая модель затрудняет понимание контекста и обработку многозначных слов. Wav2vec 2.0 благодаря обучению с самонаблюдением, может эффективно работать с вариативностью акцентов и в условиях нехватки размеченных данных. При использовании дополнительных алгоритмов диаризации может обрабатывать многоголосие, однако из-за

отсутствия встроенной языковой модели ее возможности в учете контекста и специализированной лексики ограничены. Whisper за счет своей архитектуры трансформера кодер-декодер и предобучения на огромном наборе многоязычных данных хорошо справляется с распознаванием акцентов, диалектов и специализированной лексики. Однако она требует высокой вычислительной мощности, занимает большой объем памяти, а также сильно ограничена в возможности кастомизации, что может добавлять сложности при необходимости дополнительного обучения модели под узкоспециализированную задачу.

Заключение

Подводя итог, можно отметить, что выбор модели для автоматического распознания речи зависит от конкретных задач и условий использования. Kaldi предоставляет гибкость и точность для специализированных приложений, требующих тщательной настройки. Mozilla Deep Speech привлекает своей простотой и доступностью, что делает ее подходящей для базовых сценариев и легких приложений. Wav2vec 2.0 предлагает подход к обучению с самонаблюдением, что позволяет эффективно работать с ограниченными данными, но узкая специализация ограничивает ее универсальность. Whisper демонстрирует высочайшую точность и многофункциональность, хотя ее использование связано с высокими вычислительными требованиями.

Таким образом, каждая из моделей имеет свои сильные и слабые стороны, и их выбор должен быть основан на специфике поставленной задачи. В дальнейших исследованиях ожидается практическое изучение возможностей моделей и их обработки проблем.

Литература

1. *Галунов В.И., Соловьев А.Н.* Современные проблемы в области распознавания речи // Информационные технологии и вычислительные системы. 2004. № 2. С. 41–45. URL: https://www.mathnet.ru/links/e8d1d4e4c39da5a9c7a79f7dcc0549c2/itvs652.pdf (дата обращения: 17.09.2024).

2. *Маковкин К.А.* Гибридные модели – Скрытые марковские модели / Многослойный персептрон – и их применение в системах распознавания речи. Обзор // Речевые технологии. 2012. № 3. С. 58–83. EDN UZERSP.

3. *Cho K., Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., BengioY.* Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, October 2014. P. 1724–1734. DOI: https://doi.org/10.3115/v1/D14-1179

4. *Oruh J., Viriri S., Adegun A.* Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition // IEEE Access. New York, 2022. Vol. 10. P. 30069–30079. DOI: 10.1109/ACCESS.2022.3159339

5. *Abdel-Hamid O., Mohamed A., Jiang H., Deng L., Penn G., Yu D.* Convolutional Neural Networks for Speech Recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014. Vol. 22. P. 1533–1545. DOI: 10.1109/TASLP.2014.2339736

6. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.* Attention Is All You Need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, P. 5998-6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 17.09.2024).

7. *Kipyatkova I., Karpov A.* DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi // Ronzhin A., Potapova R., Németh G. (Eds) Speech and Computer. SPECOM 2016. Lecture

Notes in Computer Science. 2016. Vol. 9811. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-43958-7_29
8. *Radford A., Jong Wook Kim, Tao Xu, Brockman G., McLeavey Ch., Sutskever I.* Robust speech recognition via large-scale weak supervision // Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202. 2023. DOI: 10.48550/arXiv.2212.04356
9. *Baevski A., Zhou H., Mohamed A., Michael A.* Wav2vec 2.0: A framework for self-supervised learning of speech representations // 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. 2022. DOI: 10.48550/arXiv.2006.11477

References

1. Galunov V.B., Soloviev A.N. (2004) Modern problems in the field of speech recognition. *Journal of Information Technologies and Computing Systems.* No. 2. Pp. 41–45. URL: https://www.mathnet.ru/links/e8d1d4e4c39da5a9c7a79f7dcc0549c2/itvs652.pdf (дата обращения: 17.09.2024). (In Russian).
2. Makovkin K.A. (2012) Hybrid Models – Hidden Markov Models / Multilayer Perceptron – and their Application in Speech Recognition Systems. Overview. *Rechevye tekhnologii* [Speech technologies]. Vol. 3. Pp. 58–83. (In Russian).
3. Cho K., Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, October. Pp. 1724–1734. DOI: https://doi.org/10.3115/v1/D14-1179
4. Oruh J., Viriri S., Adegun A. (2022) Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition. In: *IEEE Access*. New York. Vol. 10. Pp. 30069–30079. DOI: 10.1109/ACCESS.2022.3159339
5. Abdel-Hamid O., Mohamed A., Jiang H., Deng L., Penn G., Yu D. (2014) Convolutional Neural Networks for Speech Recognition. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Vol. 22. Pp. 1533–1545. DOI: 10.1109/TASLP.2014.2339736
6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2023) Attention Is All You Need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, Pp. 5998-6008. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed 17.09.2024).
7. Kipyatkova I., Karpov A. (2016). DNN-Based Acoustic Modeling for Russian Speech Recognition Using Kaldi. In: Ronzhin A., Potapova R., Németh G. (Eds) *Speech and Computer. SPECOM 2016. Lecture Notes in Computer Science*. Vol. 9811. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-43958-7_29
8. Radford A., Jong Wook Kim, Tao Xu, Brockman G., McLeavey Ch., Sutskever I. (2023) Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, Hawaii, USA. PMLR 202. DOI: 10.48550/arXiv.2212.04356
9. Baevski A., Zhou H., Mohamed A., Michael A. (2022) Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada. DOI: 10.48550/arXiv.2006.11477