ТЕХНОЛОГИИ РАСПОЗНАНИЯ РЕЧИ НА ОСНОВЕ АНАЛИЗА ВИДЕОДАННЫХ

Макар Д.А.

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

Вашкевич М.И. – доктор тех. наук

Аннотация. В работе описаны ключевые этапы обработки видеоданных в системах визуального распознавания речи. Рассмотрены различные подходы к построению таких систем, которые имеют большой потенциал для применения в промышленных условиях, биометрической идентификации и для помощи людям с нарушением речевого аппарата.

Введение.

Исследование посвящено подходу к распознанию речи на основе комплексного анализа видеоданных. В отличие от традиционных методов распознавания речи, предлагаемый фокусируется на визуальных параметрах речи – динамике артикуляции, мимике и движениях лицевых мышц. Актуальность разработки альтернативных методов распознавания речи обусловлена ограничениями систем, использующих только аудиоданные (англ. *ASR – automatic speech recognition*). Как отмечается в [1], эффективность классических ASR-систем значительно снижается при уровне шума выше 15 дБ, что делает их непригодными для работы в промышленных условиях или общественных пространствах.

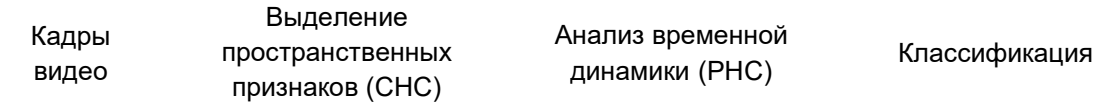
Визуальное распознание речи (англ. *VSR – visual speech recognition*) – метод, позволяющий распознавать речь по движениям губ без использования аудиосигнала. Такие системы особенно необходимы для распознавания речи в условиях акустических шумов, для помощи людям с нарушением речевого аппарата и в системах биометрической идентификации [2].

Структура системы визуального распознавания речи.

Типовая система VSR выделяет из входной видеопоследовательности либо область лица, либо изолированную область рта, после чего обрабатывает их при помощи сверточной нейронной сети (СНС) с целью извлечения признаков. На заключительном этапе признаки, полученные в СНС, классифицируются либо на уровне отдельных визем, либо на уровне слов при помощи рекуррентной нейронной сети (РНС), либо при помощи нейронной сети с архитектурой Transformer.

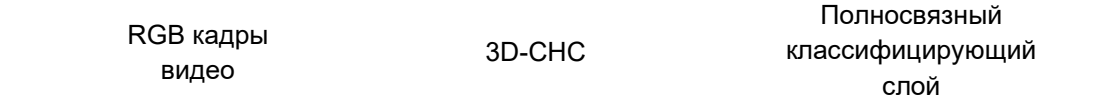
На рисунке 1, представлена схема типовой VSR, включающая три ключевых этапа:

1. Выделение пространственных признаков при помощи СНС.
2. Анализ временной динамики признаков (т.е. выявление изменений в положении губ и лицевых мышц между кадрами) при помощи РНС.
3. Классификация – интерпретация извлечения признаков для распознавания речевых единиц.



– Рисунок 1 – Типовая схема системы VSR –

Альтернативным образом VSR система может быть построена при помощи нейронной сети с трехмерной сверточной архитектурой (3D-СНС). На рисунке 2, представлена обобщенная структура системы VSR-системы на основе 3D-СНС.



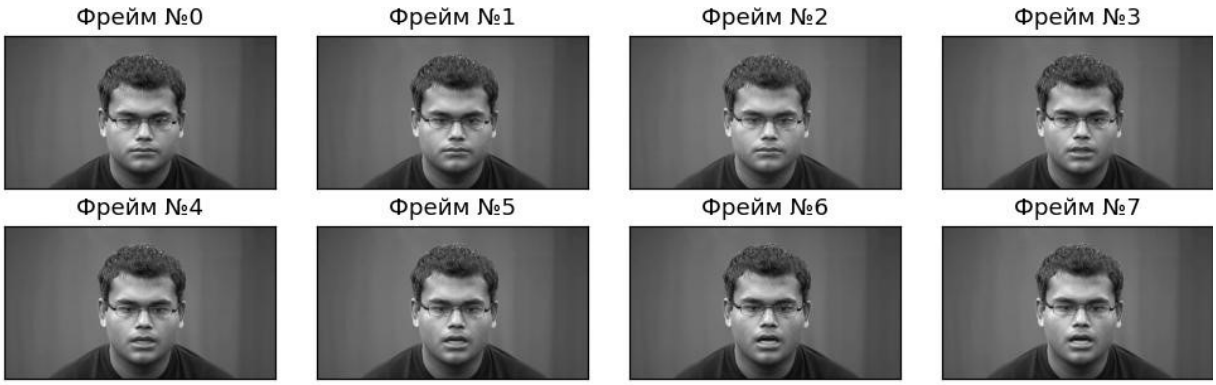
– Рисунок 2 – Структура системы VSR на основе 3D-СНС –

Использование трехмерных сверточных слоев позволяет эффективно извлекать пространственно-временные признаки из видео, что критично важно для анализа динамики артикуляции. Как видно из схемы, входные данные (RGB-кадры) последовательно обрабатываются слоями 3D-свертки, что обеспечивает сохранение временной зависимости между кадрами.

Практические аспекты построения систем визуального распознавания речи.

В данном разделе рассмотрены некоторые практические аспекты построения и обучения системы VSR. В качестве примера будет использован набор данных AVLetters2 [3]. Это набор коротких видео, где 5 дикторов на камеру произносят одну из 26 букв английского алфавита слова (каждый диктор произносит каждую букву 7 раз). На рисунке 3 показан пример начальных фреймов видео, на котором один из дикторов произносит букву «А».

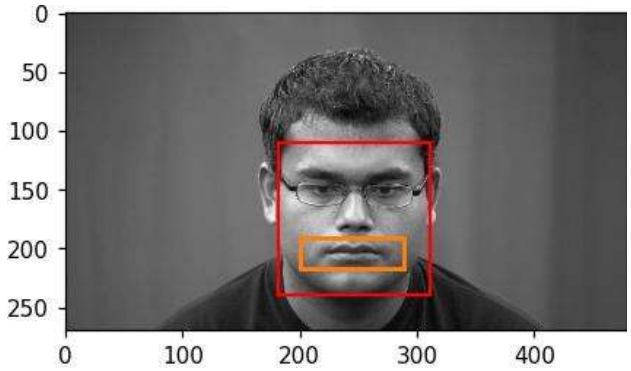
–



– Рисунок 3 – Пример кадров видеоизображения из базы AVLetters2 (бука «А»)

В системах визуального распознавания речи применяются разные стратегии выделения регионов интереса на видео: 1) захват всего лица (включая щеки, подбородок, нос) и 2) захват только области губ. Системы, которые работают только с областью губ также называют автоматизированными системами чтения по губам.

В реальных системах для предобработки видеоданных используются специальные библиотеки, как, например, Dlib [4]. На рисунке 4 представлен видеокадр изображения на котором при помощи метода dlib.get_frontal_face_detector() выделена область лица (красная рамка). Выделенную рамку лица также можно далее обработать детектором ключевых точек лица с целью выделения области губ. Для этого использовался метод dlib.shape_predictor(), результат обработки также показан на рисунке 4 (оранжевая рамка).



– Рисунок 4 – Пример распознавания области лица и губ с использованием библиотеки Dlib

Выводы и заключение

Исследование демонстрирует значительный потенциал мультимодального анализа видеоданных для совершенствования технологий распознания речи. Рассмотренные методы, основаны на анализе визуальных параметров речи, таких как динамика артикуляции и мимика и позволяют преодолеть ключевые ограничения традиционных ASR-систем, особенно в условиях высокого уровня шума. Использование 3D-CHC и алгоритмов трекинга ключевых точек обеспечивает высокую точность обработки пространственно-временных признаков, что открывает новые возможности для применения технологии в промышленных условиях, системах биометрической идентификации и помощи людям с нарушением речевого аппарата.

Таким образом, разработка и внедрение мультимодальных подходов представляет собой перспективное направление в области искусственного интеллекта и обработки естественного языка, способное значительно улучшить качество и доступность речевых технологий.

Список использованных источников:

1. Multimodal Speech Recognition Benchmarking in Noisy Environments / Y. Zhou, X. Li, H. Zhang, J. Wu // IEEE Transactions on Audio, Speech, and Language Processing (TASLP), 2022. – Vol. 30. – P. 2581–2593.

2. Petridis S. et al. End-to-end audiovisual speech recognition // IEEE intern. conf. on acoustics, speech and signal processing (ICASSP). – 2018. – P. 6548-6552.

3. Cox S. et al. The challenge of multispeaker lip-reading. // Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP). – 2008. – P 179 - 184.

4. Dlib C++ library. URL:https://dlib.net/ (дата обращения 16.04.2025)