УДК 004.9

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ВОЗРАСТА ЧЕЛОВЕКА ПО ГОЛОСУ

Зорькина А.А. (ИТМО)
Научный руководитель – к.т.н. Волохов В.А. (ИТМО)

Введение. Определение возраста человека по голосу – одна из актуальных задач в области обработки речи. С развитием методов глубокого обучения появилась возможность автоматизировать анализ голосовых характеристик, позволяя создавать системы, способные достаточно точно оценивать возраст говорящего. Данная тема важна для целого ряда практических приложений, включая безопасность, диалоговые системы и маркетинговые исследования.

Основная часть. Задача автоматического определения возраста по голосу может быть сформулирована как задача классификации или как задача регрессии. В первом случае возраст говорящего определяется как один из предопределённых классов (например, детский, взрослый или пожилой), во втором – как непрерывное значение, предсказываемое моделью. Оба подхода находят применение в современных системах анализа речи, однако классификация часто оказывается более устойчивой при ограниченном объёме данных. Современные методы решения этой задачи основаны на глубоком обучении. Первые нейросетевые подходы использовали глубокие сверточные архитектуры, однако в последние годы значительных успехов добились модели, основанные на трансформерах. Эти архитектуры, в том числе предобученные в режиме self-supervised learning (SSL), демонстрируют высокую точность в широком спектре задач, связанных с обработкой речи. SSL-модели, такие как wav2vec 2.0 [1] и HuBERT [2], позволяют извлекать информативные представления из речевого сигнала без необходимости ручной разметки, что делает их особенно ценными для задач с ограниченным количеством размеченных данных. В настоящей работе рассматриваются современные трансформерные архитектуры, в том числе на основе SSL-предобучения, применительно к задаче определения возраста говорящего. Задача формулируется как задача классификации на 3 и более классов. Тестирование проводится на мультиязычных базах, содержащих различные возрастные диапазоны, в том числе детские голоса.

Выводы. В работе были рассмотрены современные трансформерные архитектуры для решения задачи автоматического определения возраста человека по голосу. В частности, рассмотрены модели, предобученные в SSL-режиме, такие как wav2vec 2.0. Эксперименты показали, что применение подобных архитектур позволяет улучшить качество классификации по сравнению с базовым решением на основе ResNet архитектуры.

Список использованных источников:

1. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 12449-12460.
2. Hsu W.N. et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – Т. 29. – С. 3451-3460.