Интеграция признаков видео в систему распознавания речи

Старцев Михаил Леонидович

Студент

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

E-mail: mikhail.startsev@gmail.com

Под распознаванием речи понимается дословное преобразование входного сигнала в произносимый текст. В большинстве известных систем распознавания используется лишь аудио-сигнал. Однако распознавание в условиях зашумлённого аудио представляет особую трудность для таких систем. Улучшить качество распознавания позволяют, например, системы аудио-видео распознавания речи.

Обычно распознавание речи производится в несколько этапов. Сначала входной сигнал переводится в более удобное для анализа представление. В терминологии машинного обучения это называется выделеним признаков. В результате этого входному сигналу ставится в соответствие последовательность векторов-признаков. Затем с помощью акустической модели такая последовательность переводится в фонетическую транскрипцию. Для преобразования транскрипции в предложения используется языковая модель.

Использование видео в дополнение к аудио или отдельно от него для задач распознавания речи изучается достаточно давно. Для распознавания по аудио-видео данным необходимо использовать совместное признаковое описание этих данных. Видео-признаки (дескрипторы), применяемые в задачах чтения по губам традиционно делятся на три группы: основанные на пикселах региона интереса, основанные на форме губ спикера и их комбинации.

В данной работе исследуется вклад различных признаков видео в качество распознавания русской речи с помощью интеграции соответствующих признаков в популярную систему с открытым исходным кодом. Сравнение производится с системой, использующей только аудио данные, обученной на сводобно распространяемой выборке данных.

Для обучения и тестирования аудио-видео модели произведена запись эталонных наборов данных. Так, для построения обучающей базы используются фонетически репрезентативные тексты, заимствованные из базы ISABASE. Для тестовой части базы используются записи произнесения набора команд для управления приложением.

В данной работе подробно исследуются геометрические признаки, описывающие форму губ: высота открытия рта, степень его округлости и т.п. Интеграция признаков производится в систему, построенную при помощи инструментов Hidden Markov Model Toolkit (HTK).

Частота потока аудио-признаков заметно выше частоты потока видео-признаков. В работе рассматриваются вопросы синхронизации признаков аудио и видео между собой с помощью различных видов интерполяции промежуточных значений. Затрагиваются вопросы способа объединения признаковых описаний аудио- и видеоданных (от конкатенации до или после преобразования признакового пространства до нейросетевых подходов к получению общего признакового описания).

Работа выполнена в рамках проекта, поддержанного грантом «УМНИК».

Литература

1. Ahmad B. A. Hassanat Visual Speech Recognition, Speech and Language Technologies. 2014.
2. Potamianos G., Neti C. et al. Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing. MIT Press, 2004.
3. Богданов Д. А. и др. База речевых фрагментов русского языка «ISABASE». Интеллектуальные технологии ввода и обработки информации. 1998. С.74-85.