УДК 004.934.2

АНАЛИЗ АВТОМАТИЧЕСКИХ ПОДХОДОВ ДЛЯ ОЦЕНКИ КАЧЕСТВА СИНТЕЗИРУЕМОЙ РЕЧИ (TTS)

Хо Куанг Чунг (Университет ИТМО)
Научный руководитель – к.ф.-м.н, Рыбин С. В.
(университет ИТМО)

Введение. В последнее время в области синтеза интанационной речи было достигнуто зачительное продвижение благодаря применению области машинного обучения – глубокому обучению. Появилось большое число интегральных моделей, способных генерировать синтезированную речь высокого качества. Под «качеством» синтезированной речи будем понимать ее естественность – условную величину, характеризующую субъективную оценку звучания синтезированной речи по сравнению со звучанием естественной речи

Однако открытым остается важный вопрос объективной (автоматической) оценки естественности синтезированной речи. Целью исследования является поиск и анализ перспективных автоматических интегральных подходов к оценке качества синтезированной речи в задаче TTS.

Основная часть. В настоящее время наиболее надежным методом оценки качества речи является проведение субъективных тестов, в которых качество речи оценивается по среднему баллу мнений (Mean Opinion Score – MOS) группы слушателей. Однако проведение таких тестов дорого, требует много времени и выского профессионализма экспертов. Поэтому в последнее время появляется все больше технологий для автоматического предсказания этой оценки.

Здесь, как наиболее перспективные подходы, можно выделить построение моделей на основе глубокого обучения. Следует отметить, что результаты сравнения моделей существенно зависят и от диктора и от акустических характеристик записи. Кроме того не совсем корректно проводить прямое сравненение результатов, так-как обучение часто проходит на разных наборах данных. Частичным решением являются оценки естественности на данных из речевых конкурсов: Blizzard Challenge, Voice Conversion [1]. Тем не менее, кажется, что такого количества данных все еще недостаточно, особенно для таких языков, как, например, русский.

В качестве направлений развития в этом направлении можно указать нейронную сеть CNN-LSTM с mel-спектрограммой, модели с использованием смещенной подсети для компенсации субъективности слушателя и модели на основе самонаблюдаемых репрезентаций (Self-supervised Representations - SSL). В частности, результаты моделей на основе SSL показали наиболее высокую эффективность для решения задачи оценки речи [2].

Выводы. Для решения задачи автоматической оценки естественности речи наибольший интерес предстваляюти модели на основе SSL. Этот подход, безусловно, будет перспективным направлением исследований в самом ближайшем будущем.

Список использованных источников:

1. Erica Cooper, Junichi Yamagishi, How do Voices from Past Speech Synthesis Challenges Compare Today? // 11th ISCA Speech Synthesis Workshop. – 2021.
2. Erica Cooper, Wen-Chin Huang, Tomoki Toda, Junichi Yamagishi, Generalization Ability of MOS Prediction Networks // IEEE International Conference on Acoustics, Speech and Signal Processing. – 2021.