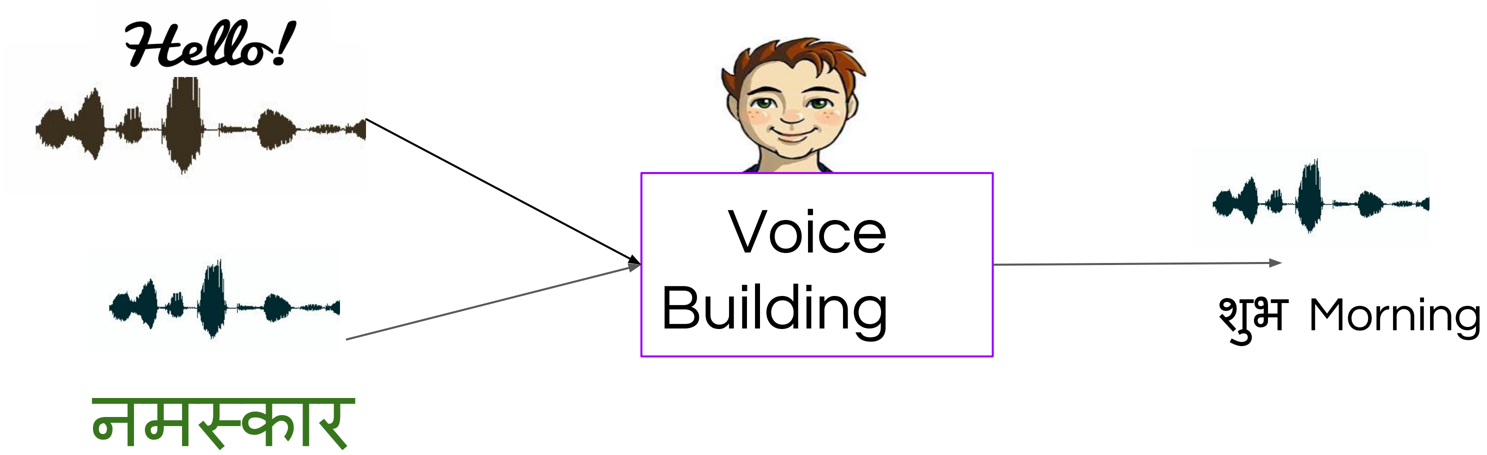
Overview of neural vocoding: Case study with WaveNet



Sai Krishna Rallabandi
Language Technologies Institute
Carnegie Mellon University

Agenda

Scope

Overview of WaveNet

- Formulation of WaveNet
- Residual Blocks
- DIlated Convolutions
- Expert Gating

Extensions to WaveNet

- Mixture Density based Loss
- Subsegmental Formulation
- Mode Sampling

Conclusion

Scope: Vocoding

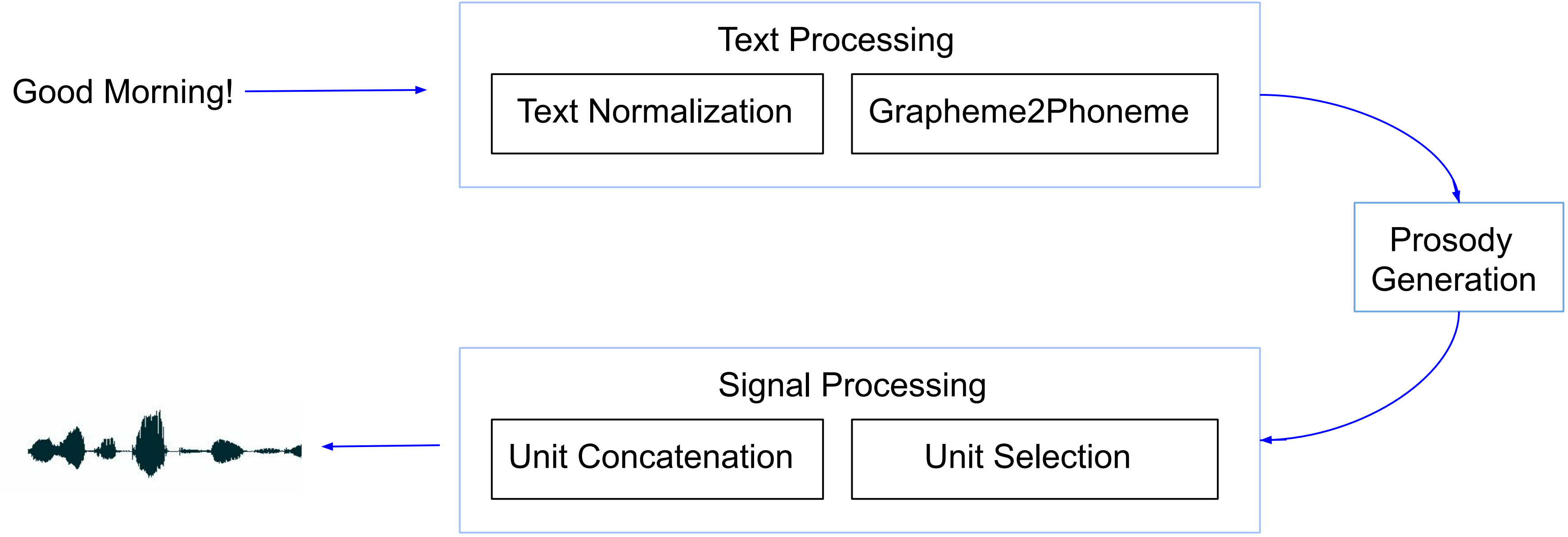


Fig: Overview of Unit Selection and Concatenation Speech Synthesis

Link about overview of Synthesis: http://cs.cmu.edu/~srallaba/Learn_Synthesis

Scope: Vocoding

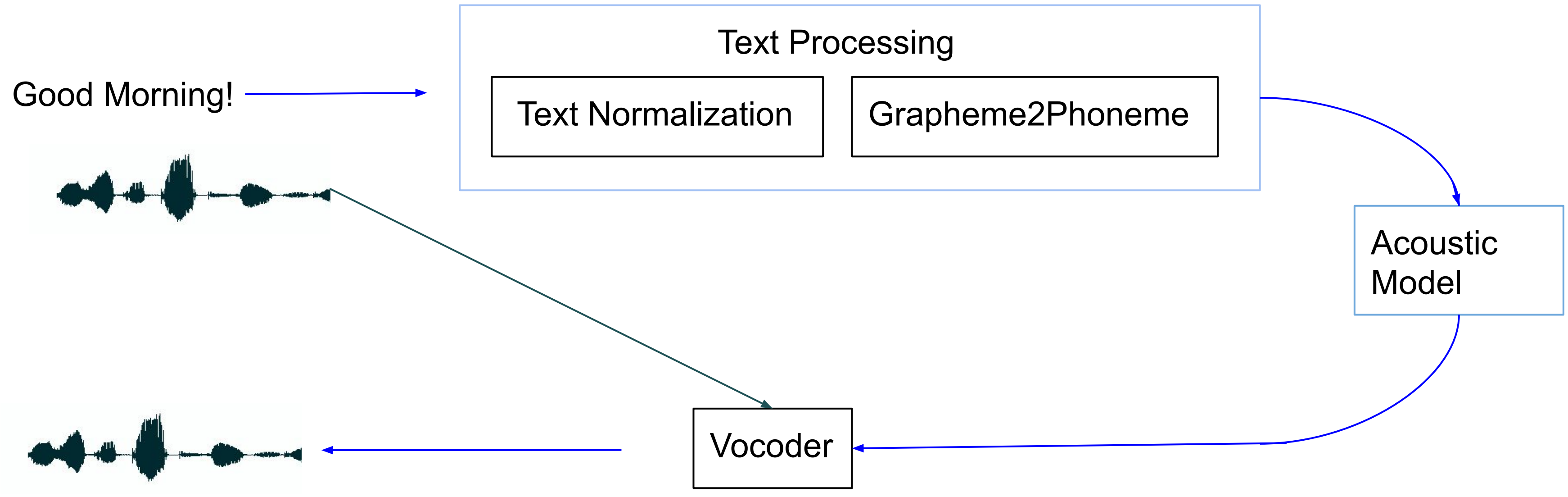


Fig: Overview of Statistical Parametric system for Speech Synthesis

Scope: Vocoding

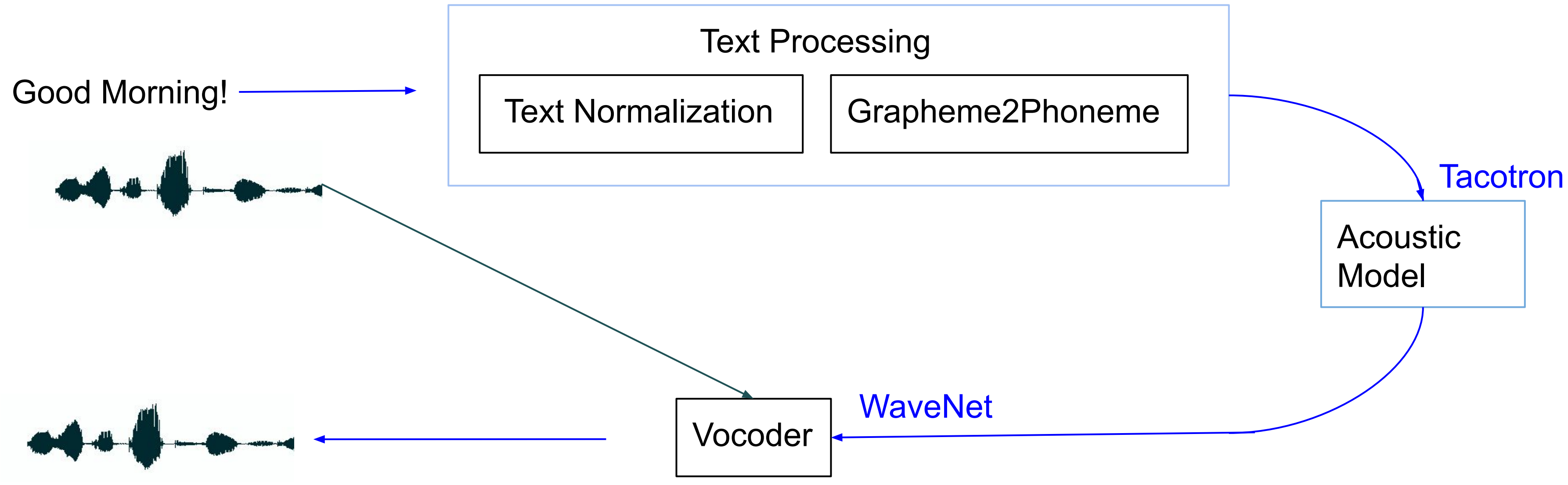


Fig: Overview of Statistical Parametric system for Speech Synthesis

Formulation of WaveNet - Probability of speech segments

- Let Ω_T denote the set of all possible sequences of length T over $\{0,1, \dots, d-1\}$.

Formulation of WaveNet - Probability of speech segments

- Let Ω_T denote the set of all possible sequences of length T.
- Let P: $\Omega_T \rightarrow [0,1]$ be a probability distribution which achieves higher values for speech sequences than for other sequences.

Formulation of WaveNet - Probability of speech segments

- Let Ω_T denote the set of all possible sequences of length T.

- Let P: $\Omega_T \rightarrow [0,1]$ be a probability distribution which achieves higher values for speech sequences than for other sequences.

- Knowledge of P enables us to test if a sequence $\{x_1 x_2 \cdots x_T\} \subset$ speech.

- Also, it allows us sample from this distribution and generate sequences that with high probability look like speech.

Formulation of WaveNet - Probability of speech segments

- Let Ω_T denote the set of all possible sequences of length T.
- Let P: $\Omega_T \rightarrow [0,1]$ be a probability distribution which achieves higher values for speech sequences than for other sequences.
- Knowledge of P enables us to test if a sequence $\{x_1 x_2 \cdots x_T\} \subset$ speech.
- Also, it allows us sample from this distribution and generate sequences that with high probability look like speech.
- But T needs to be large enough to apply this.
- As T increases, P becomes smaller and smaller.

Formulation of WaveNet - Probability of speech segments

- Let Ω_T denote the set of all possible sequences of length T.

- Let P: $\Omega_T \rightarrow [0,1]$ be a probability distribution which achieves higher values for speech sequences than for other sequences.

- Knowledge of P enables us to test if a sequence $\{x_1 x_2 \cdots x_T\} \subset$ speech.

- Also, it allows us sample from this distribution and generate sequences that with high probability look like speech.

- But T needs to be large enough to apply this.

- As T increases, P becomes smaller and smaller.

- Use conditional distribution $P(x_t \mid x_1, \ldots, x_{t-1})$

Formulation of WaveNet - The conditional probability

- The conditional probability $P(x_t \mid x_1 , \ldots , x_{t-1})$ is modelled with a categorical distribution where x_t falls into one of a number of bins (usually 256).

- WaveNet uses causal dilated convolutions to model this conditional probability on quantized raw audio.

- Raw audio is transformed to $<x_1,x_2..x_T>$ using mu-law transformation. $[-1 < x_T < 1]$

- x_t is quantized into 256 bins.

- x_t is one hot encoded.

$-2.2, -1.43, -0.77, -1.13, -0.58, -0.43, -0.67, \ldots$

signal

→

$-0.7, -0.3, 0.2, -0.1, 0.4, 0.6, 0.3, \ldots$

μ-law transformed

→

$0, 1, 2, 1, 2, 3, 2, \ldots$

quantized into 4 bins

→

bin 0	1	0	0	0	0	0	0
bin 1	0	1	0	1	0	0	0
bin 2	0	0	1	0	1	0	1
bin 3	0	0	0	0	0	1	0

...

one-hot vectors

→ ***Input to WaveNet***

Formulation of WaveNet - Dilated Convolutions

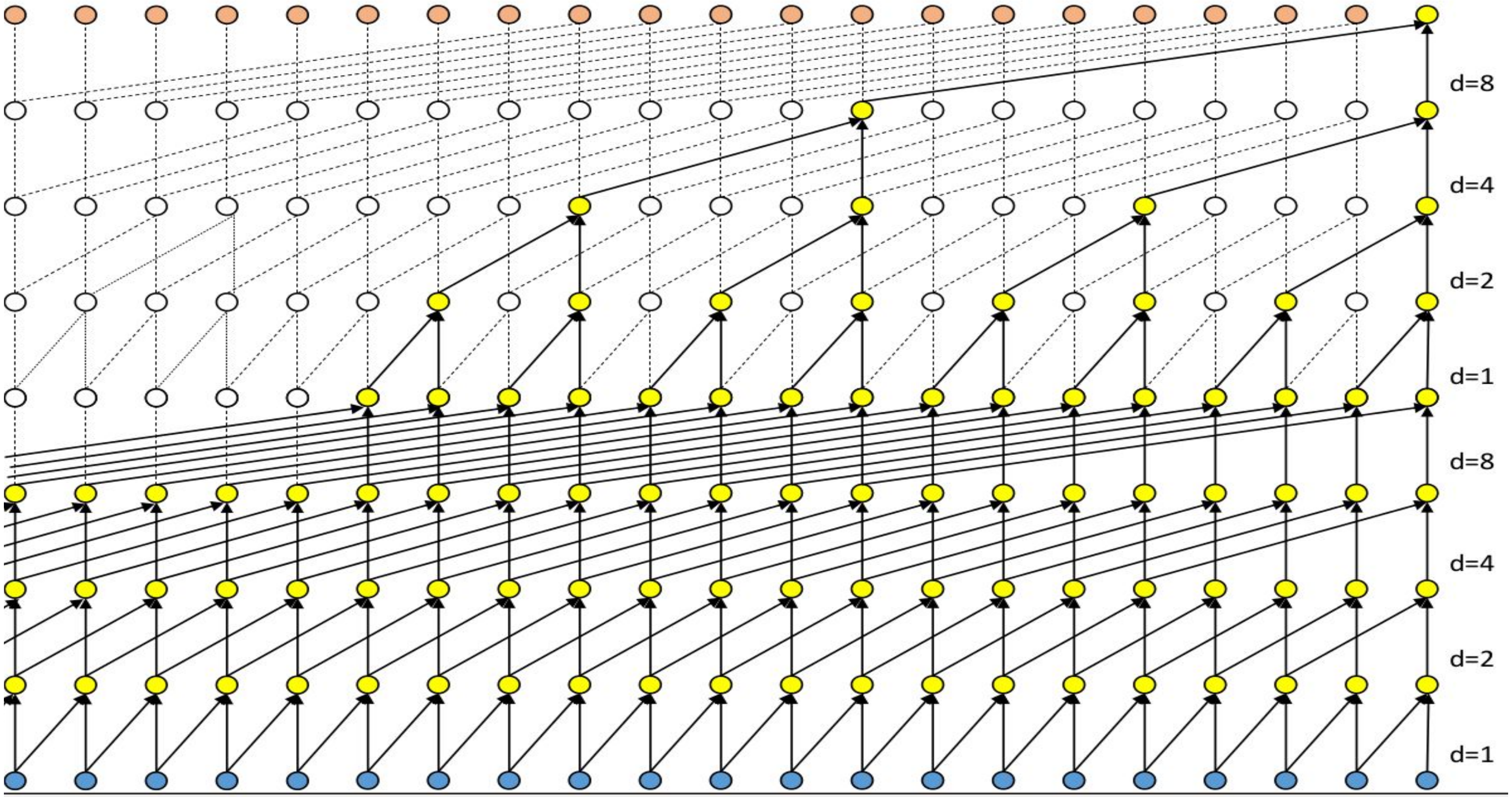


Fig: Dilated Stack with dilations 1,2,4,8,1,2,4,8

Formulation of WaveNet - Residual Connections

- WaveNet has 30 layers of dilated convolutions.

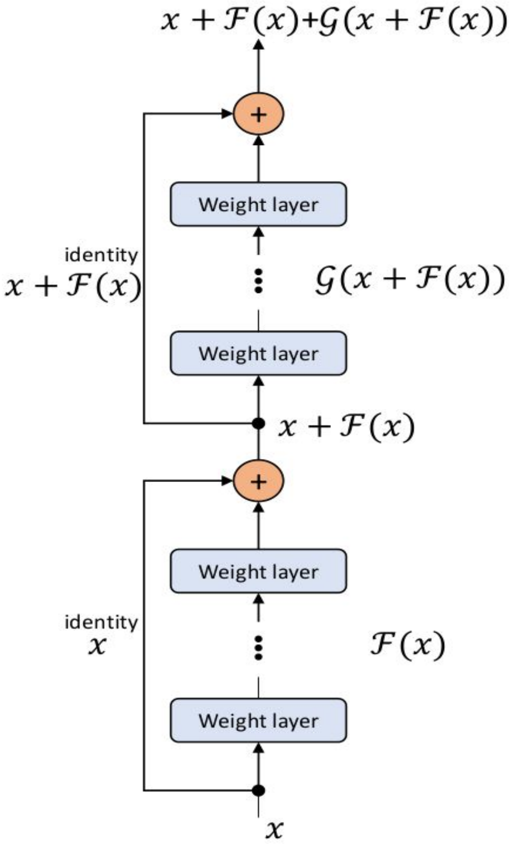
- Idea: Reformulate the mapping function $x \rightarrow f(x)$ between layers from $f(x) = F(x)$ to $f(x) = x + F(x)$.

- The residual networks have identity mappings, x, as skip connections and inter-block activations F(x).

- The residual F(x) can be easily learnt

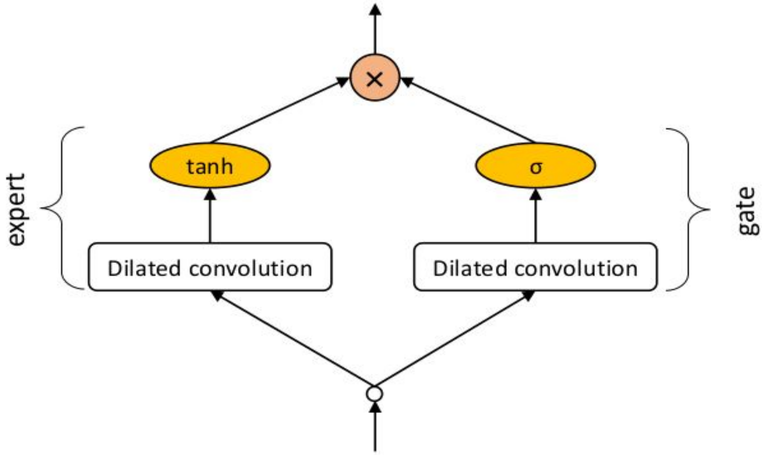
- Forward and backward signals can be directly propagated between any two blocks.

- Avoiding vanishing gradient problem



Formulation of WaveNet - Experts and Gates

- Different parts in input space might need different expertise.
- Idea: Define an expert per output channel.
- Contribution of each expert is controlled by a gating mechanism.
- The components of the output vector are mixed in higher layers, creating mixture of experts.



Formulation of WaveNet - Audio Generation

- After training, the network is sampled to generate synthetic utterances.

- At each step during sampling a value is drawn from the probability distribution computed by the network.

- This value is then fed back into the input and a new prediction for the next step is made.

Example with receptive field 3 and 4 quantization channels

Input: x_1, x_2, x_3

Output: $p_4 = Wavenet(x_1, x_2, x_3) = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.1 \end{bmatrix}$ Probability distribution over the symbols 0,1,2,3

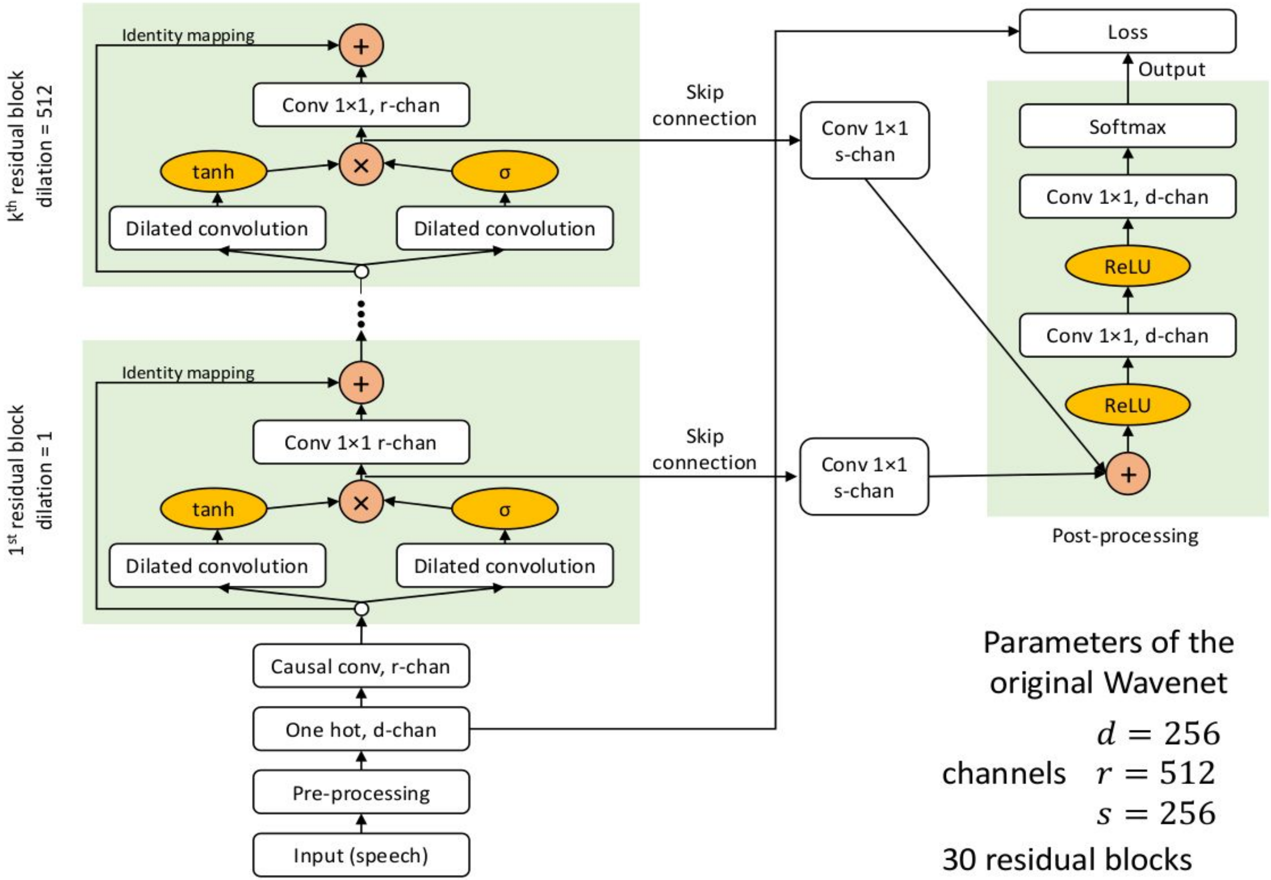
sample: $x_4 = 1$

Input: x_2, x_3, x_4

Output: $p_5 = Wavenet(x_2, x_3, x_4) = \begin{bmatrix} 0.7 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$

sample: $x_5 = 0$

Formulation of WaveNet - Basic Architecture



Parameters of the original Wavenet

$d = 256$

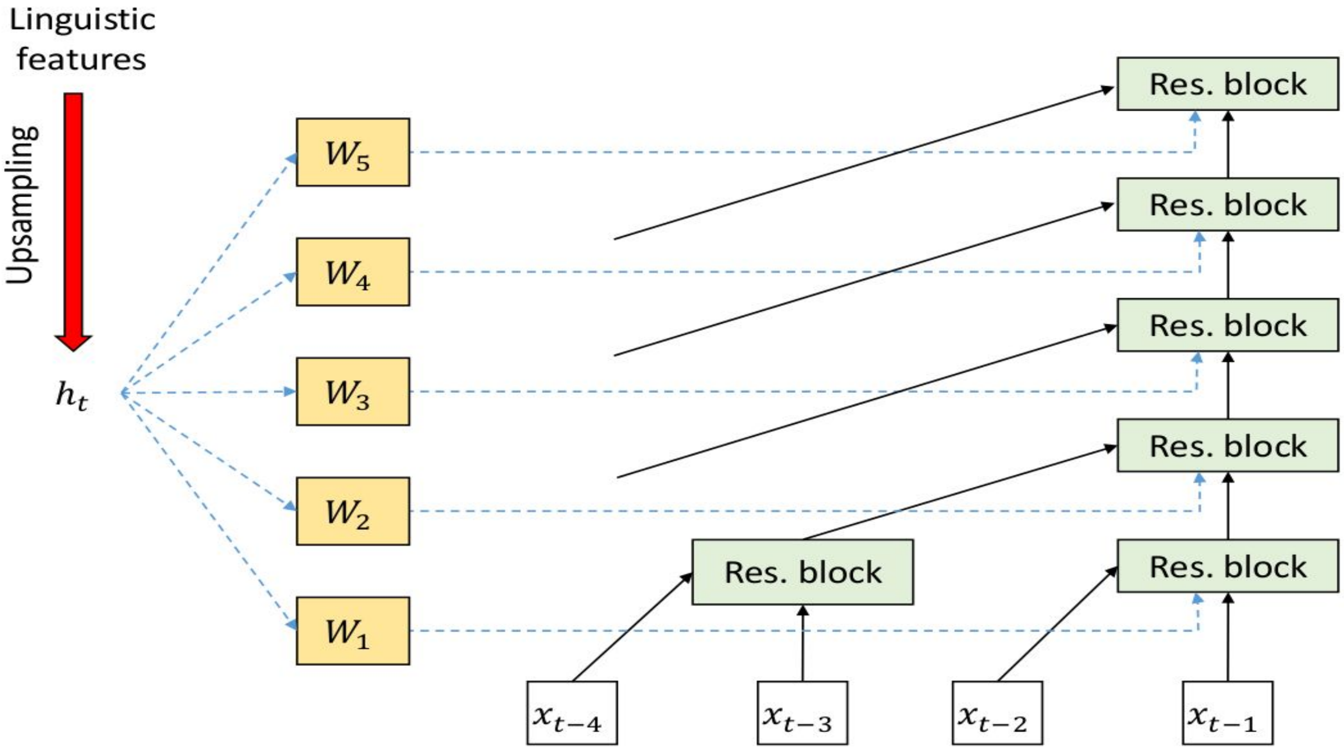
channels $r = 512$

$s = 256$

30 residual blocks

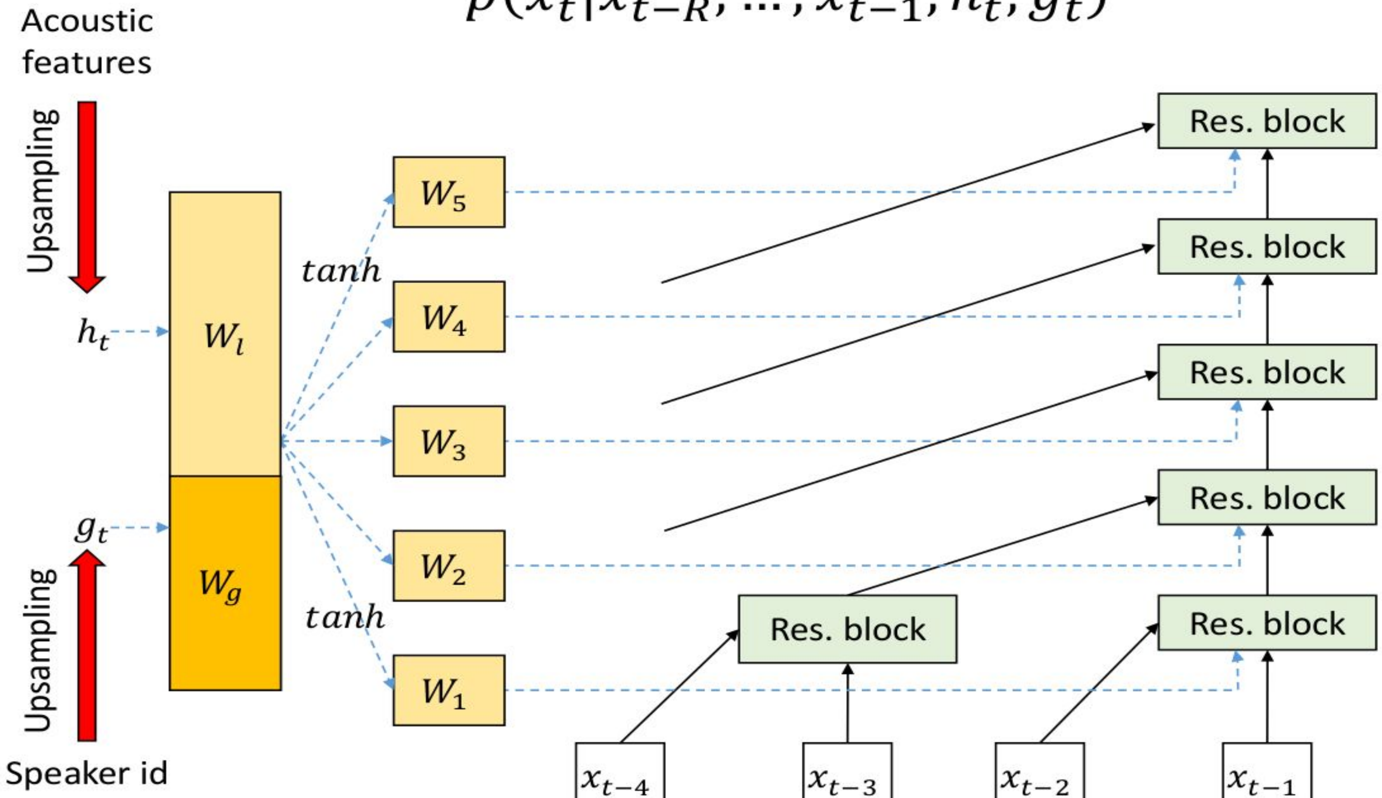
Formulation of WaveNet - Local Conditioning

$$p(x_t | x_{t-R}, \dots, x_{t-1}, h_t)$$



Formulation of WaveNet - Global Conditioning

$$p(x_t | x_{t-R}, \ldots, x_{t-1}, h_t, g_t)$$



WaveNet - Few Points

- The number of dilated modules should be ≥ 40.
- Models trained with 48 kHz speech produce higher quality audio than models trained with 16 kHz speech.
- The model need more than 300000 iterations to converge.
- The speech quality is strongly affected by the up-sampling method of the linguistic labels.
- The Adam optimization algorithm is a good choice.
- Conditioning: pentaphones + stress + continuous F0 + VUV
- If overtrained, can generate white noise.

WaveNet: http://tts.speech.cs.cmu.edu/rsk/tts_stuff/Blizzard_2018/experiments/vocoder/wavenet/test_samples/

WaveNet - Questions

- Can we improve training?

  Speed

  Time

  Data Requirement

- Can we make the model more stable?
- Can we incorporate some speech knowledge? .

Expts in Neural Vocoding - Subsegmental Formulations

- Knowledge of P enables us to test if a sequence $\{x_1 x_2 \cdots x_T\} \subset$ speech.
- Speech has long term and short term dependencies.
- WaveNet provides high fidelity.
- The building blocks of WaveNet: Dilations, Residual blocks, gating mechanism.
- In practise, the receptive field used : 500 msec
- Question: Can we provide some long term info and reduce the model complexity?
- Can we model just short term dependencies and provide long term as side information? [We any way provide spectral information]

Juvela, Lauri, et al. "Speech waveform synthesis from MFCC sequences with generative adversarial networks." *arXiv preprint arXiv:1804.00920* (2018)
http://tts.speech.cs.cmu.edu/rsk/tts_stuff/kitchen/segmental-wavenet-experiments/conditional_formulation/20August/

Expts in Neural Vocoding - Mixture Density based Loss function

- 256 way softmax makes the gradients with respect to network parameters sparse, especially early in the training: 127 is equidistant from 128 and 252

Expts in Neural Vocoding - Mixture Density based Loss function

- 256 way softmax makes the gradients with respect to network parameters sparse, especially early in the training: 127 is equidistant from 128 and 252

- If we incorporate the information that speech sounds are a continuum (which is true!) it might help the model.

Expts in Neural Vocoding - Mixture Density based Loss function

- 256 way softmax makes the gradients with respect to network parameters sparse, especially early in the training: 127 is equidistant from 128 and 252

- If we incorporate the information that speech sounds are a continuum (which is true!) it might help the model.

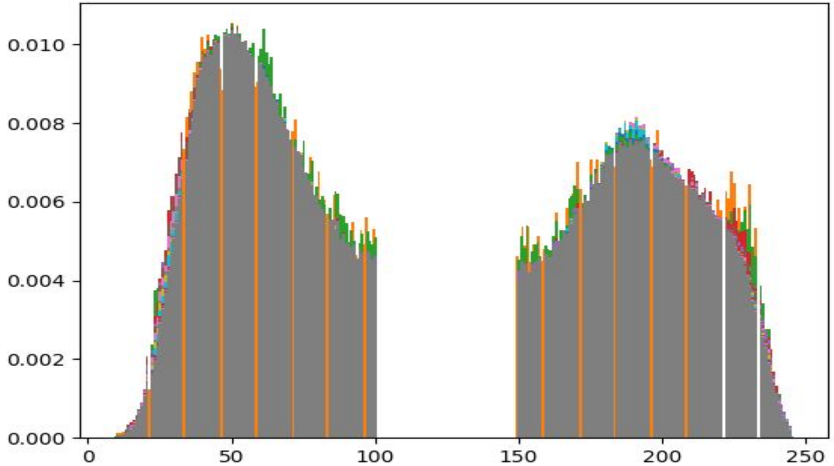


Fig: Hist plot of individual bins from natural speech

Juvela, Lauri, et al. "Speaker-independent raw waveform model for glottal excitation." *arXiv preprint arXiv:1804.09593*(2018).
Salimans, Tim, et al. "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications." *arXiv preprint arXiv:1701.05517* (2017).

Expts in Neural Vocoding - Investigating Sampling

- WaveNet uses random sampling to generate synthetic speech

Expts in Neural Vocoding - Investigating Sampling

- WaveNet uses random sampling to generate synthetic speech
- Each time WaveNet is used for inference, we obtain a different waveform. (But they sound exactly same!)

Expts in Neural Vocoding - Investigating Sampling

- WaveNet uses random sampling to generate synthetic speech

- Each time WaveNet is used for inference, we obtain a different waveform. (But they sound exactly same!)

- Observation: Running Kurtosis of the generated samples is always < 10.

Expts in Neural Vocoding - Investigating Sampling

- WaveNet uses random sampling to generate synthetic speech
- Each time WaveNet is used for inference, we obtain a different waveform. (But they sound exactly same!)
- Observation: Running Kurtosis of the generated samples is always < 10.
- Might be better to use Mode based sampling in voiced regions

Wang, Xin, et al. "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis." *arXiv preprint arXiv:1804.02549* (2018).

Expts in Neural Vocoding - Investigating Sampling

- WaveNet uses random sampling to generate synthetic speech
- Each time WaveNet is used for inference, we obtain a different waveform. (But they sound exactly same!)
- Observation: Running Kurtosis of the generated samples is always < 10.
- Might be better to use Mode based sampling in voiced regions
- Temperature sampling: Sample randomly from a distribution adjusted by a temperature θ.
- Top k: Sample from an adjusted distribution that only permits the top k samples

Wang, Xin, et al. "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis." *arXiv preprint arXiv:1804.02549* (2018).

Expts by others

- Speaker-dependent WaveNet vocoder. [Interspeech 2017]
- Multi-task WaveNet: A Multi-task Generative Model for Statistical Parametric Speech Synthesis without Fundamental Frequency Conditions [Interspeech 2018]
- Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model [Interspeech 2018]
- WaveNet Vocoder with Limited Training Data for Voice Conversion [Interspeech 2018]
- Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder [Interspeech 2018]
- High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder [Interspeech 2018]

Expts by others: Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model [Interspeech 2018]

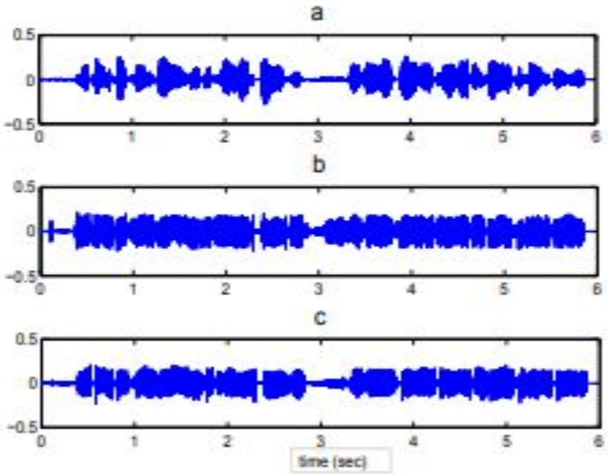


Fig: (a) Original (b) SSDRC c) wSSDRC

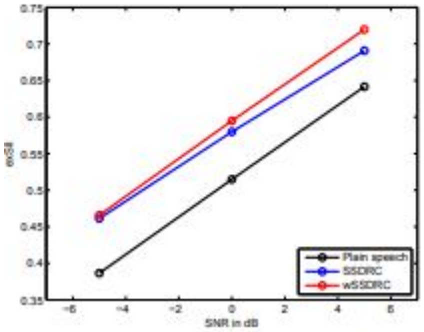
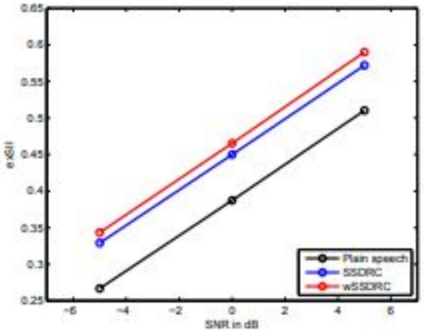


Fig: Speech Shaped Noise



Model	PT score
SSDRC	47.3%
wSSDRC	**52.7%**

Fig: Preference Test

Fig: Stationary White Noise

THANK YOU