УДК 004.934.5

АДАПТАЦИЯ И РЕДАКТИРОВАНИЕ РЕЧИ С ПОМОЩЬЮ ГЕНЕРАТИВНОГО И САМОКОНТРОЛИРУЕМОГО ОБУЧЕНИЯ

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

Введение. С использованием Voice Cloning технологий можно получать речь, которую сложно отличить от реальной человеческой, однако существует ряд приложений в различных сферах аудиоконтента, в которых требуется персонализация и внесение изменений в синтезированную речь, уже полученной TTS-системой, в частности, удаление и замена нежелательных слов, коррекция произношения и т.д. На данный момент такие преобразования тяжело достигаемы современными TTS-системами, поэтому качество моделей в задаче остается несовершенной, а данная сфера пока еще актуальна и малоизучена.

Основная часть. Основная идея предлагаемой системы состоит в автоматической разметке контент-векторов, получаемых на выходе HuBERT (Hidden-Unit BERT) [1], на графемы, а также их кластеризации. Это позволит аппроксимировать вектора каждой графемы ее центроидом для последующей стабильной трансляции входной пользовательской последовательности в необходимые контент-вектора перед их подачей в акустическую модель. В основе системы стоит модификация VITS[2], которая использует технологию Transfer learning и позволяет генерировать высококачественные аудиозаписи благодаря использованию механизма нормализующих потоков и состязательного обучения во временной области сигнала, а при обучении использует оценку ELBO (Evidence Lower Bound). В свою очередь модель HuBERT во время предварительного обучения использует подход self-supervised learning, что позволяет выявлять полезные репрезентации речи на больших объемах неаннотированных данных. Таким образом, модель HuBERT хорошо отражает фонетические характеристики речи, используя технологию внимания для анализа контекста перед оценкой выходного представления сети [3].

Выводы. Модификация VITS с использованием модели HuBERT является более высококачественной и надежной для адаптивного синтеза. Данную систему планируется использовать для аугментации в различных задачах анализа аудиоконтента. В предложенную разработку также планируется добавить возможность изменения темпа речи, эмоций диктора, а также F0 в таких задачах как расстановка интонаций и логическое выделение.

Список использованных источников

1. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units // arXiv preprint arXiv:2106.07447. – 2021.
2. Jaehyeon Kim, Jungil Kong, Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech // arXiv arXiv:2106.06103. – 2021.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need // arXiv preprint arXiv:1706.03762. – 2017.