1. ClariNet: ParallelWave Generation in End-to-End Text-to-Speech

作者：Wei Ping, Kainan Peng, Jitong Chen （Baidu Research）

时间：2018.07.30

分类：end-to-end

demo: https://clarinet-demo.github.io/

paper: https://arxiv.org/pdf/1807.07281.pdf

code: 暂时还没有官方的源代码

特点：直接采用text-to-wave的全卷积神经网络，优于text-to-spectrogram+wavenet单独训练模型

Abstract:

In this work, we propose an alternative solution for parallel wave generation by WaveNet. In contrast to parallel WaveNet (Oord et al., 2018), we distill a Gaussian inverse autoregressive flow from the autoregressiveWaveNet by minimizing a novel regularized KL divergence between their highly-peaked output distributions. Our method computes the KL divergence in closed-form, which simplifies the training algorithm and provides very efficient distillation. In addition, we propose the first **text-to-wave** neural architecture for speech synthesis, which is fully convolutional and enables fast end-to-end training from scratch. It significantly outperforms the previous pipeline that connects a text-to-spectrogram model to a separately trained WaveNet (Ping et al., 2018). We also successfully distill a parallel waveform synthesizer conditioned on the hidden representation in this end-to-end model.

Contributions:

1. We demonstrate that a single variance-bounded Gaussian is sufficient for modeling the raw waveform in WaveNet without degradation of audio quality. Our Gaussian autoregressive WaveNet is simply trained with maximum likelihood estimation (MLE).

2. We distill a Gaussian IAF from the autoregressive WaveNet by minimizing a novel regularized KL divergence between their peaked output distributions. Our method provides closed-form estimation of KL divergence, which largely simplifies the distillation algorithm and stabilizes the training process.

3. In previous studies, “end-to-end" speech synthesis actually refers to the text-to-spectrogram models with a separate waveform synthesizer (i.e., vocoder) (Sotelo et al., 2017; Wang et al., 2017). We propose the first text-to-wave neural architecture for TTS, which is fully convolutional and enables fast end-to-end training from scratch. Our text-to-wave model significantly outperforms the separately trained pipeline (Ping et al., 2018) in naturalness.

4. We also successfully distill a parallel neural vocoder conditioned on the learned hidden representation within the end-to-end architecture. The text-to-wave model with

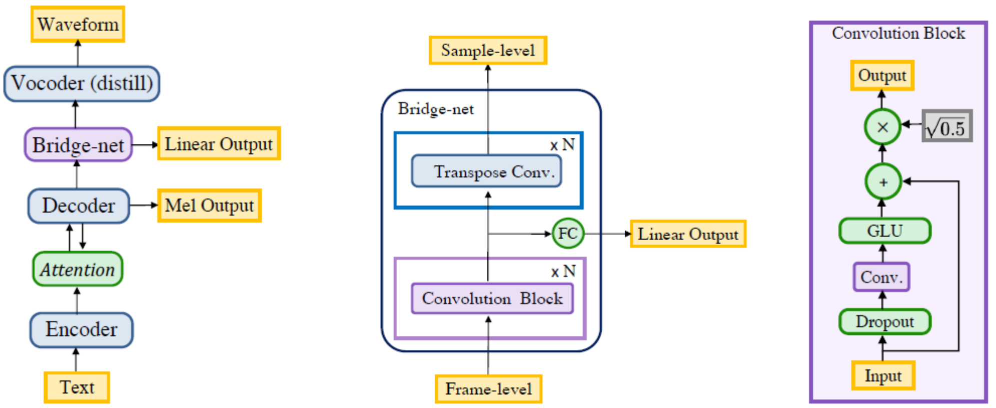
the parallel vocoder obtains competitive results as the model with an autoregressive vocoder.

Model:

In this section, we present our convolutional text-to-wave architecture (see Fig. 2 (a)) for end-toend TTS. Our architecture is based on Deep Voice 3 (DV3), a convolutional attention-based TTS

system (Ping et al., 2018). DV3 is capable of converting textual features (e.g., characters, phonemes and stresses) into spectral features (e.g., log-mel spectrograms and log-linear spectrograms). These spectral features can be used as inputs for a separately trained waveform synthesis model, such as WaveNet. In contrast, we directly feed the hidden representation learned from the attention mechanism to the neural vocoder through some intermediate processing, and train the whole model from scratch in an end-to-end manner.

注：Deep Voice3可以直接把textual特征转为频谱特征，而且带有Attention-based，可以了解一下。



(a) Text-to-wave architecture (b) Bridge-net (c) Convolution block

Figure 2: (a) Text-to-wave model converts textual features into waveform. All components feed their hidden representation to others directly. (b) Bridge-net maps frame-level hidden representation to sample-level through several convolution blocks and transposed convolution layers interleaved with *softsign* non-linearities. (c) Convolution block is based on gated linear unit.

The proposed architecture consists of four components:

Encoder: A convolutional encoder as in DV3, which encodes textual features into an internal hidden representation. * 把文本信息表示为一个隐含的状态表达

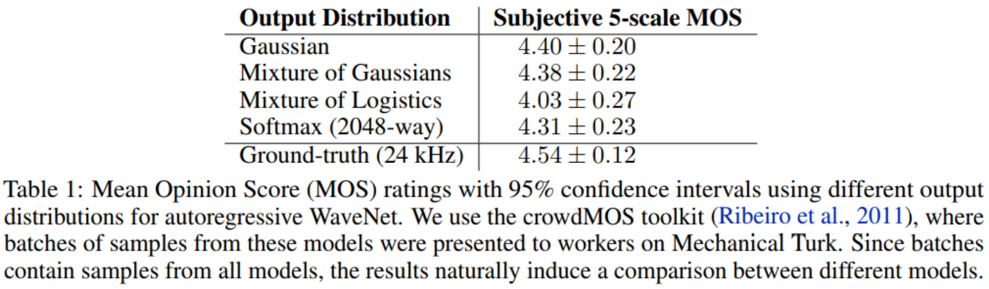
Decoder: A causal convolutional decoder as in DV3, which decodes the encoder representation with attention into the log-mel spectrogram in an autoregressive manner. * 把encoder的hidden representation 转换为带有特定物理意义的representation，如log-mel spectrogram。

Bridge-net: A convolutional intermediate processing block, which processes the hidden representation from the decoder and predict log-linear spectrogram. Unlike the decoder, it is non-causal and can thus utilize future context. In addition, it upsamples the hidden representation from frame-level to sample-level. * 生成的mel spectrogram是frame-level级别的，通过upsampling，转为sample-level，（实验中用的音频是24khz的，因此，80frame/s-->24000bits/s）。

Vocoder: A Gaussian autoregresive WaveNet to synthesize the waveform, which is conditioned on the upsampled hidden representation from the bridge-net. This component can be replaced by a student IAF distilled from the autoregresive vocoder. * 将生成的sample-level特征送到Vocoder里面进行语音合成。

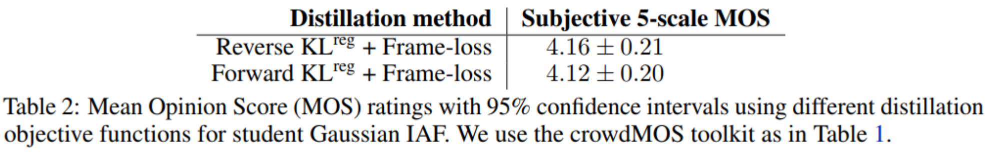
The overall objective function is a linear combination of the losses from decoder, bridge-net and vocoder; we simply set all coefficients to one in experiments. We introduce bridge-net to utilize future temporal information as it can apply non-causal convolution. All modules in our architecture are convolutional, which enables fast training 8 and alleviates the common difficulties in RNN-based models (e.g., vanishing and exploding gradient problems (Pascanu et al., 2013)). Throughout the whole model, we use the convolution block from DV3 (see Fig. 2(c)) as the basic building block. It consists of a 1-D convolution with a gated linear unit (GLU) (Gehring et al., 2017) and a residual connection. We set the dropout probability to 0.05 in all experiments. We give further details in the following subsections.

Result:



Output Distribution	Subjective 5-scale MOS
Gaussian	4.40 ± 0.20
Mixture of Gaussians	4.38 ± 0.22
Mixture of Logistics	4.03 ± 0.27
Softmax (2048-way)	4.31 ± 0.23
Ground-truth (24 kHz)	4.54 ± 0.12

Table 1: Mean Opinion Score (MOS) ratings with 95% confidence intervals using different output distributions for autoregressive WaveNet. We use the crowdMOS toolkit (Ribeiro et al., 2011), where batches of samples from these models were presented to workers on Mechanical Turk. Since batches contain samples from all models, the results naturally induce a comparison between different models.



Distillation method	Subjective 5-scale MOS
Reverse $\mathrm{KL}^{\mathrm{reg}}$ + Frame-loss	4.16 ± 0.21
Forward $\mathrm{KL}^{\mathrm{reg}}$ + Frame-loss	4.12 ± 0.20

Table 2: Mean Opinion Score (MOS) ratings with 95% confidence intervals using different distillation objective functions for student Gaussian IAF. We use the crowdMOS toolkit as in Table 1.

2. Deep Voice3: SCALING TEXT-TO-SPEECH WITH CONVOLUTIONAL SEQUENCE LEARNING

作者：Wei Ping, Kainan Peng, Andrew Gibiansky （Baidu Research）

时间：2018.02.22

分类：end-to-end

demo: https://r9y9.github.io/deepvoice3_pytorch/

paper: http://arxiv.org/pdf/1710.07654.pdf

code: https://github.com/r9y9/deepvoice3_pytorch 【pytorch实现】

特点：采用基于attention的全卷积TTS系统，提高了训练速度，并使用了大量的训练数据，其次论证了基于Attention的语音合成网络常出现的错误，并提供了解决方案。

Abstract:

We present Deep Voice 3, a fully-convolutional attention-based neural text-to- speech (TTS) system. Deep Voice 3 matches state-of-the-art neural speech synthesis systems in naturalness while training an order of magnitude faster. We scale Deep Voice 3 to dataset sizes unprecedented for TTS, training on more than eight hundred hours of audio from over two thousand speakers. In addition, we identify common error modes of attention-based speech synthesis networks, demonstrate how to mitigate them, and compare several different waveform synthesis methods. We also describe how to scale inference to ten million queries per day on a single GPU server.

Contributions:

1. We propose a fully-convolutional character-to-spectrogram architecture, which enables fully parallel computation and trains an order of magnitude faster than analogous architectures using recurrent cells (e.g., Wang et al., 2017).

2. We show that our architecture trains quickly and scales to the LibriSpeech ASR dataset (Panayotov et al., 2015), which consists of 820 hours of audio data from 2484 speakers.

3. We demonstrate that we can generate attention behavior, avoiding error modes commonly affecting sequence-to-sequence models. 【单音调】

4. We compare the quality of several waveform synthesis methods, includingWORLD (Morise et al., 2016), Griffin-Lim (Griffin & Lim, 1984), and WaveNet (Oord et al., 2016).

5. We describe the implementation of an inference kernel for Deep Voice 3, which can serve up to ten million queries per day on one single-GPU server.

Deep Voice 3 demonstrates the utility of monotonic attention during training in TTS, a new domain where monotonicity is expected. Alternatively, we show that with a simple heuristic to only enforce monotonicity during inference, a standard attention mechanism can work just as well or even better. Deep Voice 3 also builds upon the convolutional sequence-to-sequence architecture from Gehring et al. (2017) by introducing a positional encoding similar to that used in Vaswani et al. (2017), augmented with a rate adjustment to account for the mismatch between input and output domain lengths. 加入了位置编码，从而可以进行速度调节。

Model:

In this section, we present our fully-convolutional sequence-to-sequence architecture for TTS (see Fig. 1). Our architecture is capable of converting a variety of textual features (e.g. characters, phonemes, stresses) into a variety of vocoder parameters, e.g. mel-band spectrograms, linear-scale log magnitude spectrograms, fundamental frequency, spectral envelope, and aperiodicity parameters. These vocoder parameters can be used as inputs for audio waveform synthesis models.

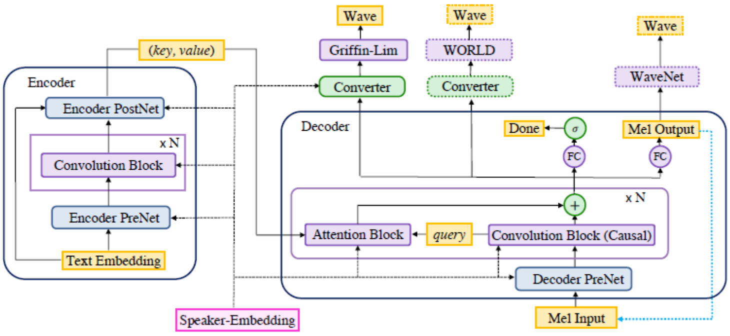


Figure 1: Deep Voice 3 uses residual convolutional layers to encode text into per-timestep *key* and *value* vectors for an attention-based decoder. The decoder uses these to predict the mel-scale log magnitude spectrograms that correspond to the output audio. (Light blue dotted arrows depict the autoregressive process during inference.) The hidden states of the decoder are then fed to a converter network to predict the vocoder parameters for waveform synthesis. See Appendix A for more details.

The Deep Voice 3 architecture consists of three components:

Encoder: A fully-convolutional encoder, which converts textual features to an internal learned representation.

Decoder: A fully-convolutional **causal decoder**, which decodes the learned representation with a multi-hop convolutional attention mechanism into a low-dimensional audio representation (mel-scale spectrograms) in an autoregressive manner.

* 去掉了RNN，加上了causal CNN.

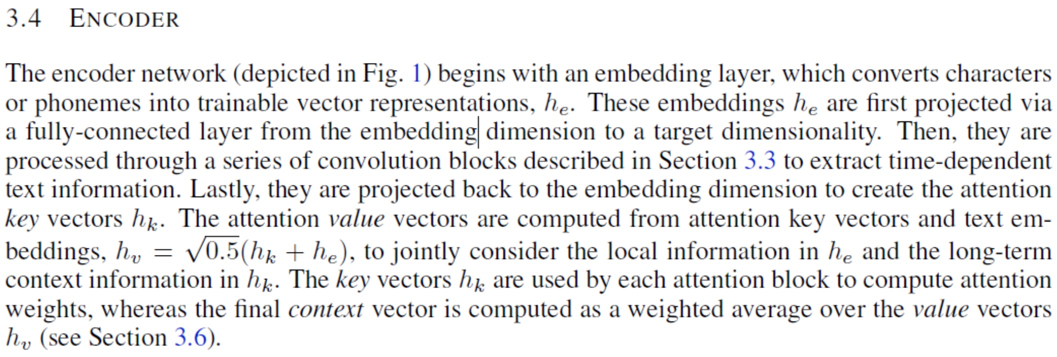
Converter: A fully-convolutional post-processing network, which predicts final vocoder parameters (depending on the vocoder choice) from the decoder hidden states. Unlike the decoder, the converter is non-causal and can thus depend on future context information.

Convolution Network及其变种（反卷积、扩展卷积、因果卷积、图卷积）此博客在介绍这四种卷积时介绍的很清楚。

关于Encoder PreNet网络：

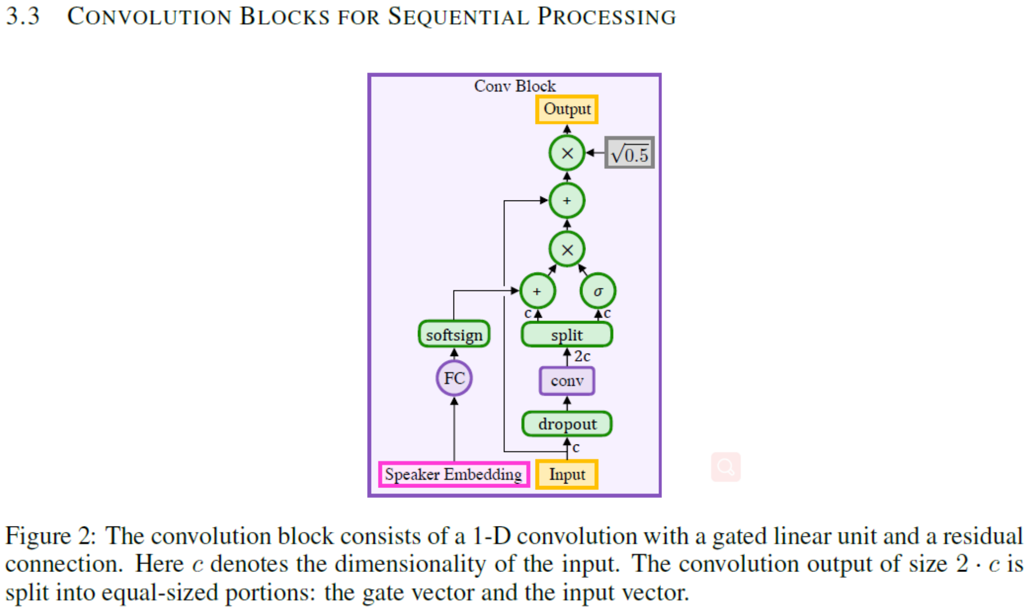
Our model can directly convert characters (including punctuation and spacing) to acoustic features, and hence learns an implicit grapheme-to-phoneme model. This implicit conversion is difficult to correct when the model makes mistakes. Thus, in addition to character models, we also train phoneme-only models and mixed character-and-phoneme models by allowing phoneme input option explicitly. These models are identical to character-only models, except that the input layer of the

encoder sometimes receives phoneme and phoneme stress embeddings instead of character
embeddings.



3.4 Encoder

The encoder network (depicted in Fig. 1) begins with an embedding layer, which converts characters or phonemes into trainable vector representations, h_e. These embeddings h_e are first projected via a fully-connected layer from the embedding dimension to a target dimensionality. Then, they are processed through a series of convolution blocks described in Section 3.3 to extract time-dependent text information. Lastly, they are projected back to the embedding dimension to create the attention *key* vectors h_k. The attention *value* vectors are computed from attention key vectors and text embeddings, $h_v = \sqrt{0.5}(h_k + h_e)$, to jointly consider the local information in h_e and the long-term context information in h_k. The *key* vectors h_k are used by each attention block to compute attention weights, whereas the final *context* vector is computed as a weighted average over the *value* vectors h_v (see Section 3.6).

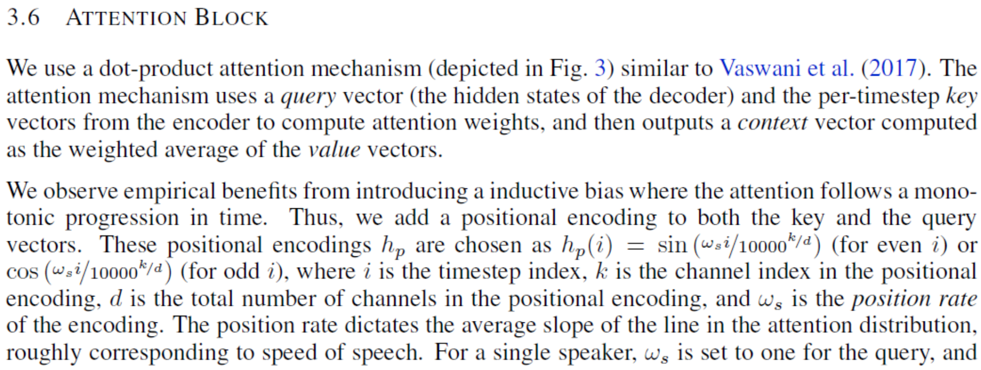


3.3 Convolution Blocks for Sequential Processing

Figure 2: The convolution block consists of a 1-D convolution with a gated linear unit and a residual connection. Here c denotes the dimensionality of the input. The convolution output of size $2 \cdot c$ is split into equal-sized portions: the gate vector and the input vector.

关于Decoder：采用因果卷积进行mel-frame2audio-frame的转换，然后送进Converter进行频谱-wav的转换。

Attention Block：



3.6 Attention Block

We use a dot-product attention mechanism (depicted in Fig. 3) similar to Vaswani et al. (2017). The attention mechanism uses a *query* vector (the hidden states of the decoder) and the per-timestep *key* vectors from the encoder to compute attention weights, and then outputs a *context* vector computed as the weighted average of the *value* vectors.

We observe empirical benefits from introducing a inductive bias where the attention follows a monotonic progression in time. Thus, we add a positional encoding to both the key and the query vectors. These positional encodings h_p are chosen as $h_p(i) = \sin\left(\omega_s i / 10000^{k/d}\right)$ (for even i) or $\cos\left(\omega_s i / 10000^{k/d}\right)$ (for odd i), where i is the timestep index, k is the channel index in the positional encoding, d is the total number of channels in the positional encoding, and ω_s is the *position rate* of the encoding. The position rate dictates the average slope of the line in the attention distribution, roughly corresponding to speed of speech. For a single speaker, ω_s is set to one for the query, and

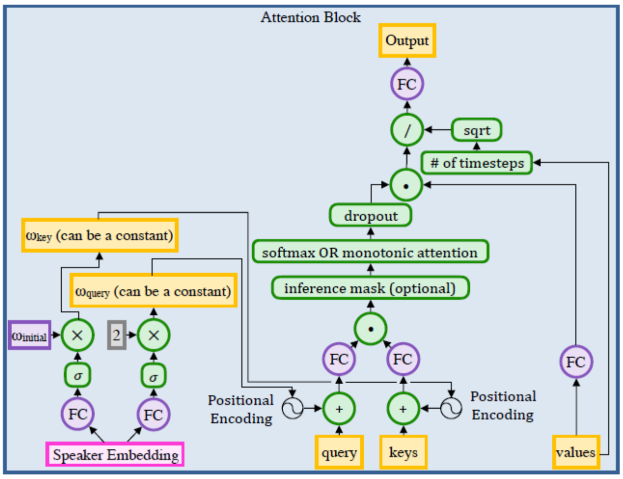
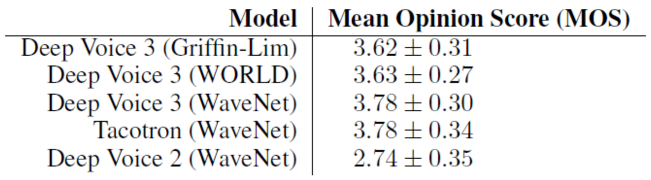


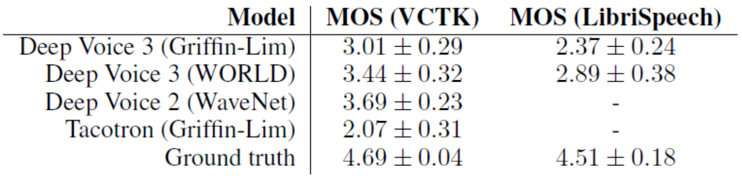
Figure 3: Positional encodings are added to both keys and query vectors, with rates of ω_{key} and ω_{query} respectively. Forced monotonocity can be applied at inference by adding a mask of large negative values to the logits. One of two possible attention schemes is used: softmax or monotonic attention from Raffel et al. (2017). During training, attention weights are dropped out.

Result:



Model	Mean Opinion Score (MOS)
Deep Voice 3 (Griffin-Lim)	3.62 ± 0.31
Deep Voice 3 (WORLD)	3.63 ± 0.27
Deep Voice 3 (WaveNet)	3.78 ± 0.30
Tacotron (WaveNet)	3.78 ± 0.34
Deep Voice 2 (WaveNet)	2.74 ± 0.35

Table 2: Mean Opinion Score (MOS) ratings with 95% confidence intervals using different waveform synthesis methods. We use the crowdMOS toolkit (Ribeiro et al., 2011); batches of samples from these models were presented to raters on Mechanical Turk. Since batches contained samples from all models, the experiment naturally induces a comparison between the models.



Model	MOS (VCTK)	MOS (LibriSpeech)
Deep Voice 3 (Griffin-Lim)	3.01 ± 0.29	2.37 ± 0.24
Deep Voice 3 (WORLD)	3.44 ± 0.32	2.89 ± 0.38
Deep Voice 2 (WaveNet)	3.69 ± 0.23	-
Tacotron (Griffin-Lim)	2.07 ± 0.31	-
Ground truth	4.69 ± 0.04	4.51 ± 0.18

Table 3: MOS ratings with 95% confidence intervals for audio clips from neural TTS systems on multi-speaker datasets. We also use crowdMOS toolkit; batches of samples including ground truth were presented to human raters. Multi-speaker Tacotron implementation and hyperparameters are based on Arık et al. (2017), which is a proof-of-concept implementation. Deep Voice 2 and Tacotron systems were not trained for the LibriSpeech dataset due to prohibitively long time required to optimize hyperparameters.

3. Tacotron 2：NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

作者：Jonathan Shen1, Ruoming Pang1, Ron J. Weiss1 （Google）

时间：2018.02.16

分类：end-to-end

demo: https://google.github.io/tacotron/publications/tacotron2

paper: https://arxiv.org/pdf/1712.05884.pdf

code: https://github.com/Rayhane-mamah/Tacotron-2 【tensorflow实现】

code2:
https://github.com/NVIDIA/tacotron2/tree/fc0cf6a89a47166350b65daa1beaa06979e4cddf
【NVIDIA pytorch实现】

特点：模型采用Seq2Seq结构，把字符特征映射到梅尔频谱特征，然后再通过wavnet进行波形的合成。其指标mean opinion score(MOS)为4.53，而专业的录音设备为4.58。论文还进一步分析了每一个结构对结果的影响，只考虑mel feature，而不考虑linguistic，duration，f0等作为condition送进wavnet，有助于模型的简化。

Abstract:

This paper describes Tacotron 2, a neural network architecture for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from those spectrograms. Our model achieves a mean opinion score (MOS) of 4:53 comparable to a MOS of 4:58 for professionally recorded speech. To validate our design choices, we present ablation studies of key components of our system and evaluate the impact of using mel spectrograms as the conditioning input to WaveNet instead of linguistic, duration, and F0 features. We further show that using this compact acoustic intermediate representation allows for a significant reduction in the size of the WaveNet architecture.

Contributions:

In this paper, we describe a unified, entirely neural approach to speech synthesis that combines the best of the previous approaches: a sequence-to-sequence Tacotron-style model [12] that generates mel spectrograms, followed by a modified WaveNet vocoder [10, 15]. Trained directly on normalized character sequences and corresponding speech waveforms, our model learns to synthesize natural sounding speech that is difficult to distinguish from real human speech.

Deep Voice 3 [11] describes a similar approach. However, unlike our system, its naturalness has not been shown to rival that of human speech. Char2Wav [16] describes yet another similar approach to end-to-end TTS using a neural vocoder. However, they use different intermediate representations (traditional vocoder features) and their model architecture differs significantly.

Model:

1. Spectrogram Prediction Network

The network is composed of an encoder and a decoder with attention. The encoder converts a character sequence into a hidden feature representation which the decoder consumes to predict a spectrogram. Input characters are represented using a learned 512-dimensional character embedding, which are passed through a stack of 3 convolutional layers each containing 512 filters with shape 5 *1, i.e., where each filters spans 5 characters, followed by batch normalization [18] and ReLU activations. As in Tacotron, these convolutional layers model longer-term context (e.g., N-grams) in the input character sequence. The output of the final convolutional layer is passed into a single bi-directional [19] LSTM [20] layer containing 512 units (256 in each direction) to generate the encoded features.

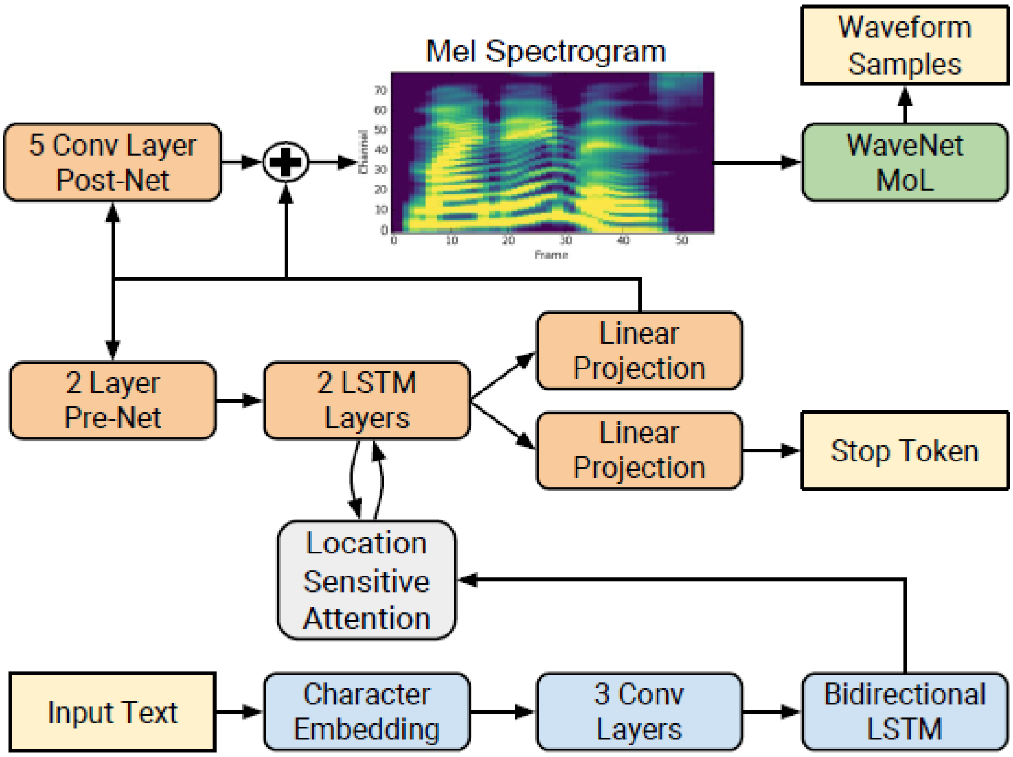


Fig. 1. Block diagram of the Tacotron 2 system architecture.

The encoder output is consumed by an attention network which summarizes the full encoded sequence as a fixed-length context vector for each decoder output step. We use the location-sensitive attention from [21], which extends the additive attention mechanism [22] to use cumulative attention weights from previous decoder time steps as an additional feature. This encourages the model to move forward consistently through the input, mitigating potential failure modes where some subsequences are repeated or ignored by the decoder. Attention probabilities are computed after projecting inputs and location features to 128-dimensional hidden representations. Location features are computed using 32 1-D convolution filters of length 31. [21 Attention-Based Models for Speech Recognition https://arxiv.org/pdf/1506.07503.pdf]

The decoder is an autoregressive recurrent neural network which predicts a mel spectrogram from the encoded input sequence one frame at a time. The prediction from the previous time step is first passed through a small pre-net containing 2 fully connected layers of 256 hidden ReLU units. We found that the pre-net acting as an information bottleneck was essential for learning attention. The prenet output and attention context vector are concatenated and passed through a stack of 2 uni-directional LSTM layers with 1024 units. The concatenation of the LSTM output and the attention context vector is projected through a linear transform to predict the target spectrogram frame. Finally, the predicted mel spectrogram is passed through a 5-layer convolutional post-net which predicts a residual to add to the prediction to improve the overall reconstruction. Each post-net layer is comprised of 512 filters with shape 5 1 with batch normalization, followed by tanh activations on all but the final layer.

 * decoder这一部分2LSTM Layers+Linear Projection+2Layer Pre-Net + Attention 构成一个自回归网络，一般的decoder网络，Liner Projection的输出+Attention的hidden state作为下一个时刻的输入，但是作者加入了2Layer Pre-Net，发现对Attention效果更好【我们可以借鉴】同时，Linear Projection 会送进一个5层的Conv Layer Post-Net网络，同时再末尾相加作为一个残差结构，生成每一帧的Mel频谱信息。 需要去了解一下Location Sensitive Attention机制。

LOSS函数采用的是MSE，但是这种自回归结构会导致Mel的每一帧都依赖于前一帧，从而导致没法并行生成，因此就会导

In parallel to spectrogram frame prediction, the concatenation of decoder LSTM output and the attention context is projected down to a scalar and passed through a sigmoid activation to predict the probability that the output sequence has completed. This “stop token” prediction is used during inference to allow the model to dynamically determine when to terminate generation instead of always generating for a fixed duration. Specifically, generation completes at the first frame for which this probability exceeds a threshold of 0.5. 【* 这里只是使用attention的context信息和lstm的输出信息结合，经过一个sigmoid函数输出一个是否结束Stop Token的概率】

2. WaveNet Vocoder

3. Evaluation

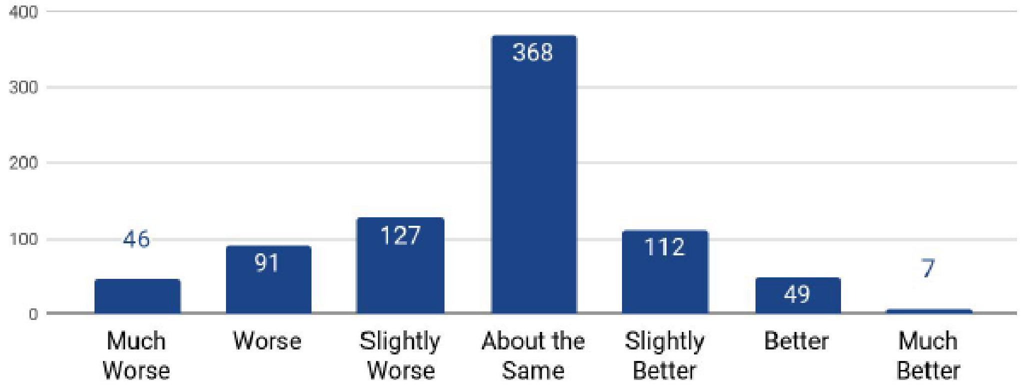
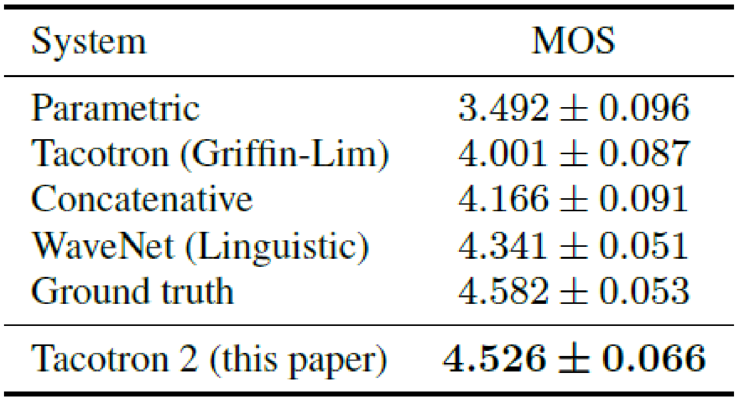


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.



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.

4. Close to Human Quality TTS with Transformer

作者：Naihan Li[*1,4], Shujie Liu[2], Yanqing Liu[3]（MSRA STCA）

时间：2018.11.13

分类：end-to-end

demo: https://neuraltts.github.io/transformertts/

paper: https://arxiv.org/pdf/1809.08895.pdf

code: 暂无

特点：1. 采用multi-head attention mechanism取代了Tacotron2的attention机制和RNN结构，使得训练效率提升，并且可以有效解决长范围的依赖问题。2. 该论文网络结构效率是Tacotron2的4.25倍，MOS比Tacotron2大0.048。

Abstract:

Although end-to-end neural text-to-speech (TTS) methods (such as Tacotron2) are proposed and achieve state-of-theart performance, they still suffer from two problems: 1) low efficiency during training and inference; 2) hard to model long dependency using current recurrent neural networks (RNNs). Inspired by the success of Transformer network in neural machine translation (NMT), in this paper, we introduce and adapt the multi-head attention mechanism to replace the RNN structures and also the original attention mechanism in Tacotron2. With the help of multi-head self-attention,

the hidden states in the encoder and decoder are constructed in parallel, which improves training efficiency. Meanwhile, any two inputs at different times are connected directly by a self-attention mechanism, which solves the long range dependency problem effectively. Using phoneme sequences as input, our Transformer TTS network generates mel spectrograms, followed by a WaveNet vocoder to output the final audio results. Experiments are conducted to test the efficiency and performance of our new network. For the efficiency, our Transformer TTS network can speed up the training about 4.25 times faster compared with Tacotron2. For the performance, rigorous human tests show that our proposed model achieves state-of-the-art performance (outperforms Tacotron2 with a gap of 0.048) and is very close to human quality (4.39 vs 4.44 in MOS).

Contributions:

Inspired by this idea, in this paper, we combine the advantages of Tacotron2 and Transformer to propose a novel end-to-end TTS model, in which the multi-head attention mechanism is introduced to replace the RNN structures in the encoder and decoder, as well as the vanilla attention network. The self-attention mechanism unties the sequential dependency on the last previous hidden state to improve the parallelization capability and relieve the long distance dependency problem. Compared with the vanilla attention between the encoder and decoder, the multi-head attention can build the context vector from different aspects using different attention heads. With the phoneme sequences as input, our novel Transformer TTS network generates mel spectrograms, and employs WaveNet as vocoder to synthesize audios. We conduct experiments with 25-hour professional speech dataset, and the audio quality is evaluated by human testers. Evaluation results show that our proposed model outperforms the original Tacotron2 with a gap of 0.048 in CMOS, and achieves a similar performance (4.39 in MOS) with human recording (4.44 in MOS). Besides, our Transformer TTS model can speed up the training process about 4.25 times compared with Tacotron2.

Model:

3.2 Scaled Positional Encoding

Transformer contains no recurrence and no convolution so that if we shuffle the input sequence of encoder or decoder, we will get the same output. To take the order of the sequence into consideration, information about the relative or absolute position of frames is injected by triangle positional embeddings, shown in Eq. 7:

$$PE(pos, 2i) = \sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}}) \tag{6}$$

$$PE(pos, 2i+1) = \cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}}) \tag{7}$$

where pos is the time step index, $2i$ and $2i+1$ is the channel index and d_{model} is the vector dimension of each frame. In NMT, the embeddings for both source and target language are from language spaces, so the scales of these embeddings are similar. This condition doesn't hold in the TTS scenarioe, since the source domain is of texts while the target domain is of mel spectrograms, hence using fixed positional embeddings may impose heavy constraints on both the encoder and decoder pre-nets (which will be described in Sec. 3.3 and 3.4). We employ these triangle positional embeddings with a trainable weight, so that these embedding can adaptively fit the scales of both encoder and decoder pre-nets' output, as shown in Eq. 8:

$$x_i = prenet(phoneme_i) + \alpha PE(i) \tag{8}$$

where α is the trainable weight.

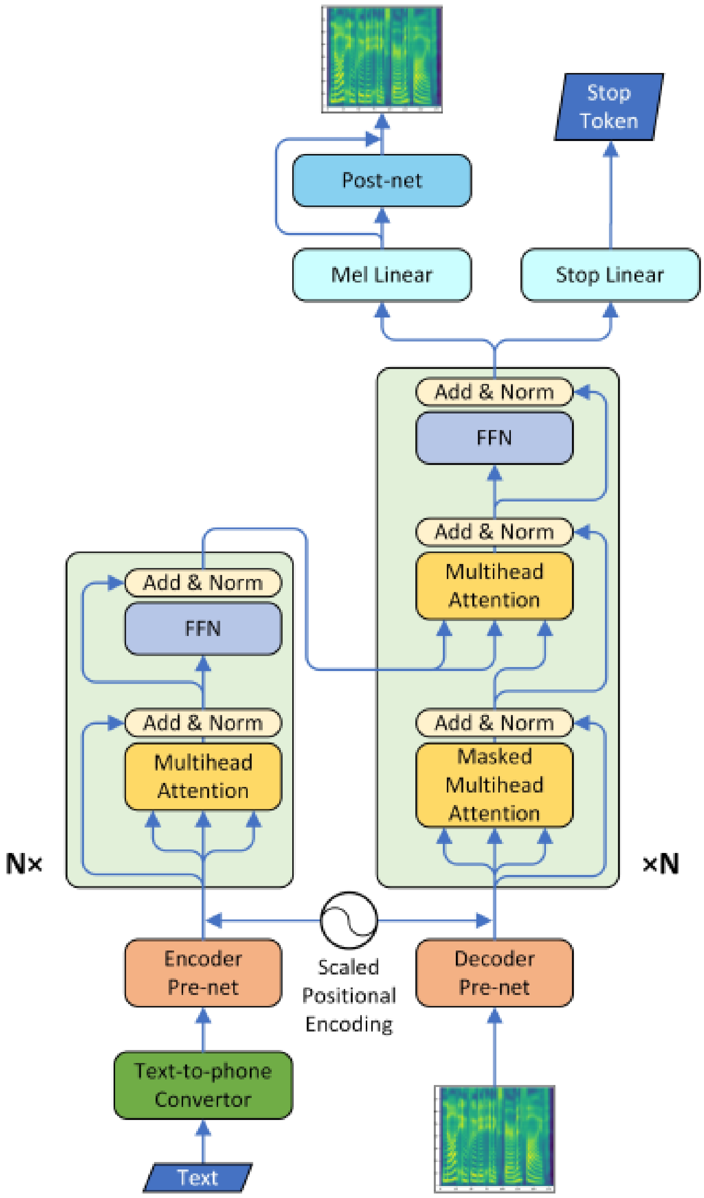


Figure 3: System architecture of our model.

Encodere Pre-net采用和Tacotron2一样的结构，3层CNN+BN+RELU。

Decoder Pre-net采用2层FC+RELU。

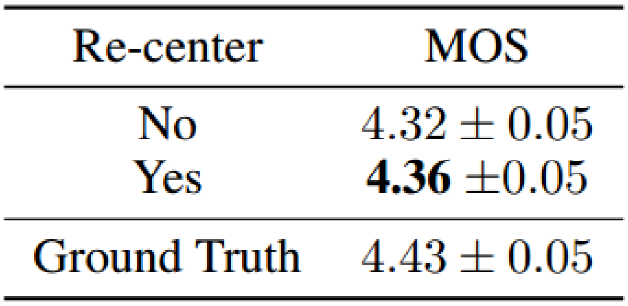
Encoder和Decoder就是采用NMT的结构。

both are built by stacks of several identity blocks. Each encoder block contains two subnetworks: a multi-head attention and a feed forward network, while each decoder block contains an extra masked multi-head attention comparing to the encoder block. Both encoder and decoder blocks have residual connections and layer normalizations. 不包含RNN结构在里面。

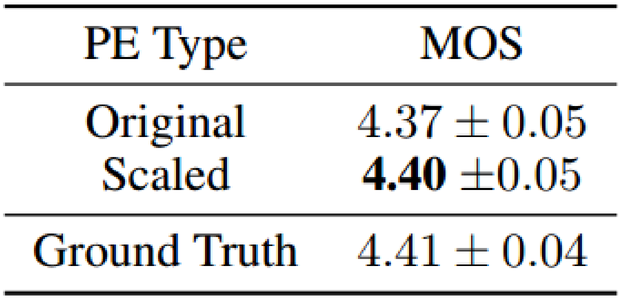
MEL Linear，Stop Linear采用的是Tacotron2的线性投影。Post Net采用的Tacotron2的5层CNN

Evaluation:

Figure 5: PE scale of encoder and decoder.

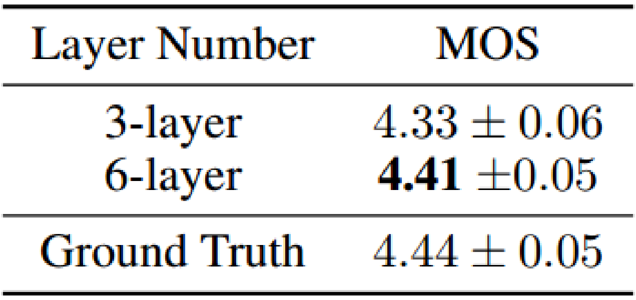


Re-center	MOS
No	4.32 ± 0.05
Yes	**4.36** ±0.05
Ground Truth	4.43 ± 0.05



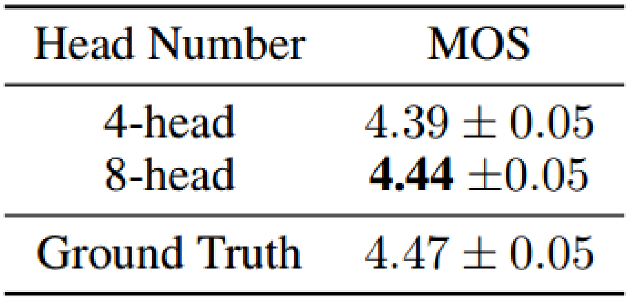
PE Type	MOS
Original	4.37 ± 0.05
Scaled	**4.40** ±0.05
Ground Truth	4.41 ± 0.04

Table 3: MOS comparison of scaled and original PE.



Layer Number	MOS
3-layer	4.33 ± 0.06
6-layer	**4.41** ±0.05
Ground Truth	4.44 ± 0.05

Table 4: Ablation studies in different layer numbers.



Head Number	MOS
4-head	4.39 ± 0.05
8-head	**4.44** ±0.05
Ground Truth	4.47 ± 0.05

Table 5: Ablation studies in different head numbers.

5. WAVEFORM GENERATION FOR TEXT-TO-SPEECH SYNTHESIS USING PITCH-SYNCHRONOUS MULTI-SCALE GENERATIVE ADVERSARIAL NETWORKS

（pitch-synchronous基音同步）

作者：Lauri Juvela1, Bajibabu Bollepalli1, Junichi Yamagishi2;3（Aalto University, Finland）

时间：2018.10.30

分类：Vocoder

demo：https://users.aalto.fi/~ljuvela/multiscale-gan/

paper：https://arxiv.org/pdf/1810.12598.pdf

code：https://github.com/ljuvela/multiscale-GAN [暂未公布]

特点：采用GAN网络来作为Vocoder进行语音合成，可以并行化训练，但是结果比Wavenet相差比较远，而glottal excitation（声门激励）模型可以实现和WaveNet相近的结果。

Abstract:

The state-of-the-art in text-to-speech synthesis has recently improved considerably due to novel neural waveform generation methods, such as WaveNet. However, these methods suffer from their slow sequential inference process, while their parallel versions are difficult to train and even more expensive computationally. Meanwhile, generative adversarial networks (GANs) have achieved impressive results in image generation and are making their way into audio applications; parallel inference is among their lucrative properties. By adopting recent advances in GAN training techniques, this investigation studies waveform generation for TTS in two domains (speech signal and glottal excitation). Listening test results show that while direct waveform generation with GAN is still far behind WaveNet, a GAN-based glottal excitation model can achieve quality and voice similarity on par with a WaveNet vocoder.

6. WAVECYCLEGAN: SYNTHETIC-TO-NATURAL SPEECH WAVEFORM CONVERSION USING CYCLE-CONSISTENT ADVERSARIAL NETWORKS

作者：Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo（NTT Corporation, Japan）

时间：2018.09.28

分类：post-processing

demo：无

paper：https://arxiv.org/pdf/1809.10288.pdf

code：无

特点：设计了一种cycle-consistent adversarial networks来解决合成音频中出现的过度平滑的问题。该方法直接对waveform进行处理，是的生成的音频听起来更加的natural。

Abstract:

We propose a learning-based filter that allows us to directly modify a synthetic speech waveform into a natural speech waveform. Speech-processing systems using a vocoder framework such as statistical parametric speech synthesis and voice conversion are convenient especially for a limited number of data because it is possible to represent and process interpretable acoustic features over a compact space, such as the fundamental frequency (F0) and mel-cepstrum. However, a well-known problem that leads to the quality degradation of generated speech is an over-smoothing effect that eliminates some detailed structure of generated/converted acoustic features. To address this issue, we propose a syntheticto-natural speech waveform conversion technique that uses cycle-consistent adversarial networks and which does not require any explicit assumption about speech waveform in adversarial learning. In contrast to current techniques, since our modification is performed at the waveform level, we expect that the proposed method will also make it possible to generate "vocoder-less" sounding speech even if the input speech is synthesized using a vocoder framework. The experimental results demonstrate that our proposed method can 1) alleviate the over-smoothing effect of the acoustic features despite the direct modification method used for the waveform and 2) greatly improve the naturalness of the generated speech sounds

model:

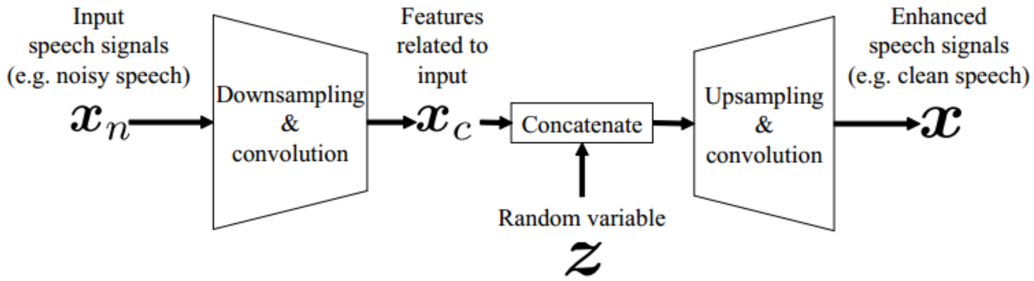
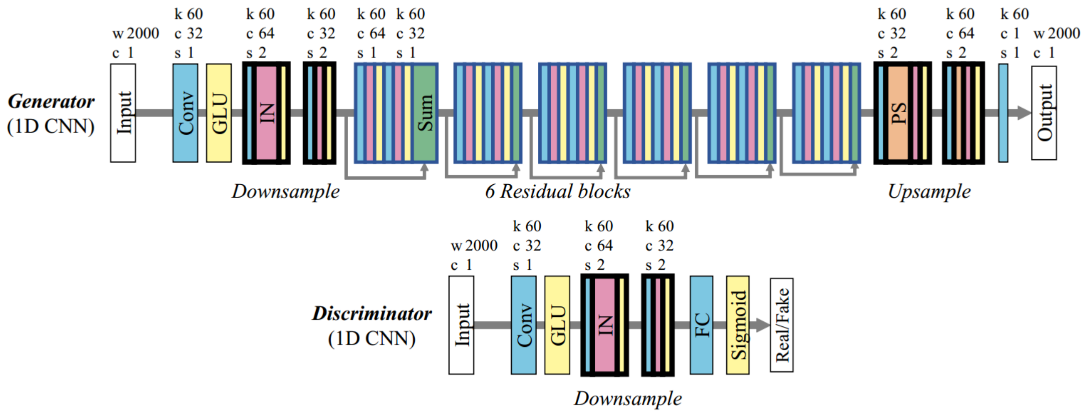


Fig. 2. Generator network for speech enhancement reported in [6]. Structure is similar to an auto-encoder.



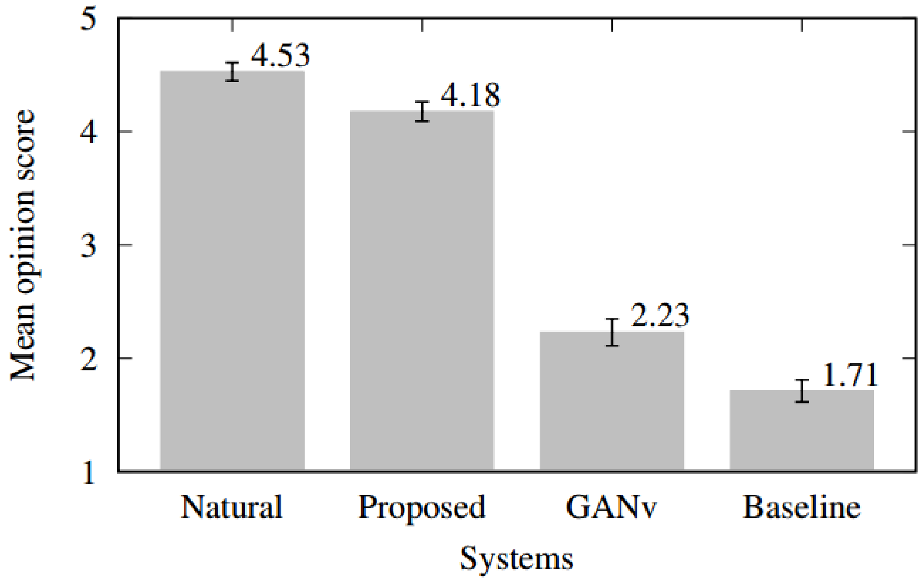


Fig. 6. Subjective 5-scale mean opinion score regarding naturalness, with 95% confidence intervals.

7. Wasserstein GAN and Waveform Loss-based Acoustic Model Training for Multi-speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder

作者：YI ZHAO[1]，SHINJI TAKAKI[2], HIEU-THI LUONG[2] (The University of Tokyo)

时间：2018.07.31

分类：acoustic model

demo: https://nii-yamagishilab.github.io/TTS-GAN-WN-MultiSpeaker/index.html

paper: https://arxiv.org/pdf/1807.11679.pdf

code: 无

特点：采用Wasserstein GAN with gradient penalty(WGAN-GP)模型结构来训练一个Multi-speaker声学模型，同时通过已经训练好的Vocoder，得到一个DML loss来优化声学模型。

Abstract:

Recent neural networks such as WaveNet and sampleRNN that learn directly from speech waveform samples have achieved very high-quality synthetic speech in terms of both naturalness and speaker similarity even in multi-speaker text-to-speech synthesis systems. Such neural networks are being used as an alternative to vocoders and hence they are often called neural vocoders. The neural vocoder uses acoustic features as local condition parameters, and these parameters need to be accurately predicted by another acoustic model. However, it is not yet clear how to train this acoustic model, which is problematic because the final quality of synthetic speech is significantly affected by the performance of the acoustic model. Significant degradation happens, especially when predicted acoustic features have mismatched characteristics compared to natural ones. In order to reduce the mismatched characteristics between natural and generated acoustic features, we propose frameworks that incorporate either a conditional generative adversarial network (GAN) or its variant, Wasserstein GAN with gradient penalty (WGAN-GP), into multispeaker speech synthesis that uses the WaveNet vocoder. We also extend the GAN frameworks and use the discretized mixture logistic loss of a well-trained WaveNet in addition to mean squared error and adversarial losses as parts of objective functions. Experimental results show that acoustic models trained using the WGAN-GP framework using back-propagated discretized-mixture-of-logistics (DML) loss achieves the highest subjective evaluation scores in terms of both quality and speaker similarity.

论文在接Introduction中提到：

Previously investigated acoustic models include the hidden Markov model (HMM) [2], the deep neural network (DNN) [3], and the recurrent neural network (RNN) [4] [5]. These are normally trained with the minimum mean squared error (MSE) criterion, and hence, the generated acoustic parameters tend to be over-smoothed regardless of the architectures. Finally, speech waveforms have been reconstructed using a deterministic vocoder based on the acoustic parameters [6] [7] [8]. However, the generated signals have artifacts and typically sound buzzy.

论文采用GAN来消除over-smoothed，通过采用Vocoder输出的模型来形成discretized mixture logistic loss，减小buzzy的声音。

Contributions:

1. The generator of GAN is conditioned on both linguistic features and speaker code, and the discriminator aiming at distinguishing the real and predicted melspectrograms is also conditioned on speaker information. The WaveNet vocoder is conditioned on both mel-spectrogram and speaker codes, as well.

2. Define a new objective function using the weighted sum of three kinds of losses: conventional MSE loss, adversarial loss（ADV）, and discretized mixture logistic loss [20]（DML） obtained through the well-trained WaveNet vocoder.[第三个loss需要考虑Vocoder后合成的波形]。

3. Simple recurrent units (SRUs) [21] are utilized as basic components since they can be trained faster than the LSTM-based RNN architecture while maintaining a performance as good as or even better than LSTM-RNN.

Model:

1、整体的模型

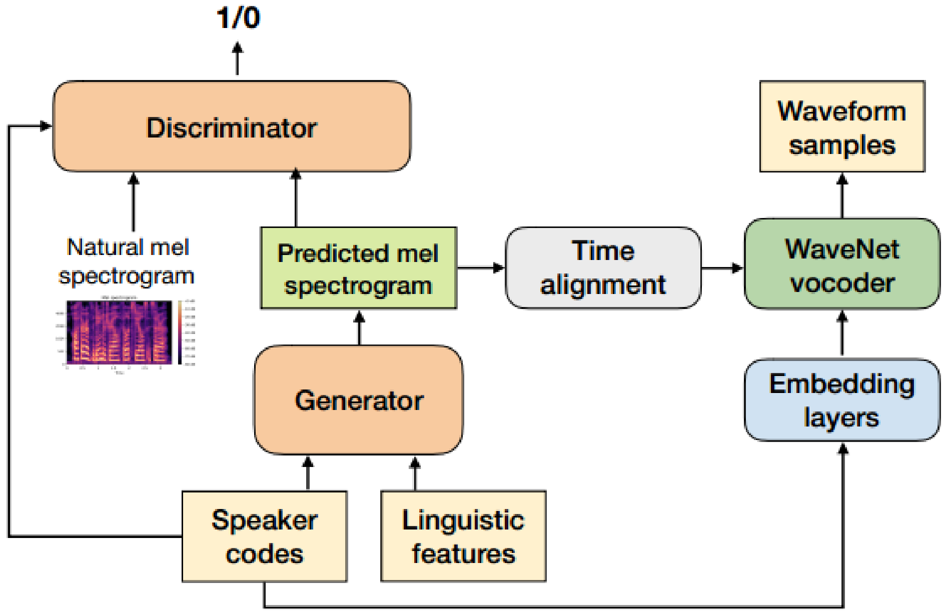
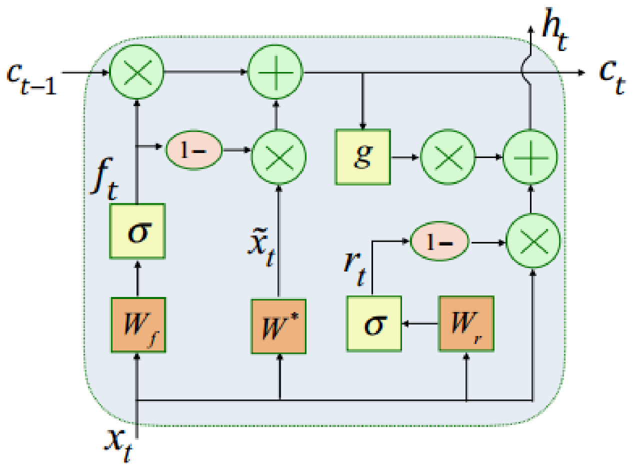


FIGURE 1: Proposed GAN-trained multi-speaker speech synthesis framework using a WaveNet vocoder.

2、SRU

SRU这里的遗忘门+加上第二列，确实可以减轻梯度消失和爆照的问题。



$$\tilde{x}_t = W^* x_t$$

$$f_t = \sigma(W_f x_t + b_f)$$

$$r_t = \sigma(W_r x_t + b_r)$$

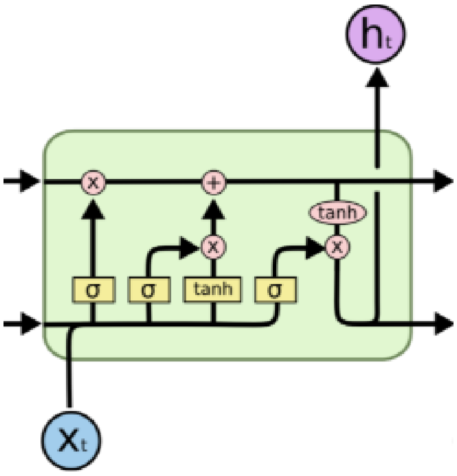
$$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot \tilde{x}_t$$

$$h_t = r_t \odot g(c_t) + (1 - r_t) \odot x_t$$

FIGURE 2: Details of the SRU cell. $\sigma(\cdot)$ and $g(\cdot)$ represent sigmoid and ReLU activation functions, respectively.

The basic form of SRU includes only a single forget gate ft to alleviate vanishing and exploding gradient problems instead of using many different gates to control the information flow. In SRU, the forget gate is used to modulate the internal state ct, which is then used to compute the output state ht. Unlike existing RNN architectures that use the previous output state in the recurrence computation, SRU completely drops the connection between the gating computations and the previous states, and this makes SRU computationally efficient and allows us to use parallelization. The complete architecture of SRU is shown in Fig. 2. The reset gate rt is computed similar to the forget gate ft and is used to compute the output state ht, which performs as a combination of the internal state g(ct) and the input xt. g(·) represents a ReLU activation function and σ(·) is a sigmoid function.

SRU相对于传统的LSTM或者RNN来说，只有两个门：遗忘门和重置门，并通过1-这个结构提高存储效率。（1-这个结构，可以使得原本等于0的节点存储新的信息）



3、GAN

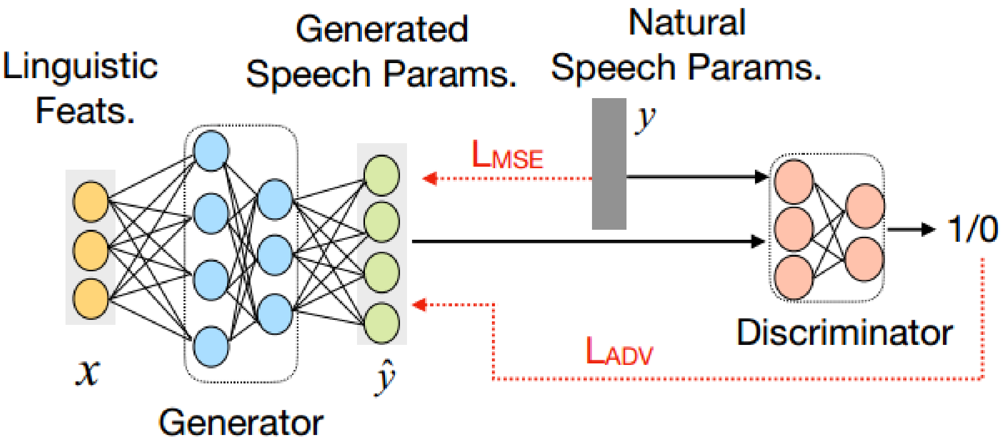


FIGURE 3: GAN-based training of TTS acoustic model. L_{ADV} indicates adversarial loss and L_{MSE} indicates L2 loss.

4、Three Loss

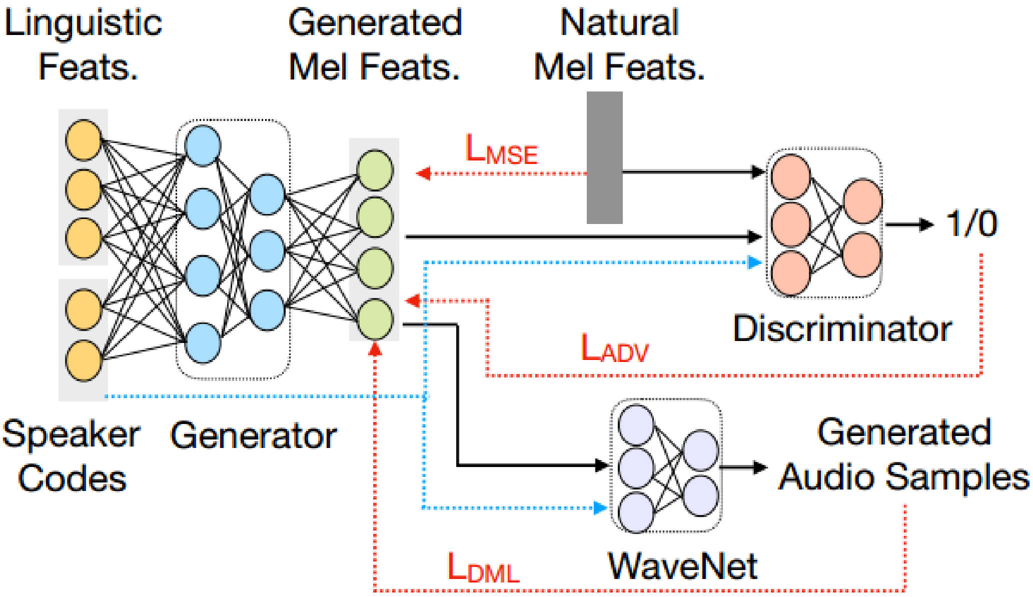


FIGURE 5: Loss functions and gradients for updating acoustic models in the proposed method. Note that neither the model parameters of WaveNet nor the discriminator are updated in this step.

8. **WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS**

作者：Ryan Prenger, Rafael Valle, Bryan Catanzaro (NVIDIA Corporation)

时间：2018.10.31

分类：Vocoder

demo: https://nv-adlr.github.io/WaveGlow

paper: https://arxiv.org/pdf/1807.11679.pdf code: https://github.com/NVIDIA/waveglow/tree/4b1001fa3336a1184b8293745bb89b177457f09b 【Pytorch实现】

特点：结合Glow+WaveNet优点，设计了WaveGlow模型，放弃了自回归结构，而是采用单个网络以及单个损失函数，使得训练更快更稳定。

Abstract:

In this paper we propose WaveGlow: a flow-based network capable of generating high quality speech from melspectrograms. WaveGlow combines insights from Glow [1] and WaveNet [2] in order to provide fast, efficient and highquality audio synthesis, without the need for auto-regression.

WaveGlow is implemented using only a single network, trained using only a single cost function: maximizing the likelihood of the training data, which makes the training procedure simple and stable. Our PyTorch implementation produces audio samples at a rate of more than 500 kHz on an NVIDIA V100 GPU. Mean Opinion Scores show that it delivers audio quality as good as the best publicly available WaveNet implementation. All code will be made publicly available online [3].

9. NEURAL SOURCE-FILTER-BASED WAVEFORM MODEL FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

作者：Xin Wang1, Shinji Takaki1, Junichi Yamagishi1∗ (National Institute of Informatics, Japan)

时间：2018.11.26

分类：Vocoder

demo: https://nii-yamagishilab.github.io/samples-nsf/index.html

paper: https://arxiv.org/pdf/1810.11946.pdf

code: https://github.com/nii-yamagishilab/project-CURRENNT-public【核心部分由C++写的】

特点：该论文提出了一个基于source-filter的模型（non-AR），该模型首先通过一个source module把声学特征转换为基于sine的激励信号，然后再通过一个filter module把此激励信号转换为声音波形。该模型速度是Wavenet的100倍，且合成语音的质量接近Wavenet。

Neural waveform models such as the WaveNet are used in many recent text-to-speech systems, but the original WaveNet is quite slow in waveform generation because of its autoregressive (AR) structure. Although faster non-AR models were recently reported, they may be prohibitively complicated due to the use of a distilling training method and the blend of other disparate training criteria. This study proposes a non-AR neural source-filter waveform model

that can be directly trained using spectrum-based training criteria and the stochastic gradient descent method. Given the input acoustic features, the proposed model first uses a source module to generate a sine-based excitation signal and then uses a filter module to transform the excitation signal into the output speech waveform. Our experiments demonstrated that the proposed model generated waveforms at least 100 times faster than the AR WaveNet and the quality of its synthetic speech is close to that of speech generated by the AR WaveNet. Ablation test results showed that both the sinewave excitation signal and the spectrum-based training criteria were essential to the performance of the proposed model.

10. **UFANS: U-SHAPED FULLY-PARALLEL ACOUSTIC NEURAL STRUCTURE FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS WITH 20X FASTER**

作者：Dabiao Ma, Zhiba Su, Yuhao Lu (Turing Robot co.ltd)

时间：2018.11.28

分类：Acoustic Model

demo:

paper: https://arxiv.org/pdf/1811.12208.pdf

code:

特点：本文提出了一种U型的声学神经架构(UFANS)，采用全卷积代替RNN，使得训练以及推理阶段快了20倍。

Abstract:

Neural networks with Auto-regressive structures, such as Recurrent Neural Networks (RNNs), have become the most appealing structures for acoustic modeling of parametric text to speech synthesis (TTS) in recent studies. Despite the prominent capacity to capture long-term dependency, these models consist of massive sequential computations that cannot be fully parallel. In this paper, we propose a U-shaped Fully-parallel Acoustic Neural Structure (UFANS), which is a deconvolutional alternative of RNNs for Statistical Parametric Speech Synthesis (SPSS). The experiments verify that our proposed model is over 20 times faster than RNN based acoustic model, both training and inference on GPU with comparable speech quality. Furthermore, We also investigate that how long information dependence really matters to synthesized speech quality.

Model:

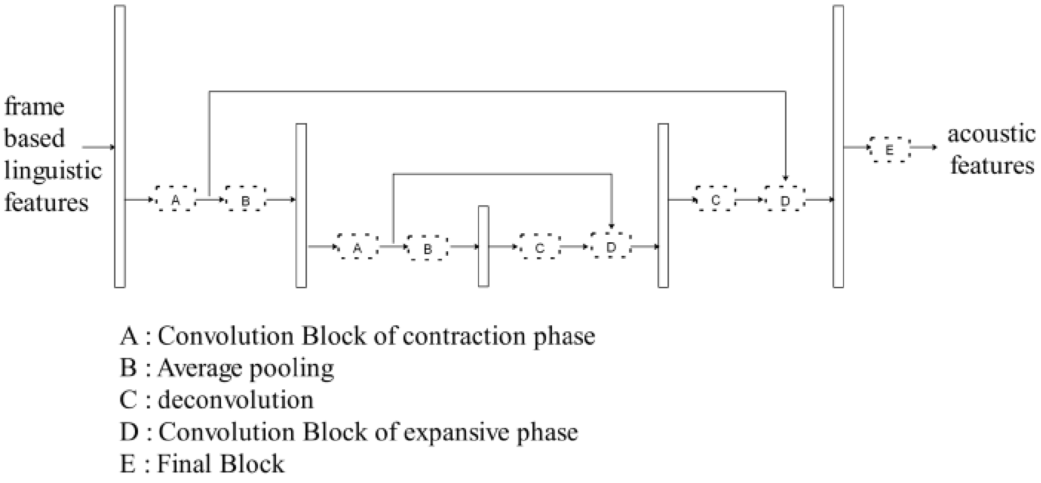


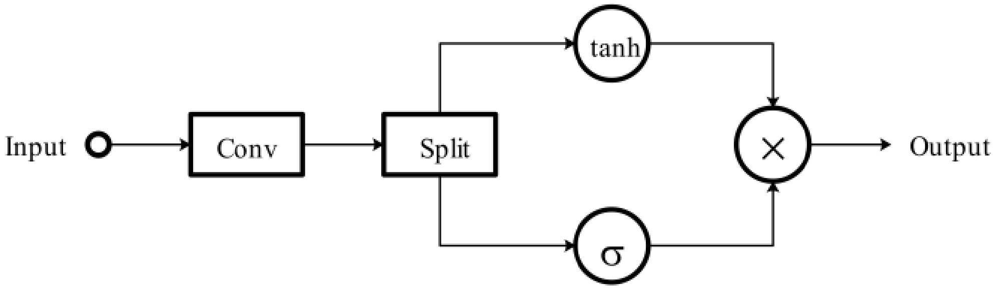
Fig. 1. UFANS, with 2 down-sampling and up-samplings.

其对应模块分别对应如下：

A：Convolution Block：用来做phase contract（压缩），同时通过把卷积网络的输出沿着channel dimension分割成两部分，可以增加网络对输入的一个自适应能力。同时通过平均池化来保存更多的信息。

$$P_1, P_2 = Split(Conv(Input)) \qquad (1)$$

$$Output = tanh(P_1) * sigmoid(P_2) \qquad (2)$$



C：Deconvolution：

D：Convolution Block：

$$P_1, P_2 = Split(Conv(Dropout(Input1 + Input2))) \quad (3)$$

$$Output = tanh(P_1) * sigmoid(P_2) \qquad (4)$$

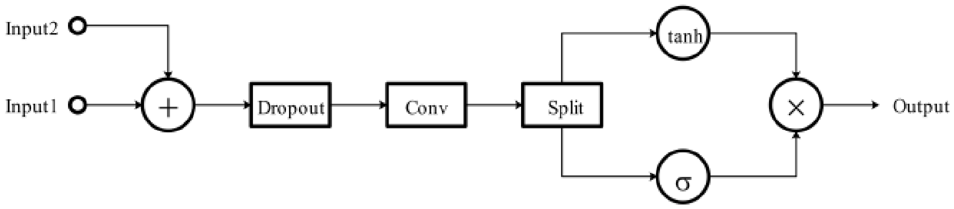
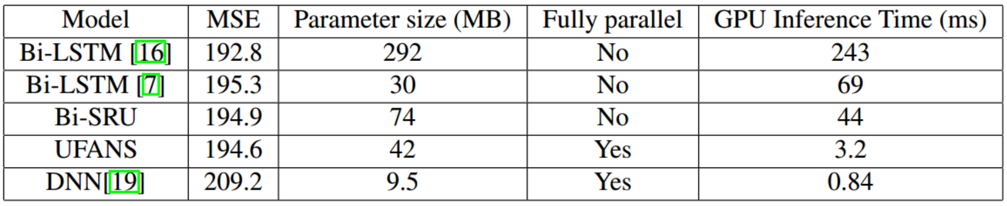


Fig. 3. Convolution block (D in Figure 1) of expansive phase, here the operation ’split’ is performed on the channel size.

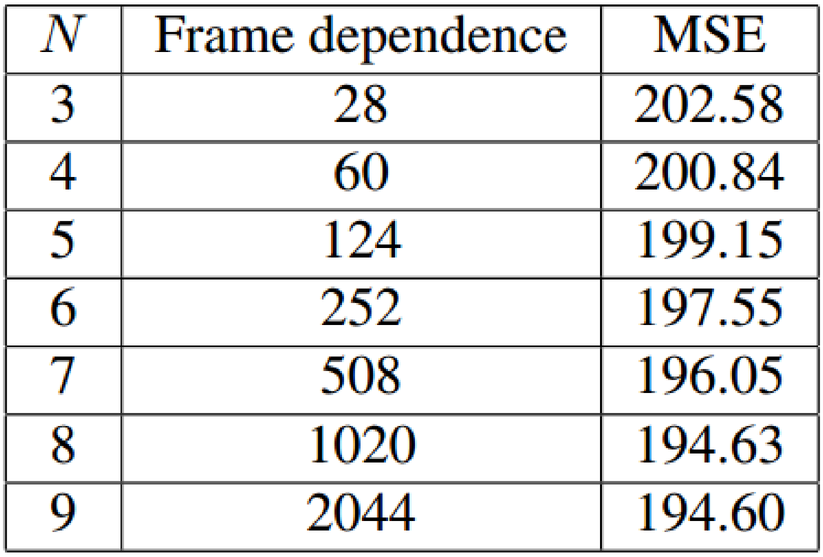
通过pooling来增加感受野，感觉会有信息损失。



Model	MSE	Parameter size (MB)	Fully parallel	GPU Inference Time (ms)
Bi-LSTM [16]	192.8	292	No	243
Bi-LSTM [7]	195.3	30	No	69
Bi-SRU	194.9	74	No	44
UFANS	194.6	42	Yes	3.2
DNN[19]	209.2	9.5	Yes	0.84

Table 2. Comparison of objective results.

还考虑了不同的感受野对MSE的影响：



N	Frame dependence	MSE
3	28	202.58
4	60	200.84
5	124	199.15
6	252	197.55
7	508	196.05
8	1020	194.63
9	2044	194.60

Table 3. MSE when N ranges from 3 to 9.

11. **SPEAKING STYLE ADAPTATION IN TEXT-TO-SPEECH SYNTHESIS USING SEQUENCE-TO-SEQUENCE MODELS WITH ATTENTION （pass)**

作者：Bajibabu Bollepalli， Lauri Juvela， Paavo Alku (Aalto University)

时间：2018.10.29

分类：Acoustic Model

demo: http://tts.org.aalto.fi/lombard_seq2seq/

paper: https://arxiv.org/pdf/1810.12051.pdf

code:

特点：提出一种迁移学习方法，基于Seq2Seq模型把正常的说话分格转换为Lombard分格（伦巴族人）。其实就是用一个已经使用正常说话分格训练好的model进行fine-tune.

Abstract：

Currently, there are increasing interests in text-to-speech (TTS) synthesis to use sequence-to-sequence models with attention. These models are end-to-end meaning that they learn both co-articulation and duration properties directly from text and speech. Since these models are entirely data-driven, they need large amounts of data to generate synthetic speech with good quality. However, in challenging speaking styles, such as Lombard speech, it is difficult to record sufficiently large speech corpora. Therefore, in this study we propose a transfer learning method to adapt a sequence-to-sequence based TTS system of normal speaking style to Lombard style. Moreover, we experiment with a WaveNet vocoder in synthesis of Lombard speech. We conducted subjective evaluations to assess the performance of the adapted TTS systems. The subjective evaluation results indicated that an adaptation system with the WaveNet vocoder clearly outperformed the conventional deep neural network based TTS system in synthesis of Lombard speech。

12. Attention Is All You Need

作者：Ashish Vaswani∗，Noam Shazeer∗， Niki Parmar∗ (Google)

时间：2017.12.06

分类：Attention

demo:

paper: https://arxiv.org/pdf/1706.03762.pdf

code:

特点：在传统的序列转换模型中，通过attention机制连接encoder和decoder，取得了最好的性能。在本文中，作者提出了一种Transformer的模型，单纯的基于attention机制，在机器翻译上取得了很好的效果。

Abstract:

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 Englishto-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Model:

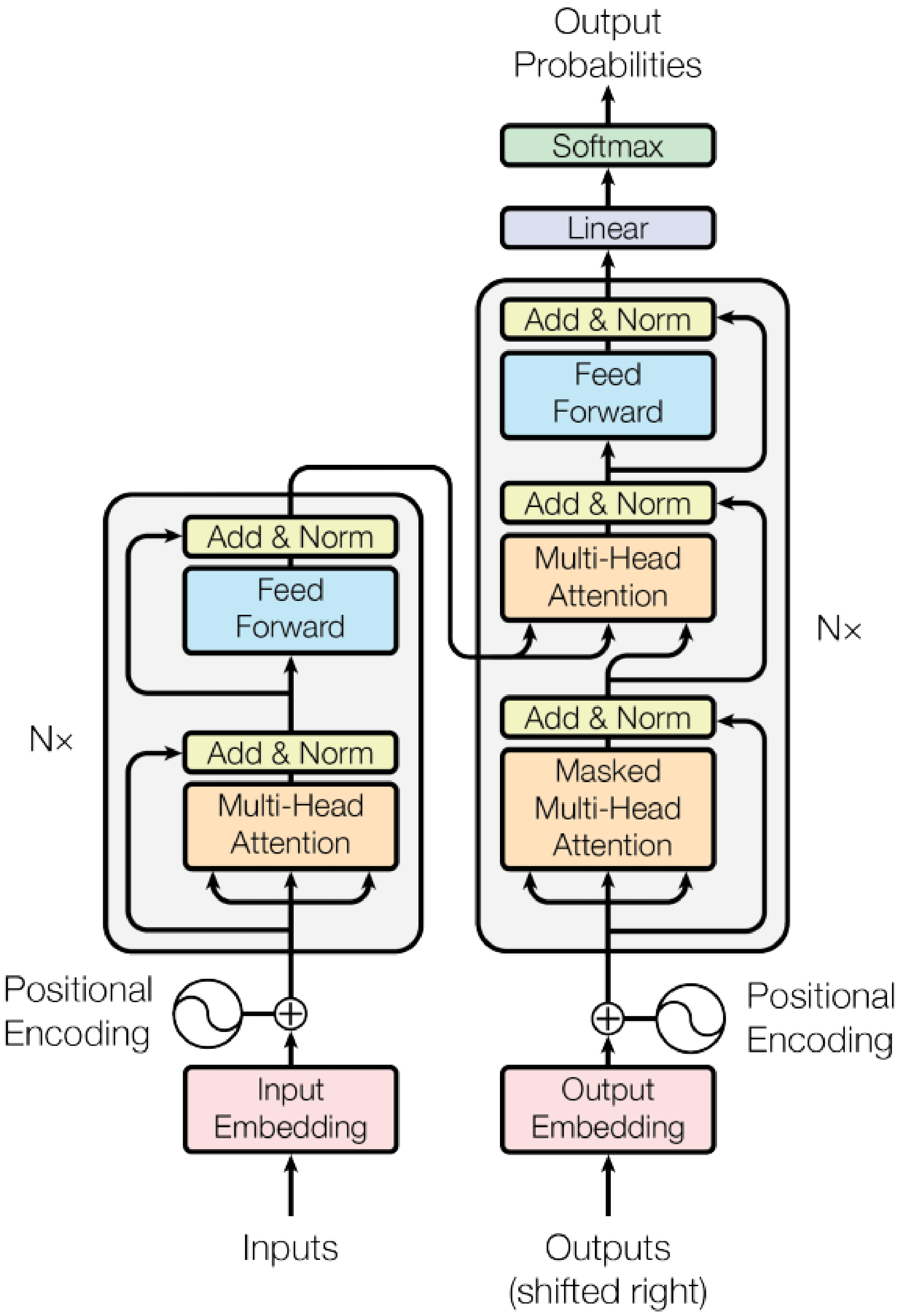


Figure 1: The Transformer - model architecture.

主要部分就是Multi-Head attention

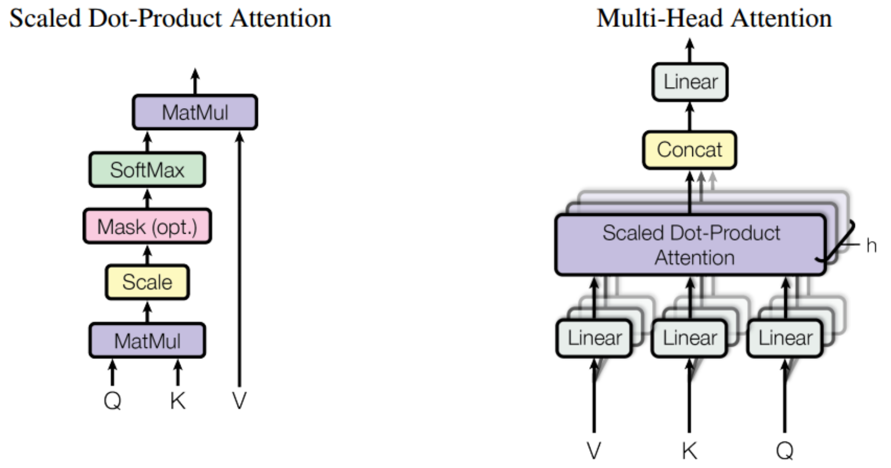
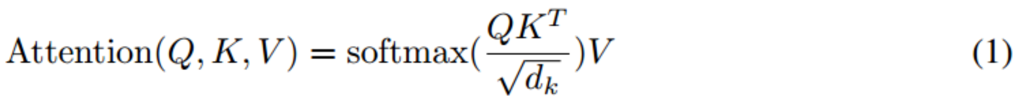
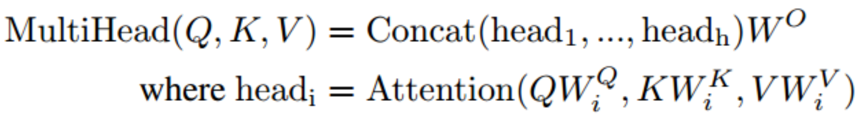


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Scaled Dot-Product Attention

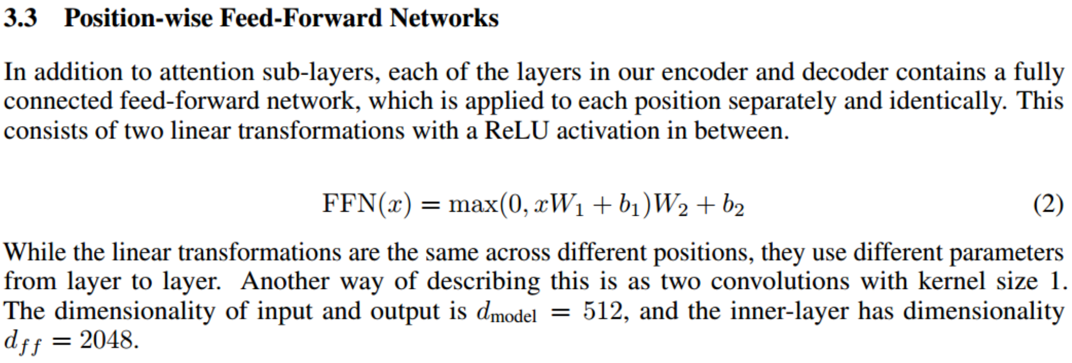


$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$



$$\begin{aligned}\mathrm{MultiHead}(Q,K,V) &= \mathrm{Concat}(\mathrm{head_1},...,\mathrm{head_h})W^O\\ \text{where } \mathrm{head_i} &= \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

所谓的FFN：



3.3 Position-wise Feed-Forward Networks

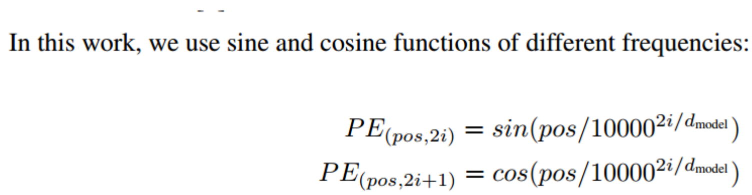
In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \qquad (2)$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{\text{model}} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$.

Positional Encoding:

因为整个网络中没有包含循环和卷积操作，因此对于两个不同时刻的相同输入，得到的结果是一样的，但是在实际中，可能会是不一样的，比如：”我爱中国“和“发展中国家”中的“中国”其实是完全不一样的性质，但是如果没有position encoding，就会得到完全一样的值，因此就有了positional encoding。



In this work, we use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

本文用实验表明该位置编码方式和学习到的位置编码方式有相似的结果。

13. SPEAKER-ADAPTIVE NEURAL VOCODERS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS SYSTEMS

作者：Eunwoo Song1;2, Jinseob Kim1, Kyungguen Byun2 (NAVER Corp)

时间：2018.11.08

分类：Vocoder（speaker-adaptive)

demo:

paper: https://arxiv.org/pdf/1811.03311.pdf

code:

特点：作者首先训练了一个speaker-independent模型，然后再此基础上进行fine-tune一个训练模型来合成指定目标说话人的风格。实验表明这种方法优于传统的source-fileter model-based vocoders and WaveNet vocoders, trained either speaker-dependently or speaker-independently.

This paper proposes speaker-adaptive neural vocoders for statistical parametric speech synthesis (SPSS) systems. Recently proposed WaveNet-based neural vocoding systems successfully generate a time sequence of speech signal with an autoregressive framework. However, building high-quality speech synthesis systems with limited training data for a target speaker remains a challenge. To generate more natural speech signals with the constraint of limited training data, we employ a speaker adaptation task with an effective variation of neural vocoding models. In the proposed method, a speaker-independent training method is applied to capture universal attributes embedded in multiple speakers, and the trained model is then fine-tuned to represent the specific characteristics of the target speaker. Experimental results verify that the proposed SPSS systems with speaker-adaptive neural vocoders outperform those with traditional source-filter model-based vocoders and those with WaveNet vocoders, trained either speaker-dependently or speaker-independently.

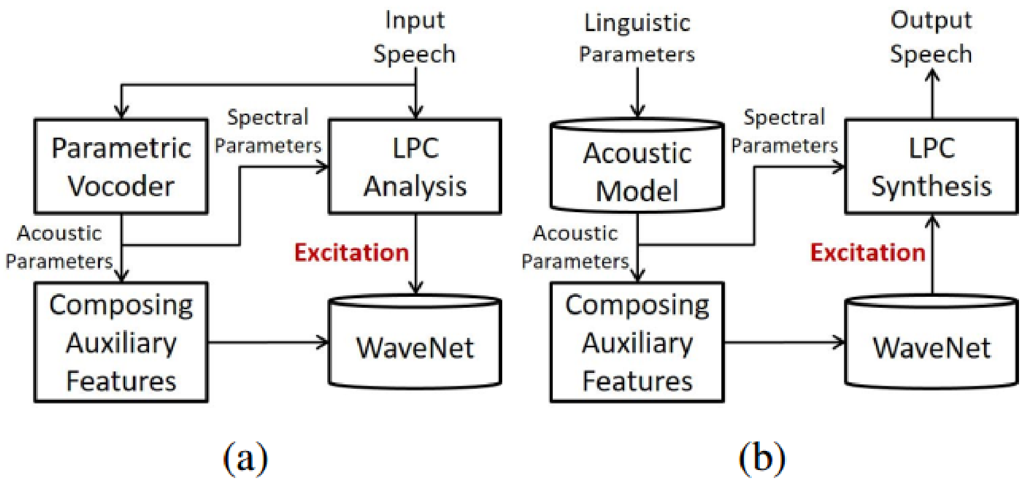


Fig. 1: ExcitNet vocoder framework for an SPSS system: (a) training and (b) synthesis.

14. Self-Attention Linguistic-Acoustic Decoder

作者：Santiago Pascual1, Antonio Bonafonte1, Joan Serra (Universitat Politecnica de Catalunya)

时间：2018.11.05

分类：Acoustic Model

demo: http://veu.talp.cat/saladtts/

paper: https://arxiv.org/pdf/1808.10678.pdf

code: https://github.com/santi-pdp/musa_tts （pytorch）

特点：采用Attention all is you need论文中decoder部分进行Linguistic-Acoustic的转换。

Abstract:

The conversion from text to speech relies on the accurate mapping from linguistic to acoustic symbol sequences, for which current practice employs recurrent statistical models like recurrent neural networks. Despite the good performance of such models (in terms of low distortion in the generated speech), their recursive structure tends to make them slow to train and to sample from. In this work, we try to overcome the limitations of recursive structure by using a module based on the transformer decoder network, designed without recurrent connections but emulating them with attention and positioning codes. Our results show that the proposed decoder network is competitive in terms of distortion when compared to a recurrent baseline, whilst being significantly faster in terms of CPU inference time. On average, it increases Mel cepstral distortion between 0.1 and 0.3 dB, but it is over an order of magnitude faster on average. Fast inference is important for the deployment of speech synthesis systems on devices with restricted resources, like mobile phones or embedded systems, where speaking virtual assistants are gaining importance.

Model:

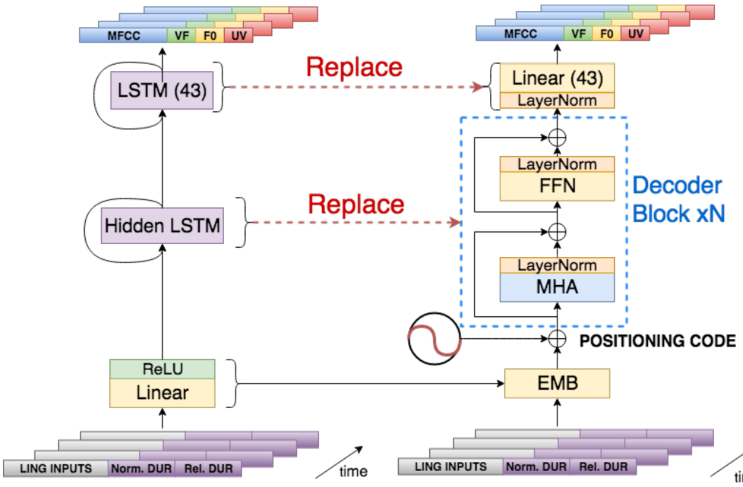


Figure 1: *Transition from RNN/LSTM acoustic model to SALAD. The embedding projections are the same. Positioning encoding introduces sequential information. The decoder block is stacked N times to form the whole structure replacing the recurrent core. FFN: Feed-forward Network. MHA: Multi-Head Attention.*

Evaluation:

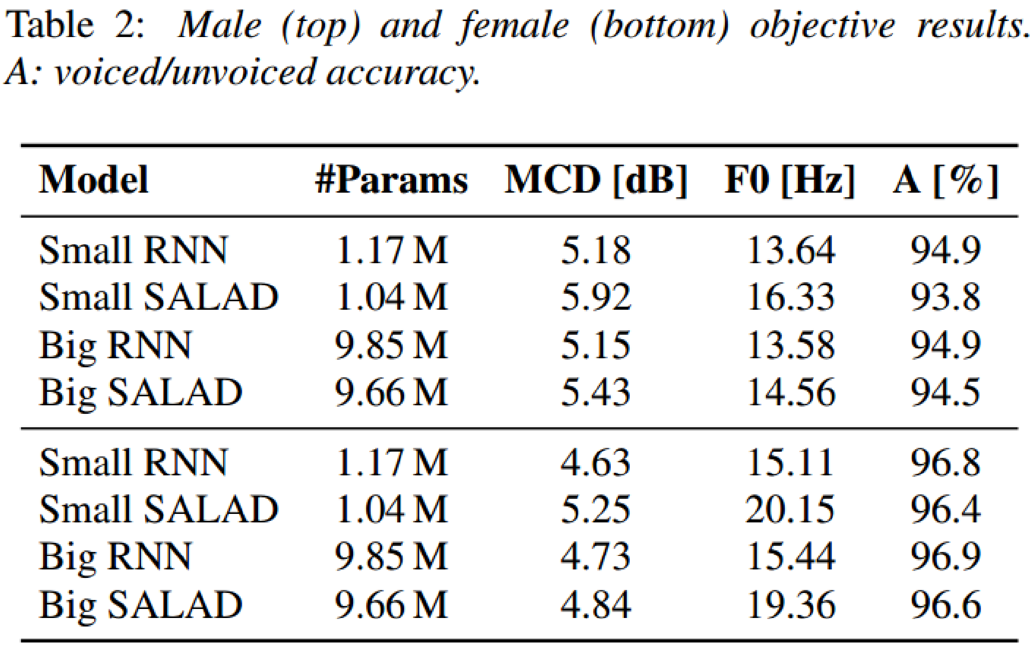


Table 2: *Male (top) and female (bottom) objective results. A: voiced/unvoiced accuracy.*

Model	#Params	MCD [dB]	F0 [Hz]	A [%]
Small RNN	1.17 M	5.18	13.64	94.9
Small SALAD	1.04 M	5.92	16.33	93.8
Big RNN	9.85 M	5.15	13.58	94.9
Big SALAD	9.66 M	5.43	14.56	94.5
Small RNN	1.17 M	4.63	15.11	96.8
Small SALAD	1.04 M	5.25	20.15	96.4
Big RNN	9.85 M	4.73	15.44	96.9
Big SALAD	9.66 M	4.84	19.36	96.6

15. **SAMPLE EFFICIENT ADAPTIVE TEXT-TO-SPEECH**

作者：Yutian Chen, Yannis Assael, Brendan Shillingford (DeepMind & Google)

时间：2018.09.27

分类：Vocoder (multi-speaker)

demo: https://sample-efficient-adaptive-tts.github.io/demo/

paper: https://arxiv.org/pdf/1809.10460.pdf code:

特点：采用meta-learning方法，基于wavenet实现multi-speaker模型的训练，可以通过很少的数据以及很少的训练步骤就可以adapt to new speakers.

Abstract:

We present a **meta-learning approach** for adaptive text-to-speech (TTS) with few data. During training, we learn a multi-speaker model using a shared conditional WaveNet core and independent learned embeddings for each speaker. The aim of training is not to produce a neural network with fixed weights, which is then deployed as a TTS system. Instead, the aim is to produce a network that requires few data at deployment time to rapidly adapt to new speakers. We introduce and benchmark three strategies: (i) learning the speaker embedding while keeping the WaveNet core fixed, (ii) fine-tuning the entire architecture with stochastic gradient descent, and (iii) predicting the speaker embedding with a trained neural network encoder. The experiments show that these approaches are successful at adapting the multi-speaker neural network to new speakers, obtaining state-of-the-art results in both sample naturalness and voice similarity with merely a few minutes of audio data from new speakers.

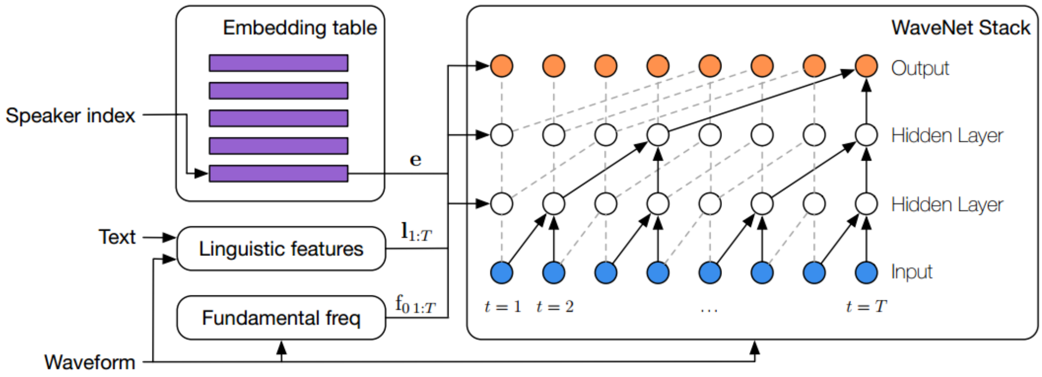


Figure 1: Architecture of the WaveNet model for few-shot voice adaptation.

16. REPRESENTATION MIXING FOR TTS SYNTHESIS

作者：Kyle Kastner, Joao Felipe Santos, Yoshua Bengio（MILA)

时间：2018.11.24

分类：Front-end

demo: https://s3.amazonaws.com/representation-mixing-site/index.html

paper: https://arxiv.org/pdf/1811.07240.pdf

code: https://github.com/kastnerkyle/representation_mixing (代码不完整)

特点：采用representation mixing, 包括character, phoneme, or mixed representations，可以提高语音合成的质量。

Atstract:

Recent character and phoneme-based parametric TTS systems using deep learning have shown strong performance in natural speech generation. However, the choice between character or phoneme input can create serious limitations for practical deployment, as direct control of pronunciation is crucial in certain cases. We demonstrate a simple method for combining multiple types of linguistic information in a single encoder, named representation mixing, enabling flexible choice between character, phoneme, or mixed representations during inference. Experiments and user studies on a public audiobook corpus show the efficacy of our approach.

17. **Reducing over-smoothness in speech synthesis using Generative Adversarial Networks**

作者：Leyuan Sheng，Evgeniy N. Pavlovskiy （Novosibirsk State University)

时间：2018.11.17

分类：

demo:

paper: https://arxiv.org/pdf/1810.10989.pdf

code:

特点：通过得到的mel-spectrogram，将其转化为image，然后通过gan网络进行训练。

Abstract:

Speech synthesis is widely used in many practical applications. In recent years, speech synthesis technology has developed rapidly. However, one of the reasons why synthetic speech is unnatural is that it often has over-smoothness. In order to improve the naturalness of synthetic speech, we first extract the mel-spectrogram of speech and convert it into a real image, then take the over-smooth mel-spectrogram image as input, and use image-to-image translation Generative Adversarial Networks(GANs) framework to generate a more realistic melspectrogram. Finally, the results show that this method greatly reduces the over-smoothness of synthesized speech and is more close to the mel-spectrogram of real speech.

18. **MODELING MULTI-SPEAKER LATENT SPACE TO IMPROVE NEURAL TTS: QUCIK ENROLLING NEW SPEAKER AND ENHANCING PREMIUM VOICE**

作者：Yan Deng, Lei He, Frank Soong （Microsoft，China)

时间：2018.11.13

分类：End-2-End (Multi-Speaker)

demo:

paper: https://arxiv.org/pdf/1812.05253.pdf

code:

特点：

Abstract:

Neural TTS has shown it can generate high quality synthesized speech. In this paper, we investigate the multi-speaker latent space to improve neural TTS for adapting the system to new speakers with only several minutes of speech or enhancing a premium voice by utilizing the data from other speakers for richer contextual coverage and better generalization. A multi-speaker neural TTS model is built with the embedded speaker information in both spectral and speaker latent space. The experimental results show that, with less than 5 minutes of training data from a new speaker, the new model can achieve an MOS score of 4.16 in naturalness and 4.64 in speaker similarity close to human recordings (4.74). For a well-trained premium voice, we can achieve an MOS score of 4.5 for out-of-domain texts, which is comparable to an MOS of 4.58 for professional recordings, and significantly outperforms single speaker result of 4.28.

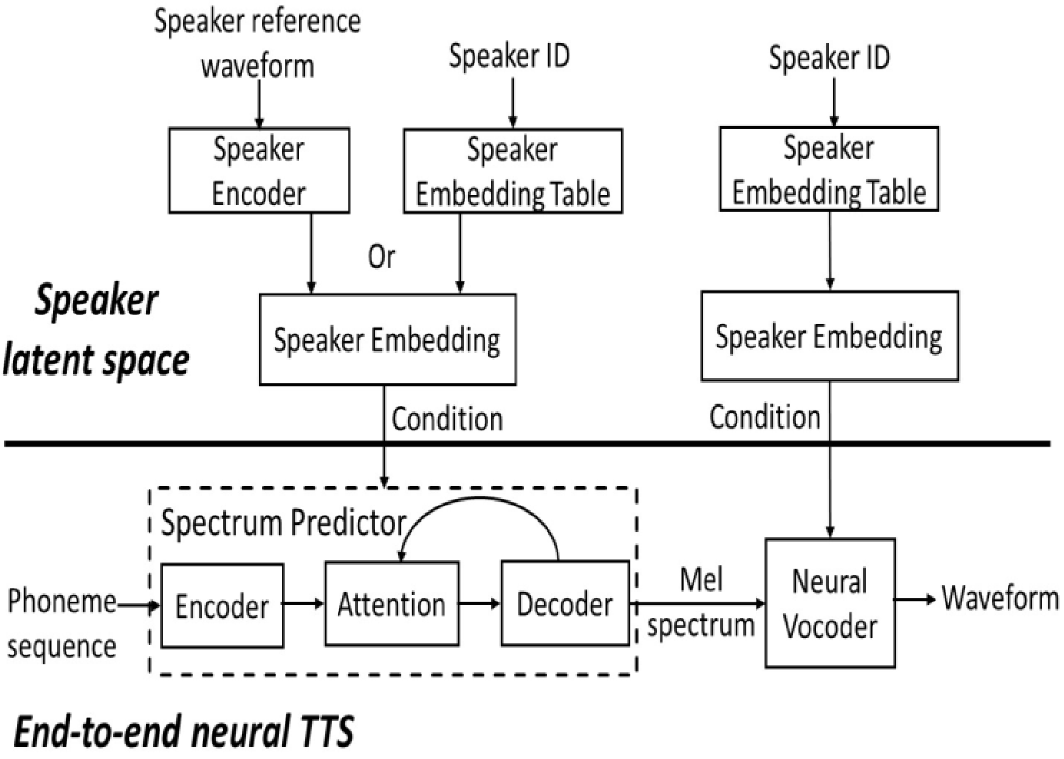


Fig. 1. Proposed multi-speaker neural TTS system.

19. **LPCNET: IMPROVING NEURAL SPEECH SYNTHESIS THROUGH LINEAR PREDICTION**

作者：Jean-Marc Valin, Jan Skoglund (Mountain View, CA, USA, Google)

时间：2018.10.28

分类：End-2-End

demo: https://people.xiph.org/~jm/lpcnet_samples/

paper: https://arxiv.org/pdf/1810.11846.pdf

code: https://github.com/mozilla/LPCNet/

特点：论文设计了一种WaveRNN的变体，通过结合线性预测和RNN网络来提升语音合成的效率，并用实验证明了该网络可以实现更高的质量，并计算难度小，有利于部署在小型的机器上。

Abstract:

Neural speech synthesis models have recently demonstrated the ability to synthesize high quality speech for text-to-speech and compression applications. These new models often require powerful GPUs to achieve real-time operation, so being able to reduce their complexity would open the way for many new applications. We propose LPCNet, a WaveRNN variant that combines linear prediction with recurrent neural networks to significantly improve the efficiency of speech synthesis. We demonstrate that LPCNet can achieve significantly higher quality than WaveRNN for the same network size and that high quality LPCNet speech synthesis is achievable with a complexity under 3 GFLOPS. This makes it easier to deploy neural synthesis applications on lower-power devices, such as embedded systems and mobile phones.

Model: LPC计算出共振峰。

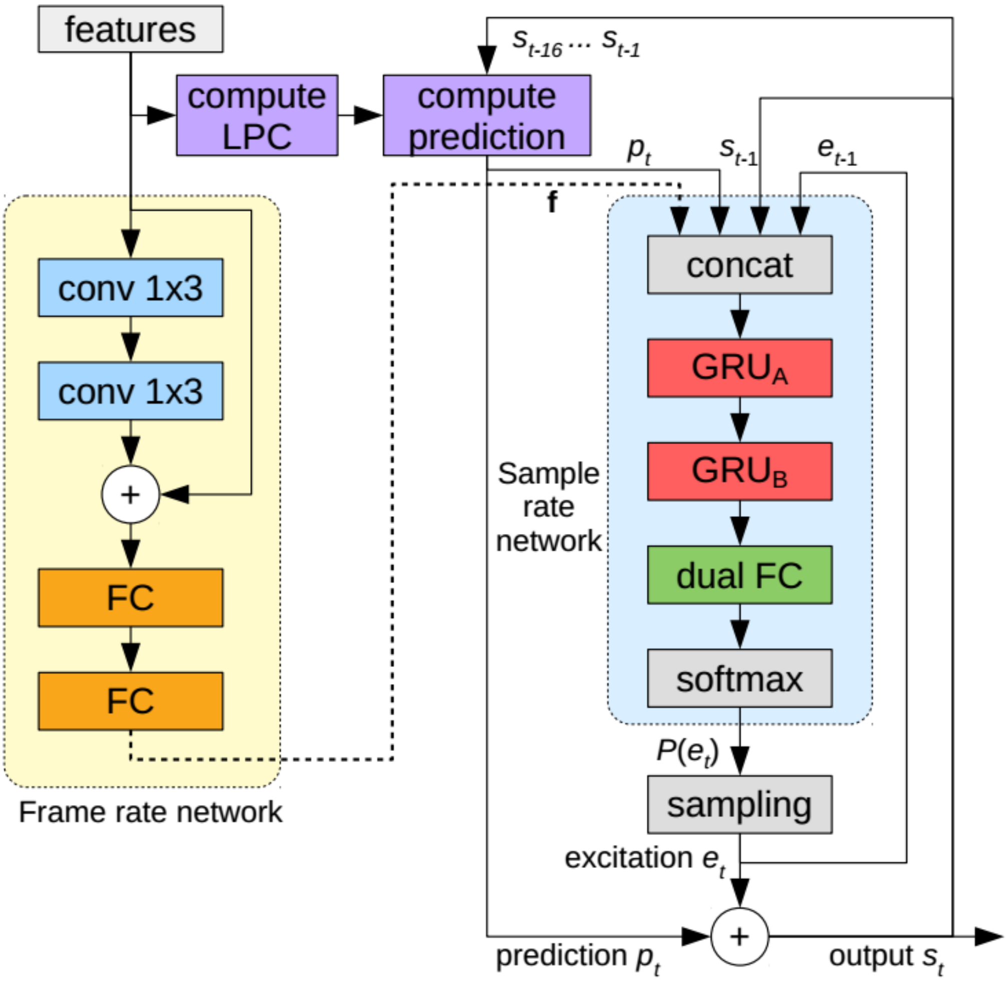


Fig. 1. Overview of the LPCNet algorithm. The left part of the network (yellow) is computed once per frame and its result is held constant throughout the frame for the sample rate network on the right (blue). The *compute prediction* block predicts the sample at time t based on previous samples and on the linear prediction coefficients. Conversions between μ-law and linear are omitted for clarity. The de-emphasis filter is applied to the output s_t.

Evaluation:

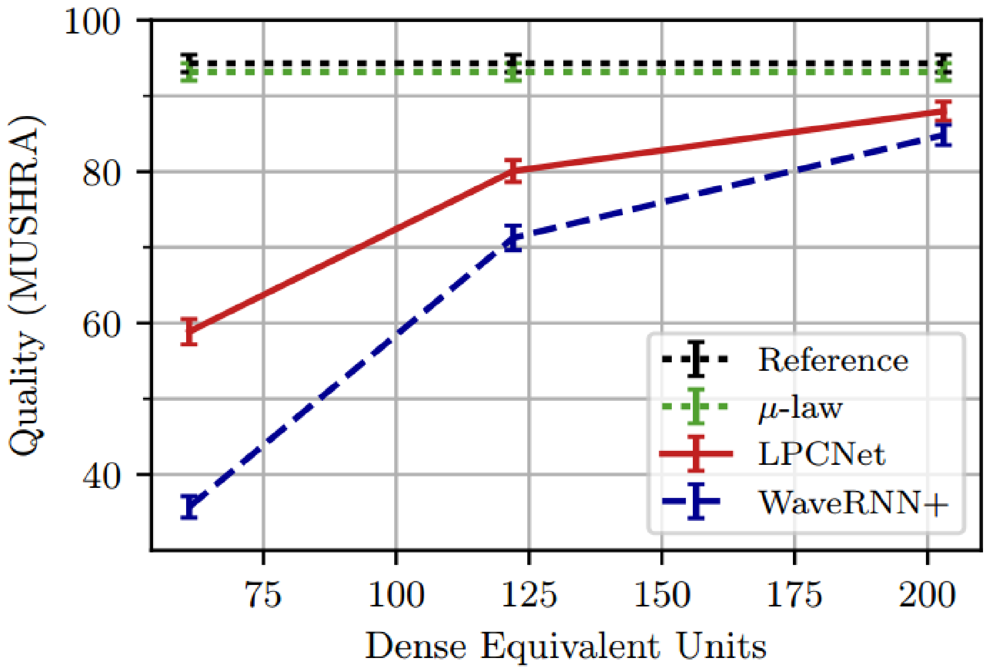


Fig. 3. Subjective quality (MUSHRA) results as a function of the dense equivalent number of units in GRU_A.

20. INVESTIGATION OF ENHANCED TACOTRON TEXT-TO-SPEECH SYNTHESIS SYSTEMS WITH SELF-ATTENTION FOR PITCH ACCENT LANGUAGE

作者：Yusuke Yasuda1, Xin Wang1, Shinji Takaki1（National Institute of Informatics, Japan)

时间：2018.10.29

分类：Acoustic Model

demo:

paper: https://arxiv.org/pdf/1810.11960.pdf

code: https://github.com/nii-yamagishilab/self-attention-tacotron

特点：采用Tacotron结构，并基于self-attention机制，实现对Japanese语音的合成

Abstract:

End-to-end speech synthesis is a promising approach that directly converts raw text to speech. Although it was shown that Tacotron2 outperforms classical pipeline systems with regards to naturalness in English, its applicability to other languages is still unknown. Japanese could be one of the most difficult languages for which to achieve end-to-end speech synthesis, largely due to its character diversity and pitch accents. Therefore, state-of-theart systems are still based on a traditional pipeline framework that requires a separate text analyzer and duration model. Towards endto-end Japanese speech synthesis, we extend Tacotron to systems with self-attention to capture long-term dependencies related to pitch accents and compare their audio quality with classical pipeline systems under various conditions to show their pros and cons. In a large-scale listening test, we investigated the impacts of the presence of accentual-type labels, the use of force or predicted alignments, and acoustic features used as local condition parameters of the Wavenet vocoder. Our results reveal that although the proposed systems still do not match the quality of a top-line pipeline system for Japanese, we show important stepping stones towards end-to-end Japanese speech synthesis.

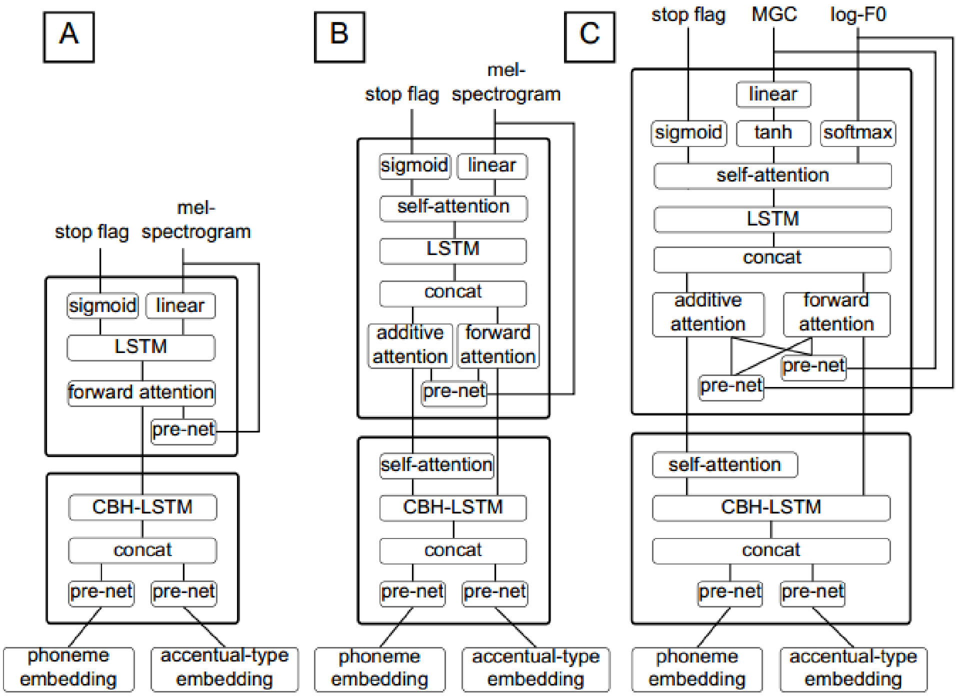


Fig. 1: Architectures of proposed systems with accentual-type embedding. A: *JA-Tacotron*. B: *SA-Tacotron*, C: *SA-Tacotron* using vocoder parameters.

效果并不咋样。

21 INVESTIGATING CONTEXT FEATURES HIDDEN IN END-TO-END TTS

作者：Kohki Mametani, Tsuneo Kato, Seiichi Yamamoto (Doshisha University, Japan)

时间：2018.11.04

分类：Acoustic Model

demo:

paper: https://arxiv.org/pdf/1811.01376.pdf

code:

特点：该论文采用Deep Voice结构，研究了中间隐含层的特性。实验结果表明，编码器的输出同时反映了语言和语音语境，如音素级的元音减少，音节级的词汇重音，词级的词性，这可能是由于语境和声学特征的共同优化。

Abstract:

Recent studies have introduced end-to-end TTS, which integrates the production of context and acoustic features in statistical parametric speech synthesis. As a result, a single neural network replaced laborious feature engineering with automated feature learning. However, little is known about what types of context information end-to-end TTS extracts from text input before synthesizing speech, and the previous knowledge about context features is barely utilized. In this work, we first point out the model similarity between end-to-end TTS and parametric TTS. Based on the similarity, we evaluate the quality of encoder outputs from an end-to-end TTS system against eight criteria that are derived from a standard set of context information used in parametric TTS. We conduct experiments using an evaluation procedure that has been newly developed in the machine learning literature for quantitative analysis of neural representations, while adapting it to the TTS domain. Experimental results show that the encoder outputs reflect both linguistic and phonetic contexts,
such as vowel reduction at phoneme level, lexical stress at syllable level, and part-of-speech at word level, possibly due to the joint optimization of context and acoustic features.

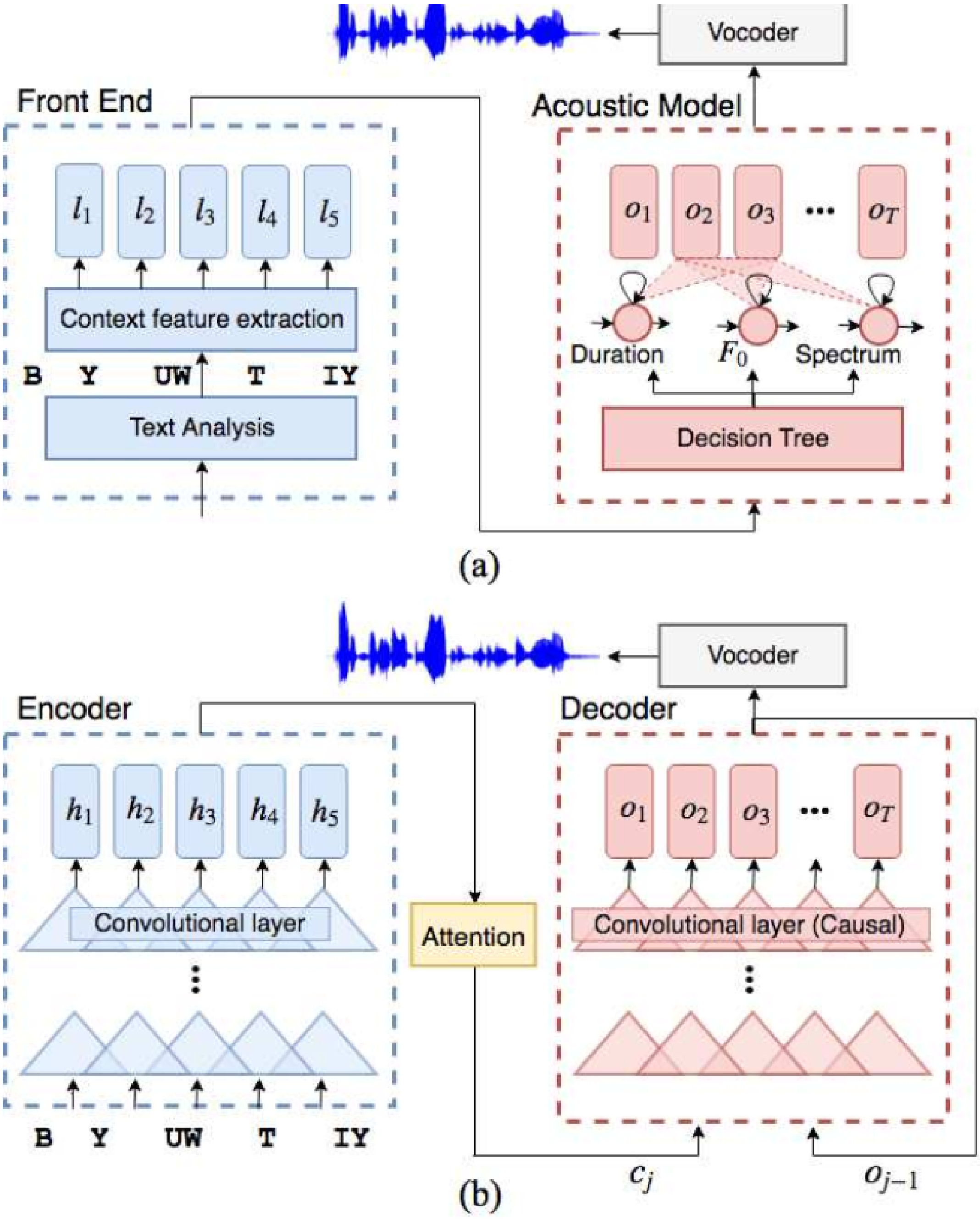


Fig. 1. Overview of speech synthesis process of (a) parametric TTS consisting of front end, acoustic model illustrated as HMMs, and vocoder and (b) end-to-end TTS model consisting of encoder, attention-based decoder illustrated as causal convolution networks, and vocoder.

22 HIERARCHICAL GENERATIVE MODELING FOR CONTROLLABLE SPEECH SYNTHESIS

作者：Wei-Ning Hsu1∗ Yu Zhang2 Ron J. Weiss2 (MIT Google Brain)

时间：2018.12.27

分类：Acoustic Model

demo: https://google.github.io/tacotron/publications/gmvae_controllable_tts/

paper: https://arxiv.org/pdf/1810.07217.pdf

code: [可以看一下demo]

特点：提出GMVAE-Tacotron模型，该模型能够控制生成的语音中很少在训练数据中标注的潜在属性，如说话风格、口音、背景噪声、录制条件等。该模型是一种基于变分自编码器(VAE)框架的条件生成模型，具有两个层次的潜在变量。第一层是一个分类变量，用于表示属于哪个组，比如（clean/noise），并且具有可解释性。第二层在第一层的基础上，用于构建特殊的属性，包括噪声等级、说话速度，或者一些更精细的控制。

Abstract:

This paper proposes a neural sequence-to-sequence text-to-speech (TTS) model which can control latent attributes in the generated speech that are rarely annotated in the training data, such as speaking style, accent, background noise, and recording conditions. The model is formulated as a conditional generative model based on the variational autoencoder (VAE) framework, with two levels of hierarchical latent variables. The first level is a categorical variable, which represents attribute groups (e.g. clean/noisy) and provides interpretability. The second level, conditioned on the first, is a multivariate Gaussian variable, which characterizes specific attribute configurations (e.g. noise level, speaking rate) and enables disentangled fine-grained control over these attributes. This amounts to using a Gaussian mixture model (GMM) for the latent distribution. Extensive evaluation demonstrates its ability to control the aforementioned attributes. In particular, we train a high-quality controllable TTS model on real found data, which is capable of inferring speaker and style attributes from a noisy utterance and use it to synthesize clean speech with controllable speaking style.

Model:

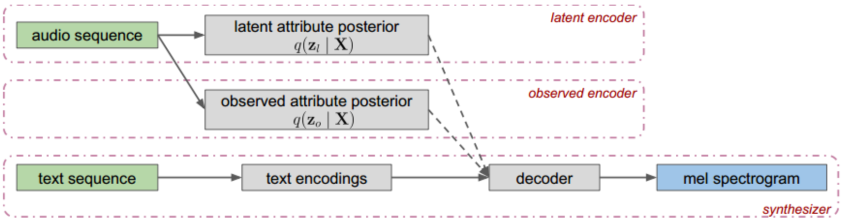

audio sequence
latent attribute posterior
$q(\mathbf{z}_l \mid \mathbf{X})$
latent encoder
observed attribute posterior
$q(\mathbf{z}_o \mid \mathbf{X})$
observed encoder
text sequence
text encodings
decoder
mel spectrogram
synthesizer

Figure 2: Training configuration of the GMVAE-Tacotron model. Dashed lines denotes sampling. The model is comprised of three modules: a synthesizer, a latent encoder, and an observed encoder.