Модель WaveNet для генерации звука

Шестакова Анна Николаевна

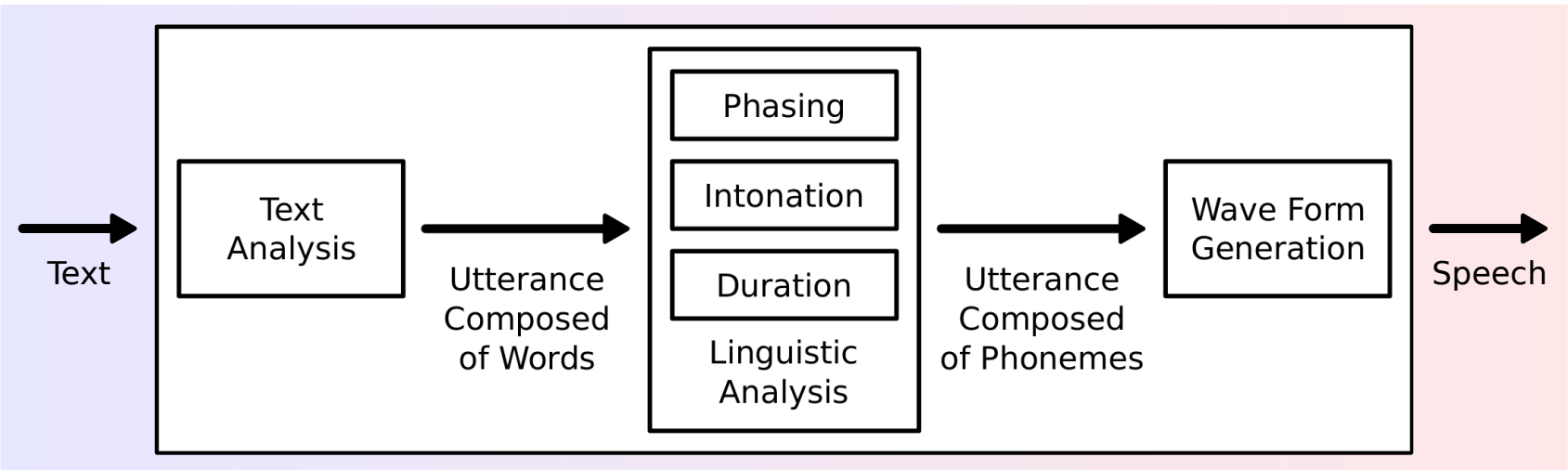
МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

17 мая 2018 г.

Постановка задачи

TTS (text-to-speech) — процесс синтезирования или генерации речи.

Примеры использования: Apple's Siri, Microsoft's Cortana, Amazon Alexa и Google Assistant.



Донейросетевые подходы

- concatenative TTS: фразы комбинируются с помощью большой базы отдельных фрагментов речи, записанных одним человеком.
- параметрический TTS: информация, необходимая для создания речи хранится в виде параметров модели. Такие модели обычно генерируют звук, прогоняя выходной сигнал через специальные обработчики, называемые вокодерами.

WaveNet

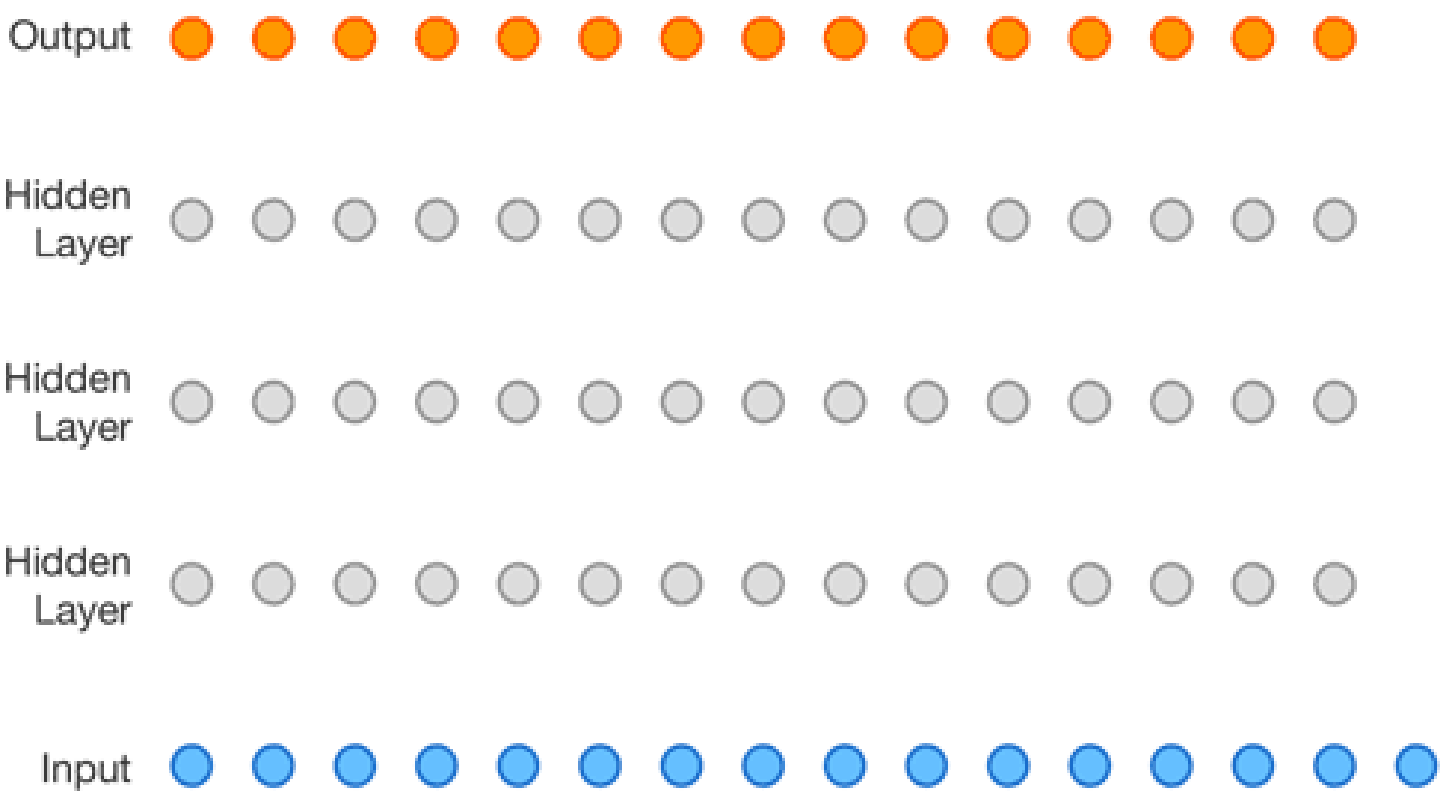
WaveNet — свёрточная нейронная сеть, предложенная компанией DeepMind в 2016 году.

Вероятность формы волны $x = \{x_1, ..., x_T\}$ получается следующим образом:

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, .., x_{t-1})$$

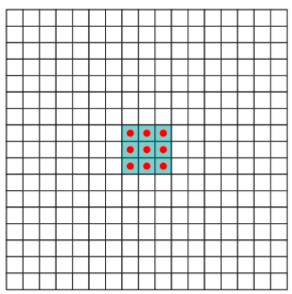
Как и в PixelCNNs условные вероятности моделируются стеком свёрточных слоев. В сети нет pooling слоев и выход модели имеет ту же размерность, что и вход.

Causal Convolutions

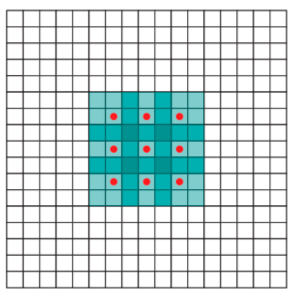


Dilated convolution

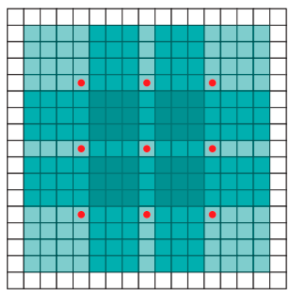
Расширенная свёртка (dilated convolution) — свёртка, где фильтр применяется к области больше его длины, пропуская входные значения с определенным шагом.



(**a**) 1-dilated convolution.



(**b**) 2-dilated convolution.



(**c**) 4-dilated convolution.

Softmax

Для моделирования вероятностей $p(x_t|x_1, ...x_{t-1})$ можно использовать смеси распределений:

$$p(x|t) = \sum_{i=1}^{m} a_i \phi_i(x|t)$$

Однако из опыта использования PixelNets известно, что softmax распределение обеспечивает лучшее качество.

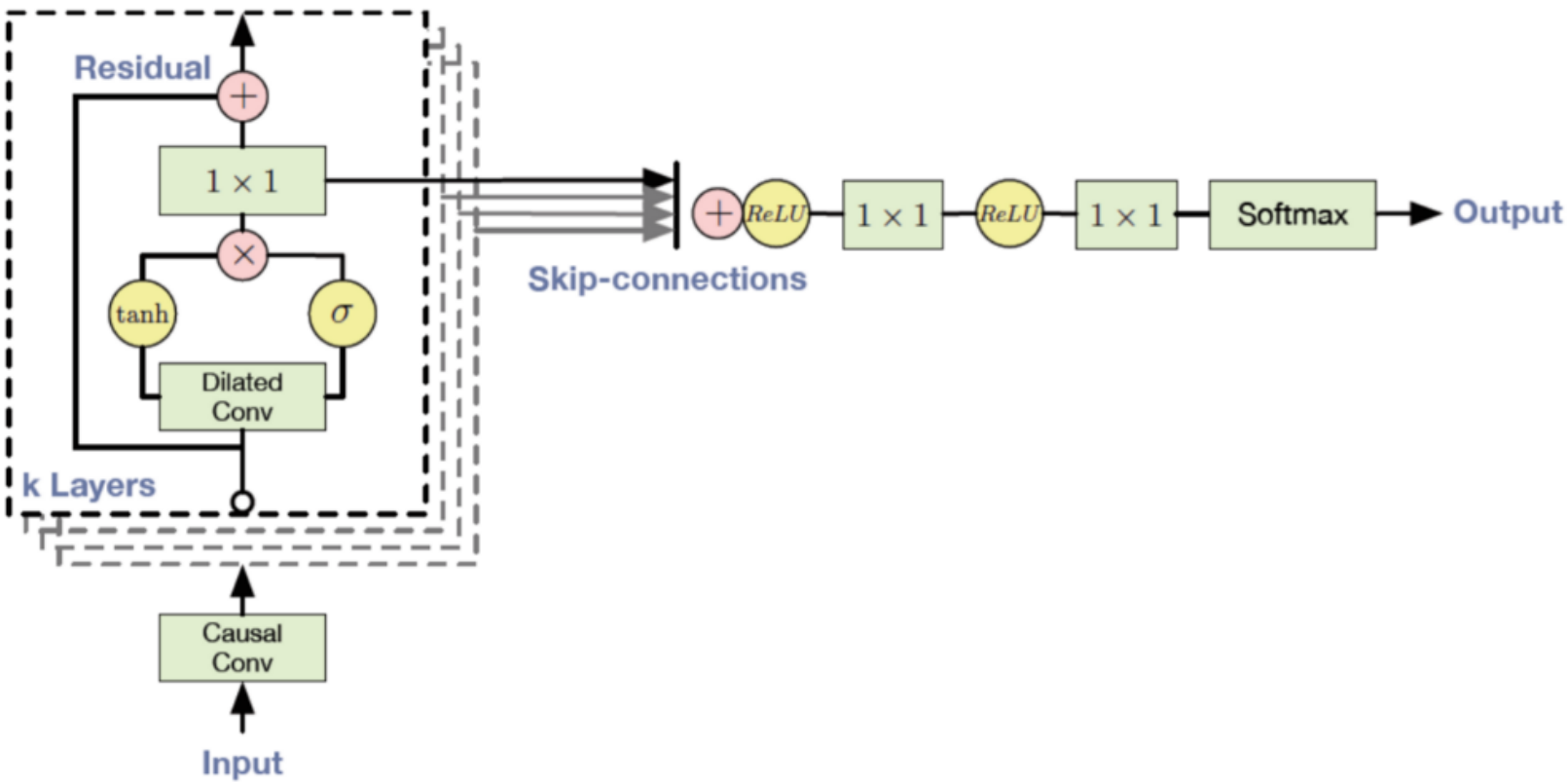
Функция активации

Используемая активация:

$$z = \tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x),$$

где $*$ — оператор свёртки, $\odot$ — поэлементное умножение, $\sigma()$ — сигмоидная функция, k — номер слоя, W — обучаемый фильтр свёртки.

Архитектура WaveNet



WaveNet

Если дан дополнительный вход h как условие, WaveNet способен моделировать условное распределение $p(x|h)$ аудио по этому входу.

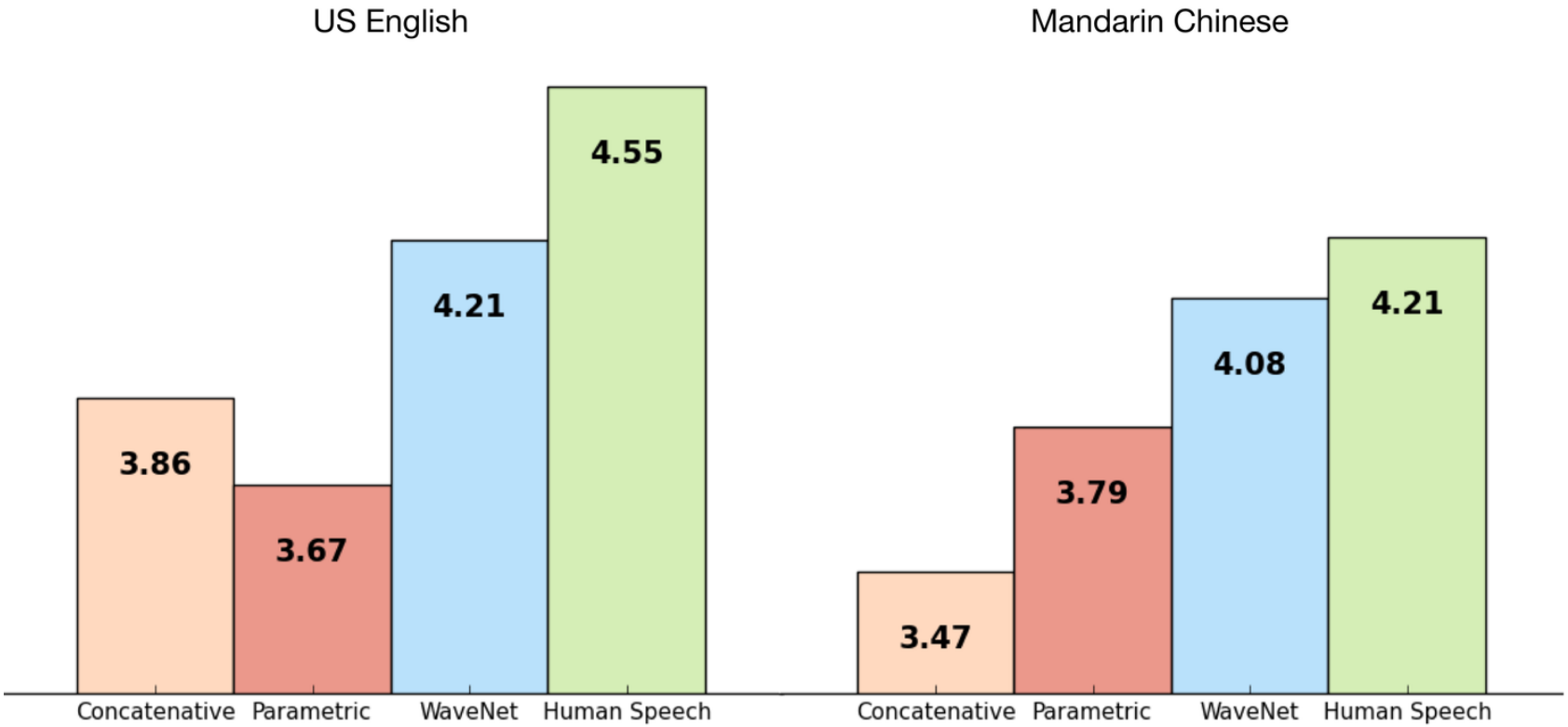
Тогда вероятность формы волны $x = \{x_1, ..., x_T\}$ примет следующий вид:

$$p(x) = \prod_{t=1}^{T} p(x_t|x_1, .., x_{t-1}, \mathbf{h})$$

Результаты

MOS (Mean Opinion Scores) — это стандартный способ делать субъективные тесты качества звука.

В тесте были использованы 100 тестовых предложений и собрано более 500 оценок.



Использование WaveNet

WaveNet использовался для генерации голосов Google Assistant для американского английского и японского языков на всех платформах Google.

WaveNet значительно сократила количество аудиозаписей, необходимых для создания речевой модели.

Ссылки

https://arxiv.org/pdf/1609.03499.pdf
https://deepmind.com/blog/wavenet-generative-model-raw-audio/
https://habr.com/company/Voximplant/blog/309648/