中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

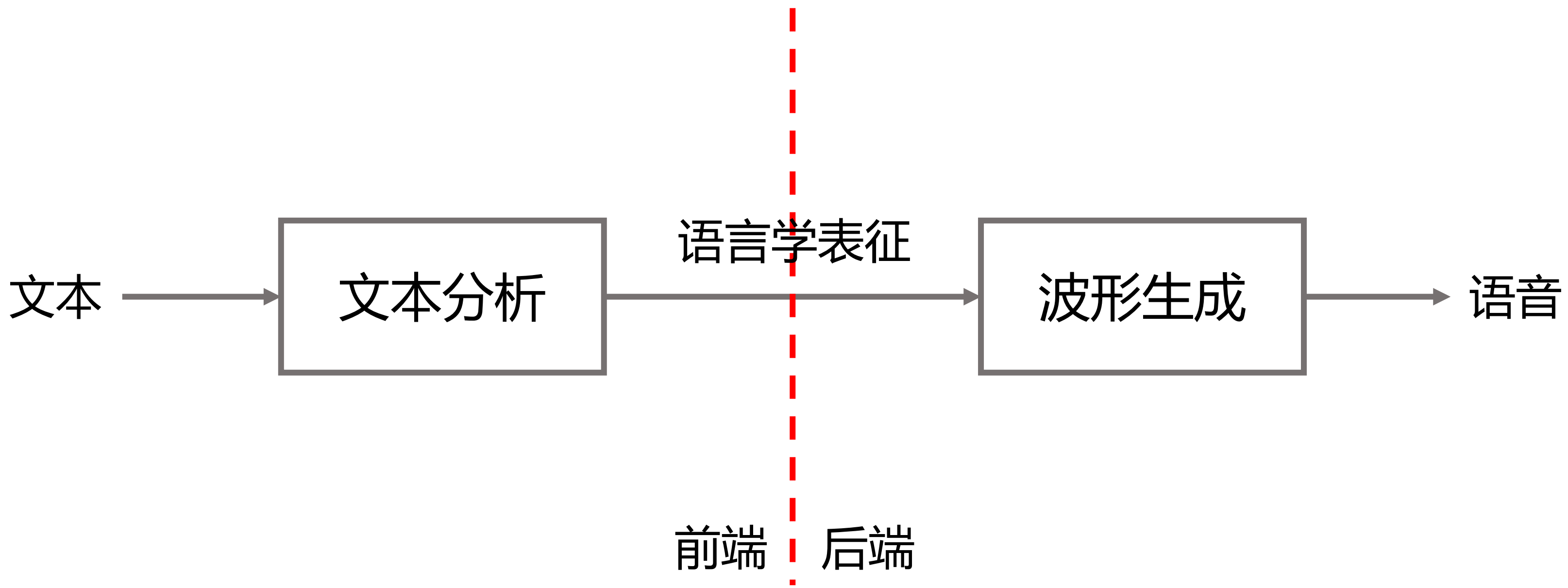
语音交互技术
——语音合成（二）

中国科学院自动化研究所
模式识别国家重点实验室

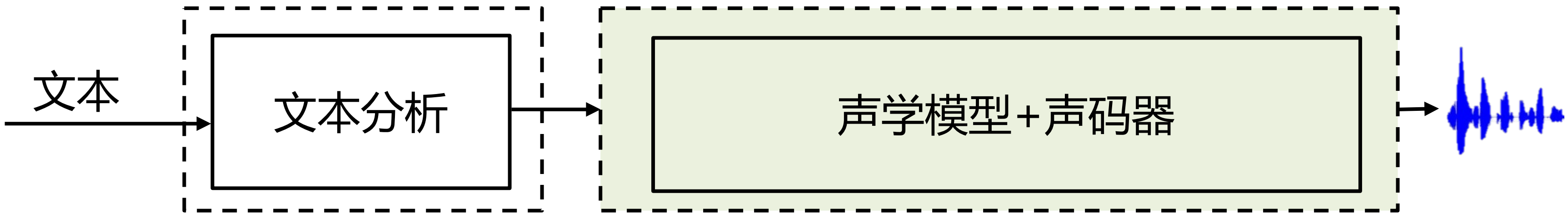
陶建华
jhtao@nlpr.ia.ac.cn

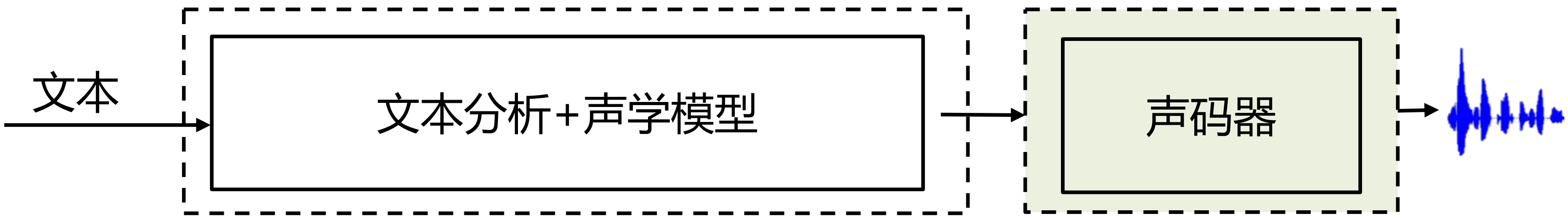
目录

语音合成

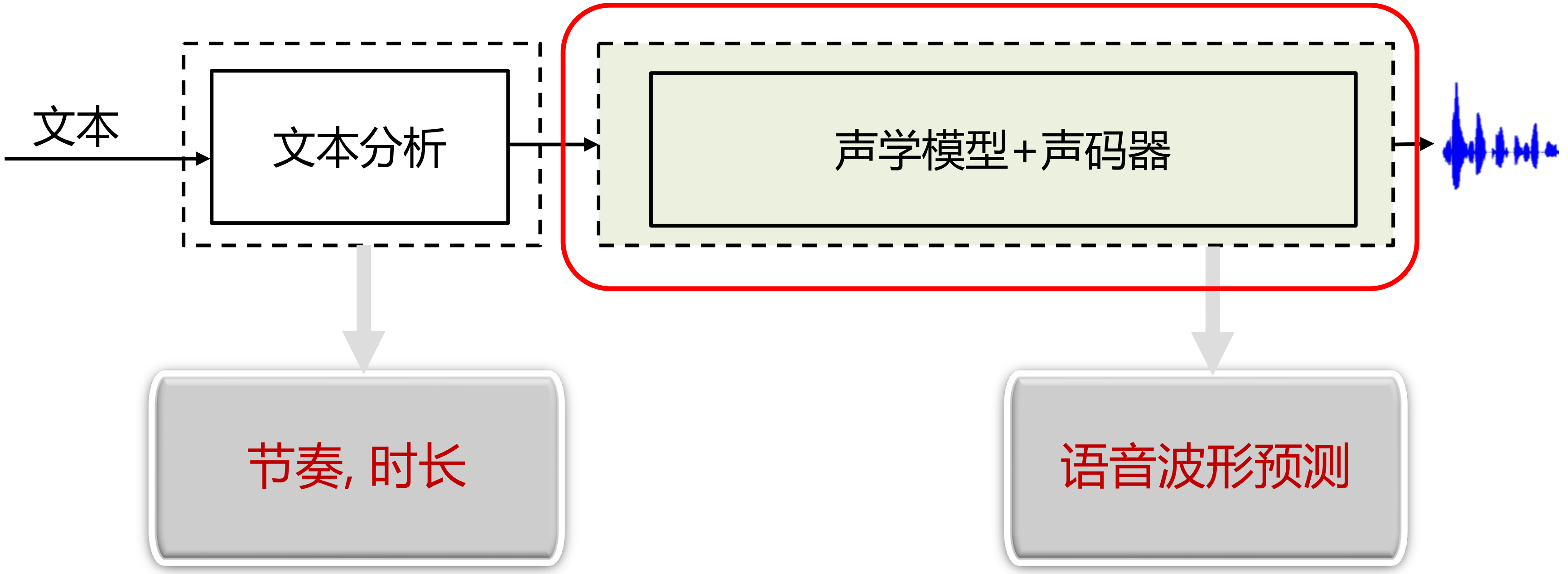


端到端语音合成

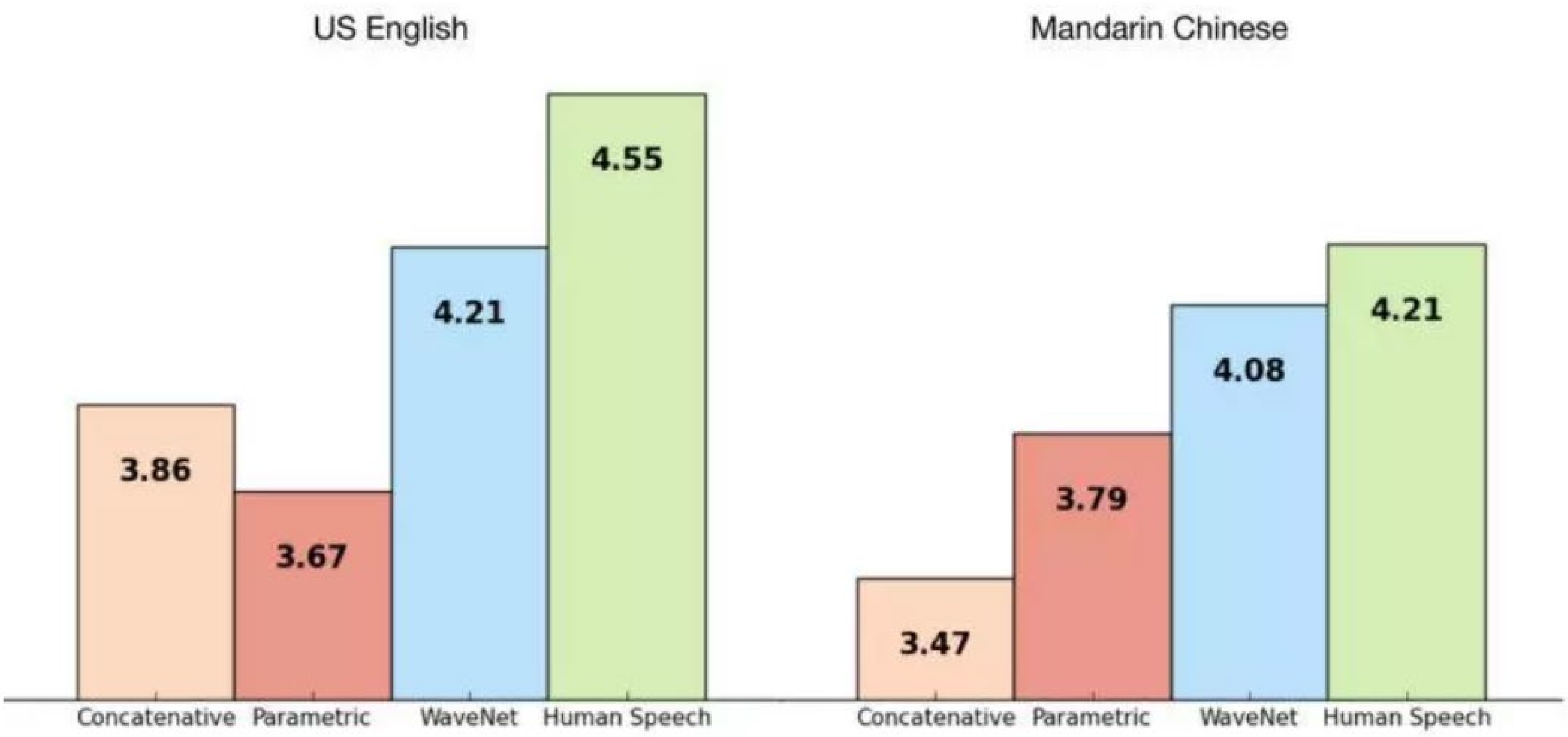




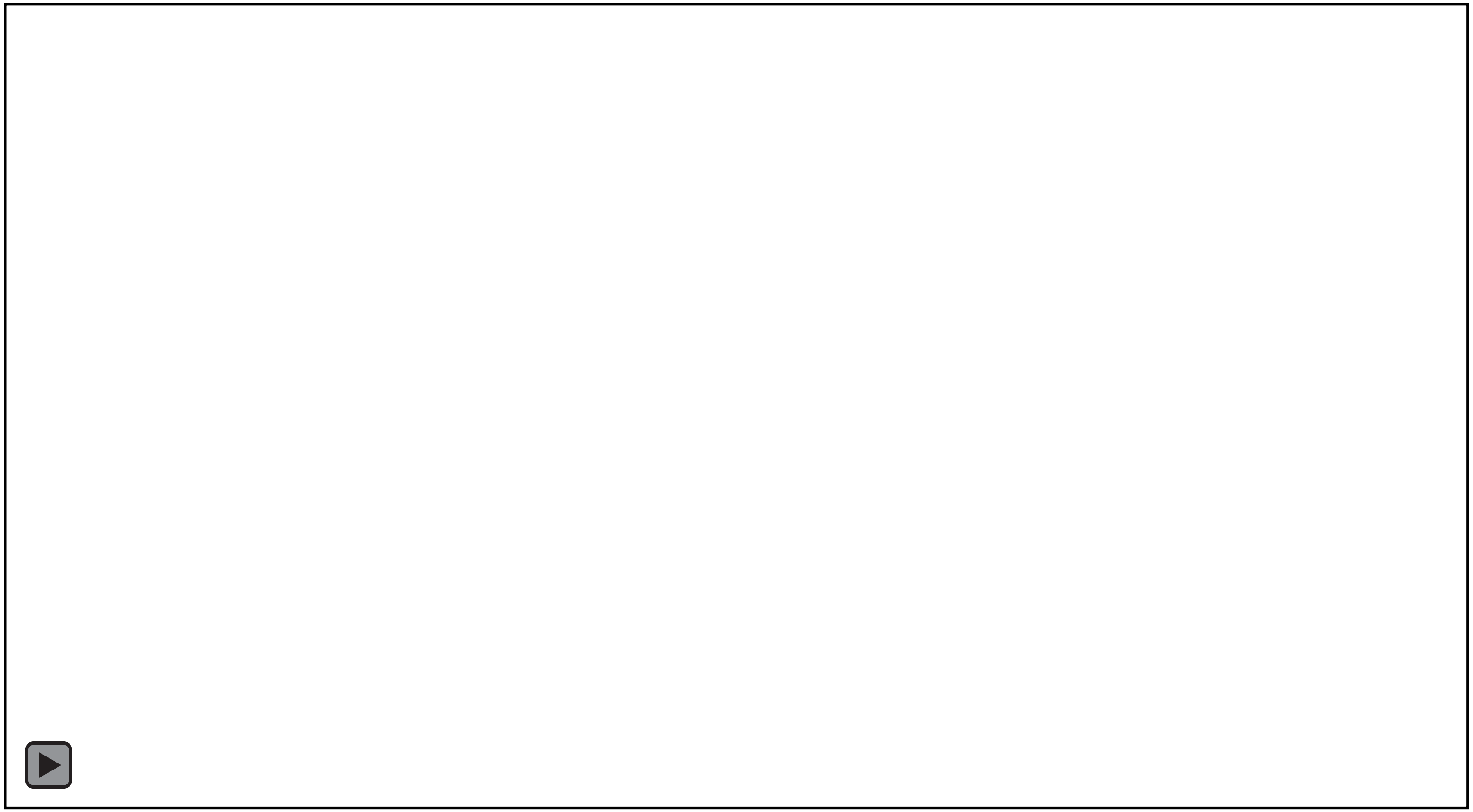
端到端语音合成



WaveNet


US English
3.86
3.67
4.21
4.55
Concatenative
Parametric
WaveNet
Human Speech
Mandarin Chinese
3.47
3.79
4.08
4.21
Concatenative
Parametric
WaveNet
Human Speech

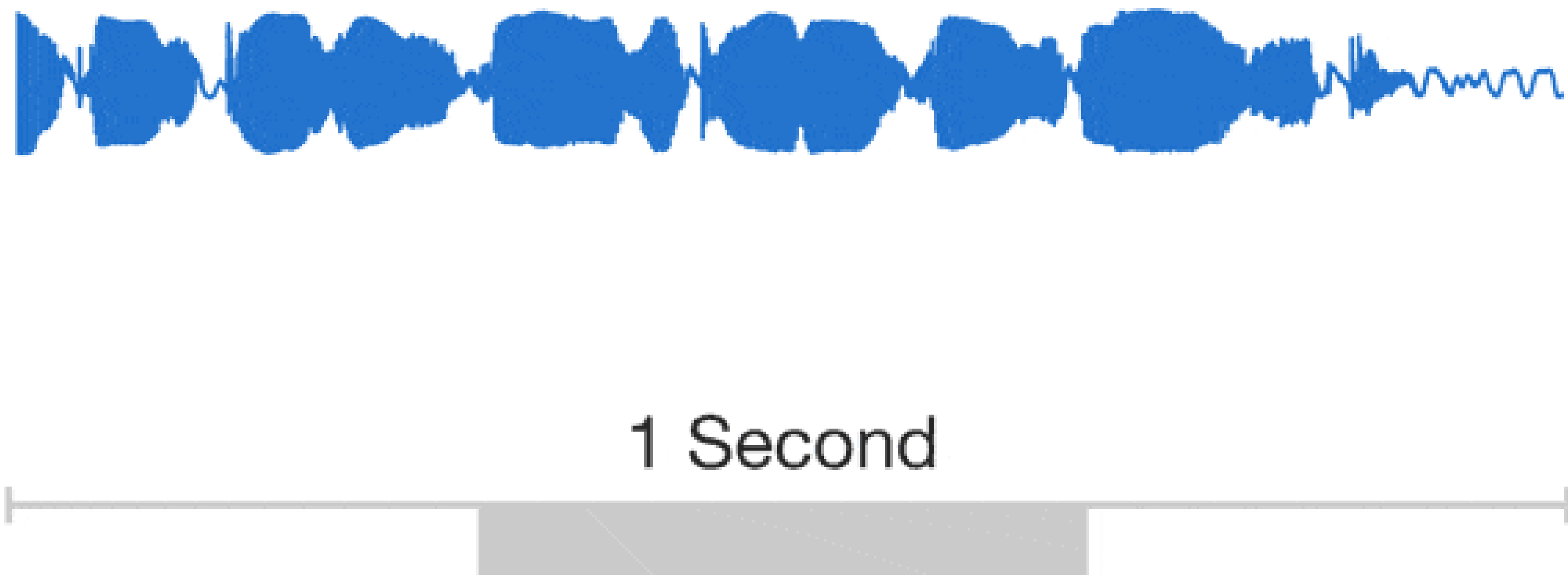
WaveNet



WaveNet by DeepMind [van den Oord 2016]

WaveNet

- 对于16kHz采样率的语音，1秒含有16,000 个采样点



Oord A V D , Dieleman S , Zen H , et al. WaveNet: A Generative Model for Raw Audio[J]. 2016.

WaveNet

- 对于一个语音波形 $\mathbf{x}=\{x_1,\ldots,x_T\}$ ，其联合概率可以分解为条件概率分布的乘积：

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

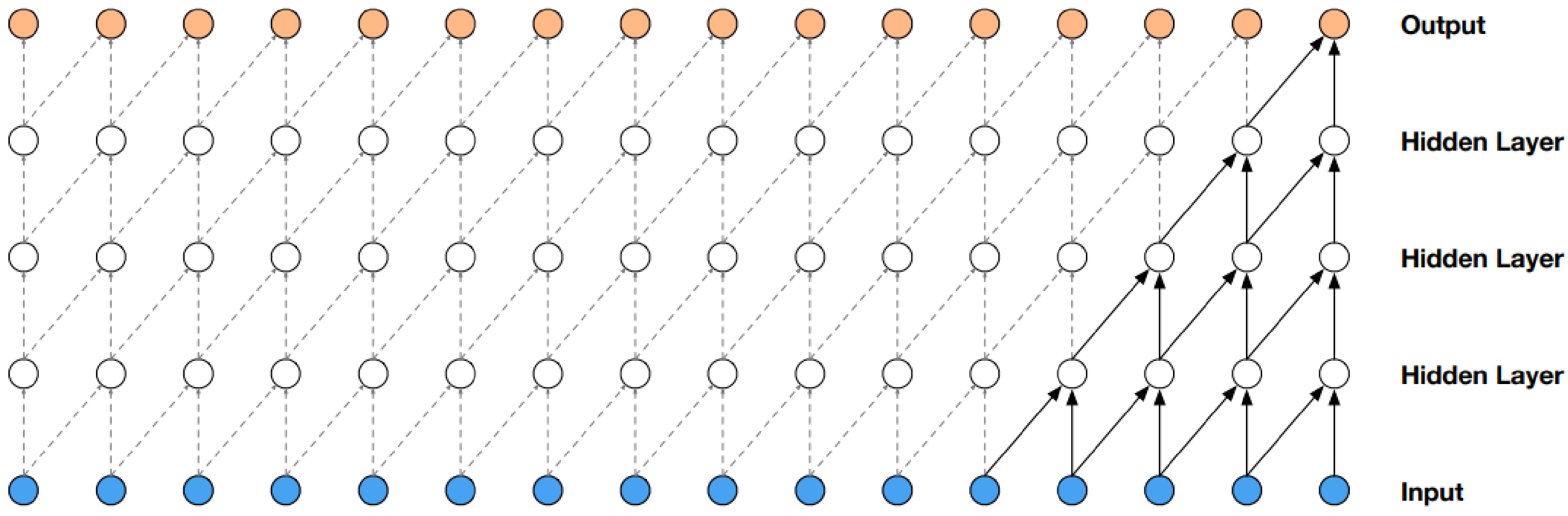
- 条件概率分布可以用堆叠的卷积层建模，思想类似于PixelCNN
  - 只有卷积层，没有池化层
  - 输入和输出的维度相同
  - 输出是离散化的类别分布
  - 优化准则最大似然估计

PixelCNNs (van den Oord et al., 2016a;b)

WaveNet

■ 因果卷积（causal convolutional layers）网络

- 为什么叫因果卷积？因为语音的时序特性，当前采样点的计算只依赖于历史的采样点信息，跟未来的信息没有关系。

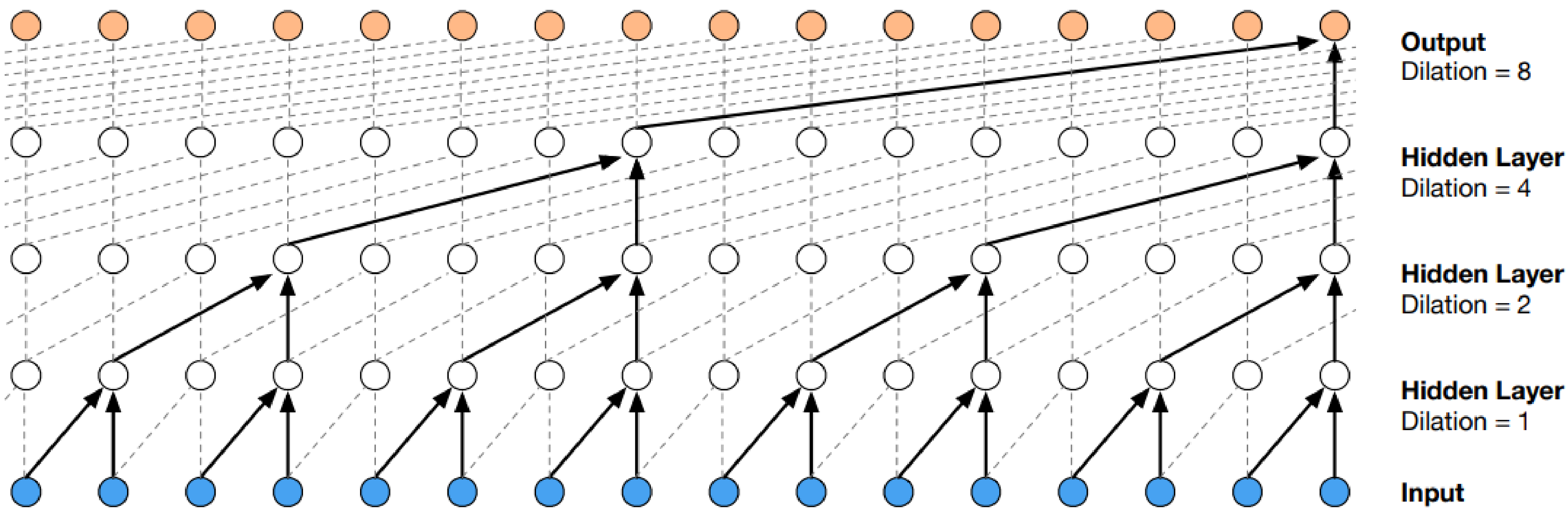


- 优点：没有循环连接，可并行计算，因此速度比RNN快，尤其是长句子
- 不足：需要很多层来增大感受野，如上图感受野为5，网络为4层；计算速度慢

WaveNet

■ 扩张因果卷积（dilated causal convolutional layers）网络

- 扩张因果卷积层也叫有洞的卷积层。

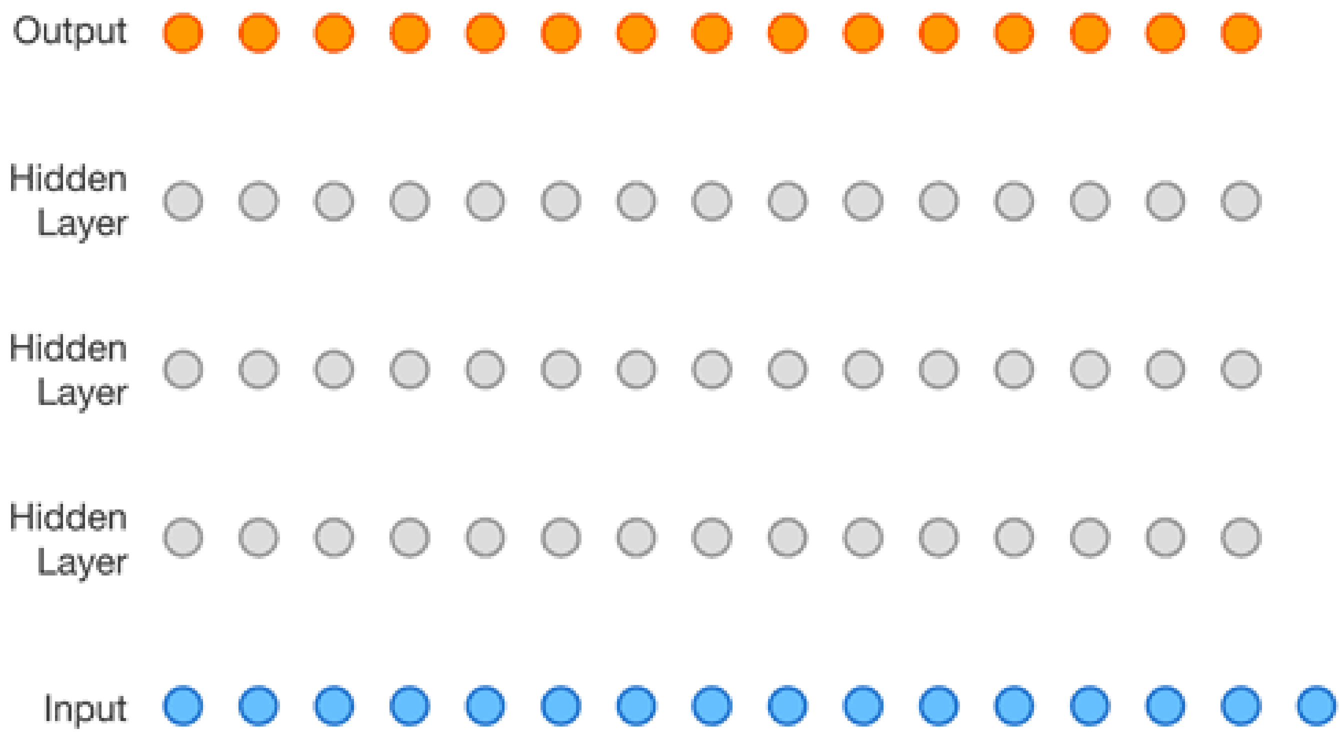


- 优点：需要很少层就能增大感受野，提高计算速度

WaveNet

■ 扩张因果卷积（dilated causal convolutional layers）网络

- 扩张因果卷积层也叫有洞的卷积层。



WaveNet

- 对于一个语音波形 $\mathbf{x}=\{x_1,\ldots,x_T\}$ ：

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

- 输出值的量化

16bit → 8bit, 65536 → 256

  - 每一个采样点用16位的整数表示，softmax层有216=65536个概率值，μ-律压缩转换(ITU-T, 1988)，量化为28=256个类别作为输出：

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

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

WaveNet

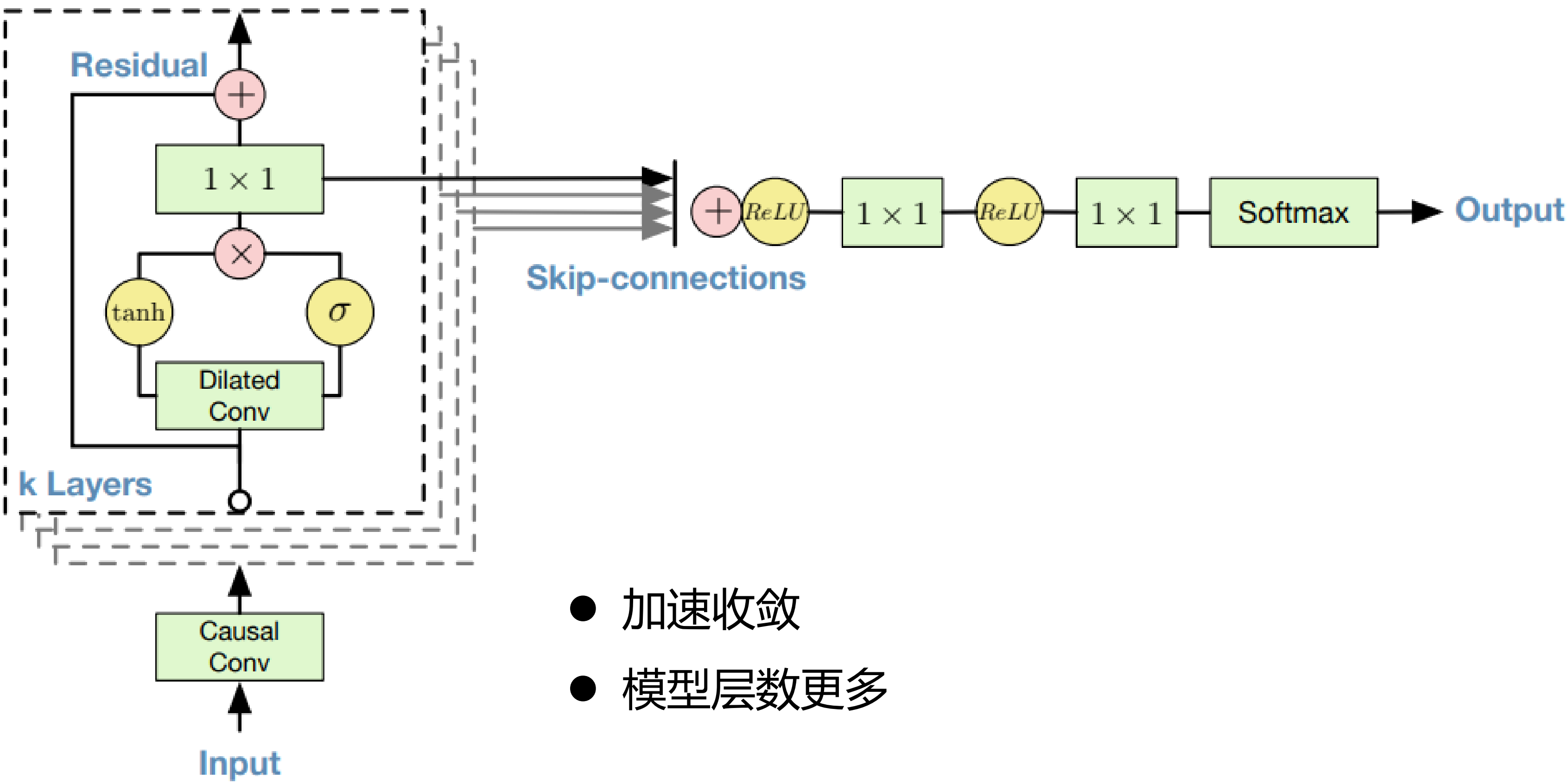
激活函数

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

- $*$ 代表卷积操作
- $\odot$ 代表点乘
- k代表层数下标
- f为滤波器
- g为门控
- W为学习参数

WaveNet

残差连接和skip连接



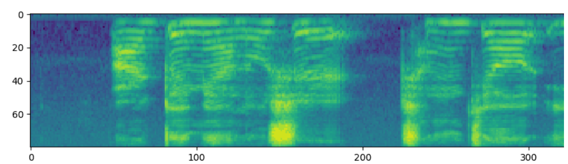
- 加速收敛
- 模型层数更多

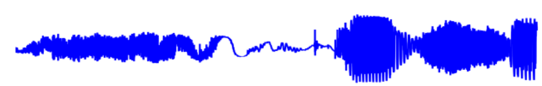
WaveNet

传统声码器 → WaveNet

信号处理 → 深度神经网络

频域 → 时域





Frame-level

Sample-level

条件WaveNet

- 对于一个语音波形 $\mathbf{x}=\{x_1,\ldots,x_T\}$ ，给定条件h，其条件分布：

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

- 全局条件激活函数

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

- 局部条件激活函数

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

条件WaveNet

- 对于一个语音波形 $\mathbf{x}=\{x_1,\ldots,x_T\}$ ，给定条件h，其条件分布：

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

- 全局条件激活函数

可学习的线性映射

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

说话人特性，全局通用

- 局部条件激活函数

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

条件WaveNet

- 对于一个语音波形 $\mathbf{x} = \{x_1, \ldots, x_T\}$ ，给定条件h，其条件分布：

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

- 全局条件激活函数

可学习的线性映射

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

说话人特性，全局通用

- 局部条件激活函数

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

文本特征，时长，基频

WaveNet性能

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)	**4.21** $\pm$ 0.081	**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

WaveNet (L) ： the WaveNet conditioned on linguistic features only

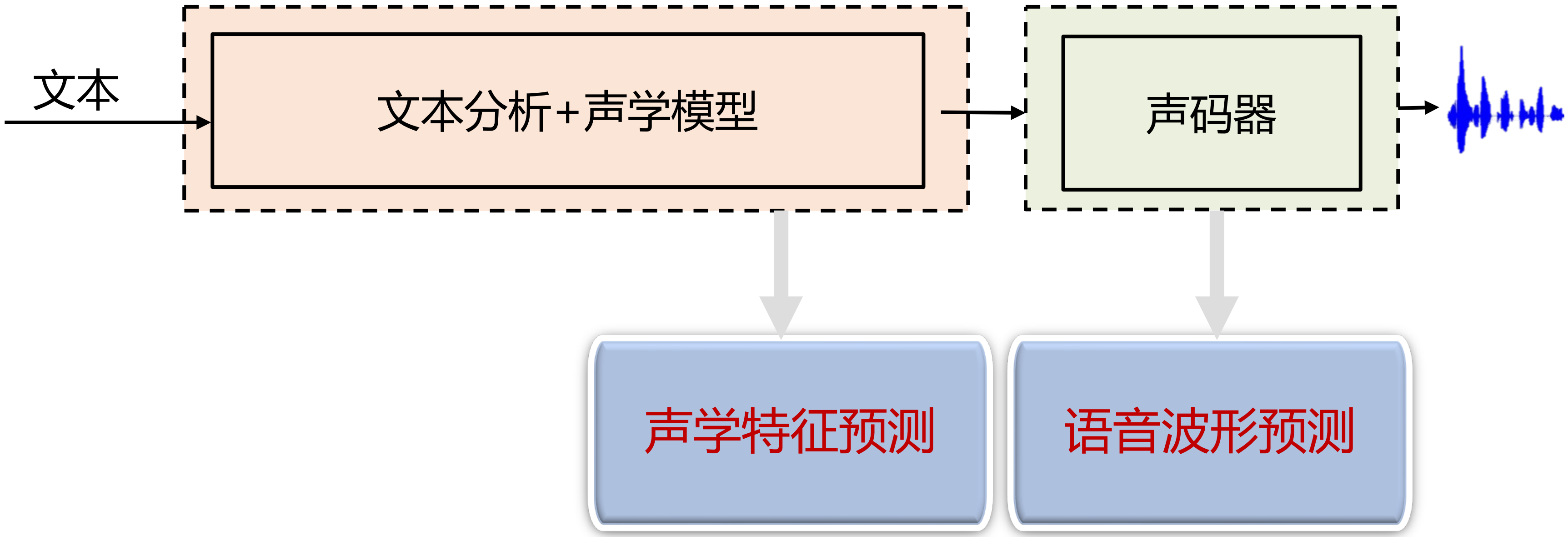
WaveNet (L+F) ： the WaveNet conditioned on both linguistic features and log F0 values

Recording	WaveNet-TTS

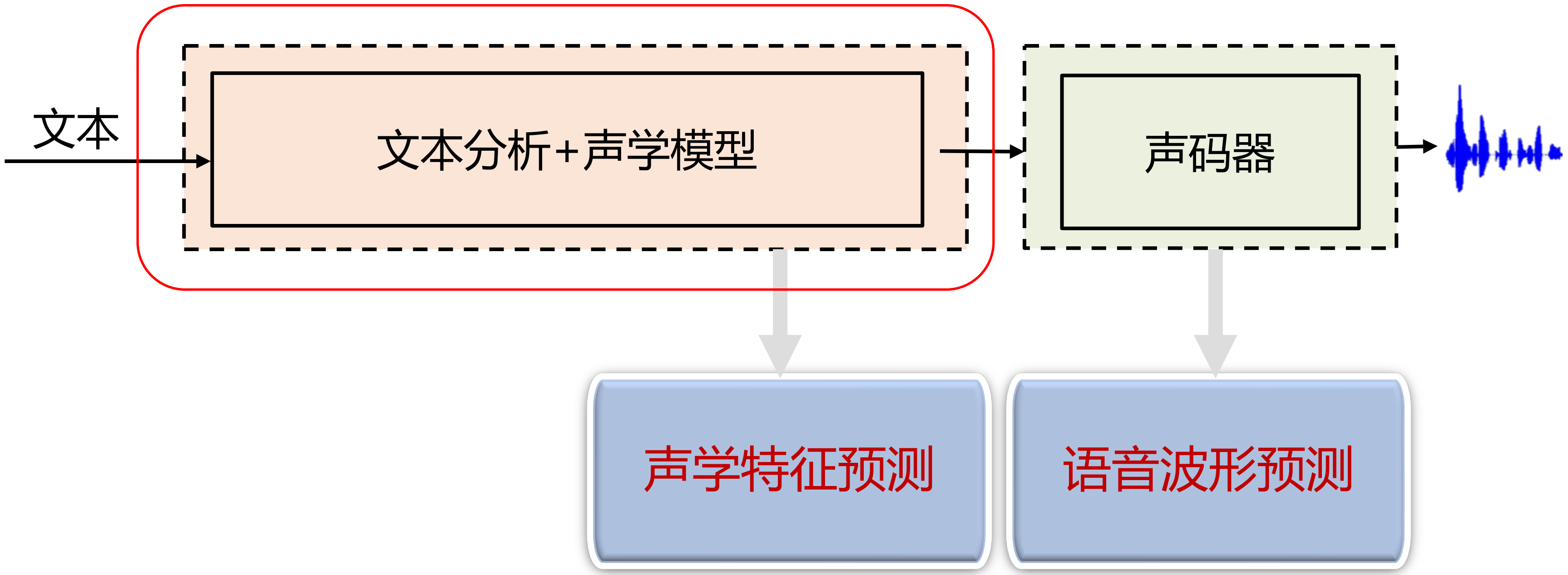
WaveNet不足

- 生成模型
  - 基于采样点的自回归模型，训练耗时
- 文本分析
  - 还需要文本分析模块，预测韵律节奏和时长等。不是真正的端到端，只是声学模型和声码器的结合

端到端语音合成

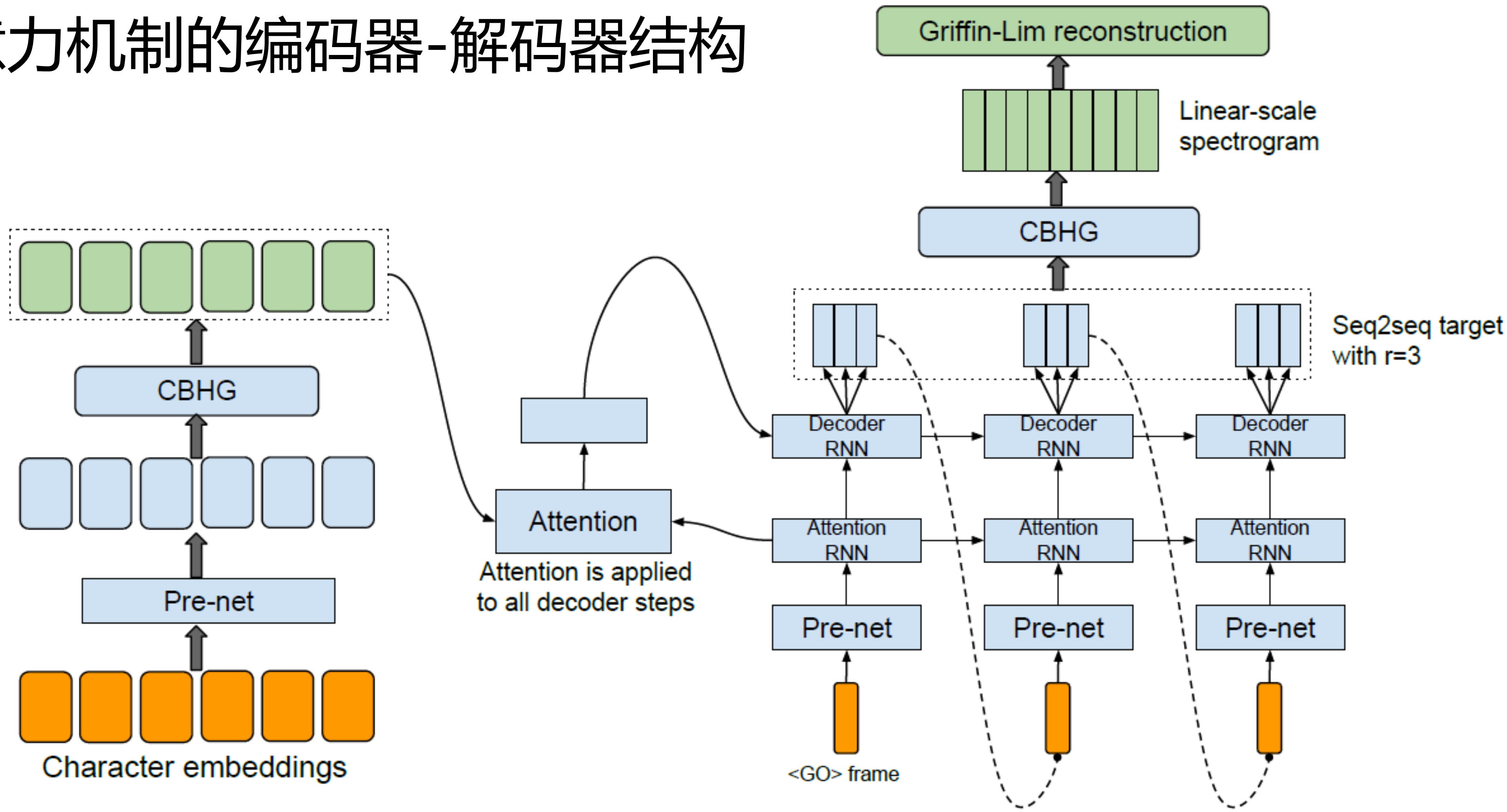


端到端语音合成



Tacotron

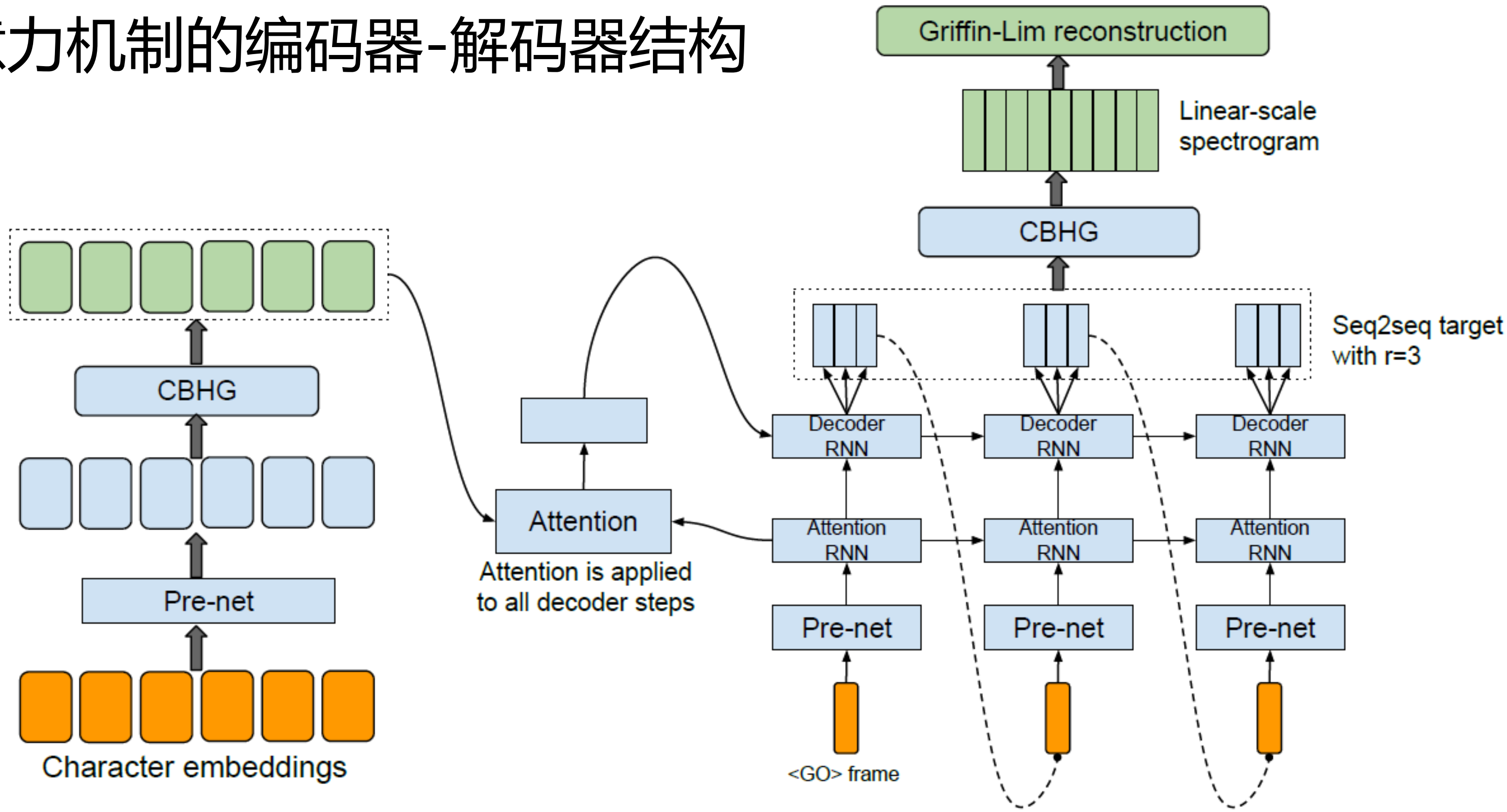
■ 注意力机制的编码器-解码器结构



Wang Y, Skerryryan R J, Stanton D, et al. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model[J]. 2017.

Tacotron

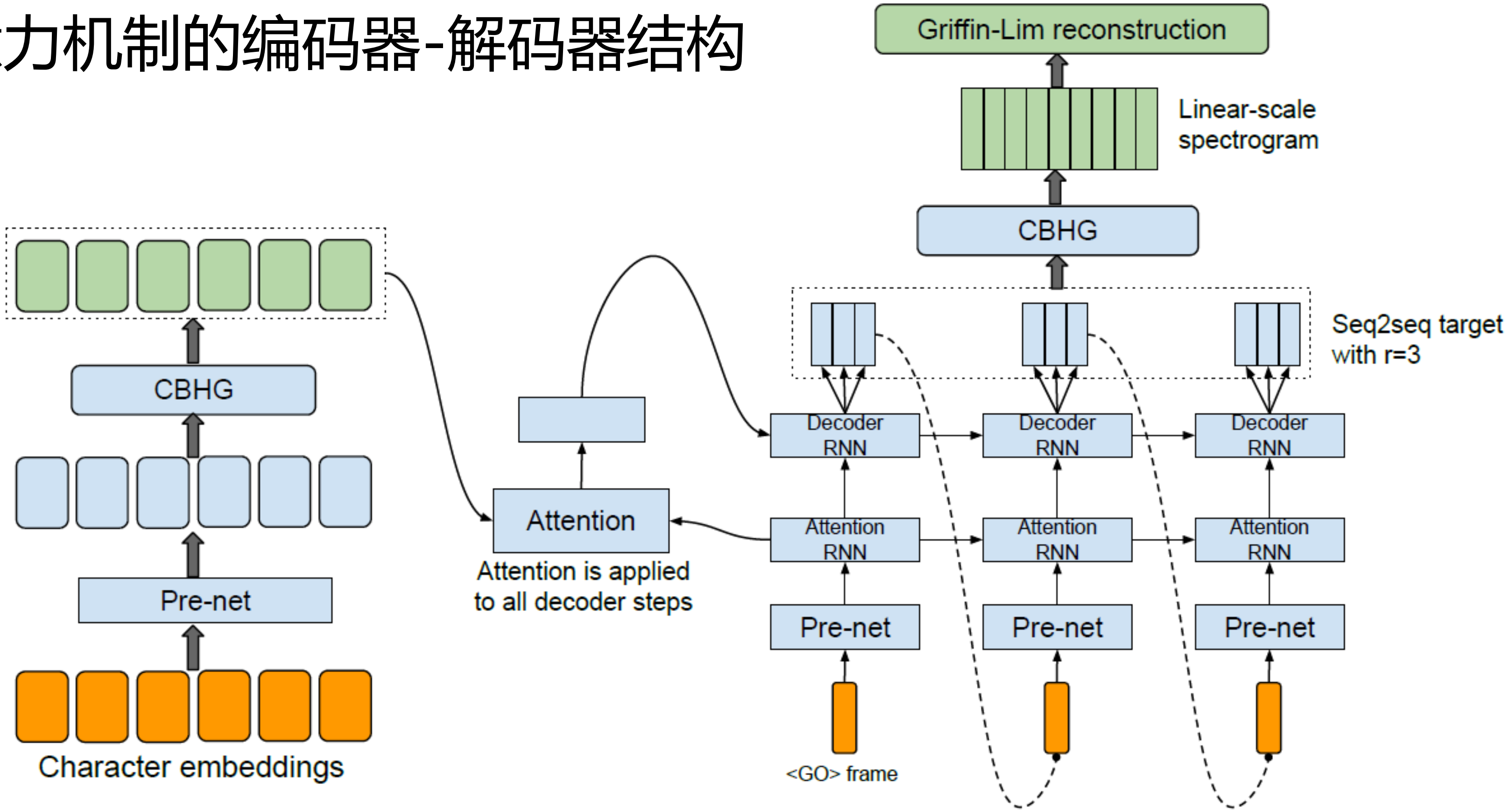
■ 注意力机制的编码器-解码器结构



模 块	输 入	输 出
seq2seq	character sequence	80-band mel-scale spectrogram
post processing network	mel-scale	linear-scale spectrogram
Griffin-Lim	linear-scale	audio

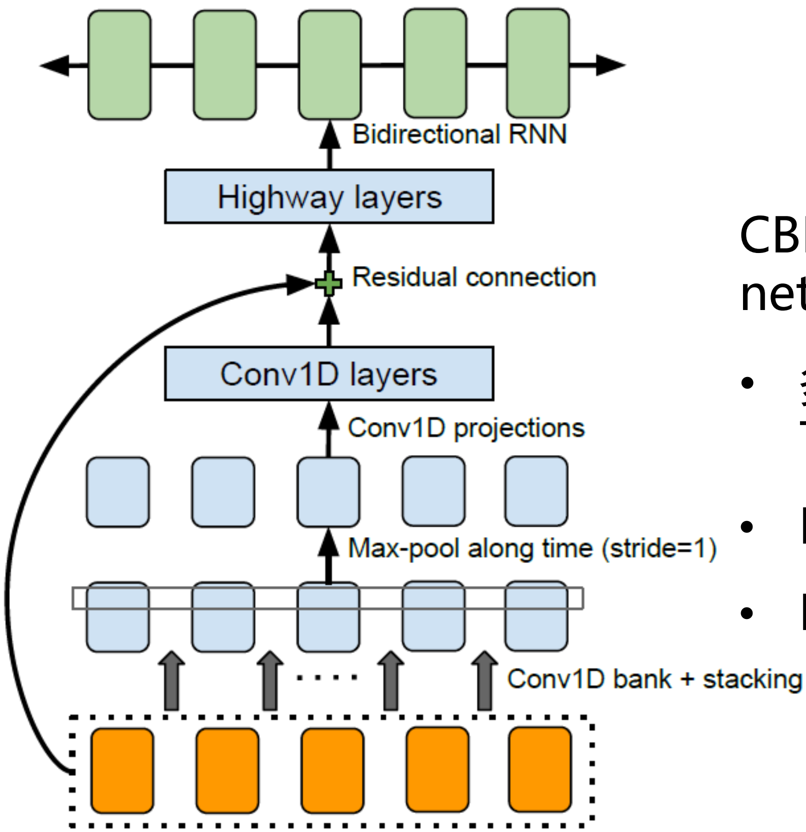
Tacotron

注意力机制的编码器-解码器结构



- **Pre-net**: a set of non-linear transformations
- **CBHG**: Convolution banks, highway, bi-GRU
- **Attention**: conventional soft attention
- **Decoder**: Stacked GRU with residual

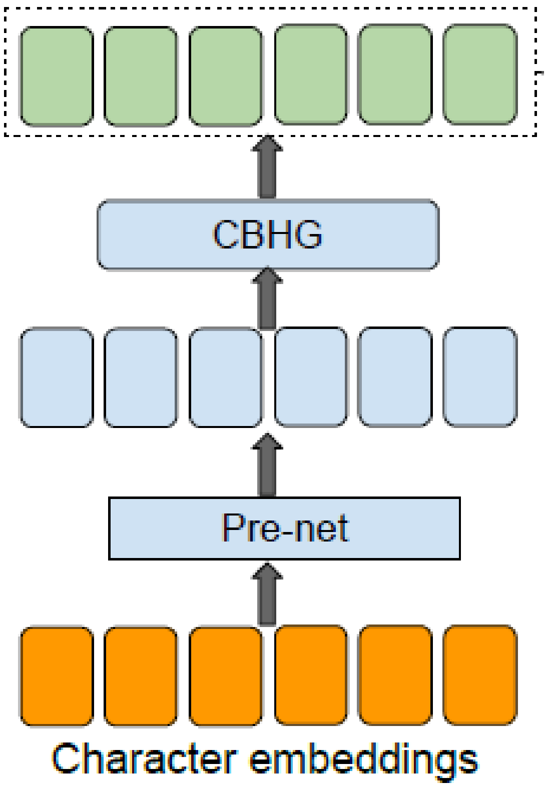
Tacotron-CBHG



CBHG（Convolutional Bank, Highway networks, GRU）

- 多个1-D 卷积滤波器：捕捉局部不变性和上下文信息
- Highway layers：抽取更高层抽象特征
- BGRU：抽取序列特征

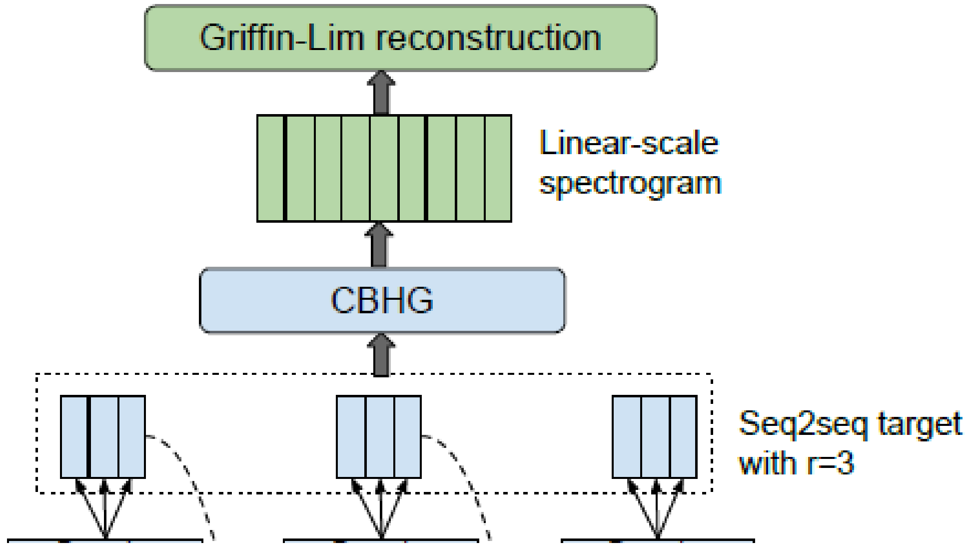
Tacotron-Encoder



Encoder

- 输入：字符序列或者音素序列（对于汉语，输入可以为带调的声韵母序列）
- Pre-net：每个字符向量进行非线性变换
- 带dropout的bottleneck layer：加速收敛和提高泛化能力。
- CBHG：缓解过拟合和发音错误

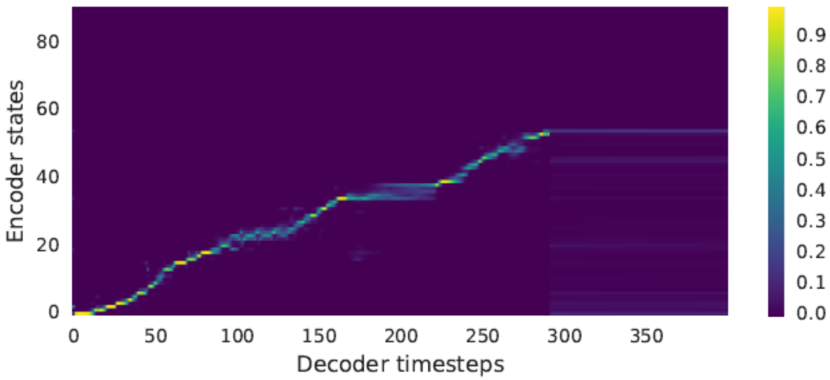
Tacotron-后处理和合成语音



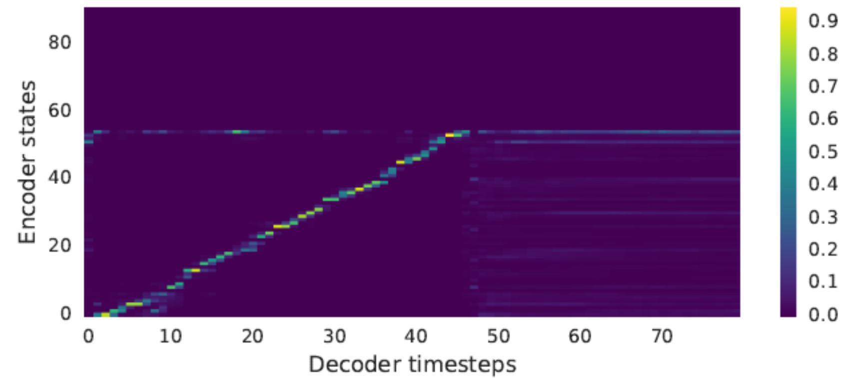
- 后处理：CBHG将mel谱映射为线性谱
- Griffin-Lim：重构相位，将线性谱，通过逆STFT转为语音波形

Tacotron

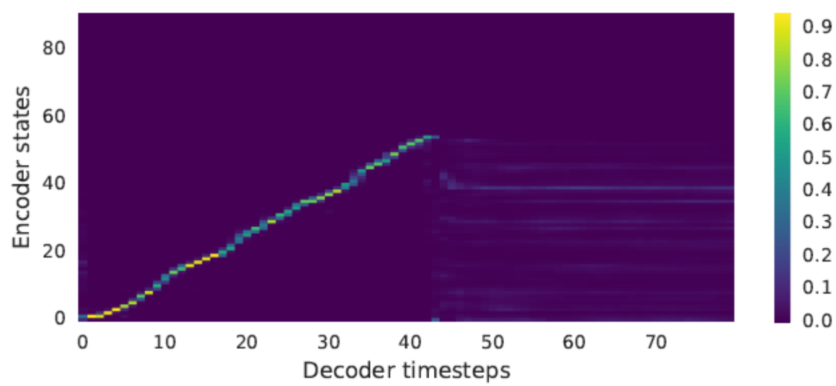
Attention alignments



(a) Vanilla seq2seq + scheduled sampling

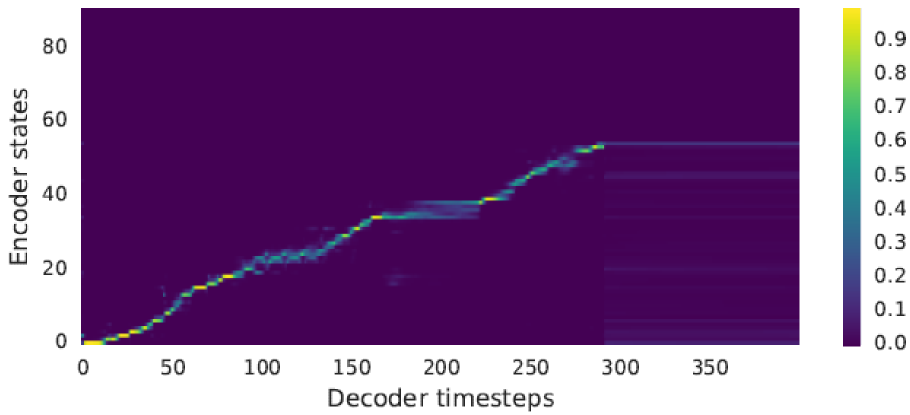


(b) GRU encoder



(c) Tacotron (proposed)

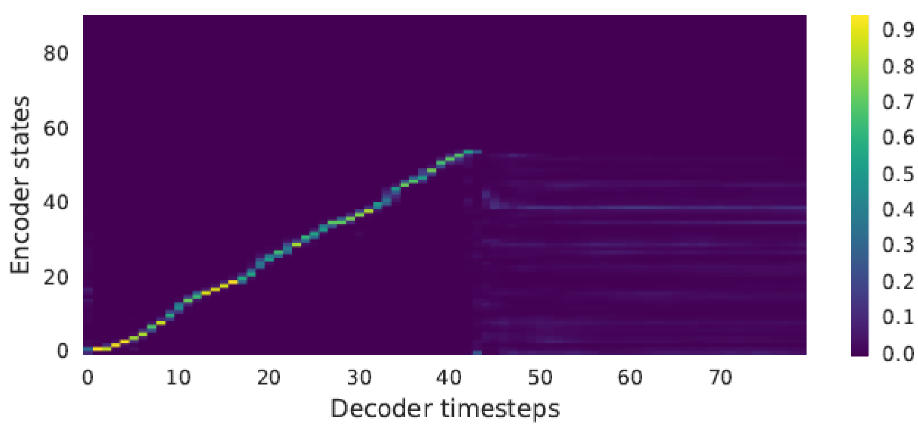
Tacotron



(a) Vanilla seq2seq + scheduled sampling

无pre or post 网络

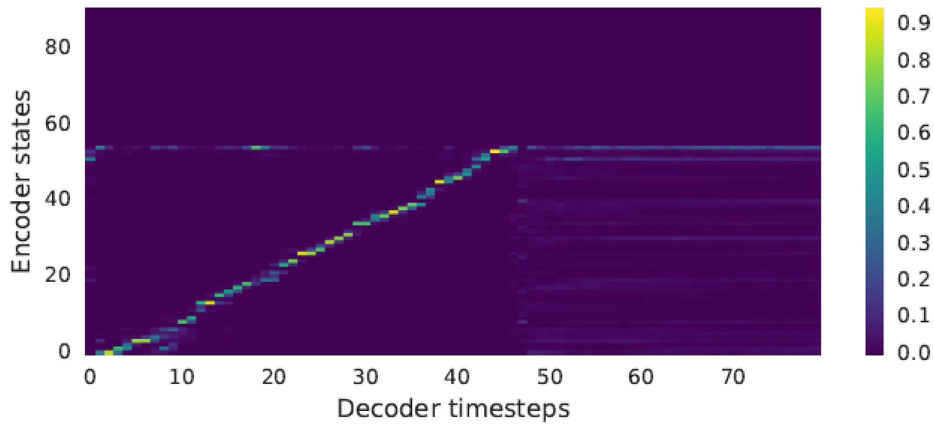
在前进之前会卡很多帧，这会导致合成信号的语音清晰度差。



(c) Tacotron (proposed)

干净、平滑的对齐

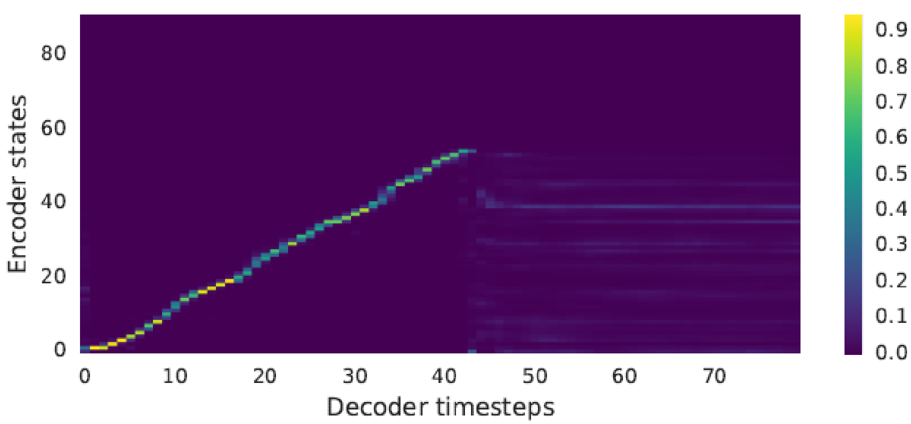
Tacotron



(b) GRU encoder

GRU替换CBHG

- 嘈杂的对齐通常会导致发音错误

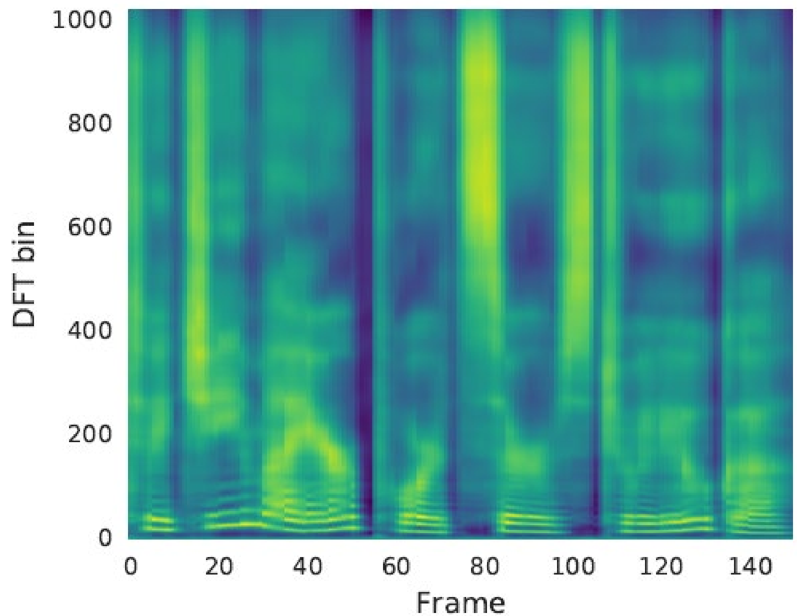


(c) Tacotron (proposed)

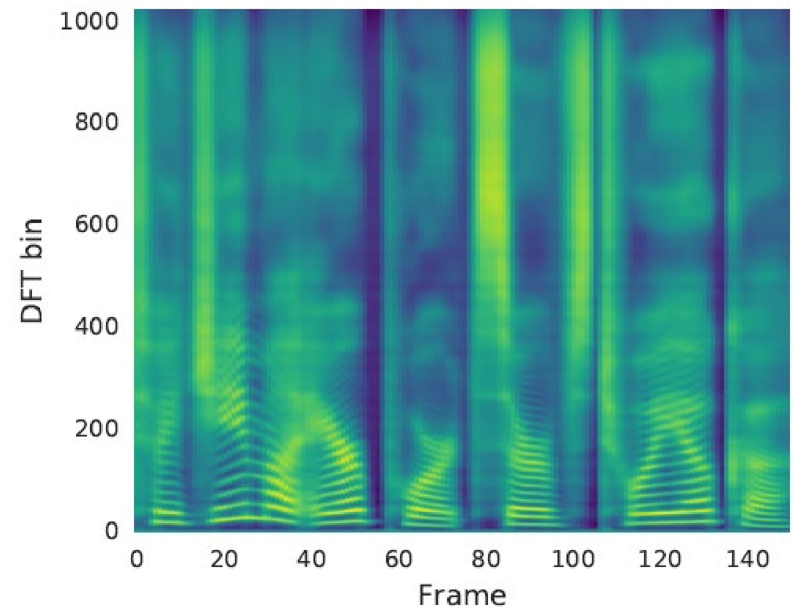
- CBHG 减少了过度拟合现象并针对长和复杂的短语具有很好泛化能力

Tacotron

后处理网络



(a) Without post-processing net



(b) With post-processing net

后处理网络减少“机器味”

- 在100 和400之间有更多的高次谐波
- 高频的共振峰结构，可减少合成语音的机器味。

Tacotron

Table 2: 5-scale mean opinion score evaluation.

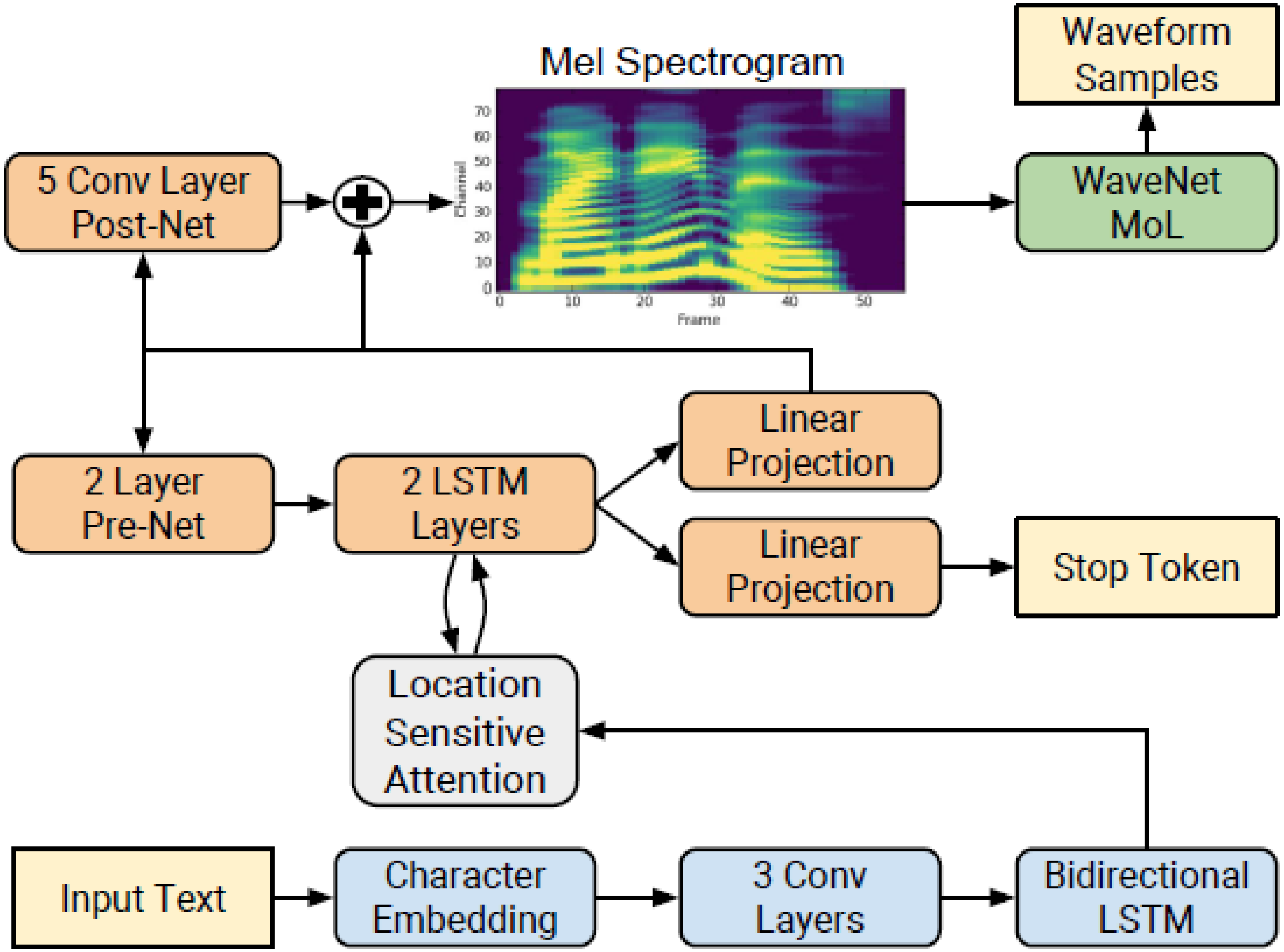
	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

- 优点
  - 有效地实现文本分析和声学模型的端到端
- 不足
  - 性能不如拼接方法
  - 与Wavenet相比，音质有“机器味”

Tacotron2

- 简洁的encoder-decoder模型
  - vanilla LSTM and convolutional layers
- 不足
  - Stop Token
  - 改进的WaveNet声码器

Tacotron2



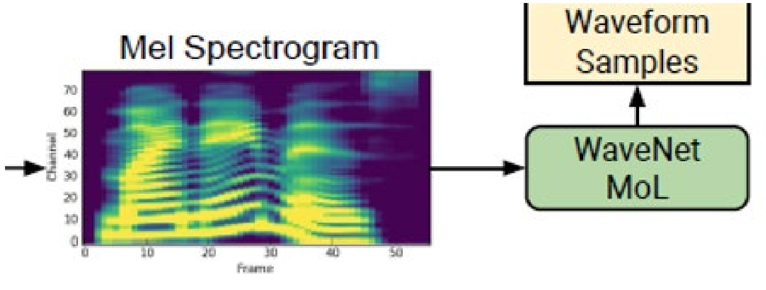
Shen J, Pang R, Weiss R J, et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions[J]. 2017.

Tacotron2-为什么输出梅尔谱

变换容易

- 梅尔谱容易转换为时域波形
- 人耳听觉特性，增强低频信息提高可懂度，减弱高频信息（摩擦音和其他噪声没有必要高保真）

更容易训练



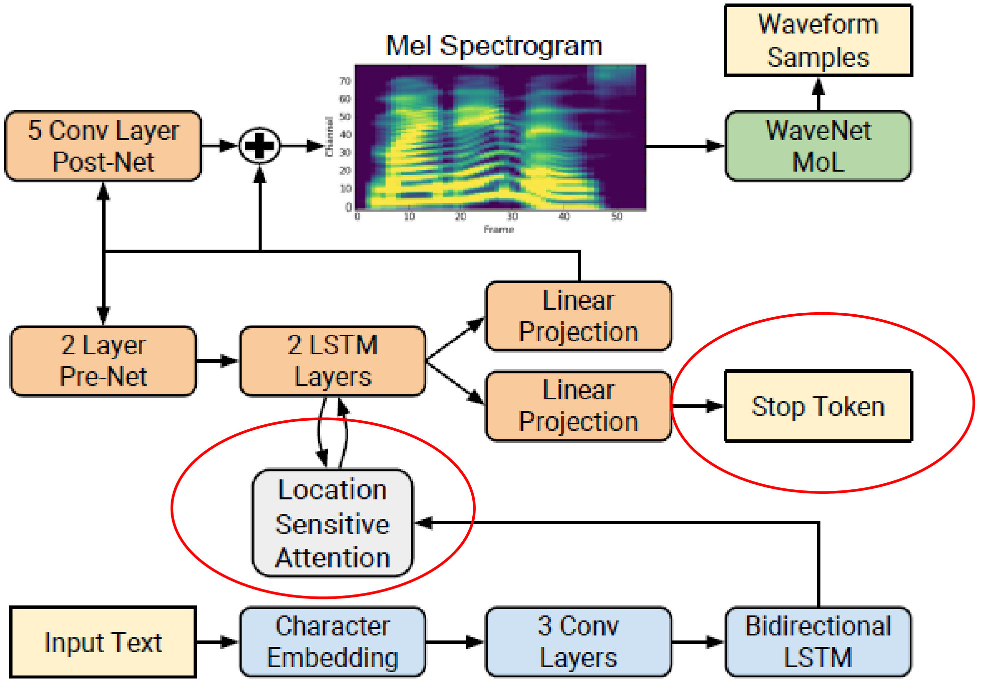
Tacotron2-谱预测网络

- 局部敏感attention

  将原始Tacotron 中的软对齐机制，替换为局部敏感注意力机制，能够有效减少漏音发生的概率。

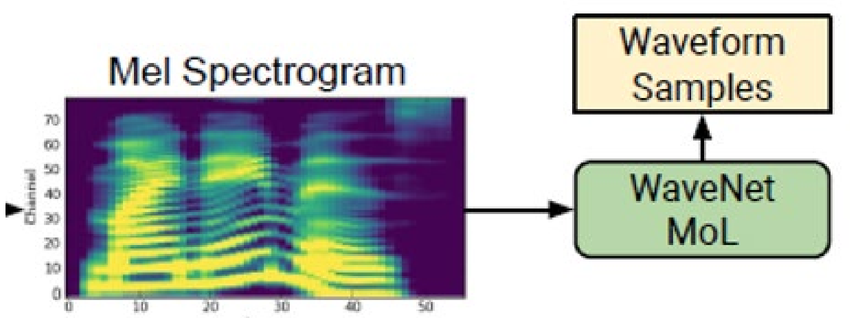
- Stop Token

  和原始Tacotron相比，增加了语音结束位置的预测损失，能缓解语音合成过程中出现尾音的问题。



Tacotron2-WaveNet声码器

- 输入：梅尔谱
- 输出
  - 采样点的分布（假设符合某种分布：如混合逻辑分布，预测其参数）
  - 分布的具体参数（对于混合逻辑分布，需要预测其mean, log scale, mixture weight）
- 损失函数
  - 预测的采样点与真实采样点的似然度



Tacotron2

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

	Synthesis	
Training	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

Table 2. Comparison of evaluated MOS for our system when WaveNet trained on predicted/ground truth mel spectrograms are made to synthesize from predicted/ground truth mel spectrograms.

Tacotron2

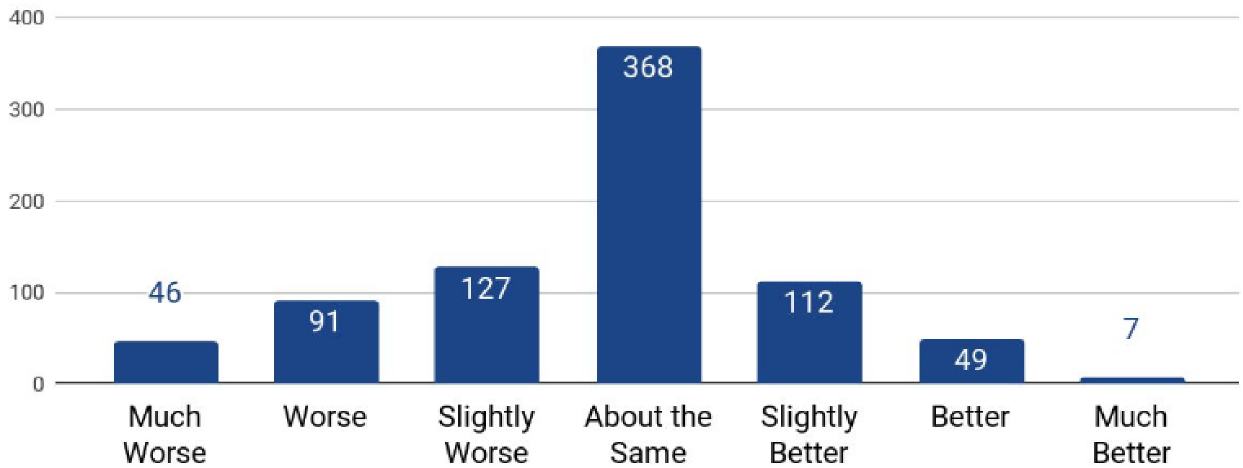
System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	$\mathbf{4.526 \pm 0.066}$

Table 3. Comparison of evaluated MOS for Griffin-Lim vs. WaveNet as a vocoder, and using 1,025-dimensional linear spectrograms vs. 80-dimensional mel spectrograms as conditioning inputs to WaveNet.

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

Table 4. WaveNet with various layer and receptive field sizes.

Tacotron2



Recording	WaveNet

Tacotron vs Tacotron2

模块	Tacotron	Tacotron2
输入层	字符系列	字符系列
文本编码器	Prenet + CBHG	Convolution + LSTM
注意力机制	软对齐	局部敏感注意力机制（减少漏音）
停顿预测	无	有（缓解尾音问题）
输出目标	线性谱	梅尔谱
声码器	GL	WaveNet

Tacotron vs Tacotron2

文本	录音	Tacotron	Tacotron2
爸爸和继母又是典型的城市人，习惯晚睡晚起。			
很像灵堂内的花圈魂幡。			
白色的浪花紧紧地追逐在船后。			

参考文献

- Alex Graves, Sequence transduction with recurrent neural networks. 2012.
- Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2014.
- Kalchbrenner N, Elsen E, Simonyan K, et al. Efficient Neural Audio Synthesis[J]. 2018.
- Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, Ming Zhou. Neural Speech Synthesis with Transformer Network. To appear in AAAI, 2019
- Oord A V D, Li Y, Babuschkin I, et al. Parallel WaveNet: Fast High-Fidelity Speech Synthesis[J]. 2017.
- Oord A V D , Dieleman S , Zen H , et al. WaveNet: A Generative Model for Raw Audio[J]. 2016.
- Shen J, Pang R, Weiss R J, et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions[J]. 2017.
- Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. 2017.
- Wang Y, Skerryryan R J, Stanton D, et al. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model[J]. 2017.

参考文献

- Zen H, Nose T, Yamagishi J, et al. The HMM-based speech synthesis system (HTS) version 2.0. [C]. In SSW. 2007: 294–299.
- Zen H , Sak H . Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C]// IEEE International Conference on Acoustics. IEEE, 2015.
- Zen H, Agiomyrgiannakis Y, Egberts N, et al. Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices[J]. 2016:2273-2277.
- Olah & Carter, "Attention and Augmented Recurrent Neural Networks", Distill, 2016. http://doi.org/10.23915/distill.00001
- Dumoulin V , Visin F . A guide to convolution arithmetic for deep learning[J]. 2016.
- Kingma D P , Dhariwal P . Glow: Generative Flow with Invertible 1x1 Convolutions[J]. 2018.
- Prenger R, Valle R, Catanzaro B. WaveGlow: A Flow-based Generative Network for Speech Synthesis[J]. arXiv preprint arXiv:1811.00002, 2018.

谢谢！