Распознавание эмоций в человеской речи для мультиязыкового набора данных

Синюков Александр Сергеевич

Студент

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

E-mail: a_sinyukov@cs.yaconnect.com

Научный руководитель *— Глазкова Валентина Владимировна*

Автоматическое распознавание речи является активной областью изучения искусственного интеллекта и машинного обучения, целью которой является создание устройств, взаимодействующих с людьми посредством речи. Речь — это информационный сигнал, который содержит лингвистическую информацию, а также паралингвистическую информацию.

Эмоции являются одним из ключевых примеров паралингвистической информации, которая передается с помощью речи. Проблема классификации эмоций имеет большой потенциал для использования в прикладных отраслях, где существуют системы с интерактивным взаимодействием с пользователем. Решение этой проблемы позволит получать отклик от пользователя естественным образом, не требуя каких-либо дополнительных действий, упрощая и ускоряя взаимодействие между приложением и человеком.

В текущем исследовании была изучена проблема классификации эмоций для набора данных, состоящего из аудиозаписей на разных языках. Решение данной задачи, поможет в создании классификатора, который не будет зависеть от конкретного языка.

Для исследования были выбраны размеченные наборы данных немецкого [1], английского [2] и итальянского [3] языков, которые были объединены в мультиязыковой набор данных. Проведена предварительная обработка записей для улучшения их характеристик: подавление шумов, нормализация громкости, выделение характерных для человеческой речи частот. Для увеличения объема набора данных с целью повышения качества классификации использовалось обогащение данных за счет синтетических образцов, полученных из первоначального набора. К исходным записям с некоторой вероятностью применялись следующие преобразования: дополнение белым гауссовским шумом, ускорение/замедление речи, случайные сдвиги. В дальнейшем аудиозаписи были преобразованы в спектрограммы с помощью оконного преобразования Фурье [4].

Для классификации эмоций предложена и реализована смешан-

ная модель сети глубинного обучения, которая использует сверточные слои для автоматической генерации признаков. Чтобы учитывать изменения речи по времени модель содержит рекуррентный слой. Точность классификации для мультиязыкового набора данных составила 67%. Также проведены эксперименты с данной моделью для кросс-языковой классификации, где один из языков не участвовал в обучении и применялся только для тестирования.

Литература

1. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. F., Weiss B. A database of german emotional speech. // Interspeech, 2005, Vol. 5, P. 1517—1520.
2. Haq S., Jackson P., Edge J. Speaker-dependent audio-visual emotion recognition. // AVSP, 2009, P. 53—58.
3. Costantini G., Iaderola I., Paoloni A., Todisco M. Emovo corpus: an italian emotional speech database. // LREC, 2014, P. 3501–3504.
4. Oppenheim A. V. Speech spectrograms using the fast Fourier transform. // IEEE Spectrum, 1970, Vol. 7, P. 57–62.