WAVE NET: A GENERATIVE MODEL FOR RAW AUDIO

План

Вводная часть:

- Хранение аудиоволн на компьютере
- Спектрограмма, мел-спектрограмма

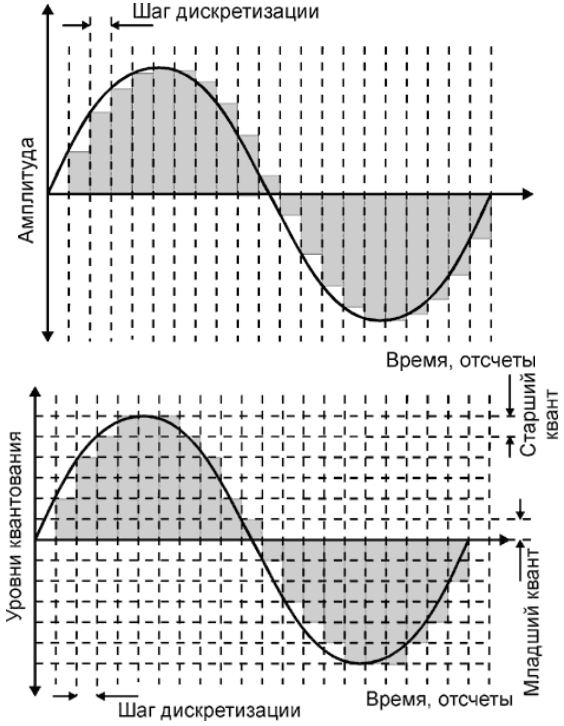
WaveNet:

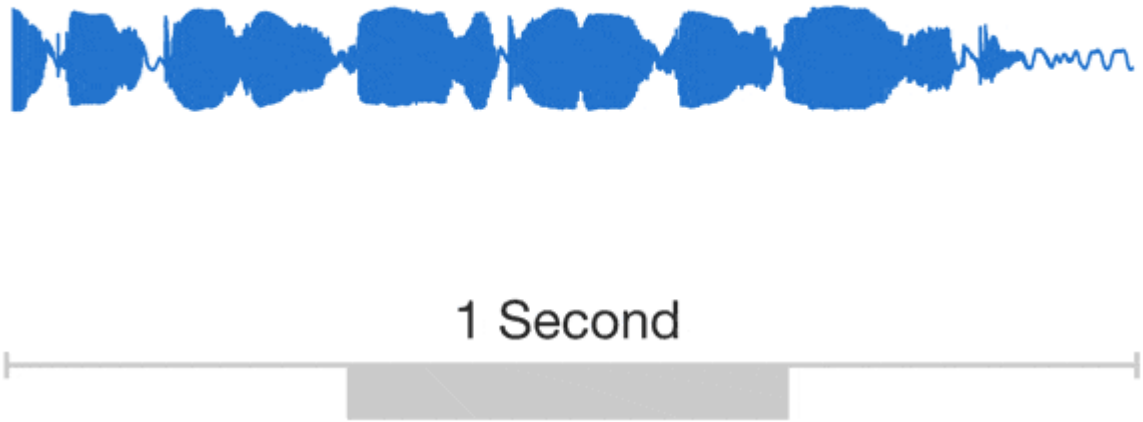
- Область применения
- Архитектура сети
- Conditional WaveNet

Эксперименты:

- Multi-Speaker Speech Generation
- TTS
- Генерация музыки
- Speech recognition

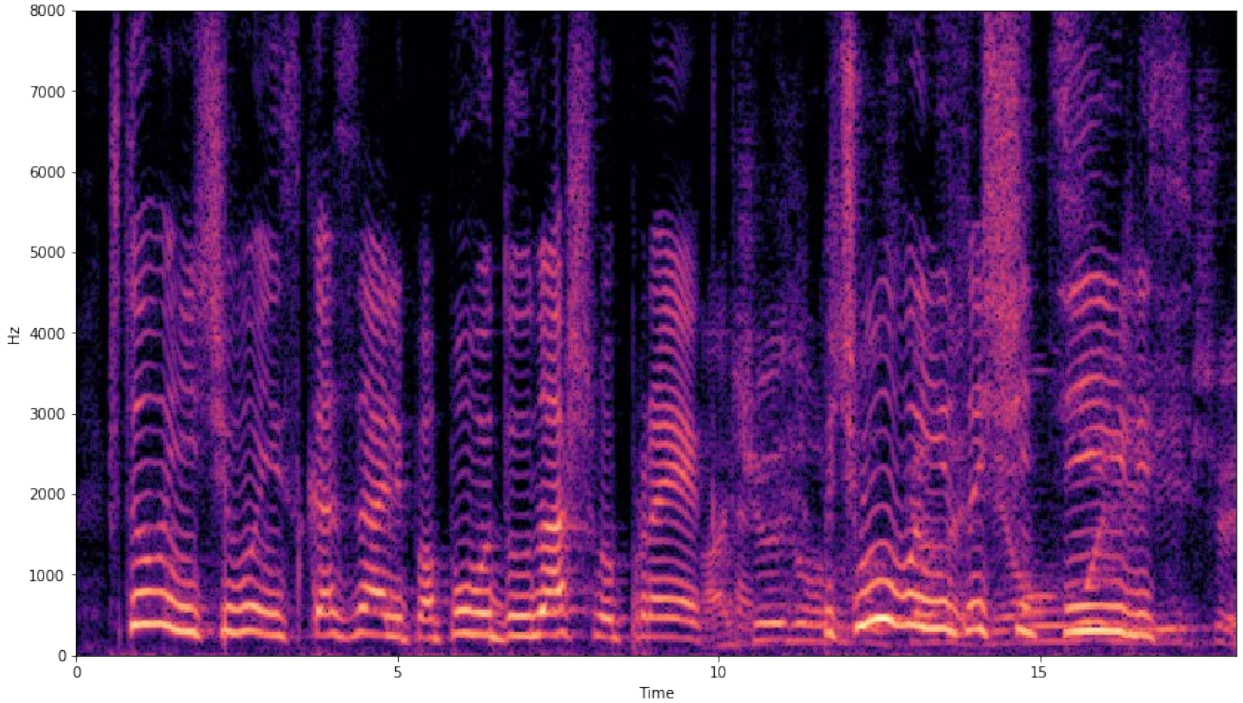
Хранение аудиоволн на компьютере





Спектрограмма

Получается с помощью преобразования Фурье на коротких фрагментах звукового сигнала



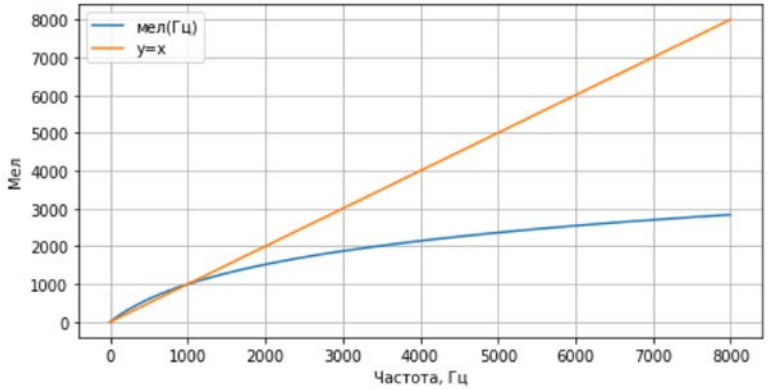
Мел-спектрограмма

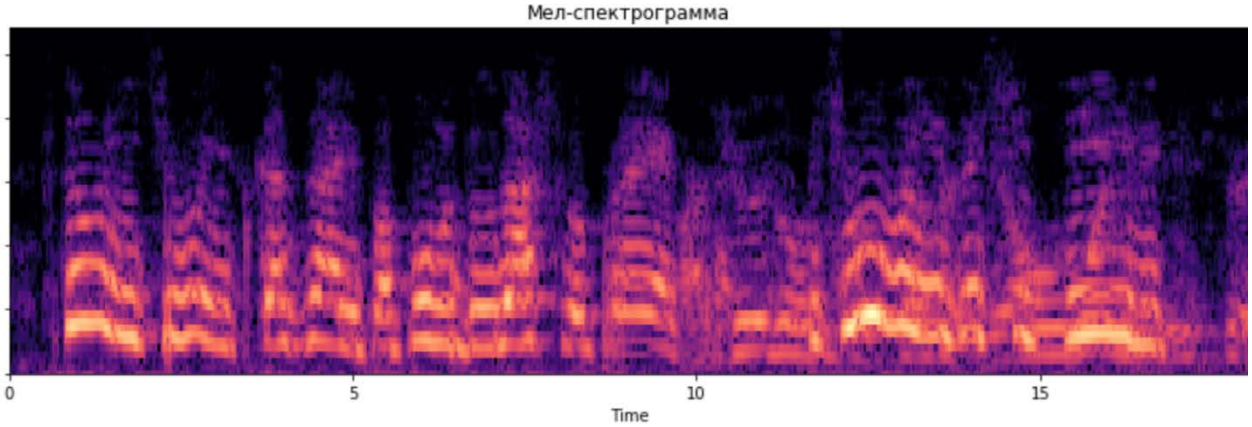
Мел – единица измерения, основана на психо-физиологическом восприятии звука человеком и логарифмически зависит от частоты.

Человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких.

Мел-спектрограмма получается из спектрограммы с помощью формулы:

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





Предыстория

До появления WaveNet существовало два основных подхода к реализации части синтеза речи:

- Конкатенативный, непараметрический подход, основанный на примерах, строит высказывание из кусочков записанной речи. Звучит довольно роботизированно
- Параметрический, основанный на моделях, известный как статистический параметрический синтез речи - использует генеративную модель.

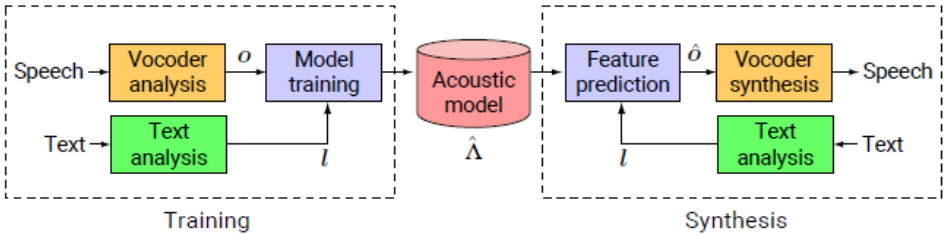


Figure 6: Outline of statistical parametric speech synthesis.

WaveNet: Область применения. Решаемые задачи

- Генерация речи (не основанная на тексте)
- TTS
- Генерация музыки
- Распознавание речи
- Преобразование голоса
- Разделение источников аудио (source separation)

WaveNet

- DeepMind (Google), 2016
- Работает напрямую с необработанным звуком
- Генеративная, авторегрессионная модель

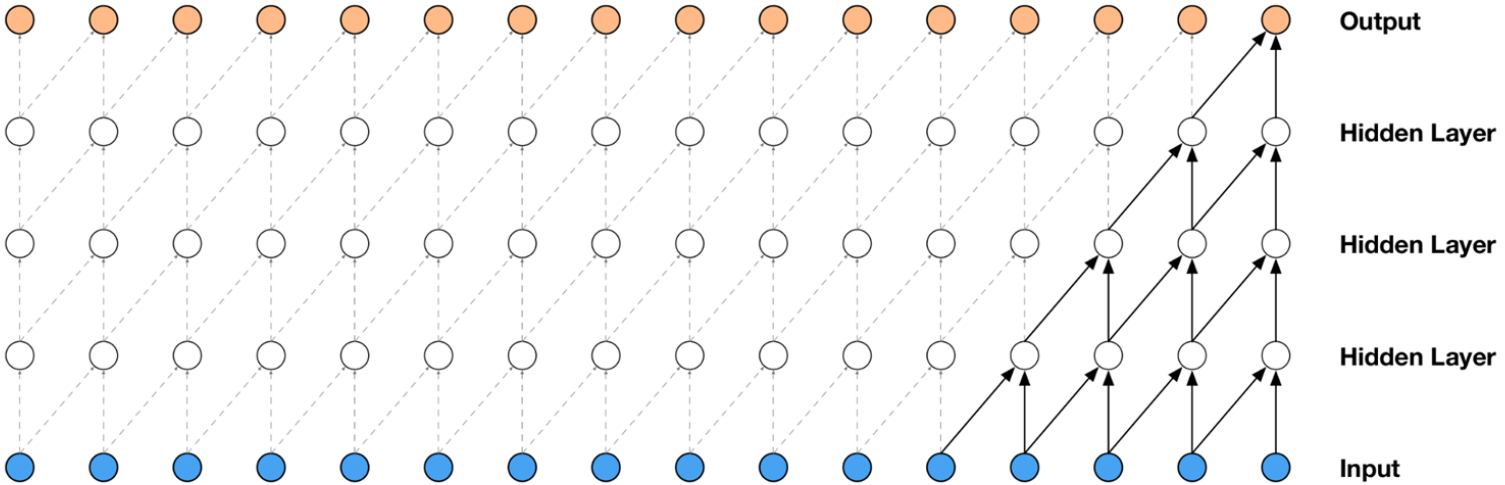
The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ is factorised as a product of conditional probabilities as follows:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \tag{1}$$

Архитектура WaveNet: Causal convolutions

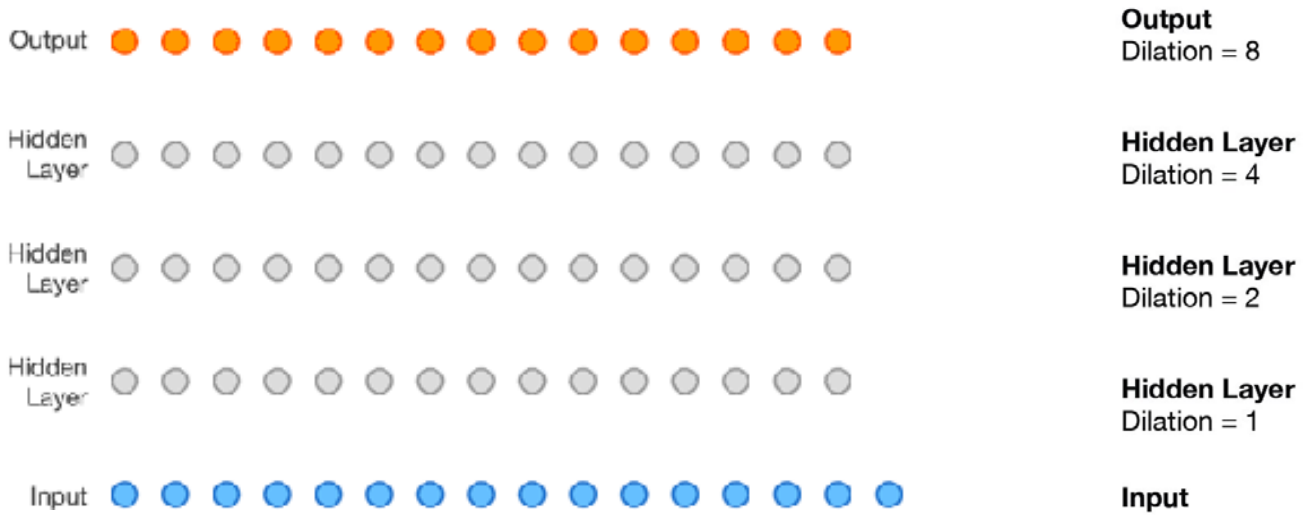
Гарантируем, что предсказания для текущего момента времени зависят только от значений предыдущих сэмплов.

В режиме обучения возможно параллельное вычисление.



Архитектура WaveNet: Dilated causal convolutions

- Главная фишечка модели
- Размер рецептивного поля растёт экспоненциально



В работе чередовались значения dilation по схеме: $1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512$

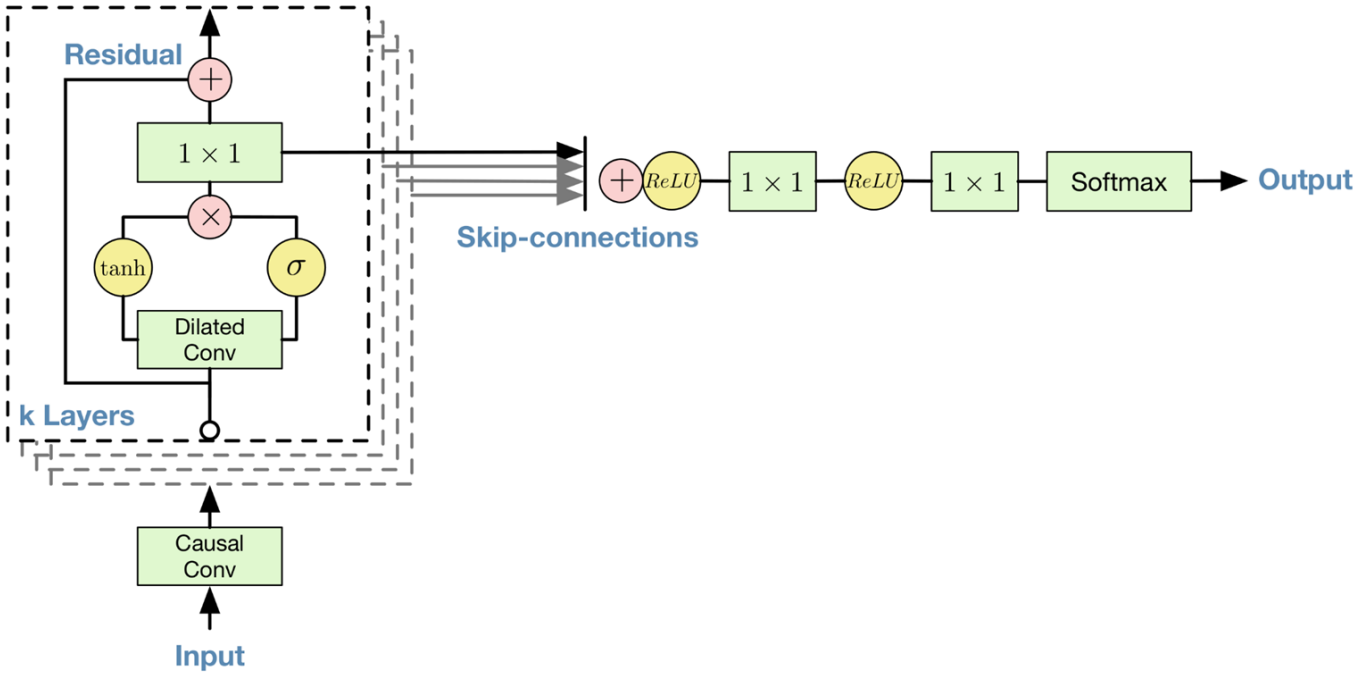
Архитектура WaveNet: Gated Activation Units

Используется активация вида:

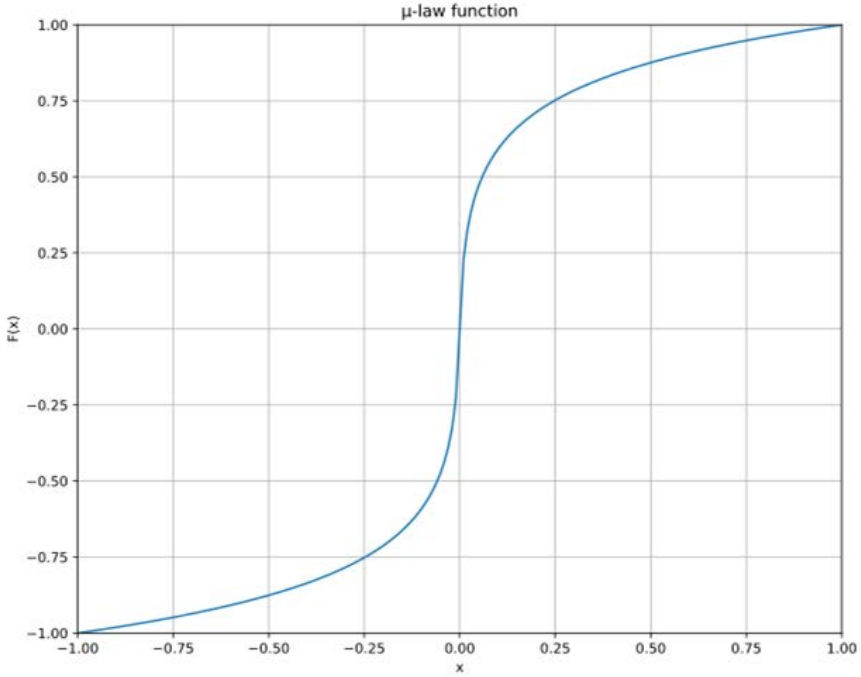
$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

Где $*$ - оператор свёртки, $\odot$ - поэлементное умножение, $W_{f,k}$ и $W_{g,k}$ - обучаемые ядра свёртки k-того слоя

Архитектура WaveNet: Residual and Skip Connections



Архитектура WaveNet: μ-law



$$f(x_t) = \text{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

where $-1 < x_t < 1$ and $\mu = 255$

Conditional WaveNets

Добавляется дополнительный параметр h.
Есть два типа conditioning: глобальный (голос спикера) и локальный (лингвистические особенности)

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

Активация для global conditioning:

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

где $V_{*,k}$ представляет собой обучаемую линейную проекцию

Активация для local conditioning:

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

где $\mathbf{y} = f(\mathbf{h})$

Context Stacks

Обрабатывает длинный фрагмент аудио, затем результат обработки используется в качестве conditional для обучаемой сети.

Неявный способ увеличить рецептивное поле.

ЭЭЭЭЭксперименты

- Multi-Speaker Speech Generation
- TTS
- Генерация музыки
- Speech recognition

Multi-Speaker Speech Generation

- Генерация речи без опоры на текст
- Модель обусловлена ID спикера (задан через one-hot вектор)






Text-To-Speech

- Локальная обусловленность лингвистическими признаками

Сравнение с лучшими бейзлайнами:

 - конкатенативная модель

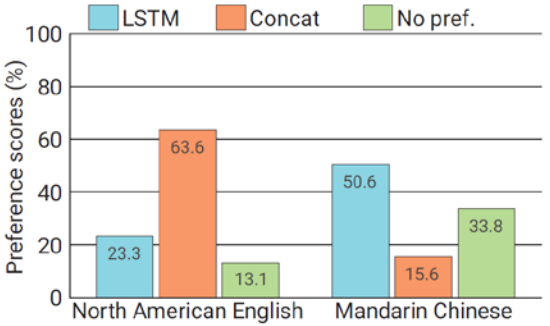
 - параметрическая модель

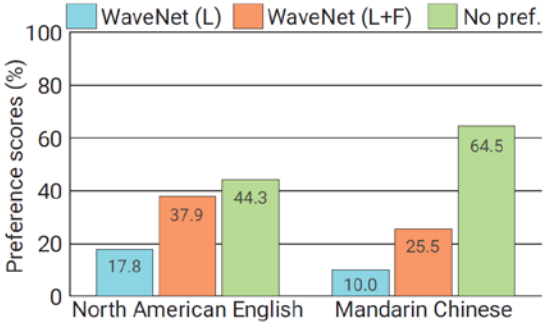
 - WaveNet

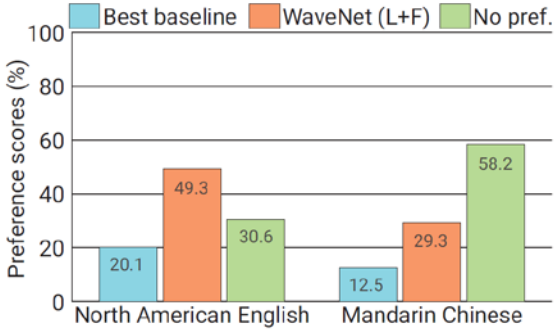
Text-To-Speech. Результаты сравнения

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

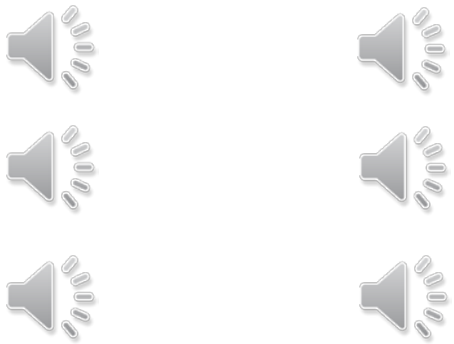
Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit μ-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.







Генерация музыки



Speech recognition

Можно модифицировать модель для распознавания речи

Использовались две функции потерь:

- одна для прогнозирования следующего сэмпла и
- одна для классификации фрагмента

Выводы

- WaveNet – авторегрессионная модель, работает со звуком непосредственно на уровне вэйвформ, совершила скачок в качестве синтеза естественно звучащей человеческой речи;
- WaveNet объединяет идеи каузальных фильтров и разреженных свёрток;
- Рецептивное поле растёт экспоненциально с увеличением числа слоёв;
- WaveNets могут быть обусловлены глобально или локально;
- WaveNet показала многообещающие результаты в различных областях, связанных с обработкой звука

Материалы

https://arxiv.org/abs/1609.03499

https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/

https://books.ifmo.ru/file/pdf/3111.pdf

https://habr.com/ru/articles/462527/