WaveNet: A Generative Model for Raw Audio

Ирина Понамарева

Higher School of Economics

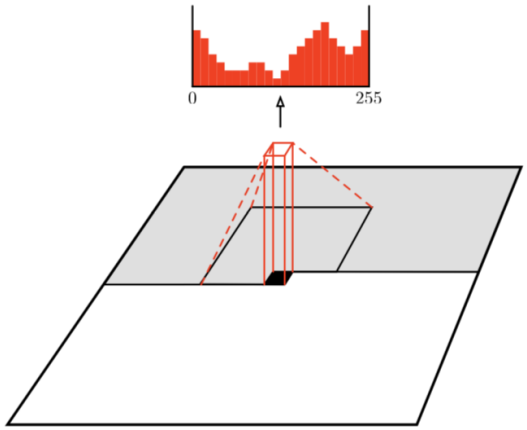
26 апреля 2019 г.

Overview

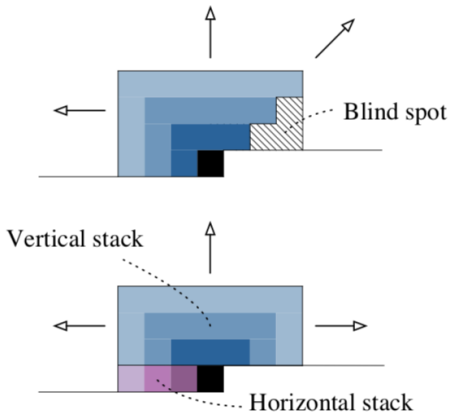
Как моделировать звук?

- Звук делится на интервалы, кодируется в одно из 16 (32, 64) значений
- $x = \{x_i\}_{i=1}^{T}$ — звук как временной ряд
- $p(X) = \prod_{i=1}^{T} p(x_i|x_{i-1}, ..., x_1)$ — каждый следующий элемент обусловлен предыдущими

Откуда взялась идея? PixelCNN



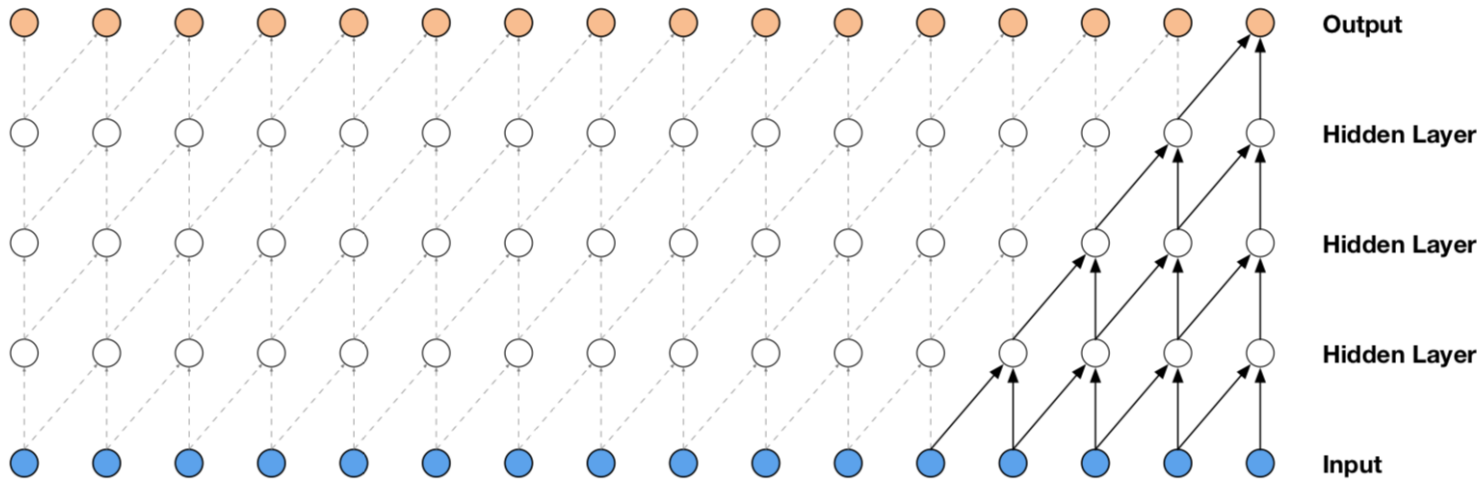
1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0



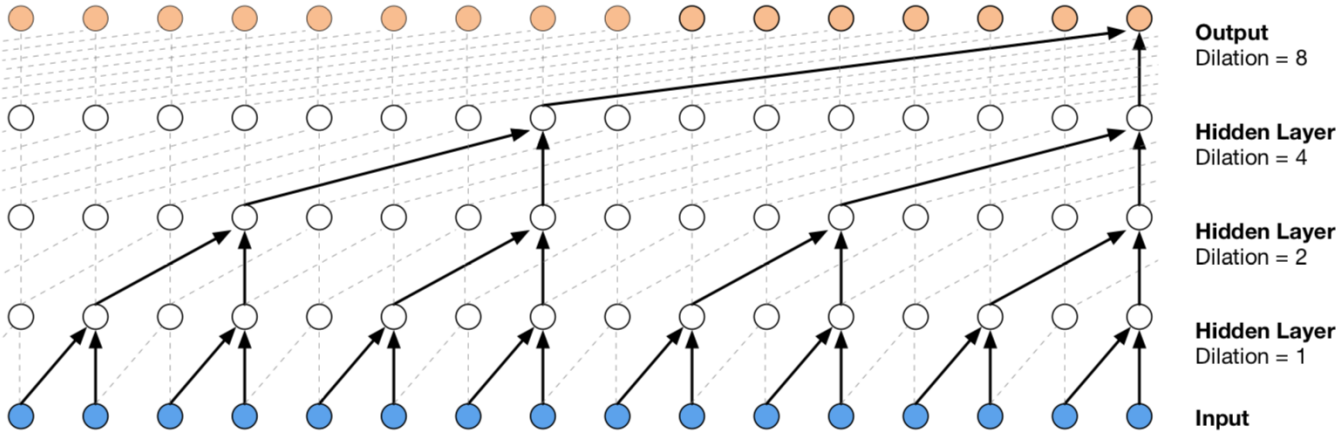
- Следующий пиксель обуславливается только предыдущими
- Для того, чтобы модель видела только то, что нужно, использовались маски
- Каждый слой по отдельности имеет свои "слепые пятна"

Устройство сети: сверточный слой

- Идея casual convolution: каждое следующее значение обусловлено предыдущими
- Для аудио: сдвиг выхода обычной конволюции на несколько шагов
- Быстрее рекуррентных!



Устройство сети: разрежаем свёрточный слой



- Таким образом, экспоненциально расширяется receptive field нейросети
- 1, 2, 4, ..., 512, 1, ...
- 1 блок (1, ..., 512) эквивалентен 1 конволюции 1 * 1024

Трансформация сигнала

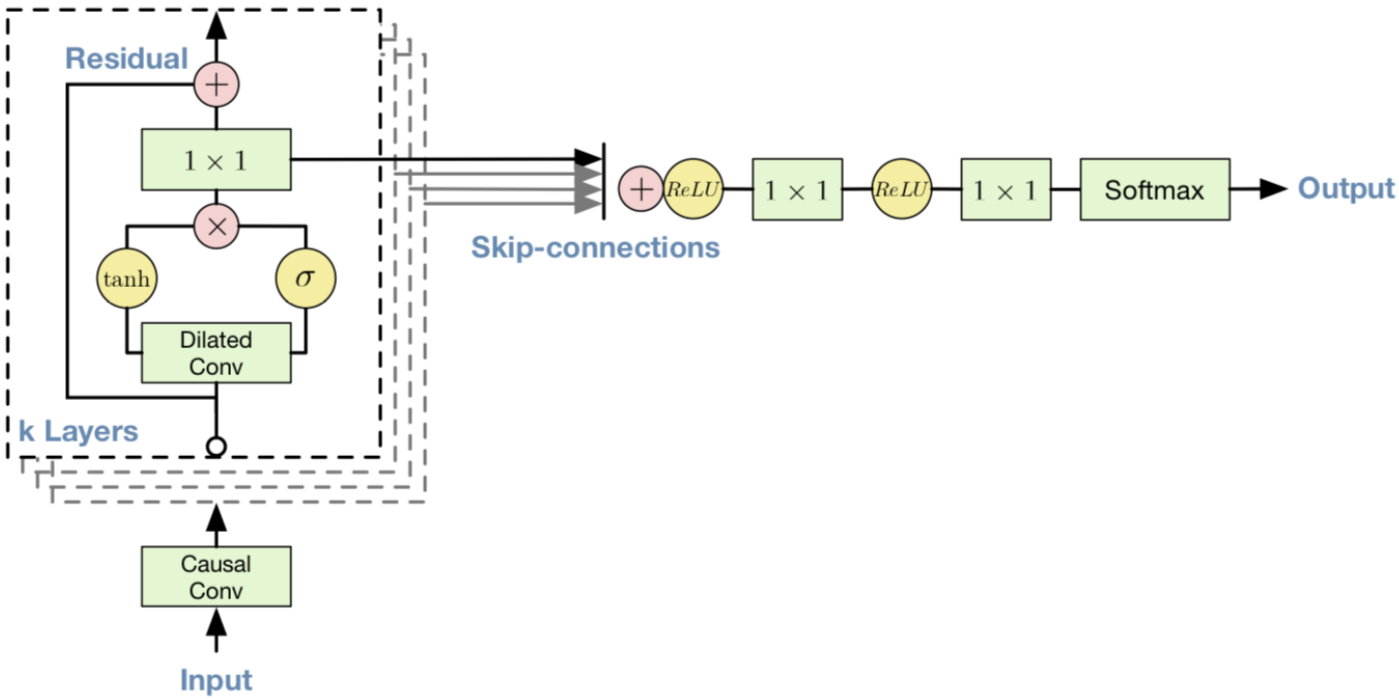
- Обычно глубина кодирования — 16, поэтому нужно было бы 65536 вероятностей (софтмаксов) для каждого временного шага. Очень дорого!
- Решение: μ-law companding transformation
- $f(x_t) = (x_t)\frac{\ln(1+\mu|x_t|)}{1+\mu}$
- $-1 < x_t < 1, \mu = 256$, затем квантуется в 256 значений
- Размерность 65536 $\rightarrow$ 256

Gated Activation Units: Как устроены слои?

Gated Activation Unit

- $\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \odot (W_{g,k} * \mathbf{x})$
- f — filter, g — gate (см. GRU)
- k — номер слоя
- W — обучаемый конволюционный фильтр

Residual and Skip connections



- Ускоряют сходимость
- Позволяют обучать более глубокие модели

Обуславливание

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

Глобальное обуславливание

- например, h — ID говорящего
- $\mathbf{z} = tanh(W_{f,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{f,k}^T * \mathbf{h})$
- V — обучаемая линейная проекция

Локальное обуславливание

- например, h_t — лингвистические фичи
- $\mathbf{y} = f(\mathbf{h})$, имеет то же разрешение, что и аудиосигнал (upsampling)
- $\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{f,k}^T * \mathbf{y})$
- V — 1 * 1 конволюция

Эксперименты

Multi-speaker speech generation

- Обуславливание ID говорящего
- Нет обуславливания текстом (поэтому результат — ненастоящие слова)
- Обучение на 44 часах речи 109 людей
- Результат — сеть выучивает речевые особенности говорящих, результат похож на речь

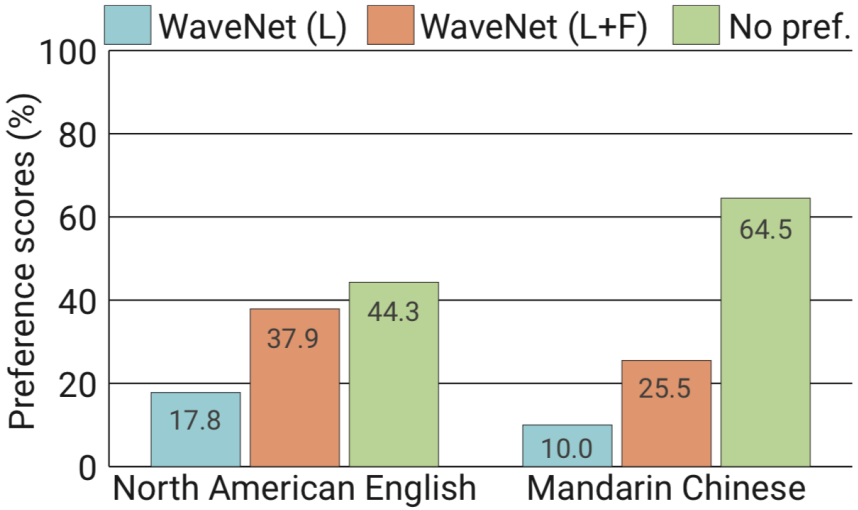
Эксперименты

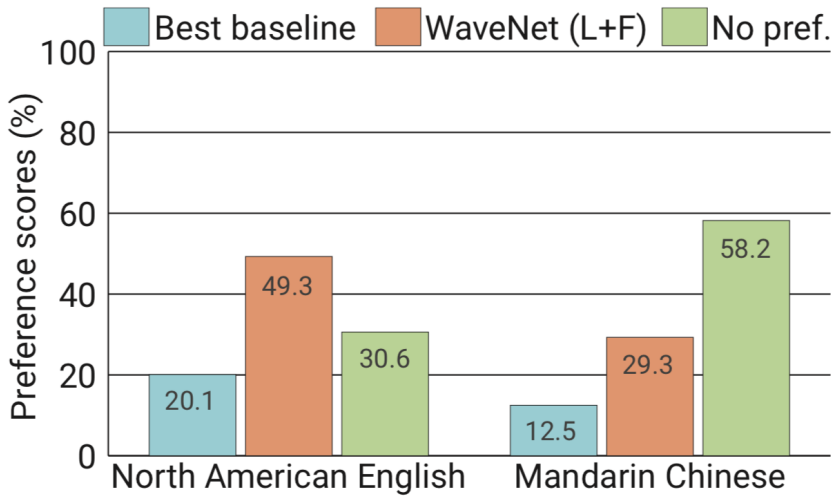
Text to Speech (TTS)

- Обуславливание ID говорящего
- Обуславливание лингвистическими фичами (+"фундаментальной частотой— определенное свойство звуковой волны)
- Сравнение с другими моделями: LSTM-RNN-based statistical parametric и HMM-driven unit selection concatenative
- Mean Opinion Score (MOS)

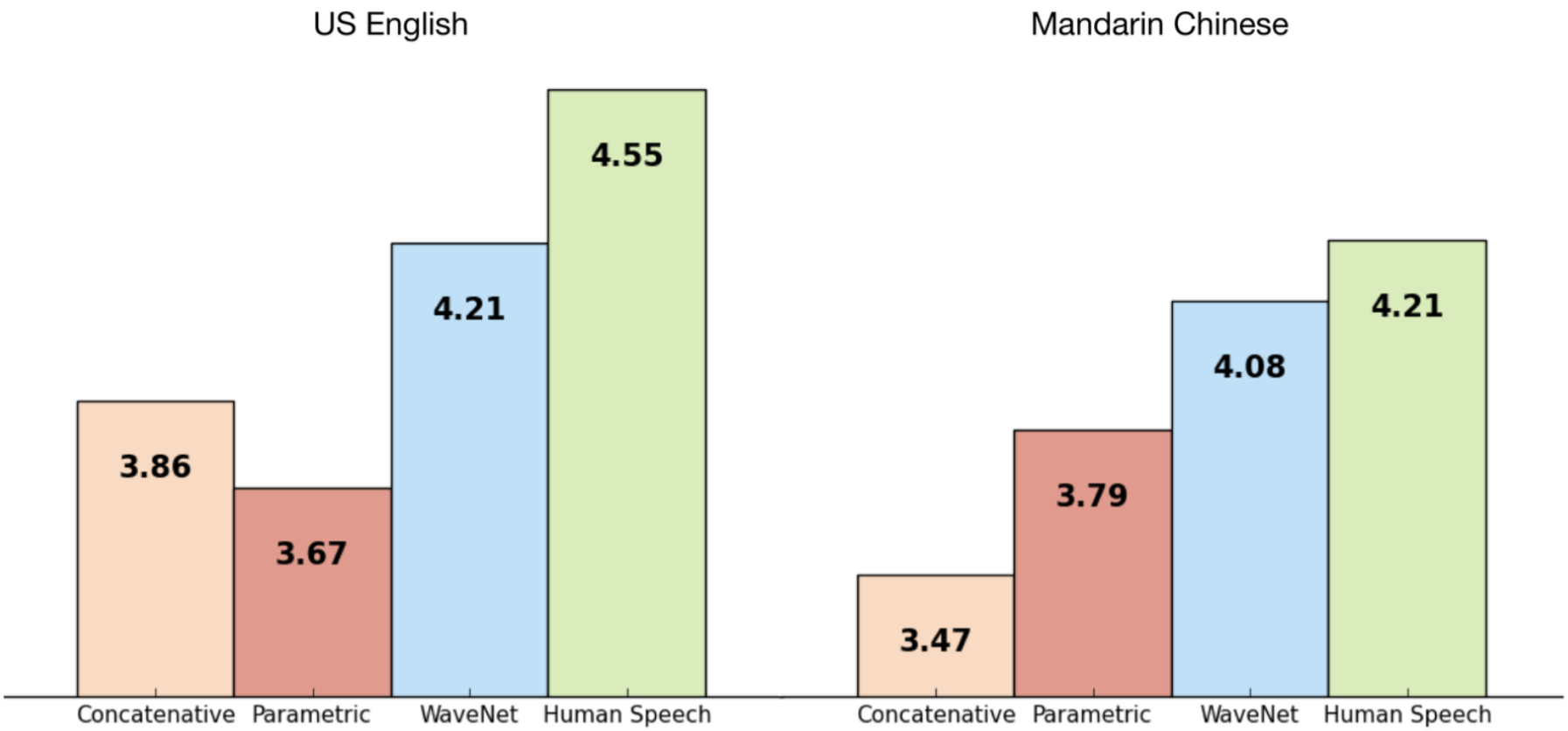
	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Эксперименты





Эксперименты



Эксперименты

Музыка

- Сложно оценить результат
- Гармонично, эстетически приятно
- Необходимо большое receptive field
- Интересно, что можно обуславливать жанром или инструментом

Выводы

- Авторегрессионная модель, моделирует звук как волну
- Сочетает casual filters и dilated convolutions
- Высокое качество результата
- Может использовать обуславливающие переменные, что даёт интересные результаты

Источники

- Van Den Oord, Aäron, et al. "WaveNet: A generative model for raw audio."https://arxiv.org/pdf/1609.03499.pdf