**Разработка модели распознавания акустических признаков акцента в английской речи носителей русского языка на основе сверточных нейронных сетей**

Чимитов Бэлиг Баирович

Студент Дальневосточного федерального университета, Владивосток, Россия

Сверточные нейронные сети CNN (convolutional neural networks) играют большую роль в технологиях распознавания. С появлением огромных репозиториев визуальных данных, а также больших вычислительных мощностей, данный вид нейронных сетей наиболее успешно реализуется в моделях распознавания изображений и видео [Simonyan: 1]. Сверточные нейронные сети имитируют работу человеческого зрения, когда на низком уровне сначала обнаруживаются базовые линии, а затем более сложные контуры [Buduma: 267]. Изначально данные архитектуры использовались для работы с двумерными изображениями. Однако, похожий принцип может быть применим и в других областях, в частности, при работе со звуковым сигналом [Buduma: 333].

При условии, что звуковая информация представлена в виде двумерных спектрограмм, подобно нахождению повторяющихся шаблонов в изображении, сверточные нейронные сети способны выделять нужную информацию для выявления акустических признаков, которые, например, отвечают за общее настроение музыки [Chowdhury: 237]. В данной же работе исследуется способность сверхточных нейронных сетей выявлять акустические признаки акцента в английской речи носителей русского языка.

В качестве материала исследования были использованы данные, свободно предоставляемые корпусом Speech Accent Archive. Архив содержит аудиофайлы, представляющие собой чтение одного и того текста дикторами 177 стран мира. Для обучения нейронной сети по распознаванию русского спикера была отобрана подвыборка, состоящая из 64 записей англоязычных дикторов (Великобритания), и 48 русскоязычных дикторов (Россия и страны СНГ).

В зависимости от быстроты, беглости, уровня владения английским языком и других индивидуальных особенностей дикторов длительность записей сильно разнится (от 18 до 66 секунд). В таких случаях целесообразно разделить аудиозаписи на одинаковые по длине сегменты, что положительно влияет на точность модели и увеличивает количество данных для обучения [Bartelds: 11]. Так, каждый файл был поделен на сегменты (далее образцы речи) длительностью 6 секунд с 3 секундным перекрытием. В результате сегментации размер выборки составляет 878 образцов речи. Из них 702 используются для обучения, 176 – для тестирования модели.

Таким образом, в результате обучения точность модели на тестовых данных достигает 63%. Низкий показатель точности может объясняться несколькими причинами. Во-первых, в исследованиях вариантов английского языка основываются на предположении, что фонетические признаки акцента наиболее выражены у дикторов со средним уровнем знания языка [Zavyalova: 575]. Поскольку в рассматриваемой выборке данных дикторы не дифференцируются по степени владения языком, модель может иметь трудности в распознавании их особенностей. Во-вторых, акцентная речь определяется как «присутствие» одного языка внутри другого и проявляется в степени проникновения структурно-функциональных фонетических признаков родного языка в изучаемый [Zavyalova: 569]. Так, задача классификации переходит в задачу регрессии, когда модель вместо того, чтобы классифицировать спикера, должна численно рассчитать, насколько особенности родного языка выражены в речи на иностранном языке. Таким образом, дальнейшая работа предполагает повышение точности модели через измерение степени «присутствия» структурно-функциональных фонетических признаков русского языка в английской речи русскоязычных носителей.

**Литература**

1. Bartelds M., Richter C., Liberman M., Wieling M. A New Acoustic-Based Pronunciation Distance Measure // Frontiers in Artificial Intelligence, 2020.

2. Chowdhury S., Vall A., Haunschmid V., Widmer G. Towards explainable music emotion recognition: the route via mid-level features // Proceedings of the 20th International Society for Music Information Retrieval Conference, 2019. P. 237–243.

3. Buduma N., Buduma N., Papa J. Fundamentals of Deep Learning. Sebastopol, 2022.

4. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // ICLR, 2015.

5. Zavyalova V.L. Tracing the roots of phonetic variation in East Asian Englishes through loan phonology // Russian Journal of Linguistics. 2020. Vol. 24. № 3. Р. 569–588.