РАСПОЗНАВАНИЕ РЕЧИ ПОСРЕДСТВОМ МОДЕЛЕЙ ИИ

Чжо Хин Аунг

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Митюхин А.И. – доцент (научн.рук.)

Представлены результаты экспериментальных исследований систем распознавания речи, построенных на применении нейросетей и использованием архитектуры Transformer.

Задача распознавания речевого сигнала считается одной из самых сложных из областей информатики, включающей в себя математику, цифровую обработку сигналов и изображений, кодирование, теорию вероятностей, лингвистику, статистику и пр. Задача сложна вследствие высокой вариативности речи, разнообразия акцентов, отсутствия контекста и т. д. Хотя решением этой задачи ученые и инженеры начинали заниматься еще в пятидесятых годах прошлого века, процент ошибок распознавания речи удалось значительно снизить только к 2010 году. К этому времени были разработаны методы глубокого обучения, на основе которых получены достаточно хорошие результаты. Решение этой задачи опирается на создание моделей машинного обучения на базе глубоких нейронных сетей (DNN). Для этого используются огромные объемы обучающих данных и высококачественные тестовые наборы.

Задачу распознавания речи можно разбить на три следующие подзадачи:

1. Анализ речевого сигнала, выделение и моделированием акустических признаков.
2. Описание зависимостей, которые существуют между словами в языке и определение возможных вероятностей следования слов друг за другом.
3. Определение наилучшего кандидата, например, слова на распознавание среди всех возможных с использованием той информации, которая создается в ходе решения первых двух подзадач.

На основании такого разделения задачи рассматриваются три основных структурных элемента системы распознавания речи:

1. Акустическая модель.

2. Языковая модель, которая представляет собой лингвистическое описания речи. Языковая модель позволяет узнать, какие последовательности слов в языке более вероятны, а какие менее.

3. Декодер.

На рисунке 1 показана обобщенная структурная схема распознавания речи. После обнаружения и захвата речевого сигнала через микрофон, предварительной фильтрации сигнала речи, АЦП, цифровой спектральной обработки и формирования опорных векторов образов речи решается задача декодирования. Надежная система декодирования текста должна содержать блок точностной обработки для улучшения распознавания.

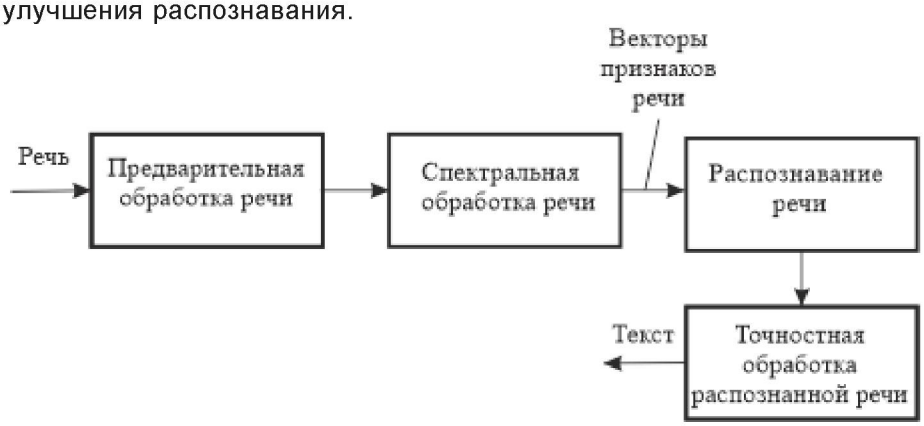

Рисунок 1 – Обобщенная структурная схема системы распознавания

Техническая реализация основывается на выполнении вычислительно затратных многоэтапных алгоритмов цифровой обработки сигналов, кодирования и методов машинного обучения с целью распознавания речевых шаблонов. На рисунке 2 показана схема блока распознавания.

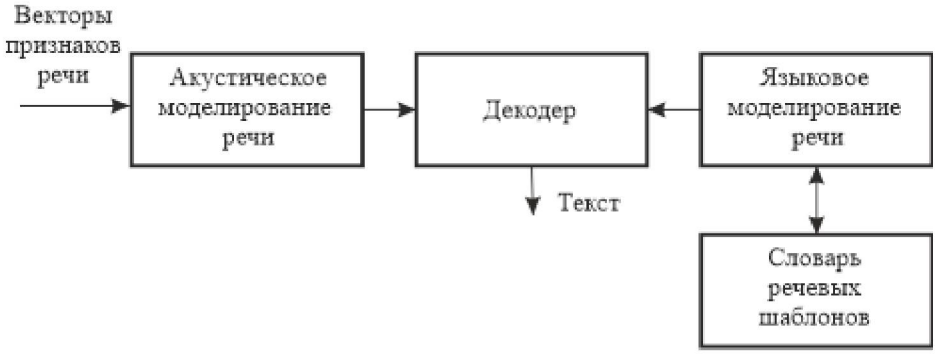

Рисунок 2 – Структурная схема блока распознавания речевого образа

Схема, рис. 2 преобразует векторы признаков в последовательности фонем слов, которые затем сравниваются с опорными образами словаря и декодируются. Основными элементами схемы, рис.2, являются:

– акустическая модель, выполняющая распознавание звуков с помощью вектора признаков образа речи;

– языковая модель, выполняет предсказание вероятной последовательности слов фразы в контекстном значении, выбирая слова из словаря речевых шаблонов.

– в словаре реализуется процесс поиска соответствия звуковых сигналов речи в виде вектора признаков образа речи и слов словаря с возможностью определения отличий слов с одинаковым звучанием, используя контекст фразы. Тем самым, выбираются наиболее вероятные варианты слов и повышается точность распознавания речи.

– декодер на основе векторных оптимальных алгоритмов сравнений выходов моделей и словаря формирует текстовый образ входной речи, рис. 1.

Исследования касались системы распознавания с векторными представлениями слов с использованием архитектуры Transformer (как в Whisper или GPT), работающей *под управлением операционной системы* без предварительного обучения языковой модели. В этом случае не требовалось использования сравнительно затратного оборудования и времени на подготовку текстов и технически не простое предварительное обучения языковой модели по заранее подготовленным аудио материалам. Использовались аудио файлы по математически насыщенной специальной дисциплине «Основы кодирования и криптографии». Кроме того, исследования проводились для варианта работы систем распознавания речи в автономном режиме (не зависящем от сети). С целью ускорения обработки

речевого сигнала для распознания ограничивалось пространство поиска речевого щаблона, соответствующего текстовому слову, т. е. не использовался поиск в словаре сети.

Экспериментальные исследования показали, что точность распознавания постепенно увеличивалась по мере временного речевого продвижения по названному курсу. Покажем это на примере анализа ошибок. Качество распознавания оценивалось коэффициентом ошибок WEP (Word Error Rate) на уровне слов

$$WER = \frac{S+D+I}{N} \cong 0{,}28, \text{ или } 28\%$$

Точность распознавания речи на математическом контенте находилась в диапазоне

$$1 - \mathrm{WER} = 70 \div 85\%.$$

Значительное число ошибок возникает из-за неправильного распознавания специальных терминов, например, «Квадратично-вычетные коды», «Спектр кода Кердока». Исследования показывают, что существуют явные проблемы при передаче математической научной информации с использованием систем автоматического распознавания речи.

Список использованных источников:

1. Митюхин, А.И. Доступность образования для людей с ограничениями слуха года. Непрерывное профессиональное образование лиц с особыми потребностями : сб. ст. VI Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 12 декабря 2025 Минск : БГУИР, 2025. – 5 с.