УДК 004.934.5

АУГМЕНТАЦИЯ АУДИОДАННЫХ НА ОСНОВЕ ТЕХНОЛОГИИ СИНТЕЗА ЭМОЦИОНАЛЬНОЙ РЕЧИ МЕТОДАМИ ГЛУБОКОГО ОБУЧЕНИЯ

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

Введение. Синтез речи, т. н. преобразование текста в речь (англ. text-to-speech, TTS), является современной технологией, предназначенной для решения широкого круга задач по человеко-машинной коммуникации и реализации систем голосовых помощников. Данная область находится на стыке ряда научных дисциплин, и сочетает в себе, в первую очередь, акустику, лингвистику и цифровую обработку сигналов [1].

Исторически основным вызовом синтеза была четкость и разборчивость синтезированной речи [2]. С развитием технологий и почти полным переходом актуальных решений на методы, основанные на глубоких нейронных сетях, в приоритете оказалось придание синтезированной речи естественности, выразительности, и, в частности, создание «эмпатичных» голосовых ассистентов [3].

Данная работа ведется в рамках НИРМА №622281 по теме «Разработка методов и алгоритмов для мультимодального распознавания валентности высказываний и доминантности дикторов в полилогах». Рассматриваемые в обзоре методы предполагается использовать с целью аугментации данных, так как доступные датасеты эмоциональной речи достаточно малочисленны. Данные на русском языке, более того, отсутствуют. В рамках НИРМА планируется сбор датасета на русском языке, и дальнейшее его расширение методами аугментации, связанными с экспрессивным синтезом.

Основная часть. Почти все опубликованные за последние 5 лет модели и методы можно условно разделить на две группы, связанные с голосовым преобразованием или непосредственно синтезом. Аугментация данных, соответственно, происходит либо с изменением разметки при голосовом преобразовании (с изменением интонации исходной записи для создания дополнительных образцов других классов), либо с сохранением разметки при использовании методов синтеза с изменением характеристик голоса (громкость, основной тон, скорость речи), но зафиксированной интонацией, заданной переменными стиля.

Голосовое преобразование (англ. voice conversion) подразумевает наличие двух аудиозаписей, одна из которых содержит целевое высказывание и голос, а вторая – целевую интонационную окраску. Целью голосового преобразования является сохранение семантической (содержательной) части голосового сообщения, равно как и его паралингвистических характеристик, при одновременном изменении другой характеристики сообщения, в рассматриваемом случае – эмоциональной или интонационной.

В ходе работы были рассмотрены одиннадцать опубликованных моделей, которые были разделены на три группы по типу базовой архитектуры: модели последовательностей (англ. sequence-to-sequence); модели на вариационном автоэнкодере; модели на вариативно-состязательных сетях (CycleGAN, CycleTransGAN, StarGAN).

Синтез, преобразование текста в речь, в общем случае описывается тремя стадиями -— обработка текста, акустическое моделирование, вокодер. Таким образом, на первом этапе создаётся некоторое фонематическое представление текста, на втором на основании этого представления формулируется спектрограмма, на третьем – генерируется звук. Для решения задачи экспрессивного синтеза в рамках разных подходов добавляется дополнительный блок, связанный с обработкой текста и/или акустическим моделированием, где в явном или скрытом виде формируется отдельный набор признаков, отвечающий за эмоциональную окраску.

В работе рассмотрены шесть опубликованных моделей, которые также были разделены на три группы по типу базовой архитектуры: модели на вариационном автоэнкодере; модели,

использующие заданные скрытые переменные стиля, модели с автоматическим определением скрытых переменных стиля.

Выводы. В ходе работы был проведен сравнительный анализ методов экспрессивного синтеза речи с точки зрения их применимости для аугментации в рамках НИРМА. Сравнение проводилось по следующим критериям: доступность исходного кода, гибкость архитектуры и оценка качества синтезированной экспрессивной речи на данных на английском языке. В результате был получен перечень архитектур, подходящих для решения задачи аугментации. В итоговый список вошли архитектуры, построенные на модели последовательностей [4], генеративной модели CycleGAN [5], модели синтеза на основе вариационного автоэнкодера VITS [6].

Список использованных источников:

1. Y. Ning. A Review of Deep Learning Based Speech Synthesis / Y. Ning [et al.] // Applied Sciences. – 2019. – Vol. 9. – № 19. – P. 4050.
2. Калиев Арман, Рыбин Сергей Витальевич Синтез речи: прошлое и настоящее // КИО. 2019. №1.
3. Emotional Speech Synthesis for Companion Robot to Imitate Professional Caregiver Speech / T. Homma [и др.] arXiv:2109.12787 [cs, eess]. – arXiv, 2021.
4. Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training. Limited Data Emotional Voice Conversion Leveraging Text-to-Speech / K. Zhou, B. Sisman, H. Li arXiv:2103.16809 [cs]. – arXiv, 2021.
5. Emotional Voice Conversion with Cycle-consistent Adversarial Network / S. Liu, Y. Cao, H. Meng arXiv:2004.03781 [eess]. – arXiv, 2020.
6. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech / J. Kim, J. Kong, J. Son arXiv:2106.06103 [cs, eess]. – arXiv, 2021.