МРНТИ 20.19.29 DOI 10.48501/3007-6986.2025.15.74.004

Сарымов Н.С., Жузбаев С.С.
Евразийский национальный университет им Л.Н.Гумилева
Казахстан, Астана
e-mail: sarym.zh@gmail.com

РАСПОЗНАВАНИЕ РЕЧИ И ПРЕОБРАЗОВАНИЕ ЕЁ В ТЕКСТ С ПРИМЕНЕНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ НА МОБИЛЬНОМ УСТРОЙСТВЕ

Аннотация

В данной статье рассматриваются современные подходы к распознаванию речи и преобразованию речи в текст с помощью глубокого обучения в мобильных приложениях. С ростом спроса на инновационные мобильные технологии распознавание речи становится важным инструментом для улучшения пользовательского опыта и создания интуитивно понятных интерфейсов. Рассматриваются ключевые технологии, такие как искусственный интеллект (ИИ), нейронные сети и методы глубокого обучения, используемые для повышения точности и скорости обработки речи в мобильных устройствах. Особое внимание уделяется требованиям к производительности мобильных приложений, безопасности данных и оптимизации алгоритмов для работы на устройствах с ограниченными ресурсами. Проанализированы существующие решения и представлены рекомендации по созданию эффективных систем распознавания речи для мобильных приложений.

Ключевые слова: искусственный интеллект, машинное обучение, алгоритм, электронное лицо, робот, цифровое право, цифровые отношения, право и информационные технологии

Сарымов Н.С., Жузбаев С.С.
Л.Н.Гумилева атындағы Евразийский национальный университет
Қазақстан, Астана
e-mail: sarym.zh@gmail.com

МОБИЛЬДІ ҚҰРЫЛҒЫДА ТЕРЕҢ ОҚЫТУДЫ ПАЙДАЛАНА ОТЫРЫП, СӨЗДІ ТАНУ ЖӘНЕ МӘТІНГЕ ТҮРЛЕНДІРУ

Андатпа

Бұл мақалада мобильді қосымшаларда терең оқыту арқылы сөйлеуді танудың және сөйлеуді мәтінге айналдырудың заманауи тәсілдері қарастырылады. Инновациялық мобильді технологияларға сұраныстың артуымен сөйлеуді тану пайдаланушы тәжірибесін жақсартудың және интуитивті интерфейстерді құрудың маңызды құралына айналуда. Мобильді құрылғыларда сөйлеуді өңдеудің дәлдігі мен жылдамдығын арттыру үшін қолданылатын жасанды интеллект (AI), нейрондық желілер және терең оқыту әдістері сияқты негізгі технологиялар қарастырылады. Мобильді қосымшалардың өнімділігіне, деректер қауіпсіздігіне және ресурстары шектеулі құрылғыларда жұмыс істеу үшін алгоритмдерді оңтайландыруға қойылатын талаптарға ерекше назар аударылады. Қолданыстағы шешімдер талданды және мобильді қосымшалар үшін сөйлеуді танудың тиімді жүйелерін құру бойынша ұсыныстар берілді.

Кілт сөздер: жасанды интеллект, машиналық оқыту, алгоритм, электрондық тұлға, робот, цифрлық құқық, цифрлық қатынастар, құқық және ақпараттық технологиялар

Sarymov N.S., Zhuzbaev S.S.
L.N. Gumilev Eurasian National University
Kazakhstan, Astana
e-mail: sarym.zh@gmail.com

SPEECH RECOGNITION AND CONVERSION TO TEXT USING DEEP LEARNING ON A MOBILE DEVICE

Annotation

This article discusses modern approaches to speech recognition and speech-to-text conversion using deep learning in mobile applications. With the growing demand for innovative mobile technologies, speech recognition is becoming an important tool for improving the user experience and creating intuitive interfaces. Key technologies such as artificial intelligence (AI), neural networks, and deep learning methods used to improve the accuracy and speed of speech processing in mobile devices are considered. Special attention is paid to the performance requirements of mobile applications, data security, and algorithm optimization for use on devices with limited resources. The existing solutions are analyzed and recommendations for creating effective speech recognition systems for mobile applications are presented.

Keywords: artificial intelligence, machine learning, algorithm, electronic person, robot, digital law, digital relations, law and information technology

Введение.

Развитие технологий искусственного интеллекта и глубокого обучения значительно расширило возможности распознавания речи в мобильных приложениях. Эти системы позволяют пользователям взаимодействовать с приложениями с помощью речи, делая взаимодействие более естественным и удобным. В данной статье рассматриваются существующие технологии разработки мобильных приложений с использованием распознавания речи и их влияние на улучшение взаимодействия между устройством и пользователем.

Основная часть

Как проводиться обучение прежде всего, имели место кратковременные эффекты адаптации: увеличение акустического сходства с предыдущим стимулом приводило к уменьшению гемодинамической активности в средне-задней STS и правых вентролатеральных префронтальных зонах. Далее, наблюдались и долгосрочные эффекты: ослабление реакции выявлено в орбитальной/островковой коре для стимулов, которые были либо наиболее, либо наименее схожи с акустическим средним всех предшествующих стимулов, а также в переднем височном полюсе, глубокой задней STS и миндалине для стимулов, имеющих наибольшее или наименьшее сходство с усреднённым обученным значением категории голосовой идентичности [1].

Основные требования к системам распознавания речи

Для успешной реализации технологии распознавания речи на мобильных устройствах важно учитывать несколько ключевых факторов:

1. Точность распознавания: мобильные приложения должны обеспечивать высокую точность преобразования речи в текст, особенно в шумной обстановке или при использовании различных акцентов.

2. Производительность и оптимизация: алгоритмы глубокого обучения, такие как конволюционные нейронные сети и рекуррентные нейронные сети, должны быть адаптированы для работы на устройствах с ограниченными ресурсами памяти и процессора.

3. Безопасность данных: поскольку распознавание речи часто связано с обработкой персональных данных, шифрование и защита передаваемой информации - важный аспект разработки мобильных приложений.

Методы исследования.

Современные системы распознавания речи часто используют гибридный подход, сочетая локальную обработку с облачными вычислениями, чтобы сбалансировать производительность и качество. Например, модель Wav2Vec 2.0, разработанная Facebook AI, может эффективно работать на мобильных устройствах за счет использования трансформаторов и предварительного обучения. Для разработки подобных приложений используются такие языки программирования, как Python с библиотеками глубокого обучения (TensorFlow, PyTorch), а также Flutter. Используются мобильные фреймворки с кроссплатформенной поддержкой.

Возможности распознавания речи

Мобильные приложения с функцией распознавания речи позволяют пользователям выполнять широкий спектр задач, таких как ввод устного текста, управление приложениями и взаимодействие с виртуальными помощниками. Эти системы также могут быть интегрированы с другими технологиями,

такими как Обработка естественного языка (NLP) и машинное обучение, для улучшения обработки естественного языка и контекстного понимания речи.

Речь — это вокализованная форма человеческого взаимодействия. На этом этапе речь говорящего принимается в форме звуковой волны. Существует множество программного обеспечения, которое используется для записи человеческой речи. Акустическая среда и оборудование для преобразования могут сильно повлиять на создаваемую речь. Вместе с речевым сигналом могут присутствовать фоновые шумы или реверберация в помещении, что является совершенно нежелательным [2].

Взаимодействие между клиентами, фронт-эндом, бэк-эндом и хранилищем данных.

1. клиент- это пользователь, который использует приложение для ввода голосовых команд.

2. Фронтенд - отвечает за запись голоса и отправку данных на сервер через API.

3. Бэкэнд обрабатывает речевые данные, применяет модель глубокого обучения для преобразования речи в текст и возвращает результаты на фронтэнд.

4. Хранилище данных- это база данных, в которой хранятся модели и данные, необходимые для обучения и совершенствования системы распознавания речи.

Таким образом, эффективное взаимодействие между всеми компонентами системы является ключевым фактором успешной реализации распознавания речи в мобильных приложениях.

Клиент будет отправлять запрос на Фронтенд часть в нашем случае это Смартфон, Смартфон передает на APIкакую либо аудиозапись, данная запись после обработки будет сохраняться на сервере, потом сервер готовые данные будет отправлять на Фронтенд

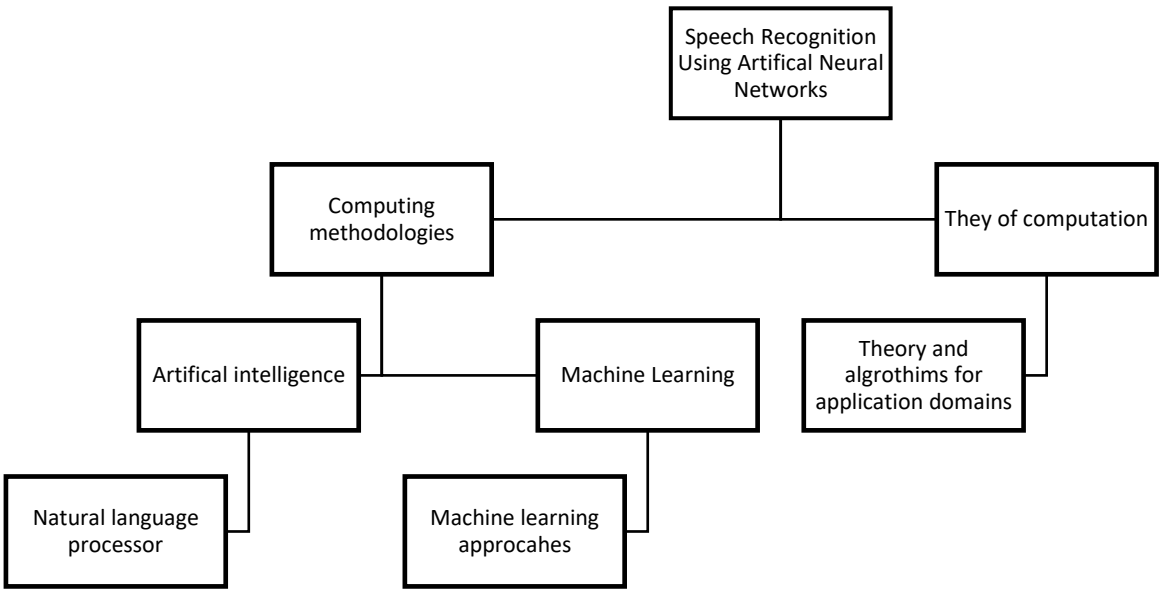

Диаграмма 1

Даннаядиаграмма была взять из литературы[3]. Эта диаграмма отображает иерархию концепций, связанных с распознаванием речи с использованием искусственных нейронных сетей. Она разделена на два основных направления: вычислительные методологии и теорию вычислений, каждая из которых подразделяется на более узкие области

Потенциал нейронных сетей

Нейронные сети для распознавания речи исследуются в рамках недавнего возрождения интереса к этой области. Исследования сосредоточены на оценке новых алгоритмов классификации образов и обучения нейронных сетей с использованием реальных данных о речи, а также на определении того, могут ли быть разработаны параллельные архитектуры нейронных сетей, которые выполняют вычисления, необходимые для важных алгоритмов распознавания речи. Большая часть работы сосредоточена на распознавании изолированных слов [4]

Сопостовление шаблонов с использованием нейронных сетей

Искусственные нейронные сети (ИНС) — это интеллектуальные системы, которые в некоторой степени связаны с упрощённой биологической моделью человеческого мозга. Они состоят из множества простых элементов, называемых нейронами, которые работают параллельно и соединены друг с другом через множители, называемые весами связей. Нейронные сети обучаются путем корректировки значений этих весов связей между элементами сети.

Нейронные сети обладают способностью к самообучению, устойчивы к ошибкам и помехам, и находят применение в идентификации систем, распознавании образов, классификации, распознавании речи, обработке изображений и других областях.

Нейронные сети с обратным распространением ошибки были использованы для идентификации видов птиц с помощью записей птичьего пения [5].

В нашем приложении ИНС используется для сопоставления шаблонов. Для данной задачи были сравнены различные архитектуры нейронных сетей, чтобы оценить их производительность.

Искусственные нейронные сети (ИНС) были обучены с использованием семи голосовых образцов, записанных в разные моменты времени от четырёх разных говорящих, которые произносили одну и ту же фразу. Во время обучения были заданы такие параметры, как начальная скорость обучения, допустимая ошибка и максимальное количество циклов обучения. Нейронная сеть была построена в среде MATLAB. На первой попытке были получены суммы квадратичной ошибки в зависимости от количества эпох в процессе обучения. Целевая сумма квадратичных ошибок была достигнута всего за 174 эпохи. Вторая попытка показывает различные скорости обучения, использованные во время обучения [6].

Уровень успеха 100% был достигнут при тестировании ИНС на обученных образцах. Однако при использовании необученных образцов уровень успеха составил лишь 66%. Это произошло из-за того, что спектральные плотности мощности (PSD) входных образцов не совпадали с образцами, использованными для обучения. ИНС была протестирована на голосовых образцах людей, которых она не знала, и успешно классифицировала эти образцы как неопознанные голоса.

Глубокое обучение

С 2006 года этот класс машинного обучения получил мощное развитие и был внедрен в сотни исследований с тех пор. Области, в которые включено глубокое обучение, варьируются от обработки информации до искусственного интеллекта. Глубокое обучение можно охарактеризовать как подполе машинного обучения, основанное на алгоритмах, которые учатся на нескольких уровнях для создания модели, представляющей сложные взаимосвязи между данными. Здесь присутствует иерархия признаков, при которой высокоуровневые признаки определяются в терминах более низкоуровневых, и именно поэтому оно называется глубокой архитектурой. Большинство моделей, включенных в этот класс, основаны на представлениях без учителя.

Глубокое обучение представляет собой точку пересечения нейронных сетей, графического моделирования, оптимизации, искусственного интеллекта, распознавания образов, а также обработки сигналов. Причины популярности глубокого обучения можно подытожить следующим образом: оно способствовало значительному увеличению вычислительных возможностей компьютерных чипов, позволило интегрировать огромные объемы обучающих данных и стало основой для недавних достижений в машинном обучении в области обработки информации и сигналов[7].

Результаты исследования

Гибридные модели искусственных нейронных сетей (ANN) и скрытых марковских моделей (HMM) обучались в три этапа[8].

1. Первый этап: была обучена базовая система GMM/HMM (смешанная модель Гаусса и скрытая марковская модель), и использовалась принудительная выравнивание для ассоциации каждого кадра данных с целевым состоянием HMM.

2. Второй этап: на акустических данных (которые могут быть векторами MFCC, логарифмическими фильтровыми банками или адаптированными к говорящему признаками, собранными вместе) была обучена глубокая версионная сеть (DBN). Веса DBN использовались для инициализации нейронной сети, которая затем обучалась предсказывать состояние HMM на основе акустических данных, используя метод обратного распространения ошибки.

3. Третий этап: поскольку система основана на контекстно-зависимых моделях, для дополнительных сведений читателю рекомендуется ознакомиться с исследованиями [9], которые также использовали контекстно-зависимые системы.

Название	Количество часов	CMLLR	WER
Voice Search	6K	Нет	16.0
YouTube	1400	Да	52.3

Таблица 1: Базовые показатели, используемые для исследования

- CMLLR (Constrained Maximum Likelihood Linear Regression): метод, используемый для адаптации моделей к различным условиям записи.

- WER (Word Error Rate): показатель, отражающий процент ошибок в распознавании слов.

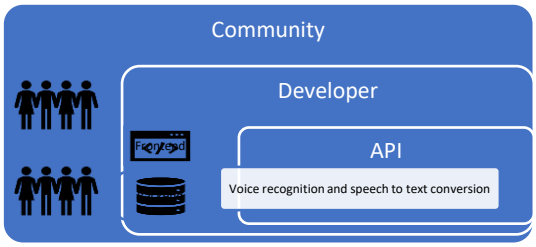

Диаграмма2

Данная диагамма показывает зависимости использования разных сервисов, пользователь использует часть разработчика а часть разработчика использует готовую платформу для отправки запроса

Основное положение

Сегодня на рынке существует множество решений для распознавания речи, использующих модели глубокого обучения для обработки речи, включая API от ведущих компаний, таких как Google Speech-to-Text и Apple Siri. Например, модель DeepSpeech компании Mozilla использует рекуррентные нейронные сети (RNN) и демонстрирует высокую точность и производительность на мобильных устройствах. Одной из ключевых проблем при разработке систем распознавания речи является необходимость поиска компромисса между производительностью и точностью обработки, особенно на устройствах с ограниченными вычислительными ресурсами [10].

Заключение

В современном мире, где технологии развиваются феноменальными темпами, интеграция технологий распознавания речи, преобразования текста в речь и глубокого обучения в мобильные приложения открывает новые горизонты взаимодействия пользователя с устройством. Эти технологии не только повышают доступность информации и упрощают выполнение задач, но и создают возможности для инновационных подходов в обучении и коммуникации

Важно отметить, что успешное внедрение систем распознавания речи требует учета множества факторов, включая точность обработки, производительность на мобильных устройствах и безопасность данных. Применение передовых технологий, таких как нейронные сети и методы глубокого обучения, способствует созданию более эффективных и удобных приложений, способных адаптироваться к потребностям пользователей.

В заключение следует отметить, что дальнейшее развитие и совершенствование технологий распознавания речи не только улучшит пользовательский опыт, но и создаст новые возможности для бизнеса в различных сферах, таких как образование, здравоохранение и обслуживание клиентов. Актуальность и востребованность таких решений в условиях цифровизации общества только возрастает и является предметом активных исследований и разработок.

Список литературы

1. Andics A. et al. Neural mechanisms for voice recognition //Neuroimage. – 2021. – Т. 52. – №. 4. – С. 1528-1540.[1]
2. Kamble B. C. Speech recognition using artificial neural network–a review //Int. J. Comput. Commun. Instrum. Eng [2]
3. Lim C. P. et al. Speech recognition using artificial neural networks //Proceedings of the First International Conference on Web Information Systems Engineering. – IEEE, 2022. – Т. 1. – С. 419-423. [3]
4. Lippmann R. P. Review of neural networks for speech recognition //Neural computation. – 2023. – Т. 1. – №. 1. – С. 1-38. [4]
5. A L Mcilraith, H C Card, "Birdsong Recognition Using Backpropagation and Multivariate Statistics", IEEE Trans on Signal Processing, vol. 45, no. 11, November 2022, [5]
6. Venayagamoorthy G. K., Moonasar V., Sandrasegaran K. Voice recognition using neural networks //Proceedings of the 2023 South African Symposium on Communications and Signal Processing-COMSIG'98 (Cat. No. 98EX214). – IEEE, 2023. – С. 29-32. [6]
7. Y. Cho and L. K. Saul, ''Kernel methods for deep learning,'' in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 22, 2019, pp. 342–350. [7]
8. Chan W. et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition //2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). – IEEE, 2022. – С. 4960-4964. [8]
9. Jaitly N. et al. Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition //Interspeech. – 2022. – Т. 2012. – С. 2578-2581. [9]
10. F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in Interspeech, 2021, pp. 437–440. [10]
11.

Сведение об авторе
Сарымов Нурсултан Сарымович
Должность: г.Астана, студент ЕНУ кафедра «Информационных систем», бакалавр
Почтовый адрес: 010000, Казахстан, Астана, Алихан Бокейхана 11
Мобильный телефон: 8 (707)-962-16-30
E-mail: sarym.zh@gmail.com
Жузбаев Серик Сулейменович
Должность: канд. физ.-мат. наук, доц., Евразийский национальный университет им. Л.Гумилева
Почтовый адрес: 010000, Казахстан, Астана, Алихан Бокейхана 17
E-mail: zhuzbayev52@gmail.com

Автор жайлы мәлімет
Сарымов Нурсултан Сарымулы
Лауазымы: Астана қаласы, ЕНУ университетінің «Ақпараттық жүйелер»бөлімінің студенты, бакалавр
Почталық мекен жайы: 010000, Қазақстан, Астана, Алихан Бокейхана 11
Ұялы телефон: 8 (707)-962-16-30
E-mail: sarym.zh@gmail.com
Жузбаев Серик Сулейменович
Лауазымы: канд. физ. - мат. ғылымдар, доц., Л. Гумилев атындағы Еуразия ұлттық университеті
Почталық мекен жайы: 010000, Қазақстан, Астана, Алихан Бокейхана 17
E-mail: zhuzbayev52@gmail.com

Information about the author
Sarymov Nursultan Sarymovich
Position: Astana, student of ENU, Department of Information Systems, bachelor

Postal address: 010000, Republic of Kazakhstan, Astana, Alikhan Bokeikhan 11
Mobile phone: 8 (707)-962-16-30
E-mail: sarym.zh@gmail.com
Zhuzbaev Serik Suleimenovich
Position: Candidate of Physical and Mathematical Sciences, Associate Professor, L.Gumilyov Eurasian National University
Postal address: 010000, Republic of Kazakhstan, Astana, Alikhan Bokeikhan 17
E-mail: zhuzbayev52@gmail.com