УДК 004.934

РАЗРАБОТКА ЛОКАЛЬНОЙ СИСТЕМЫ ДЛЯ ИЗВЛЕЧЕНИЯ ТЕКСТА ИЗ АУДИО С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Павлишина И.Р. (ИТМО)
Научный руководитель – Пантюхин И.С. (ИТМО)

Введение: Современные технологии искусственного интеллекта и глубоких нейросетевых архитектур значительно расширили возможности автоматического распознавания речи (ASR), что привело к росту точности преобразования аудиосигналов в текст. Однако, несмотря на успехи, большинство коммерческих и исследовательских систем ASR оптимизированы для английского языка и требуют значительных вычислительных ресурсов, что ограничивает их применение в условиях локальной обработки [1]. Развитие ASR для языков с высокой морфологической сложностью, таких как русский, сталкивается с рядом вызовов, включая необходимость адаптации языковых моделей и механизмов лексической обработки. Современные исследования показывают, что персонализированные модели ASR, использующие индивидуальные голосовые характеристики, могут значительно повысить точность распознавания речи [3]. В частности, адаптация моделей под особенности конкретных пользователей продемонстрировала увеличение точности распознавания до 10% по сравнению со стандартными универсальными моделями. Дополнительно, недавние исследования подчеркивают важность внедрения языковых моделей, специально разработанных для языков с ограниченными ресурсами, таких как русский. В этом контексте особую роль играют модели, использующие психоакустические принципы и адаптивные методы обучения [4]. Важным направлением исследований также является интеграция ASR в специализированные системы, например, в медицинскую диагностику и аналитические платформы. Автоматическое преобразование аудиозаписей врачебных консультаций в текстовые отчеты позволяет повысить эффективность ведения медицинской документации и ускорить анализ клинических данных [5]. Настоящее исследование направлено на разработку высокоэффективной системы автоматического извлечения текстовой информации из русскоязычного аудиоконтента, обеспечивающей баланс между точностью распознавания и вычислительной сложностью.

Основная часть: В ходе исследования была проведена сравнительная оценка моделей распознавания речи по точности, ресурсоемкости и времени обработки. На основе полученных данных выбрана модель с оптимальным соотношением качества и вычислительной эффективности. Для потокового распознавания аудио используется разбиение входного сигнала на чанки по 5 секунд. Такая сегментация позволяет обеспечить наличие связных фраз и при этом сохранить режим, близкий к “обработке в реальном времени”. Однако на стыках чанков качество распознавания снижается из-за неполных слов и недостаточного контекста. Для решения этой проблемы разработан алгоритм, позволяющий извлекать текст конкретного чанка, принимая во внимание аудио в соседних чанах (до и после). Такой подход несколько замедляет обработку (обрабатывается не последний записанный чанк, а предпоследний), но зато повышает качество извлекаемого текста.

После накопления 3–5 чанков запускается этап постобработки, включающий замену ключевых слов на соответствующие символы и преобразование числительных в цифровой формат (с учетом дробных чисел). Для корректной расстановки знаков препинания временно убираются разделители в десятичных числах, затем с помощью нейронной модели выполняется пунктуация. Поскольку модель автоматически ставит точку в конце, последнее предложение обрабатываемого текста отбрасывается и передается в следующую итерацию. Перед отдачей текста пользователю в числа возвращаются изъятые ранее разделители.

В результате алгоритм обеспечивает непрерывный поток текста с правильно расставленными знаками препинания и корректной записью символов и чисел, что позволяет получать расшифрованную речь в режиме реального времени.

Заключение: Разработанная система демонстрирует высокую точность извлечения текста из аудио с использованием нейронных сетей, оптимизированных для работы с русскоязычными данными. Экспериментальные результаты свидетельствуют о возможности дальнейшей оптимизации алгоритмов распознавания, а также перспективности интеграции системы в экосистему ePath – программно-аппаратный комплекс для проведения макроскопических медицинских исследований, разрабатываемый ООО «Гистоскан». Практическая значимость исследования заключается в обеспечении локального развертывания системы с минимальными вычислительными затратами при сохранении требуемого качества распознавания.

Список использованных источников:

1. Alharbi S., Alrazgan M., Alrashed A., Alnomasi T., Almojel R., Alharbi R., Alturki S., Alshehri F., Almojil M. Automatic Speech Recognition: Systematic Literature Review // IEEE Access. 2021. Vol. 9. P. 131858-131876. DOI: 10.1109/ACCESS.2021.3112535.
2. Malik M., Malik M. K., Mehmood K., Makhdoom I. Automatic speech recognition: a survey // Multimedia Tools and Applications. 2020. Vol. 80. P. 9411-9457. DOI: 10.1007/s11042-020-10073-7.
3. Brydinskyi V., Sabodashko D., Khoma Y., Podpora M., Konovalov A., Khoma V. Enhancing Automatic Speech Recognition With Personalized Models: Improving Accuracy Through Individualized Fine-Tuning // IEEE Access. 2024. Vol. 12. P. 116649-116656. DOI: 10.1109/ACCESS.2024.3443811.
4. Coro G., Massoli F. V., Origlia A., Cutugno F. Psycho-acoustics inspired automatic speech recognition // Computers & Electrical Engineering. 2021. Vol. 93. P. 107238. DOI: 10.1016/J.COMPELECENG.2021.107238.
5. Schultz B., Tarigoppula V. S. A., Noffs G., Rojas S., Walt A. V. D., Grayden D., Vogel A. Automatic speech recognition in neurodegenerative disease // International Journal of Speech Technology. 2021. Vol. 24. P. 771-779. DOI: 10.1007/S10772-021-09836-W.

Автор _____ Павлишина И.Р.

Научный руководитель_____ Пантюхин И.С.