ОБЗОР АЛГОРИТМОВ РАСПОЗНАВАНИЯ РЕЧИ

Миркомил Алишерович Рахманов
ассистент кафедры «Архитектура и цифровые технологии» Ташкентского международного университета финансового менеджмента и технологий.
https://doi.org/10.5281/zenodo.16924639

ANNOTATSIYA

Maqolada nutqni tanib olishning asosiy algoritmlari haqida umumiy ma'lumot berilgan bo‘lib, ularning ishlash tamoyillari, afzalliklari va kamchiliklari tavsiflangan. Klassik yashirin Markov modellari (HMM), chuqur neyron tarmoqlar (DNN), rekurrent (RNN), shuningdek, Transformer va diqqat mexanizmlariga asoslangan zamonaviy arxitekturalar ko‘rib chiqilgan. Turli algoritmlarning aniqligi, hisoblash talablari va shovqinga chidamliligi solishtirilib, ularning samarali qo‘llash sohalari qayd etilgan. Bugungi kunda eng istiqbolli va samarali algoritmlar sifatida Transformer va diqqat mexanizmlariga asoslangan modellar ta’kidlanadi, ular yuqori aniqlik bilan katta hajmdagi audio-ma’lumotlarni qayta ishlashga eng yaxshi moslashgan.

KALIT SO‘ZLAR

nutqni tanib olish, yashirin Markov modellari, chuqur neyron tarmoqlar, rekurrent neyron tarmoqlar, Transformer, diqqat mexanizmi, audio ma'lumotlarni qayta ishlash, neyron tarmoqlar, nutqni tanib olish algoritmlari.

АННОТАЦИЯ

В статье представлен обзор основных алгоритмов распознавания речи, описаны их принципы работы, преимущества и недостатки. Рассмотрены классические скрытые Марковские модели (HMM), глубокие нейронные сети (DNN), рекуррентные (RNN) а также современные архитектуры на основе Transformer и механизмов внимания. Сделан акцент на сравнении точности, вычислительных требований и устойчивости к шумам

КЛЮЧЕВЫЕ СЛОВА

распознавание речи, скрытые Марковские модели, глубокие нейронные сети, рекуррентные нейронные сети, Transformer, механизм внимания, обработка аудио, нейронные сети, алгоритмы распознавания речи.

различных алгоритмов, отмечены области их наиболее эффективного применения. Отмечается, что на сегодняшний день наиболее перспективными и эффективными являются алгоритмы, основанные на механизмах внимания (Transformer), которые обеспечивают высокую точность распознавания и лучше всего справляются с обработкой больших объемов аудио-данных.

ABSTRACT

The article provides an overview of the main speech recognition algorithms, outlining their principles of operation, advantages, and disadvantages. It discusses classical Hidden Markov Models (HMM), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) as well as contemporary architectures based on Transformer and attention mechanisms. The focus is placed on comparing accuracy, computational requirements, and noise robustness across various algorithms, highlighting their optimal application domains. It is noted that algorithms based on attention mechanisms (Transformer) are currently the most promising and effective, providing high recognition accuracy and superior performance in processing large volumes of audio data.

KEYWORDS

speech recognition, Hidden Markov Models, deep neural networks, recurrent neural networks, Transformer, attention mechanism, audio processing, neural networks, speech recognition algorithms.

Введение

В последние годы технологии распознавания речи приобретают всё большую популярность и востребованность в различных областях жизни, включая виртуальных помощников, автоматические системы перевода и интеллектуальные пользовательские интерфейсы. Современные подходы к распознаванию речи демонстрируют значительный прогресс благодаря использованию передовых алгоритмов машинного обучения и глубоких нейронных сетей. Несмотря на широкое распространение данных технологий, задача точного распознавания человеческой речи продолжает оставаться

сложной вследствие наличия множества факторов: шума окружающей среды, вариативности акцентов, индивидуальных особенностей речи и сложности контекста высказываний[1].

Исторически, развитие алгоритмов распознавания речи началось с использования статистических подходов, таких как скрытые Марковские модели (HMM), которые представляли собой надёжные, но ограниченные по точности системы. С развитием вычислительных мощностей и появлением больших массивов обучающих данных стали доступны методы глубокого обучения, прежде всего, глубокие нейронные сети (DNN), существенно улучшившие точность распознавания. Следующим этапом эволюции алгоритмов стало применение рекуррентных нейронных сетей (RNN), которые показали высокую эффективность в обработке последовательных данных и учёте временного контекста. Наряду с ними активно развивались свёрточные нейронные сети (CNN), позволяющие эффективно анализировать акустические признаки речи, представленные в виде спектрограмм и аудиоизображений[2].

Наиболее перспективными на сегодняшний день являются подходы, основанные на механизмах внимания (attention), в частности архитектуры типа Transformer. Эти модели способны учитывать важность отдельных элементов речи и обеспечивать наивысшую точность распознавания, эффективно обрабатывая большие объёмы сложных и зашумлённых аудио-данных[3]. Таким образом, цель данной статьи — рассмотреть и сравнить ключевые алгоритмы распознавания речи, описать их основные принципы работы, преимущества и недостатки, а также выделить перспективы их дальнейшего развития и применения.

Основная часть

Скрытые Марковские модели (HMM) долгое время оставались фундаментальной основой технологий распознавания речи. Они основаны на вероятностном подходе, где речь рассматривается как последовательность скрытых состояний (фонемы или слова), а наблюдаемые данные — это акустические сигналы[4].

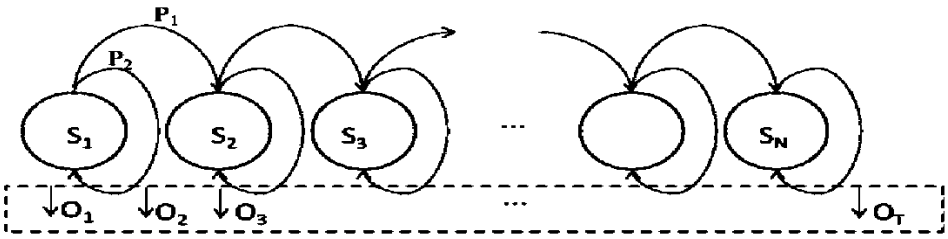
Скрытые Марковские модели описывают двухэтапный стохастический процесс. Первый этап состоит из дискретного стохастического процесса, который является статичным, каузативным и простым. Пространство состояний рассматривается как конечное. Таким образом, процесс вероятностно описывает состояние перехода в дискретность, конечное пространство состояний. Это можно наглядно представить как конечный автомат с перепадами между любыми парами состояний, которые помечены вероятностью перехода. Поведение процесса в данный момент времени t зависит только от непосредственного состояния предшествующего элемента и может быть охарактеризовано следующим образом:

$$P(S_t|S_1, S_2, \dots S_{t-1}) = P(S_t|S_{t-1}) \quad (1)$$

На втором этапе для каждого момента времени t дополнительно, путем вывода или выходных данных, генерируется O_t. Распространение ассоциативной вероятности зависит только от текущего состояния S_t, а не от каких-либо предыдущих состояний или выводных данных.

$$P(O_t|O_1 \dots O_{t-1}, S_1 \dots S_t) = P(O_t|S_t) \quad (2)$$

Эта последовательность выводных данных единственное, что можно наблюдать в поведении модели. С другой стороны, состояние последовательности принятое во время генерации данных не может быть исследована. Это и есть, так называемая, «скрытость», из которой выводится определение скрытых моделей Маркова. Если посмотреть на модель внешне – то есть понаблюдать за ее поведением – довольно часто встречаются ссылки на последовательность выводных состояний O_i, $O_1 \dots O_t$, как на причину наблюдения за последовательностью. Далее отдельные элементы этой последовательности будем называть результатом наблюдения.



1-Рис. Скрытые Марковские модели

Скрытые Марковские модели помогают вычислить, какие состояния могли привести к тем наблюдениям, которые мы видим или слышим. В распознавании речи мы видим (слышим) звуки, а модель угадывает слова или фразы, которые, вероятно, были сказаны.

DNN (Deep Neural Networks) — многослойные сети, способные эффективно моделировать сложные взаимосвязи между входными аудио-сигналами и выходными текстовыми данными. Данный алгоритм представляет собой искусственную нейронную сеть с несколькими скрытыми слоями единиц между входным и выходным уровнями. Подобно мелким нейронным сетям, DNN могут моделировать сложные нелинейные отношения. Архитектуры DNN создают композиционные модели, в которых дополнительные слои позволяют составлять элементы из нижних слоев, обеспечивая огромную учебную способность и, следовательно, потенциал моделирования сложных моделей речевых данных[5]. DNN сеть имеет входной слой x, скрытый слой s и выходной слой y. Входной слой состоит из вектора x(t), который является объединением вектора w(t), представляющим собой текущее слово, и вектора s(t – 1), который представляет собой выходные значения скрытого слоя, полученные на предыдущем шаге. Размер вектора w(t) равен размеру словаря. Выходной слой y(t) имеет тот же размер, что и w(t), и после изучения нейронной сети представляет собой вероятностное

распределение следующего слова при данном предыдущем слове и состоянии скрытого слоя в предшествующий временной шаг. Размер скрытого слоя обычно выбирается эмпирически. Все слои можно вычислить следующим образом:

$$s(t) = w(t) + s(t-1) \quad (3)$$

$$s_j(t) = f(\sum x_i(t)u_{ji}) \quad (4)$$

$$y_k(t) = G(\sum g_i(t)u_{ki}) \quad (5)$$

где f(z) – сигмоидальная активационная функция, g(z) – функция softmax.

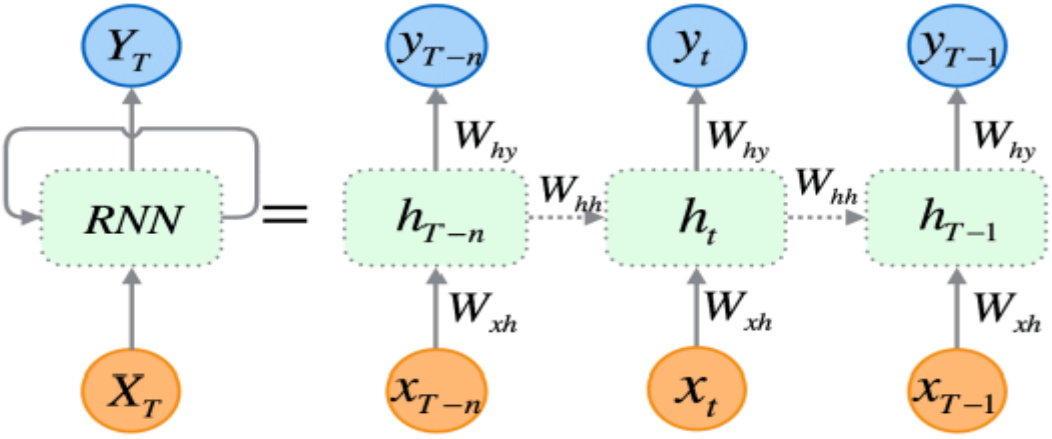
DNN произвели революцию в распознавании речи, значительно повысив точность и устойчивость систем. Они легли в основу современных гибридных и end-to-end архитектур и до сих пор используются в реальных приложениях благодаря универсальности и гибкости.

RNN (Recurrent Neural Networks) — специализированы на обработке последовательностей данных. Могут учитывать контекст предыдущих элементов последовательности благодаря наличию циклов в сети. Наиболее распространены LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Units)[6-7].

Сеть подает на вход в момент времени t выход t-1.

$$h_t = f_w(h_{t-1}, x_t) \quad (6)$$

Где h(t)-новое состояние, f(w)- функция параметрами w, h(t-1)-старое состояние, x(t)-вход в шаге t. Входом в RNN служат акустические признаки (MFCC, спектрограммы). Сеть предсказывает вероятности фонем, слов или даже целых фраз на каждом шаге. Часто используется bidirectional-RNN, когда информация учитывается не только из прошлого, но и из будущего (в обоих направлениях).В современных системах RNN часто комбинируют с CNN (для извлечения признаков) или attention/Transformer-слоями (для улучшения качества)[8].



2-Рис. Рекуррентная нейронная сеть

Рекуррентные нейронные сети сыграли огромную роль в развитии распознавания речи, особенно при переходе к end-to-end архитектурам. Благодаря способности работать с последовательными данными и учитывать длительный контекст, RNN, а особенно их варианты LSTM и GRU, долгое время были стандартом индустрии. Однако в последнее время их начинают вытеснять модели на основе внимания (attention) и Transformer, которые эффективнее обрабатывают длинные последовательности и обеспечивают ещё более высокое качество распознавания речи.

Transformer и Attention-модели: Transformer — архитектура, использующая механизм внимания (Attention) для взвешенного анализа входных данных. Не полагается на рекуррентные связи, а обрабатывает данные параллельно, выделяя ключевые фрагменты последовательности.

Основные компоненты Transformer:

1. **Self-Attention (самовнимание):** Каждый элемент последовательности сопоставляется с каждым другим элементом, чтобы понять, насколько они взаимосвязаны.
2. **Position Encoding (позиционное кодирование):**Так как Transformer не содержит рекуррентных связей, в него явно добавляется информация о позиции каждого элемента во входной последовательности.
3. **Encoder-Decoder**:Классический Transformer состоит из энкодера (кодирует входную последовательность в внутренние представления) и декодера (генерирует выходную последовательность на основе этих представлений).

Attention позволяет учитывать весь аудиоконтекст сразу, а не только предыдущее или следующее состояние, как в RNN.Это важно для речи: иногда правильное распознавание текущего звука или слова зависит от далекого контекста в предложении.

Transformer и attention-модели стали революцией в распознавании речи. Они обеспечивают более высокое качество и гибкость, чем предыдущие архитектуры, открывая путь для truly end-to-end систем, многоязычных решений и новых возможностей обработки аудиопотоков. Именно эти модели определяют сегодня и будущее развития систем автоматического распознавания речи.

Вот краткое сравнение популярных алгоритмов распознавания речи:

Характеристика	**HMM (Скрытые модели Маркова)**	**DNN (Глубокие нейронные сети)**	**RNN (Рекуррентные сети)**	**Transformer и Attention-модели**

Основная идея	Вероятностная модель последовательностей состояний и наблюдений.	Моделируют сложные функции путем обучения на больших наборах данных.	Специализированы на обработке последовательных данных.	Используют механизм внимания для выявления важных частей данных.
Точность	Средняя или умеренная	Высокая	Очень высокая	Очень высокая (лучшая)
Сложность обучения	Низкая	Высокая	Высокая	Очень высокая
Вычислительная нагрузка	Низкая	Высокая	Высокая	Очень высокая
Устойчивость к шуму	Низкая	Хорошая	Хорошая	Отличная
Способность к распознаванию длинных последовательностей	Средняя	Средняя	Высокая	Очень высокая
Примеры применения	Ранние системы распознавания речи, базовые телефоны.	Google Voice Search, Siri (начальные версии).	Siri, Alexa (промежуточные версии), системы диктовки.	Google Assistant, современные голосовые помощники.

1-Таблица: Сравнение популярных алгоритмов распознавания речи

Заключение

Современные алгоритмы распознавания речи постоянно развиваются, двигаясь от вероятностных моделей к сложным нейросетевым архитектурам. На сегодняшний день наиболее перспективными считаются модели на основе Transformer, обеспечивающие превосходную точность и широкие возможности применения.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

1. M. J. F. Gales, “Maximum likelihood linear transformations for HMMbased speech recognition,” *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98.
2. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. https://ieeexplore.ieee.org/document/18626
3. Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6645–6649. https://ieeexplore.ieee.org/document/6638947
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://arxiv.org/abs/1706.03762
5. Radford, A., et al. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *OpenAI Technical Report*. https://cdn.openai.com/papers/whisper.pdf
6. Shangguan, Y., J. Li, Q. Liang, R. Alvarez, and I. McGraw. 2019. Optimizing speech recognition for the edge. *ArXiv Prepr. ArXiv190912408*.
7. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for version 3.4). Cambridge University Engineering Department, 2009
8. M.A. Рахманов. 2024. Этапы работы с моделью распознавания речи модели kaldi: Raqamli texnologiyalar obyektlari energiya ta’minotini rivojlantirishning zamonaviy muamolari va istiqbollari. doi:/10.5281/zenodo.10723687