УДК 004.89

РАЗРАБОТКА МОДЕЛИ ПРЕОБРАЗОВАНИЯ ГОЛОСА ДЛЯ ГЕНЕРАЦИИ РЕЧИ НА РУССКОМ ЯЗЫКЕ

Кузьмин А.Д. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Шуранов Е.В.
(ИТМО)

Введение. Задача преобразования голоса (англ. Voice Conversion) – это задача, целью которой является генерация нового речевого сигнала. Такой речевой сигнал создаётся на основе двух других (называемых контентным и референсным), в котором сохранялась бы лингвистическая информация из контентного сигнала, а идентичность говорящего (тембр, манера говорения и т.п.) была бы идентичной той, что представлена в референсном. Реализация модели преобразования голоса открывает возможности для разработки систем коррекции нарушений речи, систем эмоциональной окраски речи, устройств для добавления различных вокальных эффектов, а также приложений для преобразования акцентов, способных помочь в изучении языка. Также стоит отметить, что для русского языка существует большое количество открытых данных, нацеленных на создание модели синтеза речи для одного спикера. Однако для задачи клонирования и преобразования голоса требуются наборы чистой речи, озвученные множеством различных голосов.

Основная часть. В настоящее время существует значительное количество подходов к решению задачи преобразования голоса, каждый из которых обладает своими преимуществами и недостатками. В данной исследовательской работе для решения поставленной задачи предлагается использование генеративной нейросетевой модели, основанной на архитектуре вариационного автоэнкодера (Variational Autoencoder, VAE). При этом в качестве отправной точки принимается метод и модель для синтеза речи VITS [1]. Выбор данного подхода обусловлен его высокой производительностью, конкурентным качеством генерируемой речи и относительно небольшим количеством обучаемых параметров модели, что является критически важным аспектом при решении задачи преобразования голоса с использованием небольшого объёма данных. Адаптация модели VITS к задаче преобразования голоса происходит путём замены модуля кодировки текста (англ. Text encoder) на предварительно обученную модель архитектуры HuBERT [2], полученную с помощью алгоритмов самообучения (англ. self-supervised learning, SSL) и натренированную среди прочего на кодирование лингвистической информации из речевого аудиосигнала.

Выводы. В данной работе описан процесс разработки модели преобразования голоса для генерации речи на русском языке с использованием моделей VITS и HuBERT, а также проведена оценка качества генерируемой речи.

Список использованных источников:

1. Kim J., Kong J., Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech // the 38th International Conference on Machine Learning. – 2021. – Vol. 139 – С. 5530–5540. doi:10.48550/arXiv.2106.06103
2. Hsu W., Bolte B., Hubert Y., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units – 2021. doi:10.48550/arXiv.2106.07447