УДК 004.8

АНАЛИЗ РАЗЛИЧИМОСТИ МЯГКИХ И ТВЁРДЫХ СОГЛАСНЫХ НА ОСНОВЕ ЭМБЕДДИНГОВ ФОНЕМ, ПОЛУЧЕННЫХ С ПОМОЩЬЮ HUBERTSOFT

Ананьева А. Д. (ИТМО)

Научный руководитель – кандидат технических наук, Волкова М. В. (ООО “ЦРТ”

Введение. Нейросетевые модели, обучающиеся на больших объёмах неразмеченных данных (так называемый SSL – semi-supervised learning), стали широко применяться в задачах автоматического распознавания и анализа речи [1]. Эти модели позволяют получать эмбеддинги фонем – векторные представления звуковых единиц, содержащие информацию о их акустических и фонетических характеристиках. Эти эмбеддинги могут использоваться для различных задач, включая автоматическое фонетическое транскрибирование, сегментацию речи и улучшение качества синтеза.

Основная часть. Исследование проводилось на данных корпуса CORPRES, содержащего речевые записи с фонетической разметкой. Для каждой фонемы были извлечены ее границы по ручной аннотации, а затем была использована модель HubertSoft [2], которая вычисляла эмбеддинги для каждых 20мс входного сигнала. Полученные векторы усреднялись на протяжении всей длительности фонемы, для формирования единого представления.

Эмбеддинги усреднялись для каждой фонемы, а затем вычислялась косинусная метрика для каждой пары звуков русского языка. Такой подход показал, что модель успешно кодирует некоторую фонемную информацию, хорошо разделяя звуки по месту и способу образования. Однако пары согласных фонем не разделяются по мягкости-твёрдости, косинусные расстояния между эмбеддингами для p и p’ или z и z’ составляют 0,97.

Целью настоящего исследования является реализация и анализ нейросетевого классификатора, способного различать мягкие и твёрдые согласные на основе их эмбеддинговых представлений. В обучающую выборку включались пары палатализованных и непалатализованных согласных (/p/ – /p’/, /t/ – /t’/, /k/ – /k’/ и т. д.), а также анализ проводился на уровне групп: отдельно рассматривались щелевые и смычные согласные.

В ходе работы экспериментально подбирались гиперпараметры, включая количество нейронов в слоях, функцию активации и скорость обучения, а также были рассмотрены различные модификации базовой архитектуры классификатора, для повышения качества

Выводы. Реализован классификатор для мягких и твердых согласных и проведена оценка качества.

Список использованных источников:

1. 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.

2. Van Niekerk B. et al. A comparison of discrete and soft speech units for improved voice conversion //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2022. – С. 6562-6566.