УДК 004.934.5

ГЛУБОКИЕ ПОРОЖДАЮЩИЕ МОДЕЛИ В АДАПТИВНОМ СИНТЕЗЕ РЕЧИ

Кочарян А.М. (Университет ИТМО)

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

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. Последние исследования в области TTS показывают, что качество современных моделей адаптивного синтеза уже сопоставимо с реальной человеческой речью. Во многом, этого удалось добиться путем внедрения end-to-end TTS-моделей, использующих data-driven подходы. В данной работе рассматриваются применяемые к TTS-системам подходы генеративного моделирования, улучшающие выразительные способности глубоких архитектур.

Основная часть. В общем случае, порождающие модели позволяют вычислить вероятностные модели совместного распределения скрытых переменных и данных. В одной из наиболее успешных моделей TTS – VITS [1], для соединения акустической модели и вокодера используется вариационный автокодировщик (VAE), что позволяет эффективно проводить end-to-end обучение. Также, для генерации высококачественных речевых сигналов были применены нормализующие потоки к априорному распределению латентного представления и состязательное обучение на временной области волны. Такой подход в связке со стохастическим предсказанием длительности фонем обеспечивает вариативность синтезированной речи, что можно использовать в задачах аугментации данных. Предполагается, использование языковых моделей позволит решить проблему предобработки входных текстовых данных. На основе архитектуры VITS была разработана YourTTS [2], позволившая получить SOTA-результаты в области multi-language zero-shot-TTS (ZS-TTS) и использующая текст в качестве входных данных вместо фонем.

Также, одной из перспективных направлений в области глубокого порождающего моделирования является использование концепции диффузионного вероятностного моделирования. Так, модель Grad-TTS достигает SOTA-результатов в one-shot-many-to-many-TTS [3].

Выводы. Использование генеративного моделирования в адаптивном TTS позволило достичь качественных результатов как в естественности генерируемой речи, так и в ее сходстве с диктором. Рассмотренные в работе подходы обеспечивает масштабируемое решение задачи клонирования голоса с хорошей производительностью.

Список использованных источников

1. Jaehyeon Kim, Jungil Kong, Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech // arXiv preprint arXiv:2106.06103. – 2021.
2. Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, Moacir Antonelli Ponti. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. Speech // arXiv preprint arXiv:2112.02418. – 2022.
3. Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, Jiansheng Wei. Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme // arXiv preprint arXiv:2109.13821. – 2022.