Автоматическое построение обучающего аудио-видеокорпуса применительно к задаче распознавания русской речи

Карпухин Иван Александрович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: karpuhini@yandex.ru

Системы автоматического распознавания речи давно вышли из категории узкоспециализированных инструментов и стали массовым продуктом. Качество распознавания для некоторых языков в офисных условиях приблизилось к возможностям человека. Однако существуют нерешенные задачи, стоящие на пути создания систем, способных функционировать в различных условиях повседневной жизни. К числу таких задач относится борьба с шумами звукового сигнала.

Перспективный способ повышения качества распознавания речи и уменьшения влияния аудиошумов заключается в использовании видеоданных, содержащих изображение лица диктора во время произнесения. Для этого могут учитываться движения губ, челюсти и языка говорящего. Создание мультимодальных аудио-видеосистем опирается на методы машинного обучения с учителем. При этом встает задача создания обучающего видеокорпуса данных, содержащего аудио-видеофрагменты произнесения фраз вместе с указанием текста соответствующих фраз. На ранних этапах разработки системы распознавания используют также фонетическую разметку фрагментов, в которой указаны временные интервалы произнесения отдельных фонем.

Созданием обучающих аудио-видеокорпусов традиционно занимаются команды экспертов, выполняющие видеозапись приглашенных дикторов и транскрибирование полученных фрагментов. Стоимость создания таких корпусов линейно зависит от их объема. В настоящей работе предлагается способ автоматического построения обучающих видеокорпусов с фонетической разметкой, содержащих десятки часов речи.

В качестве исходных данных используются видеозаписи чтений книг либо записи с указанием стенограммы. При этом длительность каждого фрагмента может достигать нескольких часов. Несмотря на то, что исходные данные формально удовлетворяют определению обучающего корпуса, они оказываются непригодными из-за большой

длительности фрагментов и наличия участков, не относящихся к речи (названия глав, фрагментов, различные вставки). Предлагаемая система решает следующие задачи:

- Выравнивание текста и аудиосигнала для записей многочасовой длины
- Обнаружение и фильтрация фрагментов, не относящихся к речи
- Раздление длинных фрагментов на небольшие части, пригодные для машинного обучения
- Классфикация видеофрагментов по степени поворота головы диктора и размеру области лица

Точность фонетической разметки предложенной системы достигает 75%. По мимо этого, в работе получена теоретическая оценка качества построения видеосистемы распознавания по полученному обучающему видеокорпусу.

Работа поддержана грантом по программе «УМНИК».

Литература

1. Baker J. Developments and directions in speech recognition and understanding, part 1 // Signal processing magazine, IEEE, 2009, Т. 26, № 3. P. 75–80.
2. Карпов А. А. Реализация автоматической системы многомодального распознавания речи по аудио- и видеоинформации // Автоматика и телемеханика. 2014. № 12. С. 125–138.
3. Stewart D. Robust audio-visual speech recognition under noisy audio-video conditions // IEEE Transactions on Cybernetics, 2014, Т. 44, № 2. P. 175–184.