0.1. *Болдаков В.С.* Восстановление аудиосигнала из скрытых представлений глубоких нейронных сетей

В настоящее время широкое распространение получили глубокие нейронные сети обработки аудиосигнала общего назначения, предназначенные для решения широкого спектра задач: распознавание речи, детекция ключевых слов, разделение речи дикторов, идентификация диктора. Например, низкие показатели WER (word error rate) получила нейронная сеть WavLM [1], обученная исследователями Microsoft. Из высокой эффективности подобных моделей при решении вышеуказанных задач можно сделать вывод о качестве и информативности скрытых представлений, полученных при помощи данных моделей.

Следующим шагом в использовании скрытых представлений является решение двух других известных задач: улучшение качества аудиосигнала и воспроизведение оригинального аудиосигнала из скрытых представлений. В данной работе исследуется возможность восстановления аудиосигнала из скрытых представлений на основе генеративно-соревновательного подхода обучения нейронных сетей.

Основной мотивацией для использования скрытых представлений сигнала вместо хорошо зарекомендовавших себя подходов с использованием спектрограмм является высокая устойчивость скрытых представлений к различного рода шумам и помехам. В данной работе предлагается рассмотреть возможность восстановления аудиосигнала из скрытых состояний модели WavLM, предоставляющей скрытые состояния из множества $\mathbb{R}^{768}$ с размером окна 25 мс, при помощи расширения архитектуры восстановления аудиосигнала из мелспектрограмм HiFi-Gan [2], обученной при помощи генеративно-состязательного подхода. Функция потерь для обучения генератора имеет следующий вид:

$$L = E[D(G(\boldsymbol{h})) + (mel(G(\boldsymbol{h})) - mel(\boldsymbol{x}))^2],$$

где $\boldsymbol{h}$ — скрытое состояние сигнала, полученное из WavLM, D — дискриминатор, G — обучаемый генератор, mel — функция вычисления мелспектрограммы, $\boldsymbol{x}$ — истинный аудиосигнал.

В результате данной работы получено решение, позволяющее восстанавливать аудиосигнал из скрытых характеристик WavLM. Данный подход может быть использован для восстановления сигнала из более широкого спектра подобных архитектур, например из wav2vec 2.0 [3]. Также данный подход может быть использован для улучшения качества аудио, увеличения частоты дискретизации. Возможно использования в качестве второго этапа, вокодирования, в задаче синтеза речи.

Научный руководитель — к.т.н. Ракитский А. А.

Список литературы

[1] Chen S., Wang C., , Chen Z., et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing // IEEE Journal of Selected Topics in Signal Processing. 2022. Vol. 1. N. 1. P. 1–14.

[2] Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis // Proc. Intern. Conf. «2020 Conference on Neural Information Processing Systems». Vancouver, BC, Canada: NeurIPS, 2020.

[3] Ren Y., Hu C., Tan X., et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. [Электронный ресурс]. URL: https://arxiv.org/abs/2006.11477 (дата обращения 12.09.2022).