Wavenet:
A Generative Model for Raw Audio

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan,
Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu

2022. 06. 21. (TUE)

Youngwon Choi
모두의연구소

풀잎스쿨 **Hands-on TTS**

Contents

1. Introduction

Introduction

□ Wavenet은 image나 text 기반의 neural autoregressive generative models에서 영감을 받아 만들어진 raw audio generation technique 이다.

□ Wavenet 논문의 main contributions은 다음과 같다.

- Wavenet은 기존 보고된 TTS 모델들보다 subjective naturalness가 훨씬 좋은 raw speech signals을 생성할 수 있음.
- Raw audio generation을 위해 필요한 long-range temporal dependencies를 처리하기 위해, Wavenet에는 굉장히 넓은 receptive fields를 가지는 dilated causal convolutions이란 새 구조가 적용됨.
- Wavenet은 speaker identity에 대한 정보가 conditional variable로 들어갔을 때, 여러가지 voice를 생성해 낼 수 있음.
- Wavenet은 음성 생성 뿐만 아니라 speech recognition, 그리고 music과 같은 다른 오디오 양식(modality)를 생성하는 task에도 유용함.

2. Wavenet

Wavenet

□ Waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ 의 joint probability 는 다음과 같이 표현된다.

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

– 즉 각 오디오 샘플 x_t는 이전의 모든 샘플들에 대해 conditioned 됨.

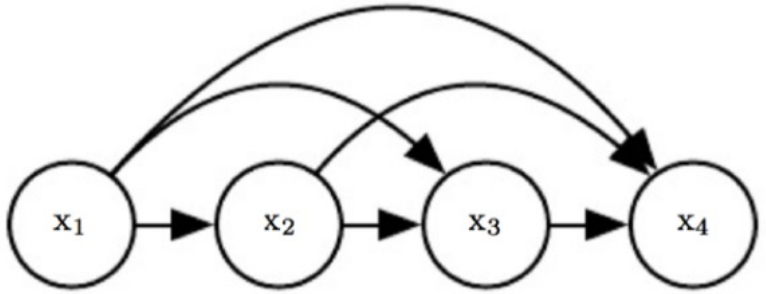
□ Wavenet에서는 이전의 샘플들을 기반으로, softmax layer를 이용하여 그 다음 샘플 x_t의 categorical distribution을 output으로 내놓게 된다.

– Model parameter에 대해, predicted data의 log-likelihood를 maximize하도록 optimize됨.

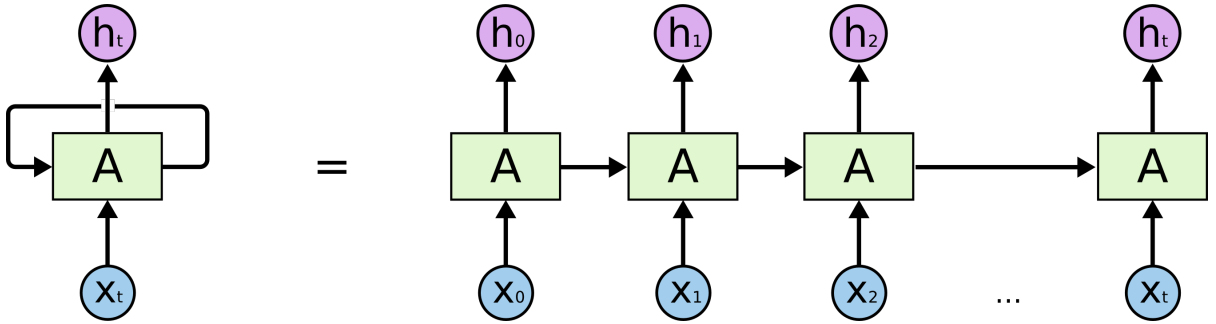
– Log-likelihood는 tractable하기 때문에, validation set을 이용하여 model이 overfitting되고 있는지, underfitting되고 있는지 쉽게 확인 가능하다고 함.

Autoregressive model

□ Autoregressive model이란 자기 자신을 입력으로 하여 자기 자신을 예측하는 모형을 의미함.



□ 대표적인 사례로 RNN 계열 모델들이 있음.



Wavenet

*2.1 Dilated **Causal** Convolution ← Wavenet의 핵심!*

□ Wavenet의 main ingredients는 causal convolution이다.

- Causal convolution을 사용함으로써, prediction $p(x_{t+1}|x_1, \ldots, x_t)$가 미래 샘플 $x_{t+1}, \ldots, x_T$의 영향을 받지 않도록 할 수 있음.
- 1-D data에서는, normal convolution의 output을 timesteps shifting하는 것으로 causal convolution을 쉽게 구현할 수 있음.
- Noncausal 1D convolution에서는 prediction이 미래 샘플의 영향을 받음.

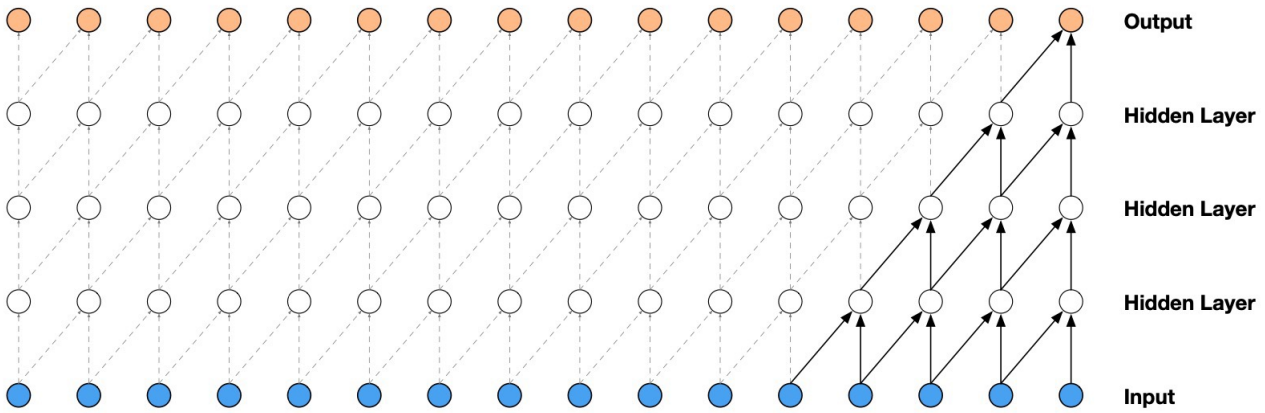
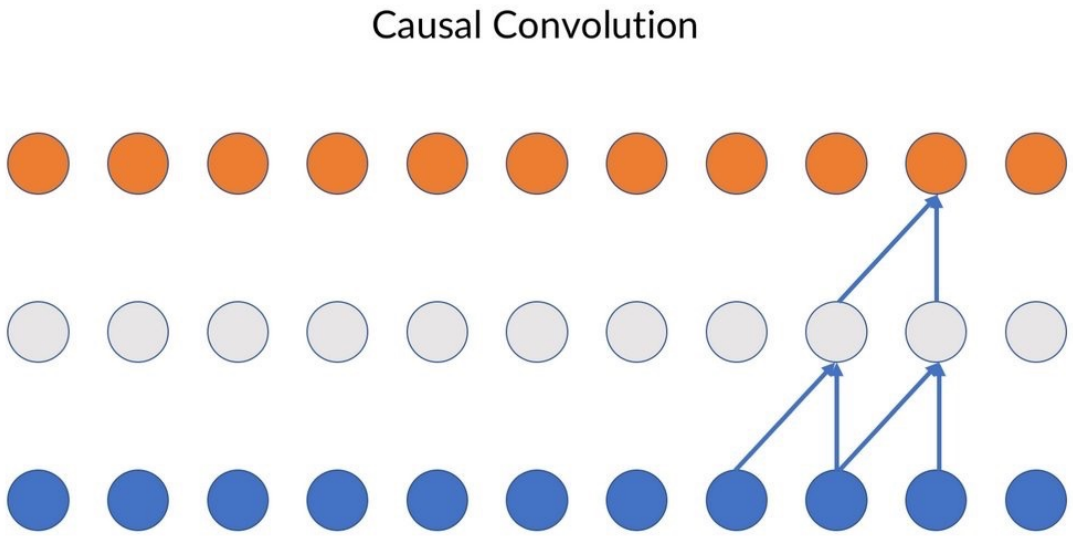


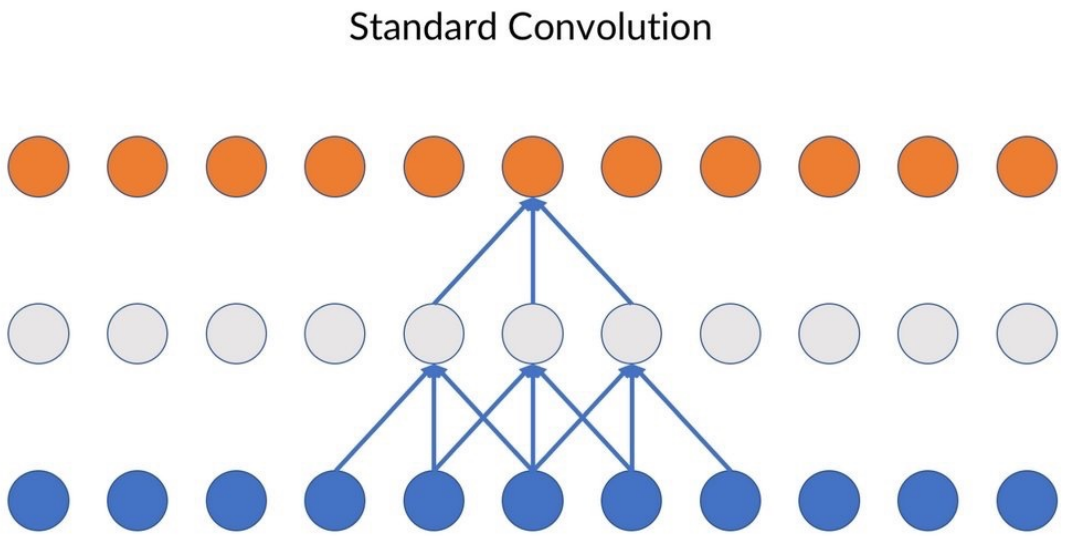
Figure 2: Visualization of a stack of causal convolutional layers.

Wavenet

*2.1 Dilated **Causal** Convolution ← **Wavenet의 핵심!***

□ Causal convolution vs Noncausal convolution





Wavenet

*2.1 Dilated **Causal** Convolution*

□ Causal Convolution의 Training 시에는 ground truth $\mathbf{x}$를 알고 있기 때문에 parallel하게 prediction이 가능하다.

– Causal convoution은 recurrent connections가 없기 때문에, RNN 계열 모델들보다 학습이 빠름.

□ 모델을 활용하여 Generate할 때에는, prediction이 sequential하게 이루어진다.

– Sample이 predicted되면, 그 다음 sample의 prediction을 위해 fed back됨.

Wavenet

*2.1 **Dilated** Causal Convolution*

□ Causal Convolution의 문제점은 receptive field를 늘리기 위해서 많은 layer, 혹은 각 convolution layer에서 큰 filter를 필요로 한다는 점이다.

– 따라서, 본 모델은 computational cost를 크게 늘리지 않고도 order of magnitude(크기 척도)로 receptive field를 늘릴 수 있는 dilated convolution을 사용함.

□ Dilated Convolution이란 input values 내에서 몇 step을 skipping 함으로 필터가 넓은 영역에 걸쳐 작용할 수 있도록 하는 convolution layer이다.

– 효과적으로 넓은 receptive field를 갖는 모델을 구현 가능.

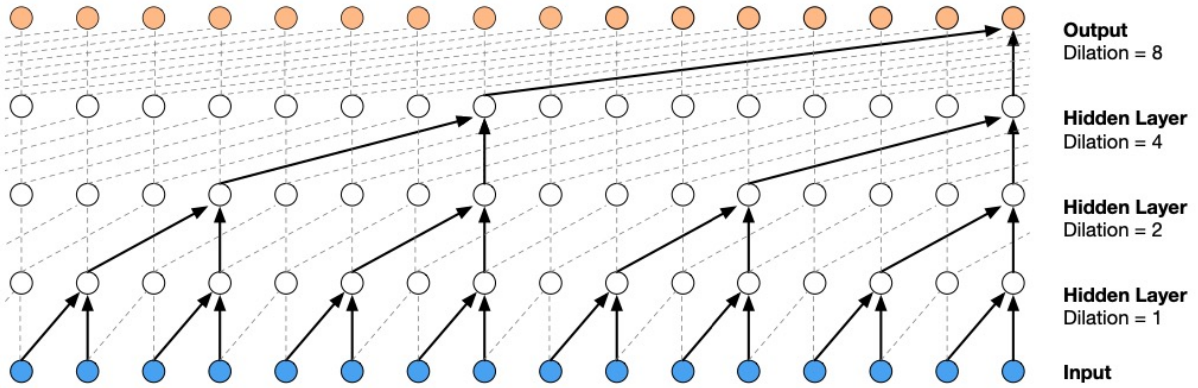


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

Wavenet

*2.1 **Dilated** Causal Convolution*

□ 본 논문에서는 dilation configuration을 limit 까지 2배로 늘리고, limit까지 늘어나면 반복하도록 함.

- 즉 1, 2, 4, …, 512, 1, 2, 4, …, 512, 1, 2, 4, …,512

□ 위 configuration에 대한 Intuition은 다음과 같다.

- Exponential하게 dilation factor을 키우면, Receptive field도 exponential하게 커짐.
  - 1, 2, 4, …, 512 로 dilation factor를 키우면, receptive field의 크기는 1024 가 됨.
  - 1 * 1024 convolution보다는 위와 같은 configuration이 더 효과적이고 non-linear함.
- Block들을 쌓아 올리는 것은 차후에 모델 크기나 receptive field 크기를 키워야 할 때 용이함.

Wavenet

2.2 Softmax Distribution

□ Raw audio 는 보통 한 timestep에 16-bit integer value로 quantization 되어 있다.

- 이 때, 각 value에 대해 모든 probabilities 값을 갖는 모델을 설계하려면, softmax layer의 output이 65,536이어야 함.
- 이는 사실상 설계하기 어려움.

□ 따라서, 16bit quantization을 8bit quantization으로 줄이기 위해 μ-law companding transformation 를 사용함.

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

  - $-1 < x_t < 1, \mu = 255$
- Linear하게 quantization하는 것 보다, 위와 같이 non-linear한 quantization 방법이 더 효과적이라고 함.

Wavenet

2.3 Gated Activation Units

□ 본 논문에서는 gated PixelCNN 논문에서 사용한 gated activation unit을 사용했다.

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

- $*$: convolution operator, $\odot$: element-wise multiplication operator, σ: sigmoid function
- k : layer index, f, g : filter and gate respectively, W: learnable convolution filter

□ ReLU보다 Gated Activation이 성능이 좋다고 함.

- 두개의 activation function을 복합적으로 사용해서 그런게 아닐까… 하는 추측이 있음.

Wavenet

2.4 Residual and Skip Connections

□ Causal convolution을 거친 후 k번의 residual block을 거침.

– Convergence를 빠르게 하고 deep한 모델의 학습을 위해, residual skip connection과 parameterised skip connection이 본 residual block에서 활용됨.

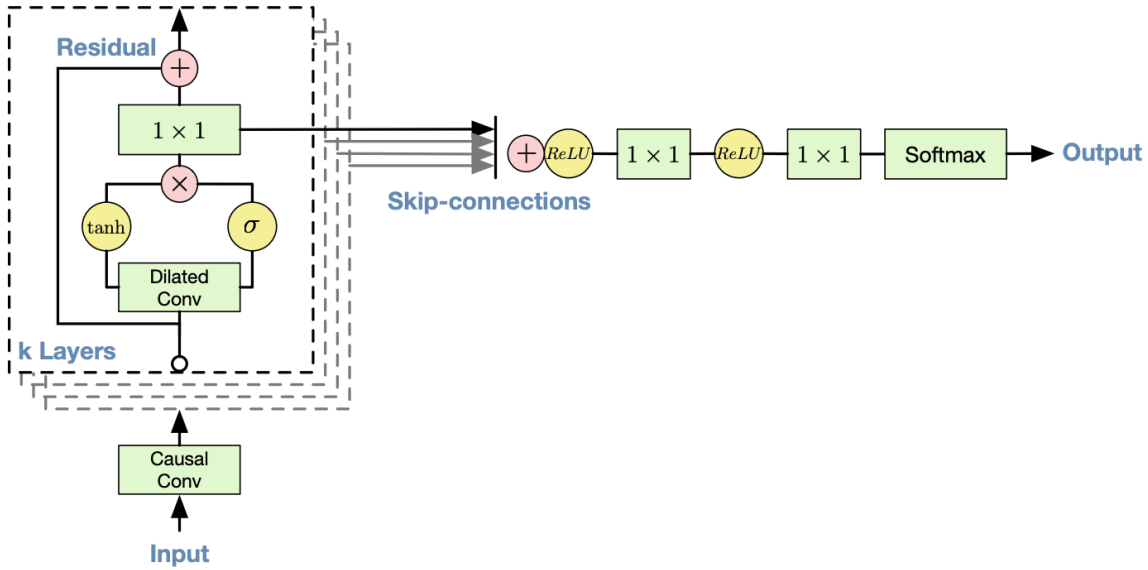


Figure 4: Overview of the residual block and the entire architecture.

Wavenet

2.5 Conditional Wavenets

□ Wavenet은 additional input h가 주어졌을 때 conditional distribution $p(\mathbf{x}|h)$를 모델링할 수 있다.

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

□ Conditional variables를 입력함으로, wavenet이 required characteristics를 가지는 audio를 만들게 하도록 학습시킬 수 있다.

- For multi-speaker setting, conditional variable <- speaker identity
- For TTS, conditional variable <- information about text

□ 저자는 두 가지 방법으로 model을 condition했다.

- Global conditioning, for speaker embedding
- Local conditioning, for TTS.

Wavenet

2.5 Conditional Wavenets

□ Global Conditioning은 모든 timstep의 output distribution에 영향을 미치는 single latent representation $\mathbf{h}$로 characterized 된다.

- e.g. a speaker embedding in a TTS model
- 2.3 에서 나온 activation funtion은 아래와 같이 표현됨.

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

  - $V_{*,k}$: Learnable linear projection

□ Local condition에서는 timeseries 데이터 h_t를 이용한다.

- h_t는 sampling frequency가 audio signal보다 낮음.
- e.g. linguistic features

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

  - $\mathbf{y} = f(\mathbf{h})$, f: transposed convolution network that maps $\mathbf{h}$ with the same resolution as the audio signal.

Wavenet

2.6 Context Stacks

□ 저자는 모델의 receptive field를 늘리는 다른 방법으로 context stacks를 제안하고 있다.

□ 작지만 receptive field가 넓은 wavenet의 output을 synthesis에 사용할 wavenet의 conditional variable로 사용한 것으로 보인다.

- TTS 실험에서는 사용하지 않은 것으로 보임.

3. Experiment

Experiment

🙌 Speech Synthesis 실험 결과만 설명드리겠습니다. 🙌

3.1 Multi-Speaker Speech Generation – Global condition으로 화자 정보를 주는 케이스

□ Dataset: VCTK
- 44 hours of data from 109 speakers.
- Feeding the speaker ID in model in the form of a one-hot vector.

□ Wavenet은 non-existent, but human language-like words를 생성한다.
- Language나 image를 다루는 Generative models들도, 추론을 통해 나온 샘플들이 대충 볼 때는 좋아보이나, 자세히 보면 unnatural해 보이는 경향이 있다고 함.
- Sample의 long range coherence가 부족한 이유는 300 milliseconds정도의 한정된 model의 Receptive field 때문.

□ 하나의 WaveNet은 데이터셋 내의 모든 speaker에 대한 Sample를 만들 수 있다.
- 하나의 모델이 109명의 화자에 대한 특징을 다 표현가능할 정도로 강력함을 증명.
- 1명의 화자 데이터로 학습시키는 것 보다 여러명의 화자 데이터로 학습을 진행하는 것이 더 좋은 성능을 보이며, 이는 모델의 internal representation이 다수의 화자들 사이에서 공유되기 때문인 것으로 보임.

□ 화자 정보뿐만 아니라, recording quality, 숨소리 등 다른 오디오 정보들도 모방하였다고 함.

Experiment

3.2 TTS – Local condition으로 텍스트 관련 정보를 주는 케이스

□ Dataset
- North American English dataset: 24.6 hours, female speakers.
- Mandarin Chinese dataset: 34.8 hours, female speakers.

□ Local condition: linguistic(언어적) features, log F0
- Log F0는 외부 모델을 이용하여 predict함.
  - 200Hz 이하의 F0 값 특성상, wavenet의 작은 receptive field값은 이를 반영하기 충분하지 않다고 함.
- 어떤 방식으로 모델에 input하였는지에 대한 설명은 없음.

□ Evaluation: Subjective paired comparison tests, MOS
- Subjective paired comparison tests: 두 개의 음원 샘플을 듣고, 더 선호하는 쪽을 선택
- MOS: 음원을 듣고 naturalness를 기준으로 1부터 5 사이에서 품질을 평가함.

Experiment

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

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

Experiment

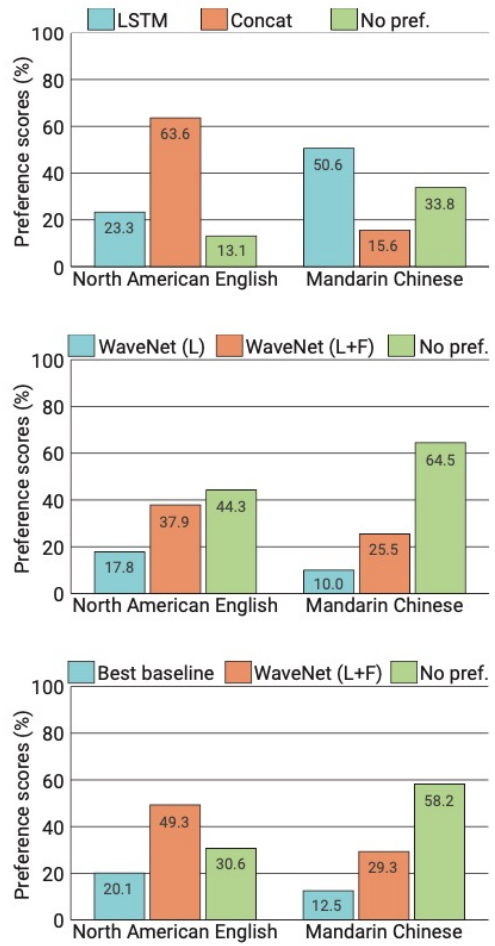


Figure 5: Subjective preference scores (%) of speech samples between (top) two baselines, (middle) two WaveNets, and (bottom) the best baseline and WaveNet. Note that LSTM and Concat correspond to LSTM-RNN-based statistical parametric and HMM-driven unit selection concatenative baseline synthesizers, and WaveNet (L) and WaveNet (L+F) correspond to the WaveNet conditioned on linguistic features only and that conditioned on both linguistic features and $\log F_0$ values.

Thank you for listening!
Q&A