УДК 004.934.5

МНОГОПОЛОСНЫЙ НЕЙРОННЫЙ ВОКОДЕР ДЛЯ СИСТЕМЫ СИНТЕЗА РЕЧИ

Свищев А.Н. (Университет ИТМО)
Томилов А.А. (ООО «ЦРТ-Инновации»)
Научный руководитель – к. ф.-м. н. Рыбин С.В.
(Университет ИТМО)

В докладе предлагается оригинальный способ распараллеливания и ускорения генерации речевого сигнала из его акустических признаков в системах синтеза речи на основе нейронных сетей.

Современные подходы в области параметрического синтеза речи основываются на применении глубинных нейронных сетей. Суть задачи сводится к неоднозначному преобразованию текстовой последовательности к превосходящему ее на несколько порядков по количеству информации речевому сигналу. При этом промышленные решения ограничены строгими требованиями к производительности и к совместимости с имеющимся вычислительным оборудованием. Вычислительно требовательные сквозные нейросетевые подходы в этом случае уступают место декомпозиции задачи, в которой выделяют два ключевых этапа: генерацию промежуточного представления речевого сигнала низкого разрешения и его последующее преобразование в конечный сигнал высокого разрешения. На втором этапе в качестве основного решения, позволяющего добиваться высокого перцептивного качества синтезируемого речевого сигнала, применяются нейронные вокодеры.

По способу генерации сигнала нейронные вокодеры можно разделить на два семейства: последовательные и параллельные. В последовательных нейронных вокодерах решается задача авторегрессии — моделирование условного распределения каждого отсчета сигнала на основе предыдущих отсчетов, акустических признаков и параметров модели. Такой процесс подразумевает синтез конечного результата шаг за шагом. Это позволяет добиваться высокого перцептивного качества, но накладывает ограничения на производительность и не позволяет эффективно отображать вычисления на вычислительные ресурсы. Напротив, параллельные нейронные вокодеры позволяют добиваться высокой утилизации мощностей используемого оборудования, но синтезируют сигнал менее приемлемого перцептивного качества. Так же большинство представителей этого семейства нейронных вокодеров для эффективной работы требуют высокопроизводительных графических вычислителей.

Одной из определяющих характеристик цифрового сигнала является частота дискретизации. Согласно теореме Котельникова, именно она определяет частотную однозначность содержащейся в спектре такого сигнала информации. С этой точки зрения, чем больше частота дискретизации — тем шире полоса однозначного сигнала и выше его перцептивное качество, если мы говорим о речевом аудиосигнале. Можно разделить исходный сигнал на несколько сигналов во временной области так, чтобы каждый содержал спектральную информацию из своего участка спектра, при этом для каждого можно кратно уменьшить частоту дискретизации.

Разложение состоит из двух этапов: свертки с фильтром и децимации. Восстановление сигнала из частотных подобластей также состоит из двух этапов: интерполяция и фильтрация. В зависимости от ядра, с которым производиться свертка на этапе разбиения на частотные подобласти можно выделить два основных типа: квадратурные зеркальные фильтры (QMF) и сопряженные зеркальные фильтры (CMF). Оба типа ядер позволяют произвести идеальную реконструкцию сигнала, добавив лишь временной сдвиг, но второй имеет большую вариабельность по параметрам.

В этой работе предлагается обучать последовательный нейронный вокодер задаче синтезирования компонент разложенного сигнала. Это позволяет на каждом шаге

авторегрессии вычислять по одному значению из каждой частотной подобласти, что кратно сокращает общее количество шагов синтеза и позволяет добиться большего параллелизма в вычислениях.

В работе продемонстрировано, что такой подход не только осуществим, но и позволяет добиться лучшего перцептивного качества конечного речевого сигнала. Также показано, что средние вычислительные затраты на каждый отсчет конечного сигнала снижаются с увеличением числа компонент разбиения. При этом последовательность вычислений становится более эластичной для отображения на вычислительные ресурсы.

Предлагаемый подход частично устраняет недостатки последовательных нейронных вокодеров без кардинальных изменений их архитектур и процесса обучения. Это позволяет рекомендовать его к внедрению в соответствующие системы синтеза речи.