Неавторегрессионные методы синтеза речи и их модификации

Филимонов Владислав Аскольдович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ifilin45@gmail.com

Научный руководитель *— Дьяконов Александр Геннадьевич*

Современные подходы синтеза речи разбивают исходную задачу на две независимо решаемых подзадачи:

- преобразование последовательности символов $(x_1, \ldots, x_{T_1})$ в последовательность аудио признаков $(f_1, \ldots, f_{T_2})$;
- преобразование последовательности аудио признаков в аудио сигнал.

Первая подзадача хорошо решается авторегрессионными методами, которые моделируют распределение на текущем выходе f_t при условии уже сгенерированной последовательности $(f_1, \ldots, f_{t-1})$:

$$p_{\text{AR}}(F|X;\theta) = \prod_{t=1}^{T_2} p(f_t|f_{1:t-1}, x_{1:T_1}).$$

Однако такой подход при тестировании требует T_2 запусков модели для получения последовательности длины T_2, что является очень затратным по времени.

При неавторегрессионном подходе отдельно моделируются продолжительности $(d_1, \ldots, d_{T_1})$ символов $(x_1, \ldots, x_{T_1})$, таким образом, что $\sum_{t=1}^{T_1} d_t = T_2$, где T_2 - длина выходной последовательности:

$$p_{\text{NA}}(F|X) = p(f_{1:T_2}|x_{1:T_1}, d_{1:T_1}) \cdot p(d_{1:T_1}|x_{1:T_1}).$$

Неавторегрессионные модели способны сгенерировать выходную последовательность за один запуск, но требуется дополнительное исследование для обобщения уже известных методов повышения качества авторегрессионных моделей и разработки новых методов, специфичных для неавторегрессионных моделей.

В качестве неавторегрессионной модели рассматривается модель FastPitch [1] и предлагается ряд модификаций, позволяющих:

- обучаться на данных от разных дикторов, используя обучаемые векторные представления для каждого диктора и оценку скорости речи speaking rate [2];

- эффективнее строить необходимые для обучения продолжительности, используя Montreal Forced Aligner [3];

- контролировать скорость речи как всей синтезируемой фразы, так и отдельных букв. Для контроля скорости всей фразы предлагается изменять оценку speaking rate [2], а для контроля скорости отдельных букв корректировать предсказанные моделью продолжительности;

- использовать для обучения записи, содержащие небольшие шумы или паузы в начале и в конце (например, включение микрофона). Для этого вводятся специальные символы, которые соответствуют этим шумам, их наличие и продолжительности оценивается с помощью Montreal Forced Aligner [3];

- произносить сложные слова, используя их фонемное представление вместо буквенного. Для построения этого представления использовался словарь CMUDict [4] и система правил построения фонем для слов, отсутствующих в нем.

Метод с описанными модификациями превосходит авторегрессионный метод Tacotron 2 [5] в скорости работы в 121 раз на графическом ускорителе nVidia GeForce GTX 2080 и позволяет генерировать более естественно звучащие аудио, чем оригинальный метод FastPitch [1].

Литература

1. Lancucki A. et al. FastPitch: Parallel Text-to-speech with Pitch Prediction // arXiv preprint arXiv:2006.06873, 2020.
2. Bae. J. et al. Speaking Speed Control of End-to-End Speech Synthesis using Sentence-Level Conditioning // arXiv preprint arXiv:2007.15281, 2020.
3. McAuliffe M. et al. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Interspeech, Stockholm, Sweden, 2017, pp. 498-502.
4. The CMU Pronouncing Dictionary `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`
5. Shen J. et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, ICASSP, Calgary, AB, Canada, 2018, pp. 4779-4783.