УДК 004.934.5

МЕТОДЫ ПРОСОДИЧЕСКОЙ ОБРАБОТКИ ДЛЯ СИНТЕЗА ЭМОЦИОНАЛЬНОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ

Зорькина А. А. (Университет ИТМО)
Научный руководитель – к.ф.-м.н. Рыбин С.В. (Университет ИТМО)

В данной работе рассматриваются различные подходы к просодической обработке для задачи синтеза эмоциональной речи, основанные на алгоритмах глубокого обучения.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №622281 «Разработка методов и алгоритмов для мультимодального распознавания валентности высказываний и доминантности дикторов в полилогах».

Введение. Просодическая обработка заключается в предании озвучиваемому тексту интонационного оформления и является одной из важнейших составляющих в разработке системы синтеза речи. То, насколько естественно и эмоционально будет звучать искусственная речь во многом определяется именно на этапах просодической обработки, что делает эту задачу особенно актуальной для современных систем синтеза речи (англ. text-to-speech, TTS).

Основная часть. На данный момент существует большое количество различных подходов к извлечению и моделированию просодических характеристик. Одним из первых таких подходов является алгоритм просодической обработки по лингвистическим правилам, однако подобный подход не является масштабируемым и не подходит для задачи синтеза эмоциональной речи, что делает необходимым рассмотрение более актуальных методов, основанных на использовании машинного и глубокого обучения. В данной работе в первую очередь рассматриваются методы, основанные на алгоритмах глубокого обучения - предсказание просодических признаков (границ синтагм, пауз, длительности фонем) с использованием нейронных сетей и добавление их в модели TTS для получения спонтанной речи [1], подходы, основанные на излечении просодии из аудио [2], а также подходы к учету просодической информации с использованием предварительно обученных языковых моделей [3].

Выводы. В ходе работы был проведен аналитический обзор актуальных статей, раскрывающих описанные методы просодической обработки и позволяющий рассмотреть их с точки зрения возможности применения в задаче синтеза эмоциональной речи.

Список использованных источников:

1. Yan Y. et al. Adaspeech 3: Adaptive text to speech for spontaneous style //arXiv preprint arXiv:2107.02530. — 2021.
2. Wang Y. et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis //International Conference on Machine Learning. – PMLR, 2018.
3. Hayashi, T., Watanabe, S., Toda, T., Takeda, K., Toshniwal, S., Livescu, K. (2019) Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis. Proc. Interspeech. — 2019.