УДК 004.934

УЛУЧШЕНИЕ КАЧЕСТВА ТЕКСТОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ ДИКТОРА

Газизуллина А.Р. (Университет ИТМО)
Научный руководитель - к.т.н. Новоселов С.А.
(Университет ИТМО)

Распознавание диктора по индивидуальным акустическим характеристикам его голоса является актуальной задачей рассматриваемой в сфере информационной безопасности, а также процессах, связанных с высоким уровнем защищенности. Голос является уникальным для каждого человека, что делает его идеальным элементом для идентификации. Однако качество современных методов верификации по голосу зависит от множества факторов, таких как длительность звукозаписи, особенности акустической обстановки, частоты дискретизации, качества звукозаписи, индивидуальных характеристик микрофона и записывающего устройства, языка, качества обучающей выборки, намерений говорящих выдать себя за других дикторов. Перечисленные факторы формируют пространство гиперпараметров систем распознавания говорящего по голосу большой размерности, что оправдывает постоянную потребность в улучшения качества верификационных систем.

В настоящей работе рассматривается решение задачи построения модели распознавания диктора по голосу на основе x-векторной системы в фреймворке общего назначения. Нами было принято взять за основу исследований архитектуру X-векторов так как она дает наиболее стабильную картину точности распознавания диктора на разных наборах данных. X-вектора появились в процессе эволюции статистических систем верификации, основанных на модели гауссовых смесей таких как i-вектора. X-векторная модель это метод представления голосового сегмента аудиозаписи в сжатой и в то же время богатой индивидуальными для говорящего признаками форме. Простейшая ее конфигурация представляет собой 3 TDNN (нейронная сеть с временной задержкой) слоя с 512 фильтрами и размерами контекста 5, 5, 7, двумя слоями сверток 1x1, статистическим пуллингом, двумя полносвязными слоями и Softmax слой. Изначально X-вектора были реализованы в kaldi, но данный фреймворк неудобен для исследований, ограничен в возможностях кастомизации. Однако, перенос системы в фреймворки общего назначения является нетривиальной задачей. Это связано с тем что методы, с помощью которых kaldi достигает state-of-the-art результатов, при их воспроизведении для обучения систем в фреймворках общего назначения не дают прироста в качестве. К таким методам можно отнести натуральный градиент, метод ансамблирования моделей, метод формирования батчей. В данной работе мы исследуем различные конфигурации X-векторной модели и рассматриваем влияние различных факторов на деградацию качества при переносе системы из Kaldi в Pytorch.

Разработанная нами система позволяет уменьшить значение EER на 1.5% относительно эквивалентной системы в Kaldi по тестам на закрытой корпоративной базе в телефонном канале. Тесты на телефонной части базы NIST 2019 и на базе VOiCES содержащей записи в микрофонном канале демонстрируют, что рассматриваемая в данной работе модель дает результаты сопоставимые с результатами полученными в Kaldi.