УДК 004.93

Применение генеративных состязательных сетей в синтезе речи

А. Калиев
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к. ф.-м. н., доцент, С.В. Рыбин
(Университет ИТМО, г. Санкт-Петербург)

Исследования выполнены за счет стартового финансирования университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей»

Нейронные сети широко применяются в технологии синтеза речи. Можно выделить несколько явных проблем, где применение нейронных сетей сделали весомый прорыв в этой области. Первое, и в тоже самое время самое несложное, это применение НС для решения проблемы предсказание длины фонемы. Второе, это применение для решения проблемы акустического моделирования речи каждой фонемы предложения или фразы. В обоих этих случаях, НС сделали большой прорыв. В конечном счете улучшив качество синтезированной речи. Однако стоить отметить применение НС имеет свои серьезные недостатки. Среди них можно выделить следующие, как наиболее проблемными, это требование больших данных для обучения НС и требование к вычислительным мощностям. Решение обоих этих проблем являются дорогостоящими в финансовом плане.

Однако, не решенной проблемой остается замена вокодера на НС. Создание НС способного генерировать естественную человеческую речь. Сложность создание такой модели, заключается в оценочных функциях ошибки НС. Современные применяемые оценочные функций ошибки для обучения НС не способны качественно отделить шипящие звуки, как например «ш» или «ж», от простого шума. Как следствие ученые вынужденные применять вокодеры, такие как WORLD [1] или STRAIGHT [2] для генерации речевого сигнала.

Самой распространенной функцией ошибки НС для акустического моделирования и предсказания длины фонем, основан на вычислении среднеквадратического отклонения. К сожалению, такой подход несет явные сложности в генерации человеческой речи. Очевидным является проблема сглаженности генерируемого речевого сигнала. Функция среднеквадратичного отклонения способствует генерации сглаженного сигнала речи. Такая сканированная речь легко определяется слушателями как не естественная.

Решить эту проблему призваны генеративные состязательные сети (ГСС). ГСС состоят из двух НС. Одна генерирует последовательность акустических параметров, вторая проверяет их на естественность распределения. Первая НС называется генератор, вторая дискриминатор. Во время обучения дискриминатору поточно подается два вектора акустических параметров. Первая получена из базы данных записанных речей, вторая сгенерирована с помощью генератора. Дискриминатору ставится задача определить какая из них естественная и какая из них фальшивая. В тоже время генератору ставится задача генерировать распределение акустических параметров так, что дискриминатору было трудно определить их фальшивость.

Литература

1. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications // IEICE transactions on information and systems. –2016. –V. E99-D. PP. 1877-1884
2. Kawahara H., Masuda-Katsuse I., de Cheveigne A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds // Speech Communication. – 1999. –V. 27. PP. 187-207