WaveNet
a generative model for raw audio

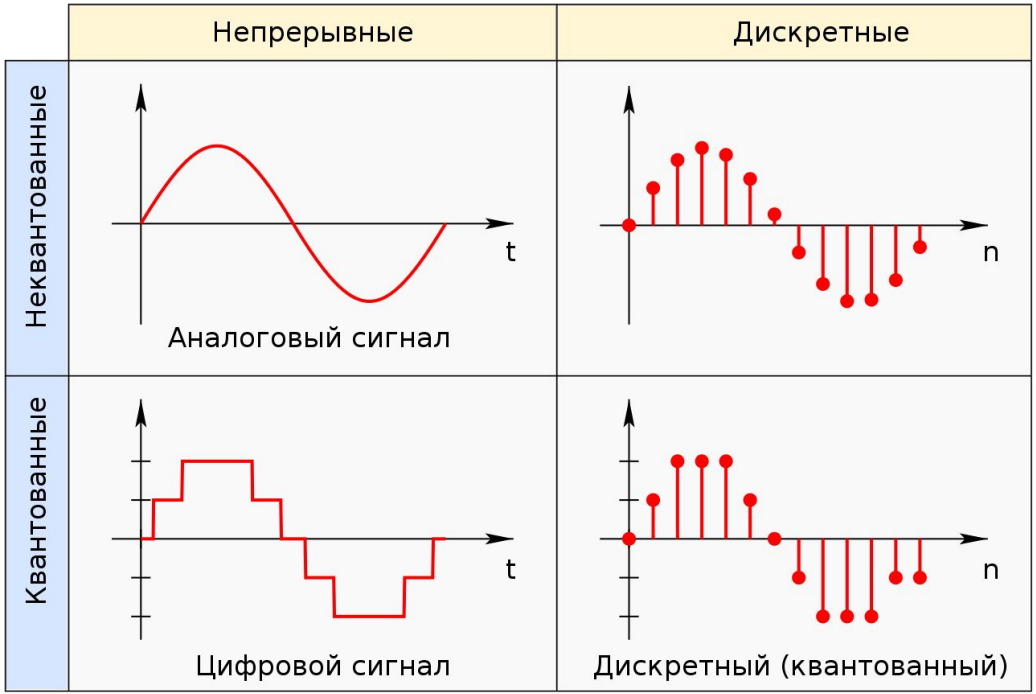
Пономарев Евгений, БПМИ211

План

- Хранение аудиоволны на компьютере
- Спектрограмма и Mel-спектрограмма
- Text-to-speech problem. Компоненты синтеза речи
- Архитектура WaveNet
- Conditional WaveNet
- Эксперименты
- Заключение

Хранение аудиоволны на компьютере

- Дискретизация
- Квантование
- Кодирование



Параметры записанной аудиодорожки

- Sample rate - частота дискретизации (8kHz, 44.1kHz, 48kHz, 96kHz...)
- Sample size - количество бит на один сэмпл
- Количество каналов

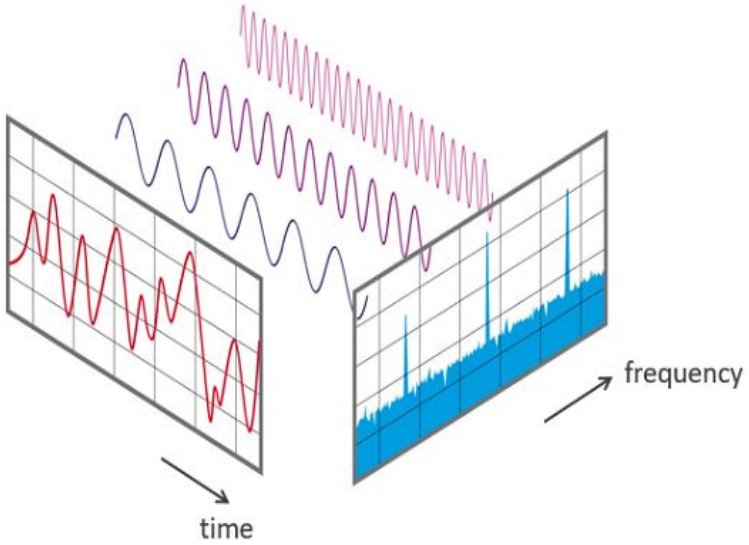
Частотный спектр сигнала

- Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

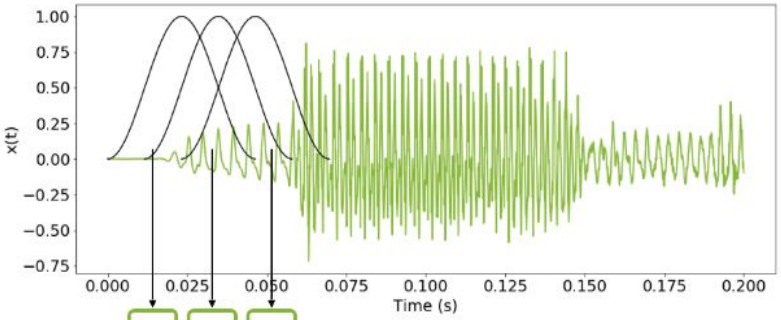
Frequency — Original signal

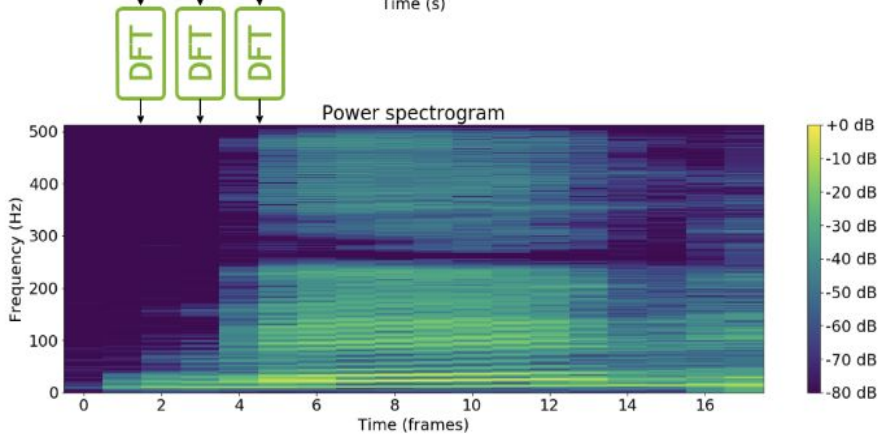
$$f(x) = A_1 * sin(freq_1 x + \phi_1) + ...$$
$$...$$
$$... + A_n * sin(freq_n x + \phi_n)$$

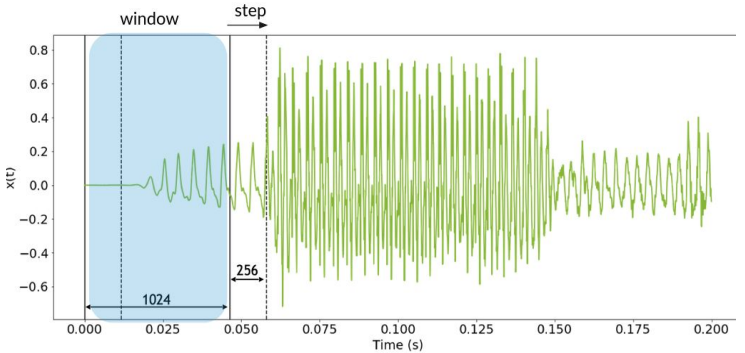


Теорема Котельникова:
Любой непрерывный сигнал, содержащий частоты до f, можно без потерь передать используя частоту дискретизации не меньше 2f

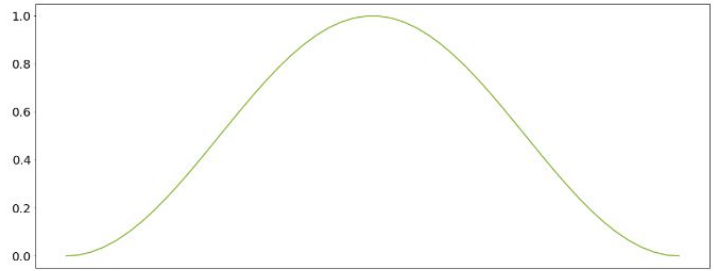
Спектрограмма и Mel-scale





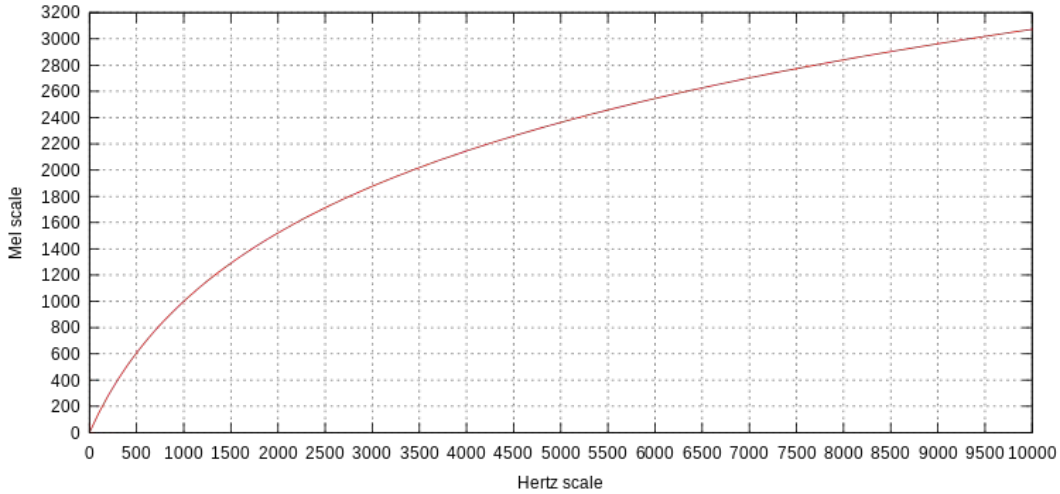


Hann's window:

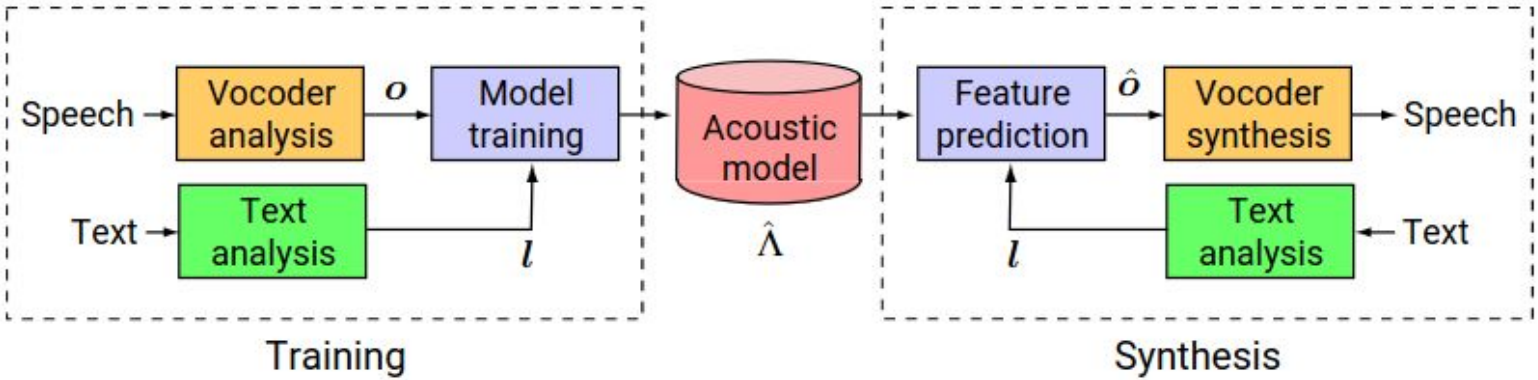


- Человеческое ухо распознает разницу
  в низких частотах лучше, чем в высоких
- Для человека:
  500 Hz << 600 Hz
  но 5000 Hz ~ 5100 Hz

$$m(f) = 2959 \log_{10}(1 + \frac{f}{700})$$



Text-to-speech problem



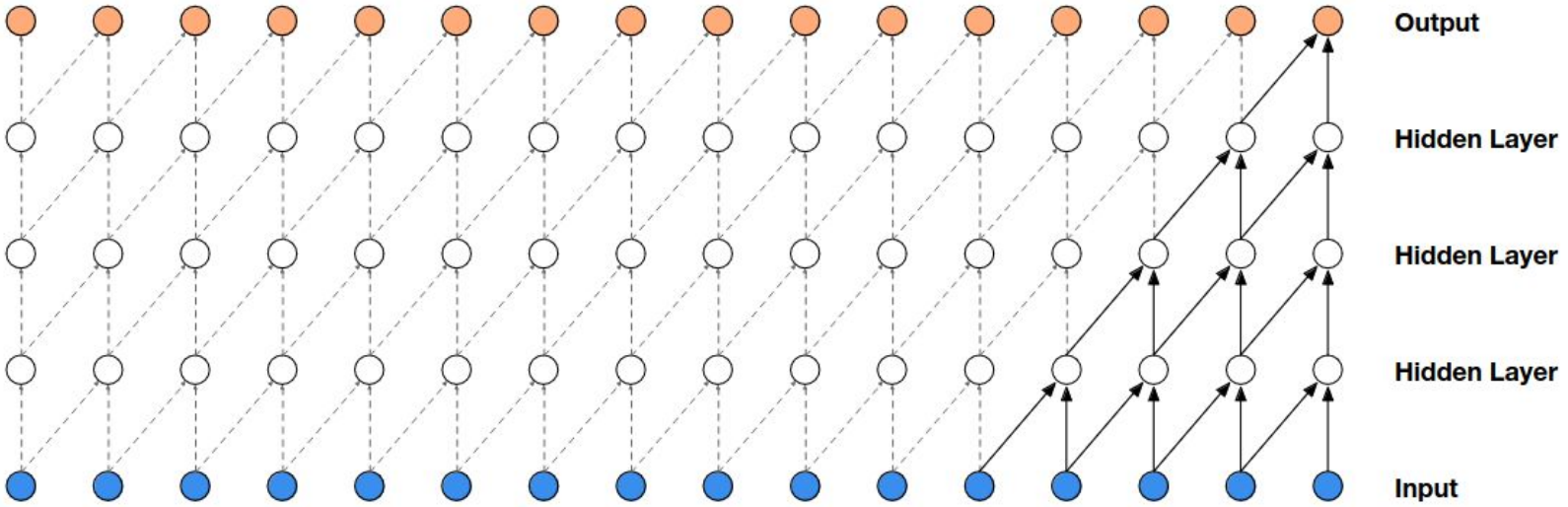
- Hidden Markov Model (HMM)
- LSTM

WaveNet

- Работает напрямую с сырым аудиосигналом, не переходя в частотный диапазон
- Позволяет обрабатывать очень длинные последовательности с помощью Dilated Causal Convolutions
- Показывает SOTA результаты для 2016 года в генерации речи и TTS
- Применима для генерации музыки

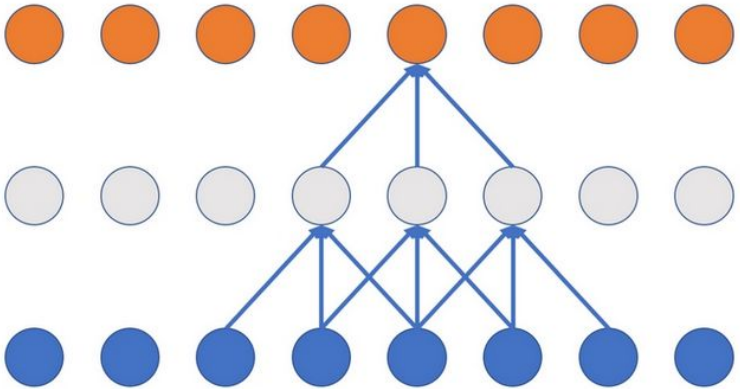
Causal Convolutions

- На каждом слое предсказания для момента времени t зависят только от от 1, ..., t - 1 предсказаний предыдущего слоя

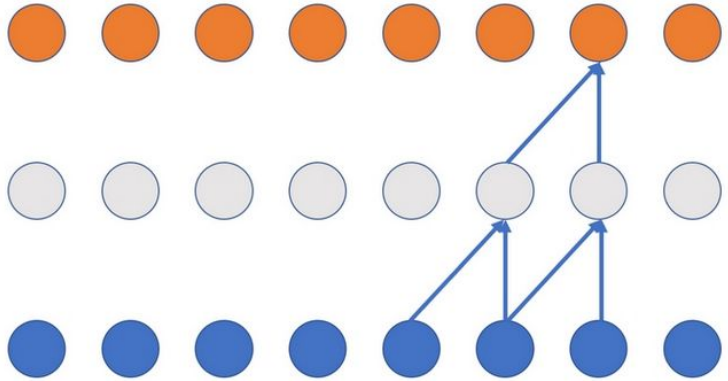


Разница с обычными свертками

Standart convolitions

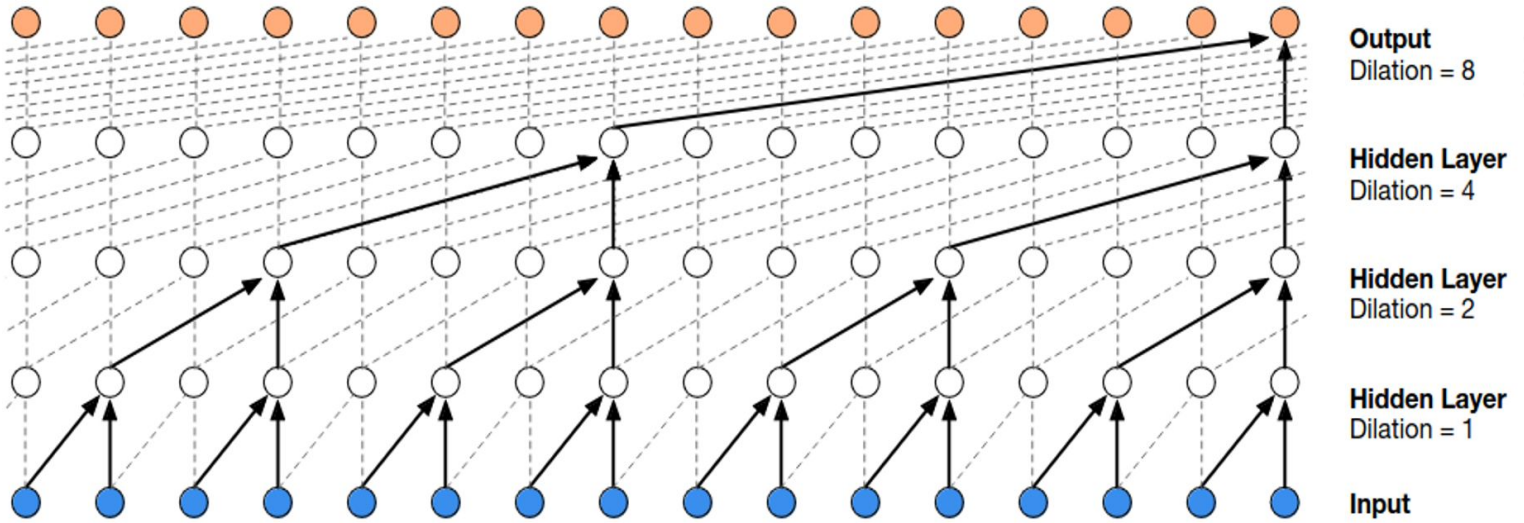


Causal convolutions



Dilated causal convolution

- Увеличили receptive field



SoftMax distribution

Хотим предсказывать вероятность сэмпла в момент времени t, обусловленного на все предыдущие предсказания

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

Квантованный 16 битный сигнал требует моделирования 65536 вероятностей для каждого значения амплитуды.

С помощью μ-law сэмплы сжимаются до 256 возможных значений вероятности

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

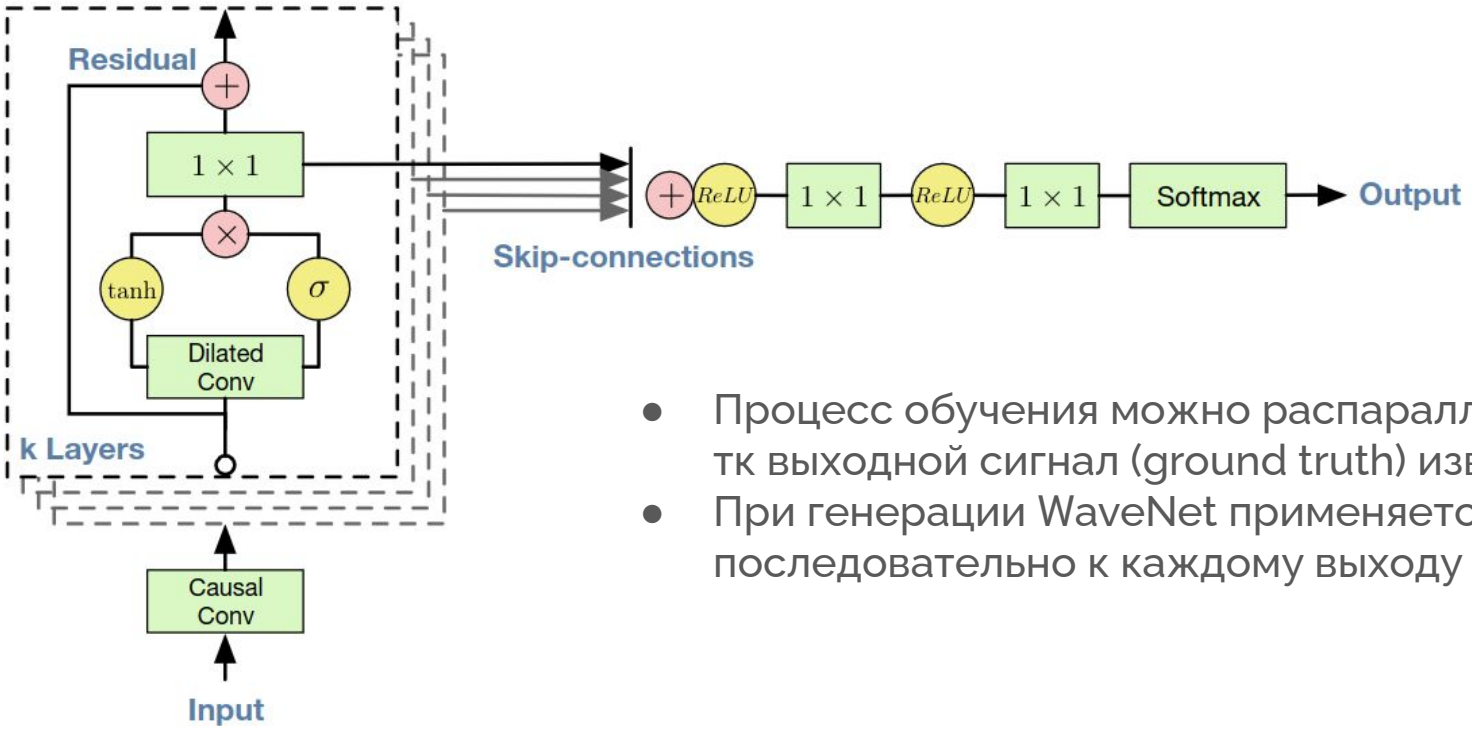
Gated activation units

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

$W_{f,k}, W_{g,k}$ - обучаемые матрицы
$\odot$ - операция поэлементного умножения
$*$ - операция свертки

Функция активации, используемая вместо ReLU между слоями dilated causal convolutions

Архитектура WaveNet



- Процесс обучения можно распараллелить, тк выходной сигнал (ground truth) известен
- При генерации WaveNet применяется последовательно к каждому выходу

Conditional WaveNet

Можно добавлять дополнительную информацию к последовательности сэмплов

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

Например:

- ID спикера в задаче генерации речи
- текст в задаче Text-to-speech
- различные лингвистические характеристики текста

Существуют два способа обусловить генерацию:

- *Global conditioning*

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

- *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 y - последовательность условий размера, равного размеру генерируемой последовательности сэмплов

Multi-Speaker Speech Generation

- Датасет: 44 часа аудио от 109 различных спикеров
- Модель обуславливалась только на ID говорящего
- ID кодировался в форме one-hot вектора
- Результат: модель научилась генерировать несуществующие слова, по звучанию похожие на человеческую речь.
- Причина: модель не обусловлена на сам текст

Text-to-speech

- Данные: 24.6 часов английской речи и 34.8 - северокитайской
- Все модели, обученные для сравнительного анализа, обуславливались на лингвистических особенностях текста и фундаментальных частотах
- Метрика 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

Примеры сгенерированной речи

English

Mandarin

Генерация музыки

- Необходимы большие значения рецептивного поля
- Модель постоянно меняет стиль музыки, состав инструментов, громкость
- Тем не менее, генерируемые фрагменты звучат музыкально
- Сложно объективно оценить качество модели


Распознавание речи

- Добавлен mean-pooling слой после сверток, который аггрегирует информацию о небольших фрагментах аудио длиной в 10ms
- Две функции потерь
  - Loss для предсказания следующего сэмпла
  - Loss для классификации фрагмента
- Получили **18.8** PER, что являлось лучшим результатом на датасете TIMIT для моделей, работающих с сырым аудио