Andreas Pirchner Stephanie Autherith

universität innsbruck

Machines Talking

The Natural Way

Outline

1. The Art Of Speech Synthesis
2. WaveNet’s Dawn
3. Text To Speech
4. Performance
5. Summary

Outline

Application Of Speech Synthesis

Speak now

Hey Cortana

Speech Synthesis

Applications

Speech Synthesis

Applications

- Present in every smart device

Speech Synthesis

Applications

- Present in every smart device
- Assistive Technology

Speech Synthesis

Applications

- Present in every smart device
- Assistive Technology
- Infotainment and Entertainment

Speech Synthesis

Applications

- Present in every smart device
- Assistive Technology
- Infotainment and Entertainment
- Text To Speech

Speech Synthesis

Applications

- Present in every smart device
- Assistive Technology
- Infotainment and Entertainment
- Text To Speech
- Aiding People With Disablities

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

- hours of speaker recordings

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

- hours of speaker recordings
- sliced into fragments - phonemes

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

- hours of speaker recordings
- sliced into fragments - phonemes
- concatenation of phonemes to form new phrases

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

- hours of speaker recordings
- sliced into fragments - phonemes
- concatenation of phonemes to form new phrases
- example based model

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Methods

Statistical Parametric Speech Synthesis[1]

- predict waveform for each sound
- via HMM - prediction based on current state
- via NN - based on interaction with layers and output
- generative model

Concatinative Speech Synthesis

- hours of speaker recordings
- sliced into fragments - phonemes
- concatenation of phonemes to form new phrases
- example based model

Combination

Apple’s latest iteration of Siri combines those two approaches

[1] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

Speech Synthesis Issues

Prosody

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

- noticeable longer phrases/sentences

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

- noticeable longer phrases/sentences
- causes speech to sound off and unnatural

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

- noticeable longer phrases/sentences
- causes speech to sound off and unnatural
- makes it less intelligible

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

- noticeable longer phrases/sentences
- causes speech to sound off and unnatural
- makes it less intelligible

Can we do better?

Speech Synthesis Issues

Prosody

- rhythm and intonation in natural speech
- conveys emotion and linguistic cues
- gives context and meaning

Lack of Prosody

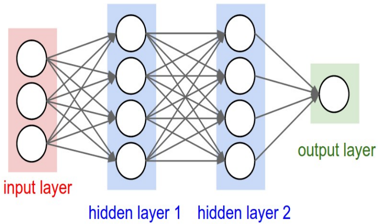
- noticeable longer phrases/sentences
- causes speech to sound off and unnatural
- makes it less intelligible

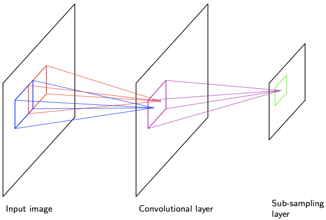
Can we do better?

- WaveNet

Outline

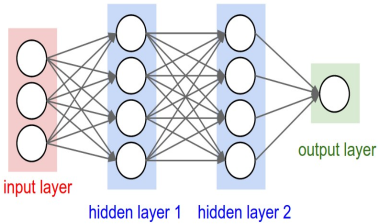
WaveNet

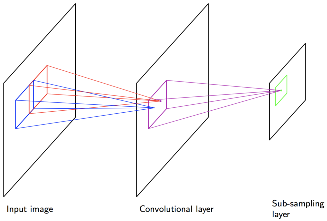




General

WaveNet

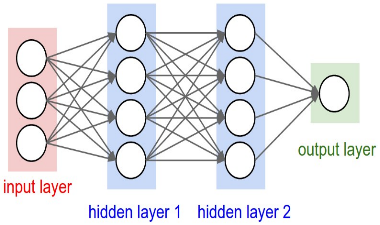


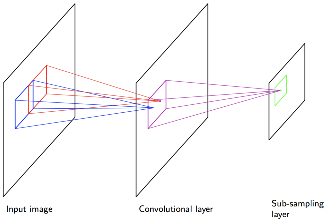


General

- Convolutional Deep Neural Network by DeepMind

WaveNet

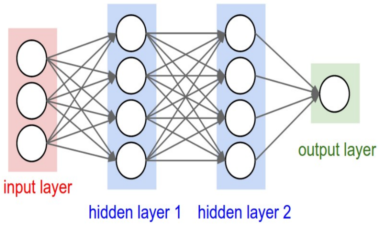


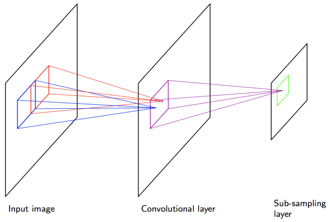


General

- Convolutional Deep Neural Network by DeepMind
- generates raw audio waves via predictive distributions

WaveNet

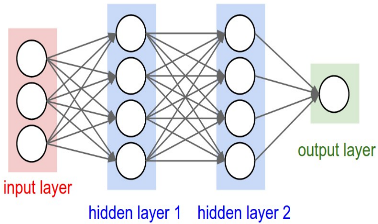


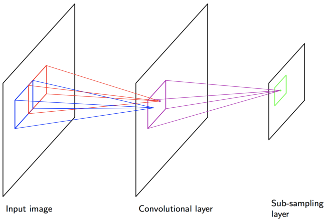


General

- Convolutional Deep Neural Network by DeepMind
- generates raw audio waves via predictive distributions
- all outputs influenced by previously generated samples

WaveNet

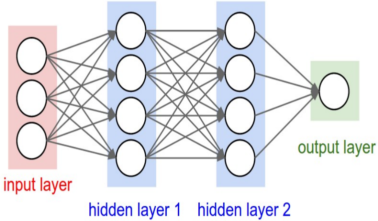


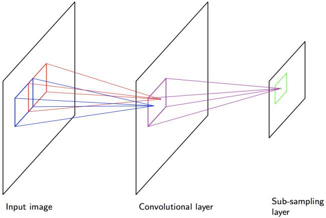


General

- Convolutional Deep Neural Network by DeepMind
- generates raw audio waves via predictive distributions
- all outputs influenced by previously generated samples

WaveNet

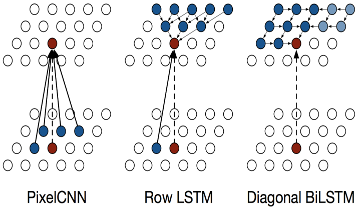




General

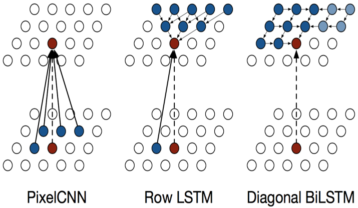
- Convolutional Deep Neural Network by DeepMind
- generates raw audio waves via predictive distributions
- all outputs influenced by previously generated samples
- inspired by DeepMind's PixelCNN

PixelCNN/RNN



Inspiration

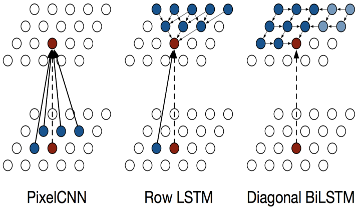
PixelCNN/RNN



Inspiration

- capable of producing natural appearing images

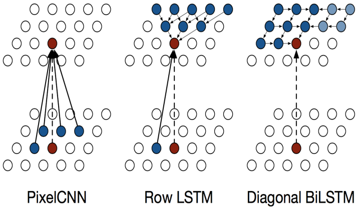
PixelCNN/RNN



Inspiration

- capable of producing natural appearing images
- pixel per pixel and one color channel at a time

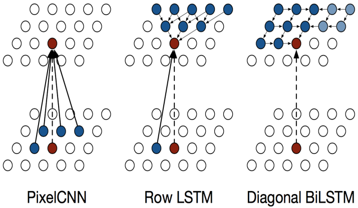
PixelCNN/RNN



Inspiration

- capable of producing natural appearing images
- pixel per pixel and one color channel at a time
- RNN: convolution of LSTM layers for calculation along one dimension

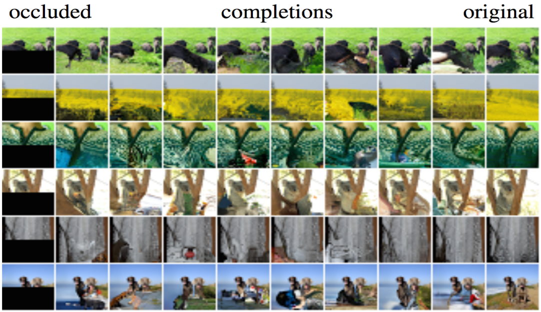
PixelCNN/RNN



Inspiration

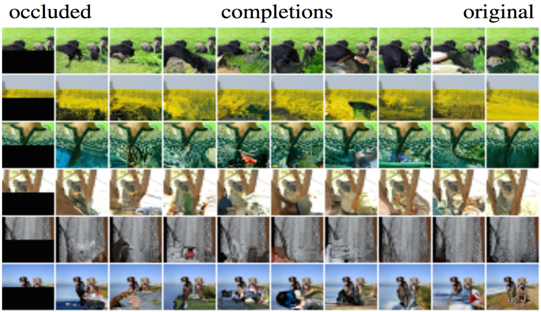
- capable of producing natural appearing images
- pixel per pixel and one color channel at a time
- RNN: convolution of LSTM layers for calculation along one dimension
- CNN: fully convolutional with fixed dependency range (masks)

PixelCNN Connection



Similar Challenges

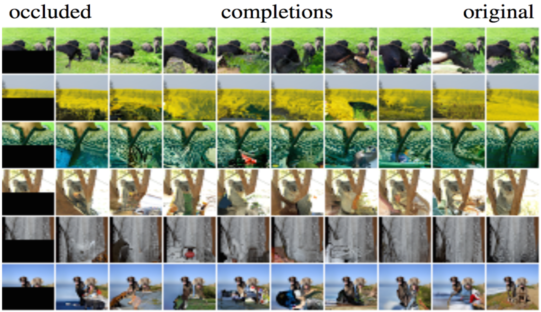
PixelCNN Connection



Similar Challenges

- for images and audio i.r.t. sample-inter-dependencies and sizes

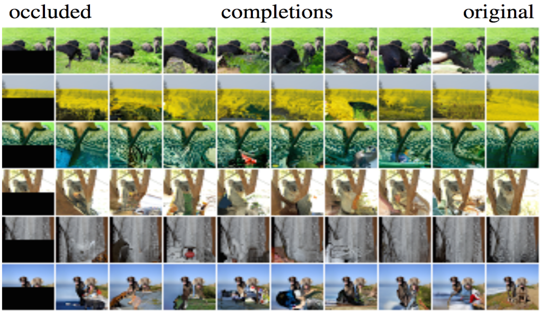
PixelCNN Connection



Similar Challenges

- for images and audio i.r.t. sample-inter-dependencies and sizes
- PixelCNN can accommodate thousands prediction per image

PixelCNN Connection



Similar Challenges

- for images and audio i.r.t. sample-inter-dependencies and sizes
- PixelCNN can accommodate thousands prediction per image
- WaveNet needs to predict 16000 samples per second

Inside WaveNet

Model

Inside WaveNet

Model

- fully probabilistic and autoregressive

Inside WaveNet

Model

- fully probabilistic and autoregressive
- predicts its outcome assuming the current value depends on previous ones

Inside WaveNet

Model

- fully probabilistic and autoregressive
- predicts its outcome assuming the current value depends on previous ones

More formally:

Inside WaveNet

Model

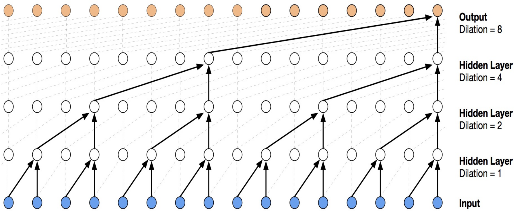
- fully probabilistic and autoregressive
- predicts its outcome assuming the current value depends on previous ones

More formally:

- The joint probability $p(\mathbf{x})$ - with $\mathbf{x} = (x_1, \ldots, x_T)$ being the waveform - is the product of all probabilities for x_t conditional on $x_1, \ldots, x_{t-1}$:

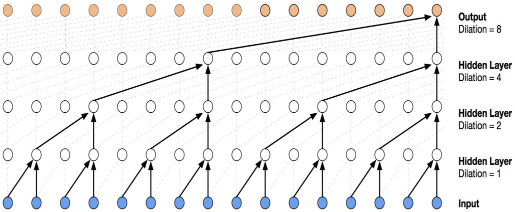
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

Inside WaveNet



Architecture

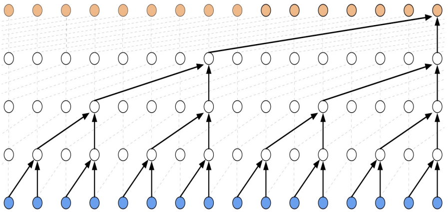
Inside WaveNet



Architecture

- models distribution via stacking causal convolutional layers

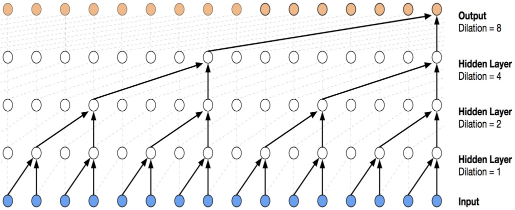
Inside WaveNet



Architecture

- models distribution via stacking causal convolutional layers
- no pooling layers - hence no downsampling - no information loss

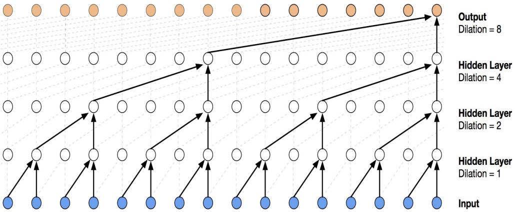
Inside WaveNet



Architecture

- models distribution via stacking causal convolutional layers
- no pooling layers - hence no downsampling - no information loss
- causal convolutions prevent predicting based on future values

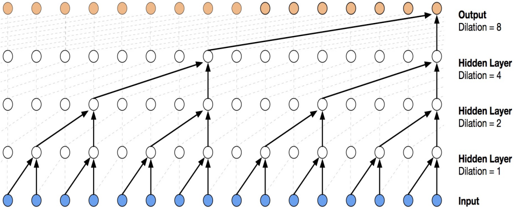
Inside WaveNet



Architecture

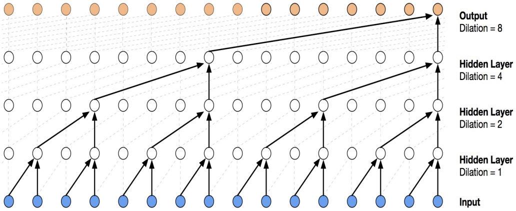
- models distribution via stacking causal convolutional layers
- no pooling layers - hence no downsampling - no information loss
- causal convolutions prevent predicting based on future values
- shifting the output of convolutions by a few timestamps

Inside WaveNet



Architecture

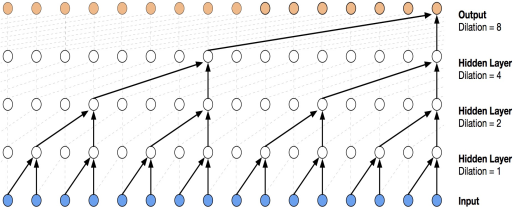
Inside WaveNet



Architecture

- CNN faster at training then RNN

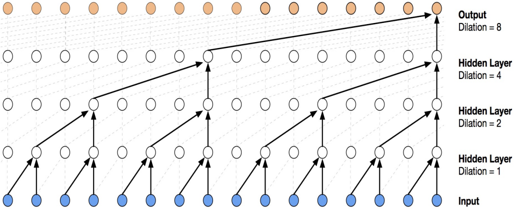
Inside WaveNet



Architecture

- CNN faster at training then RNN
- Additional layers/larger filters requried to keep receptive field

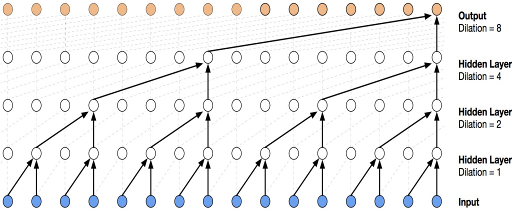
Inside WaveNet



Architecture

- CNN faster at training then RNN
- Additional layers/larger filters requried to keep receptive field
- Dilations keep computation expenses low while growing receptive field exponentially

Inside WaveNet



Architecture

- CNN faster at training then RNN
- Additional layers/larger filters requried to keep receptive field
- Dilations keep computation expenses low while growing receptive field exponentially
- Stacks of repeated 1,2,4,8..512 dilation steps

Inside WaveNet

Retrieving Outputs

Inside WaveNet

Retrieving Outputs

- Modelling a categorical distribution over all samples individually

Inside WaveNet

Retrieving Outputs

- Modelling a categorical distribution over all samples individually
- Normally used: Mixture Models
    - representing subpopulation of whole data
    - not used since inferring data's shape

Inside WaveNet

Retrieving Outputs

- Modelling a categorical distribution over all samples individually
- Normally used: Mixture Models
  - representing subpopulation of whole data
  - not used since inferring data's shape
- Fully Connected Layer: Softmax combined with μ law companding transformation

Inside WaveNet

Retrieving Outputs

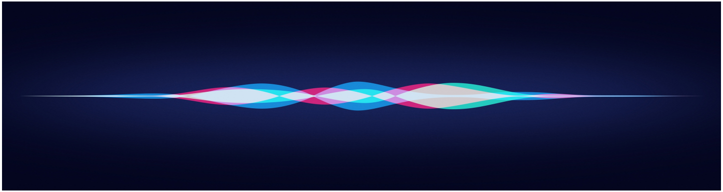
- Modelling a categorical distribution over all samples individually
- Normally used: Mixture Models
  - representing subpopulation of whole data
  - not used since inferring data's shape
- Fully Connected Layer: Softmax combined with μ law companding transformation
- Shrinks probabilities per 16 bit. sequence from 65k to 256

Inside WaveNet

Retrieving Outputs

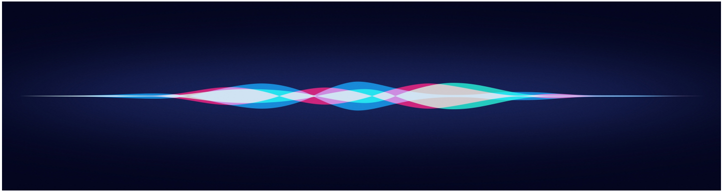
- Modelling a categorical distribution over all samples individually
- Normally used: Mixture Models
  - representing subpopulation of whole data
  - not used since inferring data's shape
- Fully Connected Layer: Softmax combined with μ law companding transformation
- Shrinks probabilities per 16 bit. sequence from 65k to 256
- Almost lossless reconstruction possible

Inside WaveNet



Features

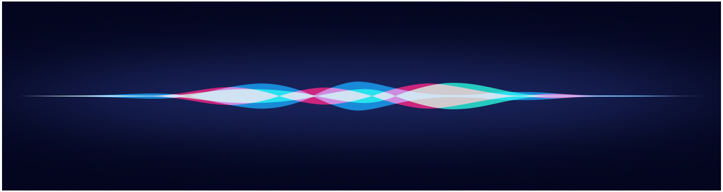
Inside WaveNet



Features

- Capable of training on local and global features

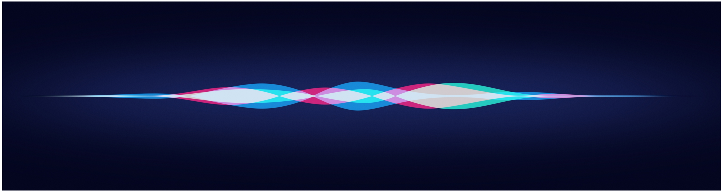
Inside WaveNet



Features

- Capable of training on local and global features
- Allows synthesis of samples with characteristics

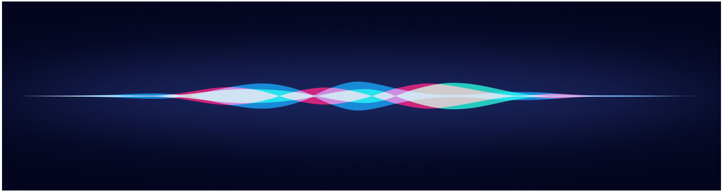
Inside WaveNet



Features

- Capable of training on local and global features
- Allows synthesis of samples with characteristics
- Global: influences all timestamps
  direct application to activation functio e.g.: Speaker’s identity

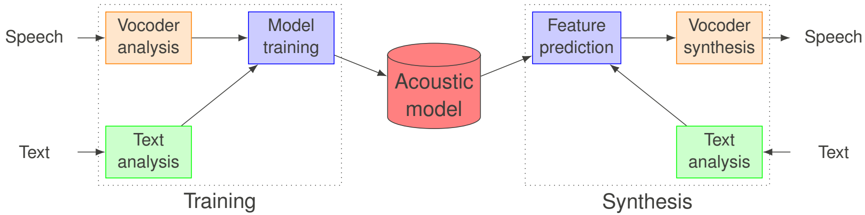
Inside WaveNet



Features

- Capable of training on local and global features
- Allows synthesis of samples with characteristics
- Global: influences all timestamps
  direct application to activation functio e.g.: Speaker's identity
- Local feature: limited timespan
  applied as transformed time series e.g.: emotion

Text To Speech



Vocoder (vocal encoder)...device for analyzing and synthesizing human voice signals

Performance

Experiment by Van den Oord, Dieleman, et al. (2016):
Subjective preference (%) in naturalness rated by paid native speakers

North American English

SPSS	Concat	WaveNet	No pref.
7.6		82.0	10.4
	20.1	49.3	30.6

Mandarin Chinese

SPSS	Concat	WaveNet	No pref.
12.5		29.3	58.2
	7.6	55.9	36.5

Summary

WaveNet

- is a new approach to model and synthesize natural speech.
- utilizes a Convolutional Neural Network to model temporal dependencies in speech.
- outperforms existing methods like statistical parametric speech synthesis.
- is computationally very expensive.