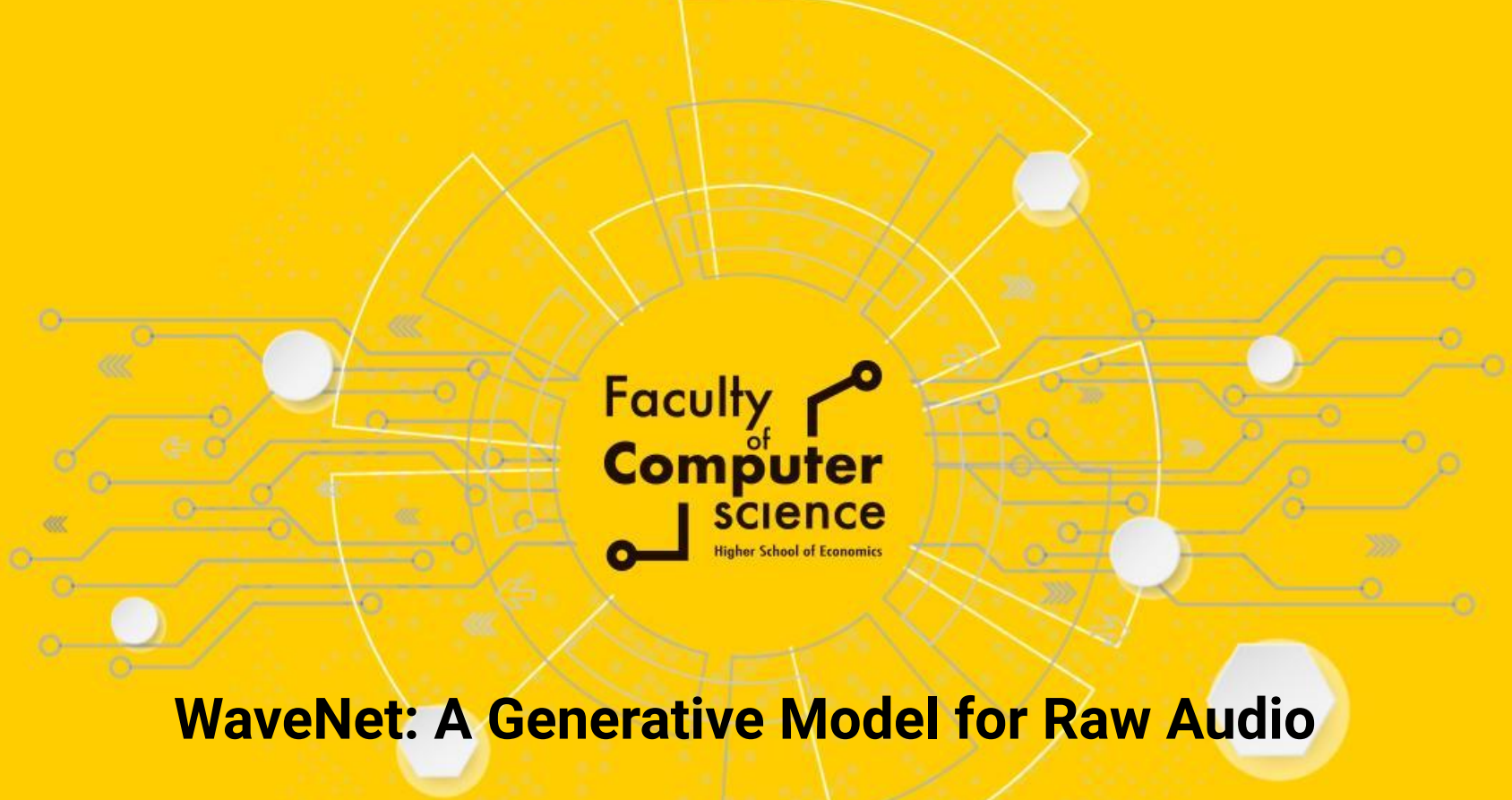


WaveNet: A Generative Model for Raw Audio

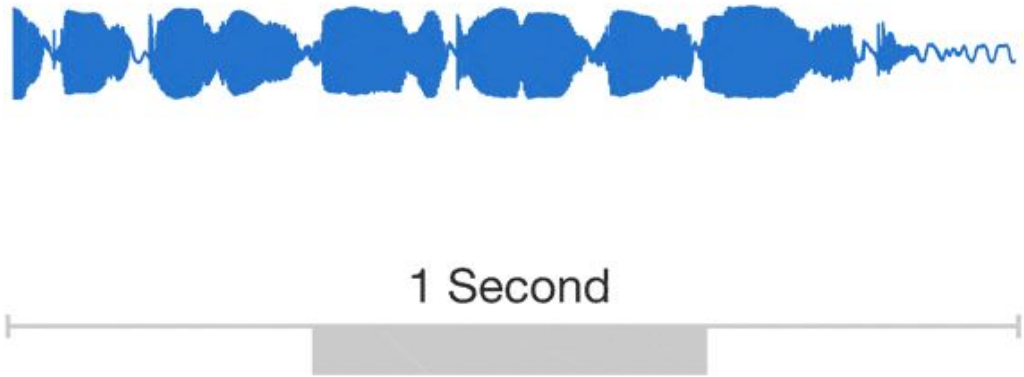
Подготовила:
Иванова Алеся Александровна, БПМИ202

План

1. Как звук хранится в компьютере, спектрограмма и мел-спектрограмма
2. Text-To-Speech
3. WaveNet:
    - архитектура
    - эксперименты

Хранение звука в компьютере

- Фиксируется амплитуда звукового сигнала через равные промежутки времени, хранится последовательность 16-битных чисел

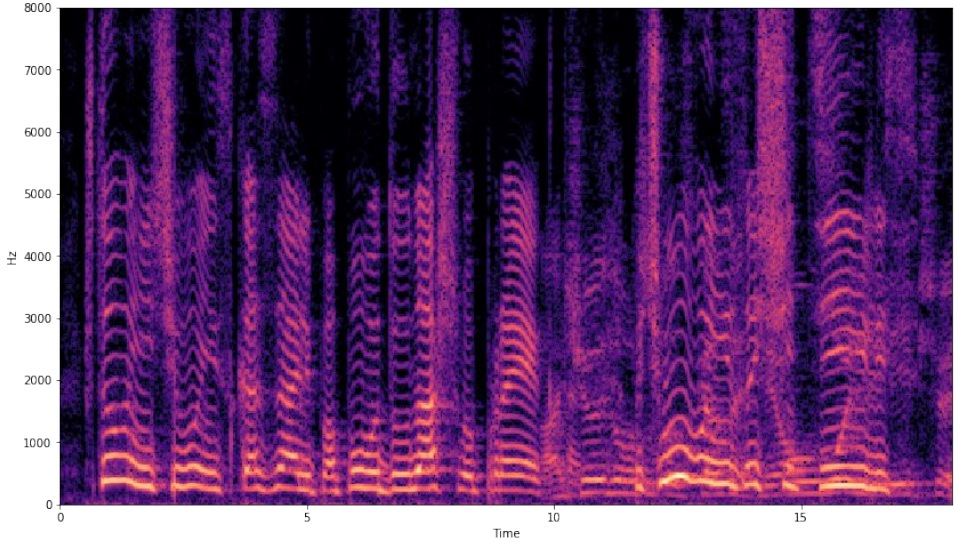


Спектрограмма

- Применение преобразования Фурье к коротким фрагментам звукового сигнала:

$$F(k,m) = \sum_{n=0}^{L-1} x[n+m]w[n]e^{-i\frac{2\pi}{L}kn}$$

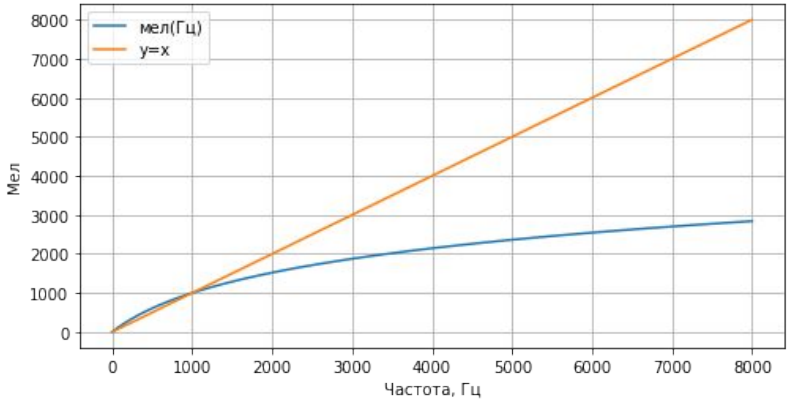
- Показывает зависимость амплитуды от времени и частоты

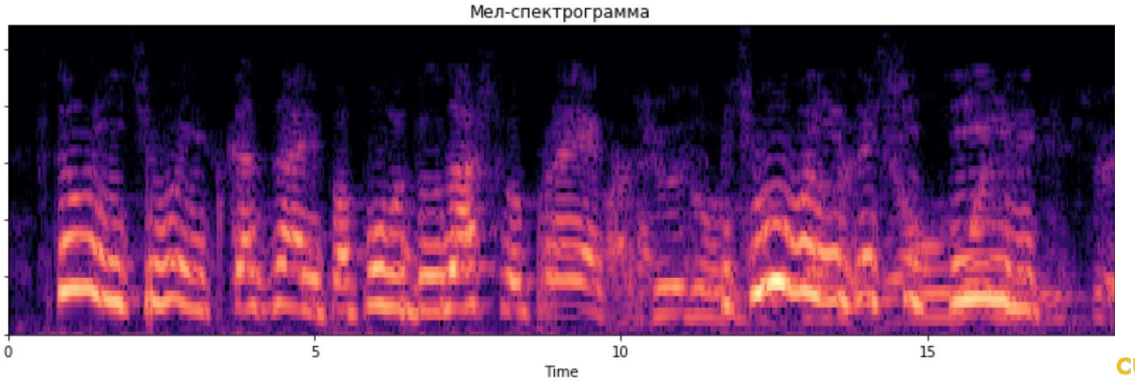


Мел-спектрограмма

- Человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких
- Мел — психофизическая единица высоты звука

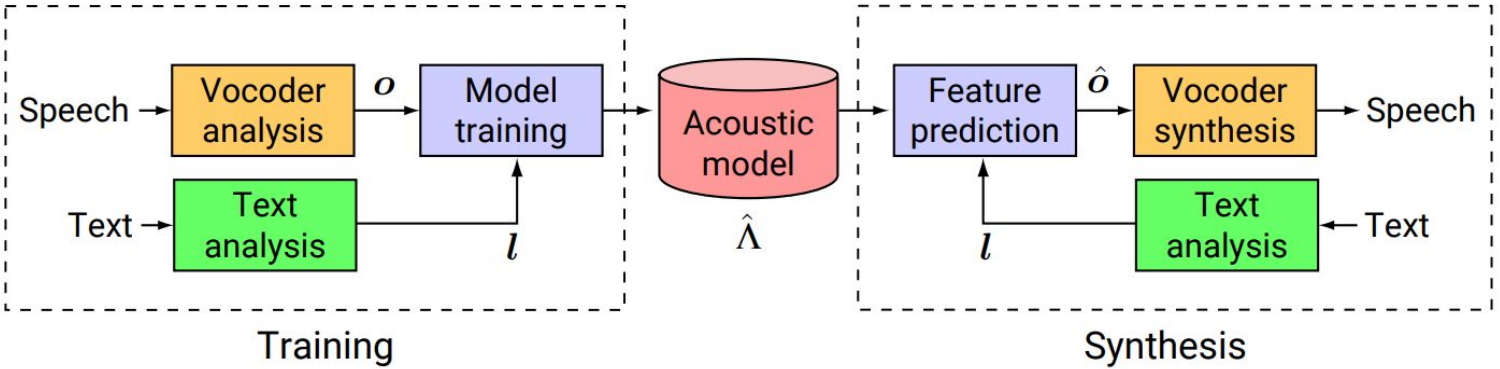
$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right)$$





Text-To-Speech

- Две компоненты: анализ текста и синтез речи



- Используются модели:
  - hidden Markov models (HMMs)
  - feed-forward neural networks
  - recurrent neural networks

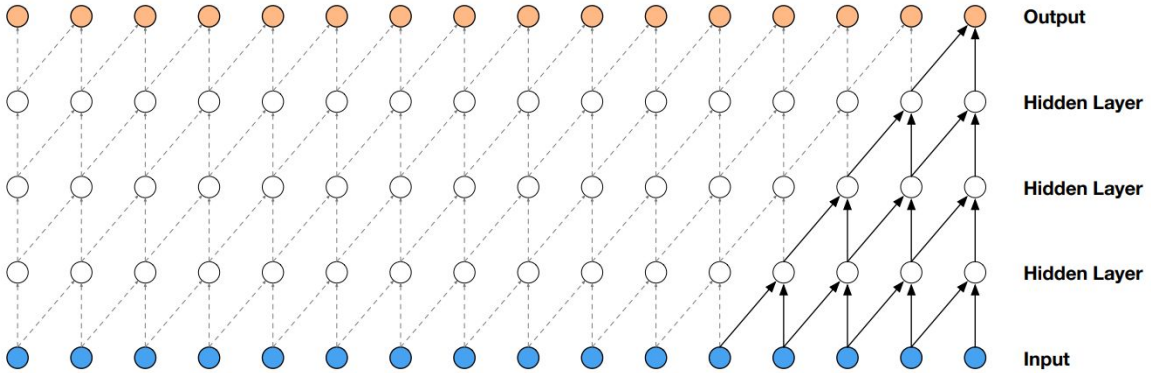
WaveNet

- Работает напрямую с сырыми амплитудами
- Предсказывает вероятности, что амплитуда в момент времени t примет каждое из возможных значений, если известны значения амплитуд в предыдущие моменты времени:

$$p\left(x_t \mid x_1, \ldots, x_{t-1}\right)$$

Dilated Causal Convolutions

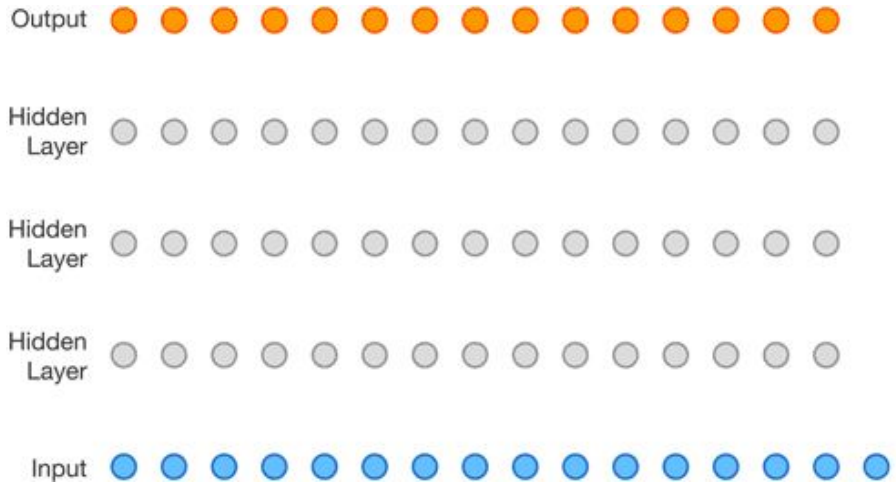
- Causal Convolutions - основная составляющая модели
- При их использовании можно гарантировать, что предсказания для момента времени t зависят только от значений в предыдущие моменты



- При обучении предсказания для всех моментов времени можно вычислять параллельно
- При генерации предсказания для нового семпла передаётся в модель для получения предсказаний для следующих семплов

Заголовок

- Проблема casual convolutions: нужно много слоёв или большая длина фильтра для достаточно большого рецептивного поля
- Решение: dilated convolution



- В WaveNet шаг фильтра удваивается для каждого слоя до предела, а затем повторяется. Например: 1, 2, 4, . . . , 512, 1, 2, 4, . . . , 512, 1, 2, 4, . . . , 512.

Softmax Distributions

- Для предсказания итоговых вероятностей $p\left(x_t \mid x_1, \ldots, x_{t-1}\right)$ используется SoftMax слой:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

- Проблема: для 16-битных значений амплитуд нужно генерировать 65,536 вероятностей в каждый момент времени
- Решение: используется сжатие значений амплитуд с помощью μ-law companding transformation (μ-закона), получается 256 различных значений:

$$f\left(x_t\right) = \operatorname{sign}(x_t) \frac{\ln\left(1 + \mu\left|x_t\right|\right)}{\ln\left(1 + \mu\right)} \qquad \text{where } -1 < x_t < 1 \text{ and } \mu = 255$$

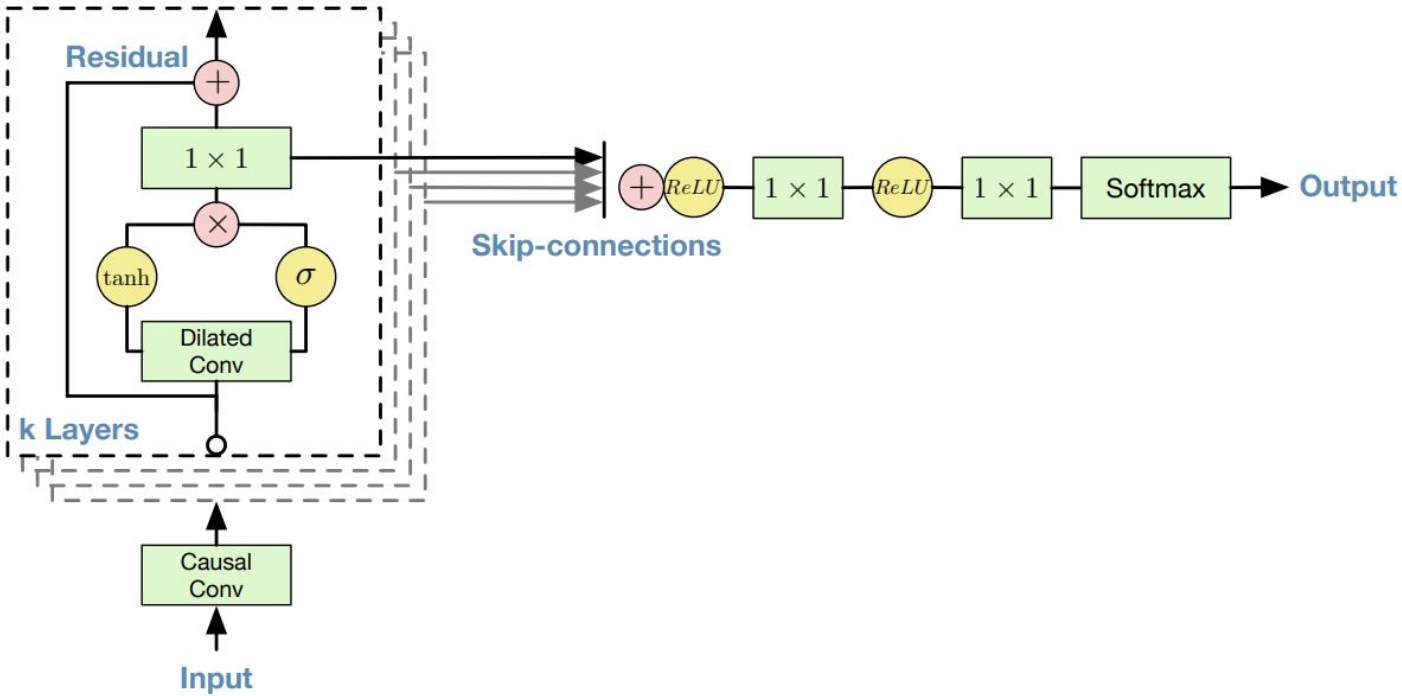
Gated Activation Units

- Используются слои активации:

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

* - оператор свёртки
$\odot$ - поэлементное умножение
k - номер слоя
W - обучаемый фильтр

Residual and Skip Connections



Conditional WaveNets

- Можно добавить дополнительный параметр для генерации аудио с определёнными характеристиками:

$$p\left(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}\right)$$

- Пример: выбор спикера при генерации речи
- Типы параметризации:
  - глобальная: один параметр, который влияет на все предсказания
  - локальная: последовательность параметров, разные параметры для разных моментов времени
- Параметр добавляется на слое активации:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right)$$

Context Stacks

- Ещё один способ увеличения рецептивного поля
- Context Stack обрабатывает длинный фрагмент аудио и подаёт выход как локальный параметр для WaveNet

Эксперименты

- Multi-speaker speech generation
- Text-to-speech
- Music
- Speech recognition

Multi-speaker speech generation

- Генерация речи без опоры на текст
- Параметр: номер спикера (задаётся one-hot вектором)

- Генерирует несуществующие, но похожие на человеческие звуки, воспроизводит характерные черты спикера
- Отсутствие связности речи из-за ограниченного рецептивного поля (~300 мс, 2-3 фонемы)

Text-to-speech

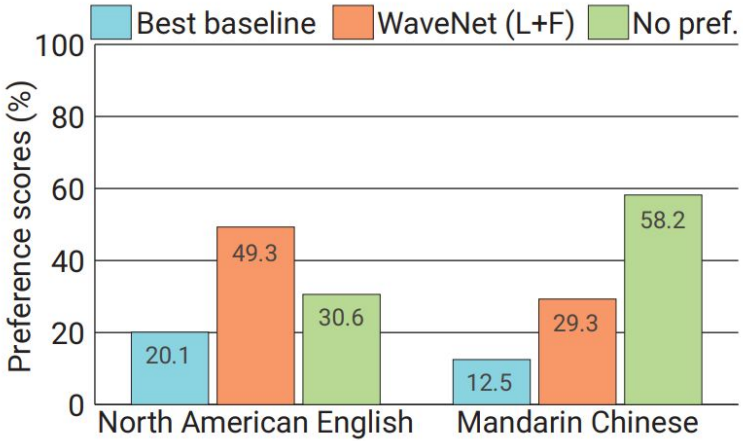
- Локальная параметризация лингвистическими и фонетическими признаками (текущий звук, слог, слово)

Сравнение с лучшими моделями от Google до WaveNet:



Parametric *Concatenative* *WaveNet*

Text-to-speech



Speech samples	Subjective 5-scale MOS in naturalness	
	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

Text-to-speech

- Можно добавить номер спикера в качестве параметра




Music

Speech recognition

- Можно адаптировать модель для распознавания речи
- 2 компоненты функции потерь:
  - предсказание следующего семпла
  - классификация фрагмента

Материалы

- https://arxiv.org/pdf/1609.03499v2.pdf
- https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio

Вопросы?