Adapting Wavenet for Speech Enhancement

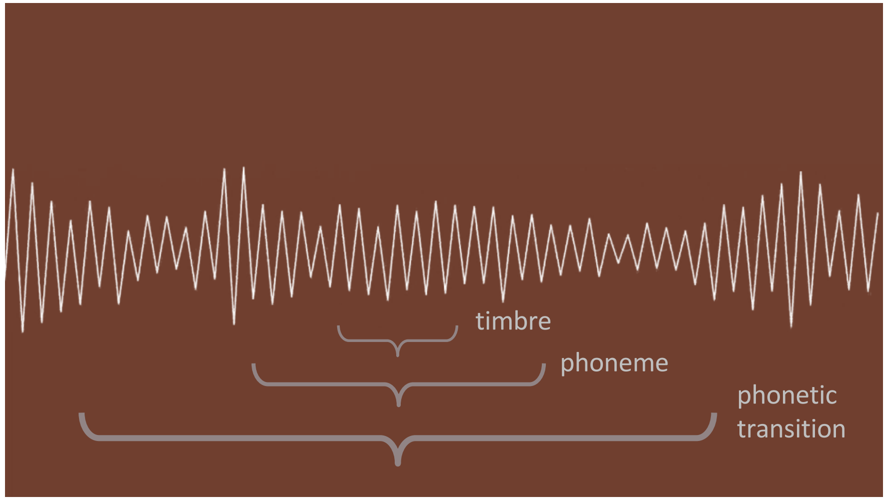
DARIO RETHAGE | JULY 12, 2017

I am

- Master Student
- 6 months @ Music Technology Group, Universitat Pompeu Fabra
- Deep learning for acoustic source separation
- With Jordi Pons, Audio Signal Processing Lab

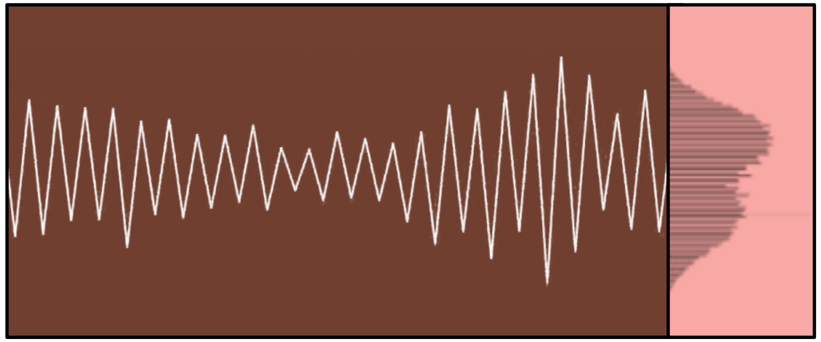
Learning from raw audio

- High dimensionality
- Many levels of structure
- No hand crafted feature extraction
- No discarding of information (phase)
- Until recently computationally intractable



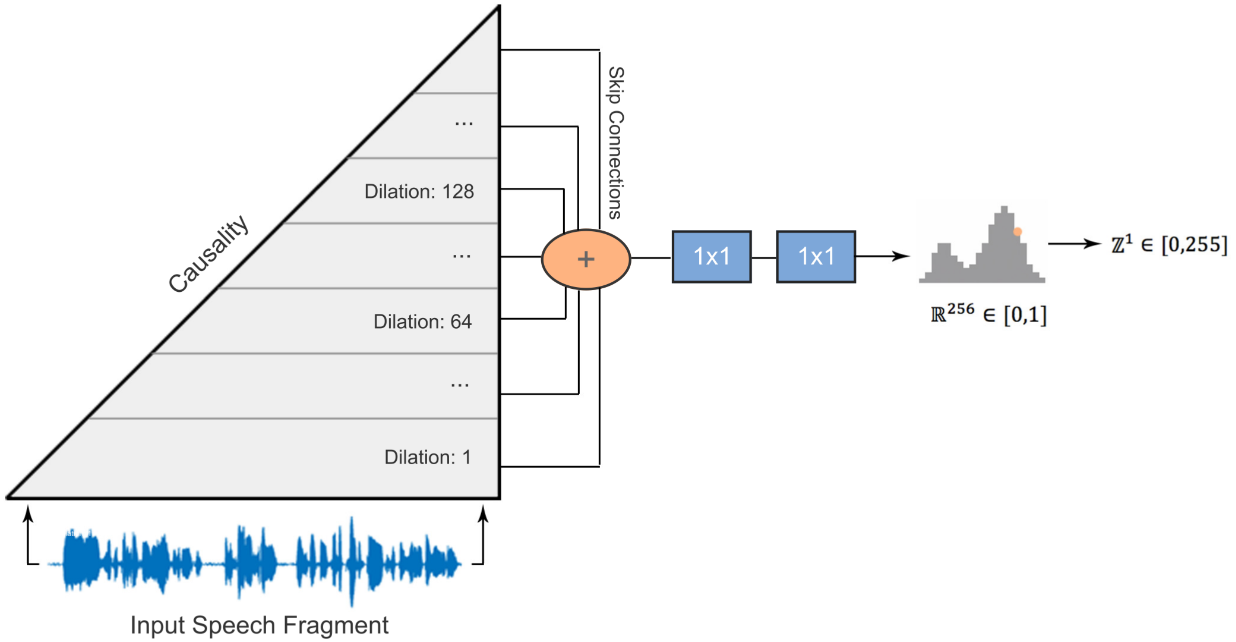
Wavenet: A Generative Model for Raw Audio

- Speech synthesis on waveform level using auto-regressive, generative model
- Generates 8-bit (256 values) probability distribution
- Sample output distribution (probabilistic task)
- Considerable parameter savings
  - Small filters
  - Large dilations
- 16kHz sampling rate (wide-band)
- Very slow
- Not strictly end-to-end



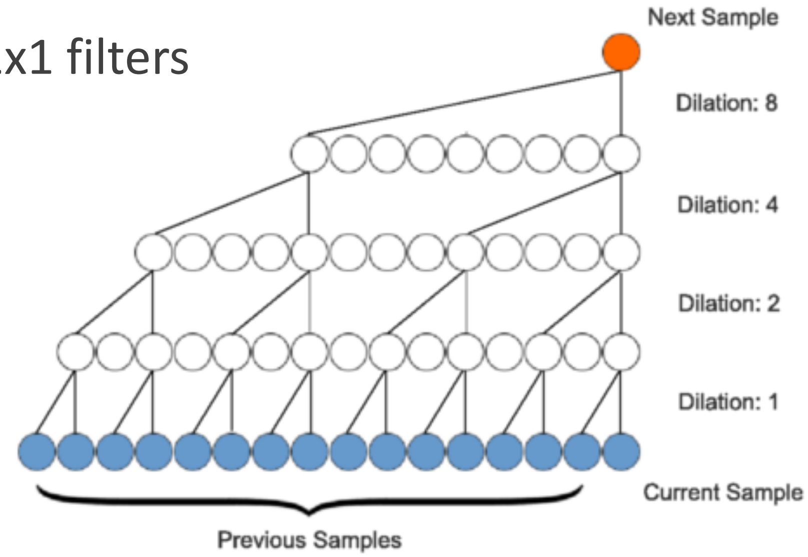
Wavenet: Key Features

- Causality
- Gated Units
- Softmax Output
- µ-law Quantization
- Dilation
- Stacks



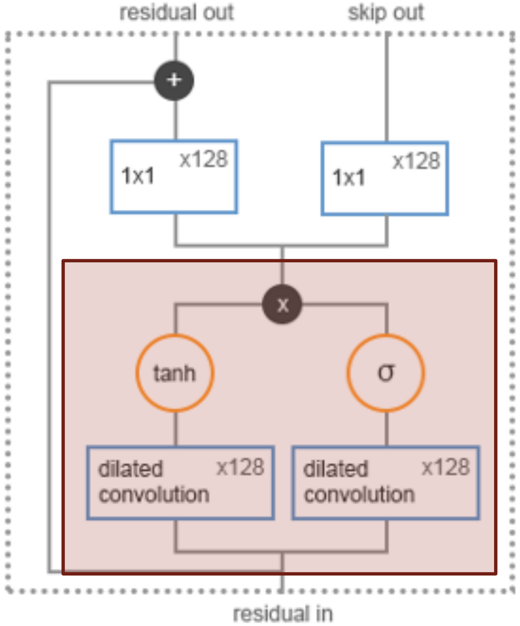
Causality

- Only previous and current sample inform prediction of sample t + 1
- Asymmetric padding
- 2x1 filters



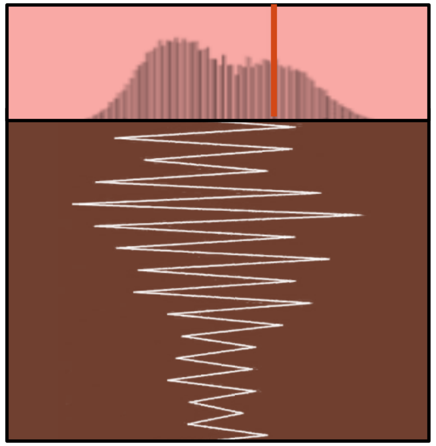
Gated Units

- Control contribution of each layer



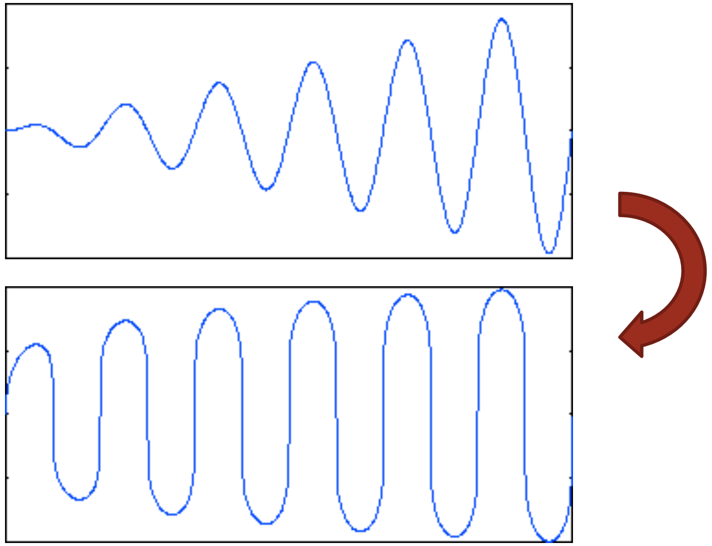
Softmax

- No assumptions about output distribution
- Well suited for multi-modal distributions
- Requires discretization of output



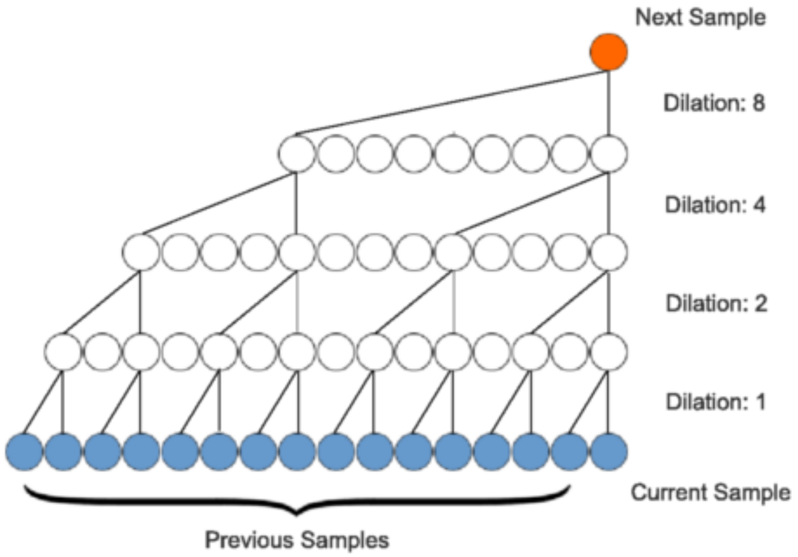
μ-law quantization

- Non-linear companding
- Better use of 8-bit quantization space



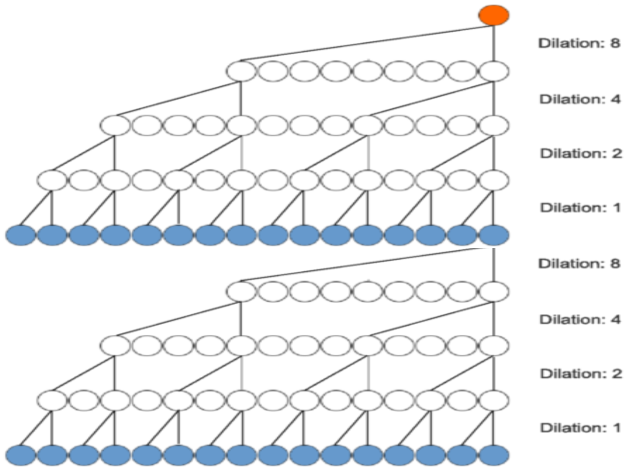
Dilation

- Larger receptive field, same parameters
- By powers of 2



Stacks

- Repeat dilation pattern
- More depth, less width



Wavenet: Reimplementation

- Many open questions
  - Filter Depths
  - Number of Layers
- Trained on VCTK, 109 native speakers of English, good phonetic coverage
- Proof of concept
- ~600k parameters

Speech Enhancement

- Within acoustic source separation
- Deterministic
- Goal: Improve intelligibility and/or overall perceptual quality of speech signal
- Until recently, greatest successes in the frequency domain
  - e.g. estimating spectral mask

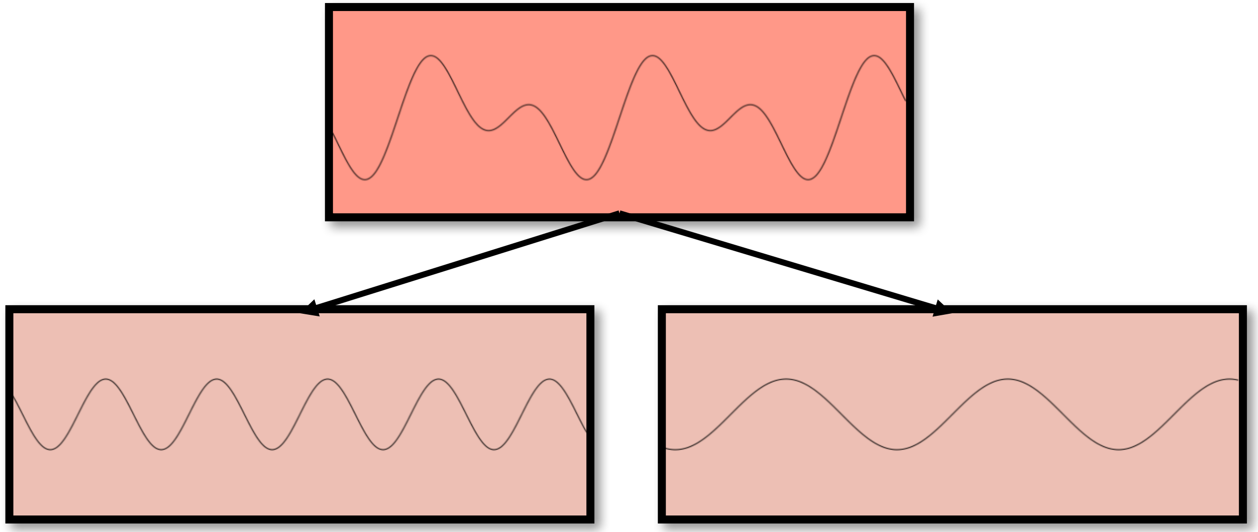
$$m_t = s_t + b_t$$

m: mixture
s: speech
b: background

Either estimate $\hat{\boldsymbol{s}}$ given $\boldsymbol{m}$ directly or $\widehat{\boldsymbol{b}}$ given $\boldsymbol{m}$, since $\boldsymbol{s} = \boldsymbol{m} - \boldsymbol{b}$

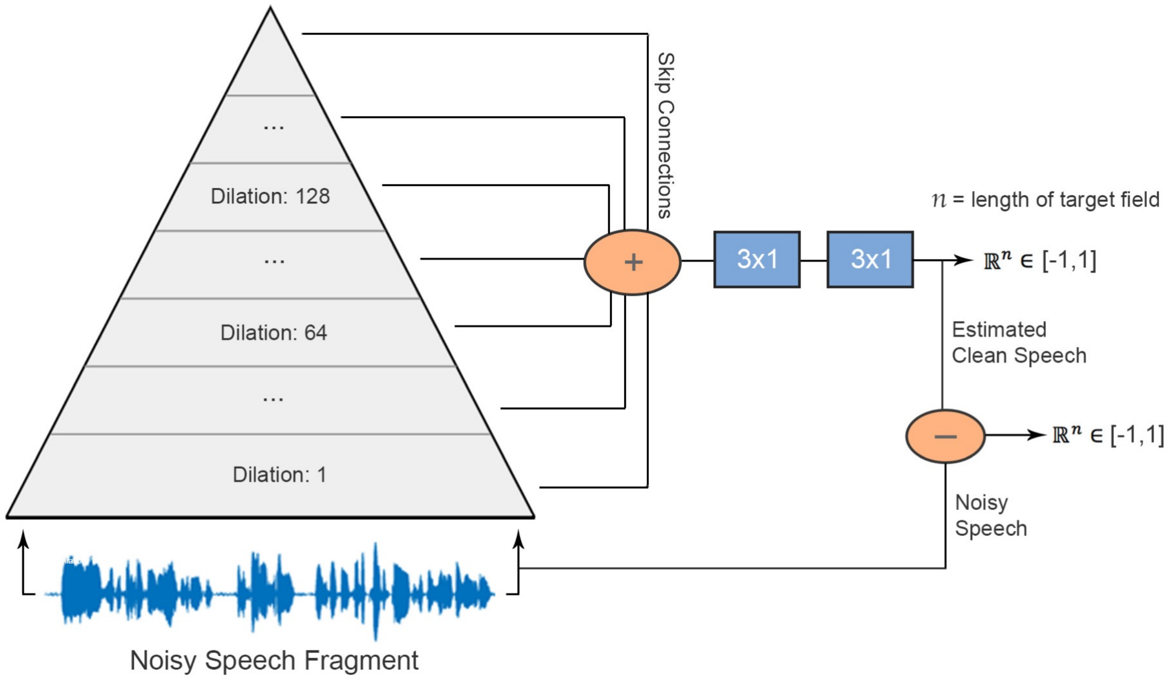
A Wavenet For Source Separation

- Generic architecture, suitable for any acoustic source separation
- Blind two-source separation
- Discriminative
- End-to-end
  - Time-domain input/output
  - No pre/post-filtering
  - No quantization
- 16kHz sampling rate (wide-band)
- Flexible



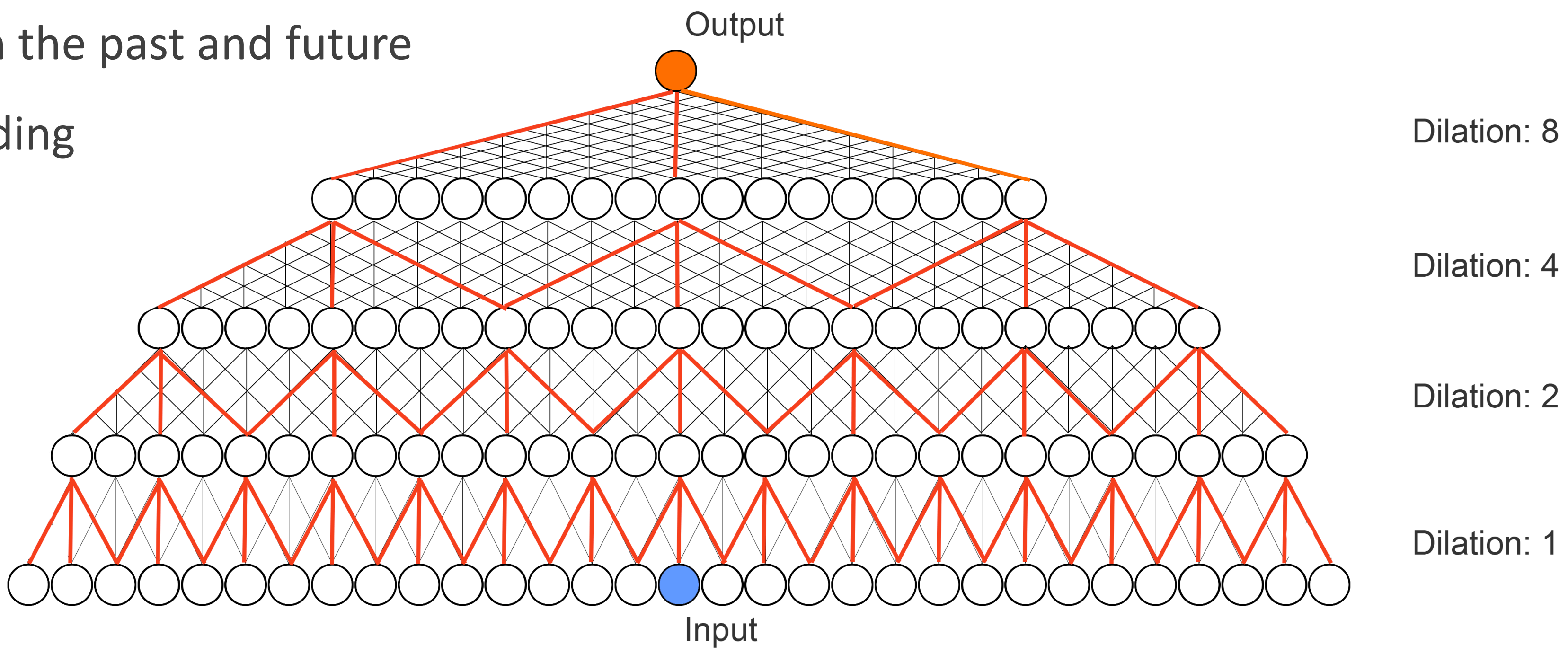
Key Contributions

- Non-causality
- Real-valued predictions
- Non-autoregressive
- Target fields
- Enforces time continuity
- Energy-conserving loss



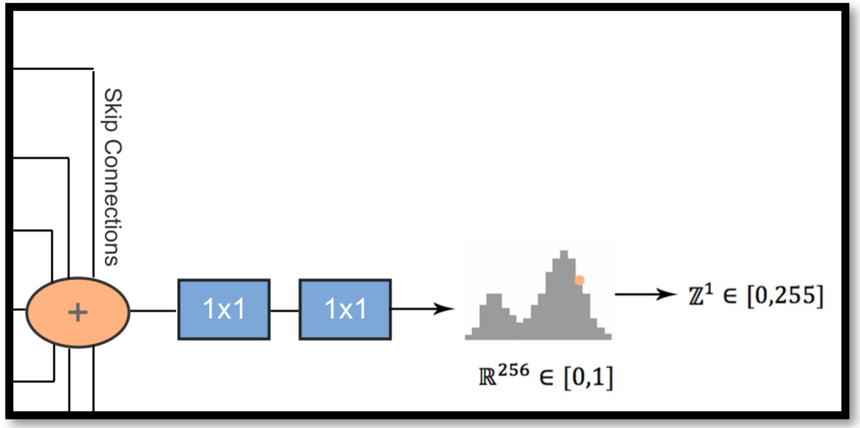
Non-causality

- Equal context in the past and future
- Symmetric padding
- 3x1 filters

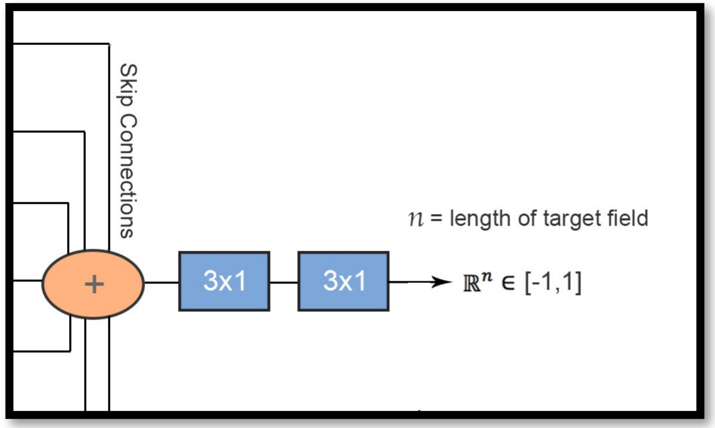


Real-valued Predictions

- Assumes Gaussian output distribution
- No quantization error
- One output unit per output sample
- µ-law companding disadvantageous

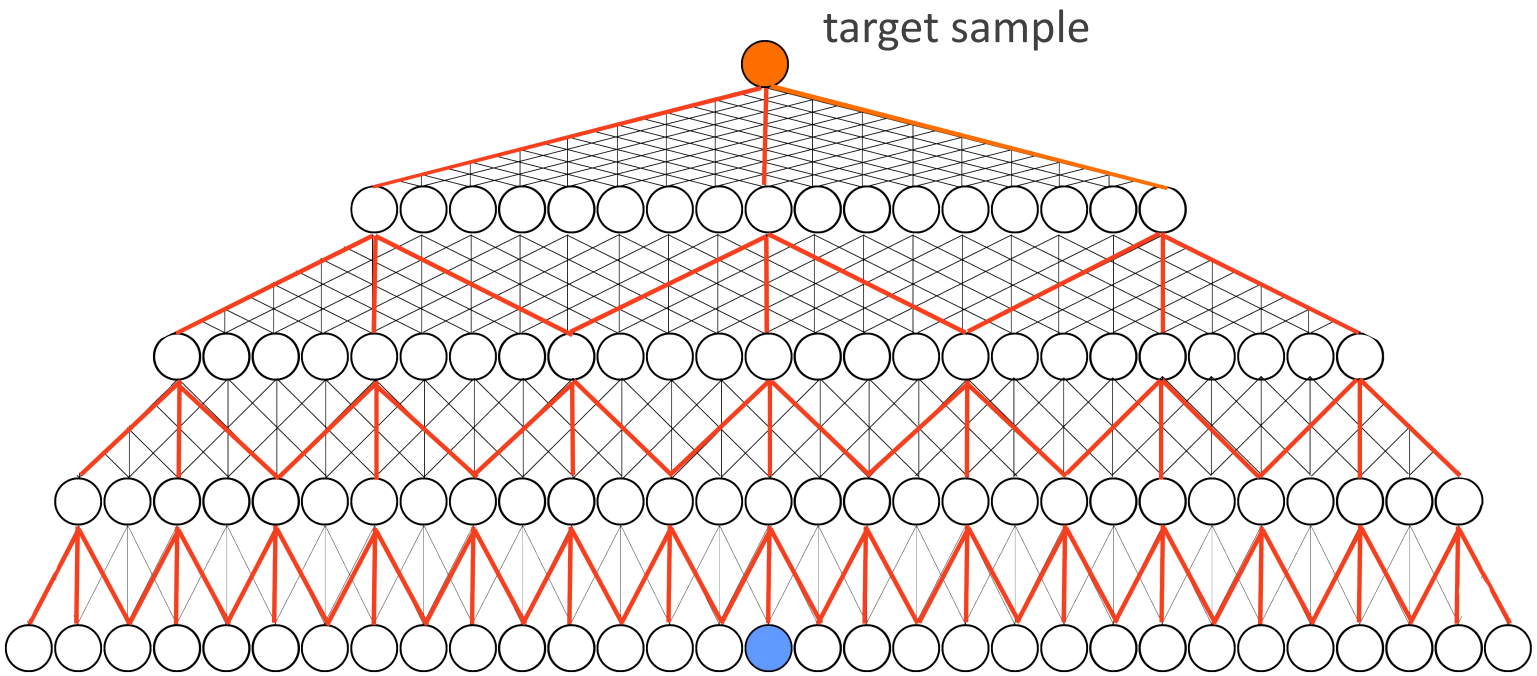


Wavenet

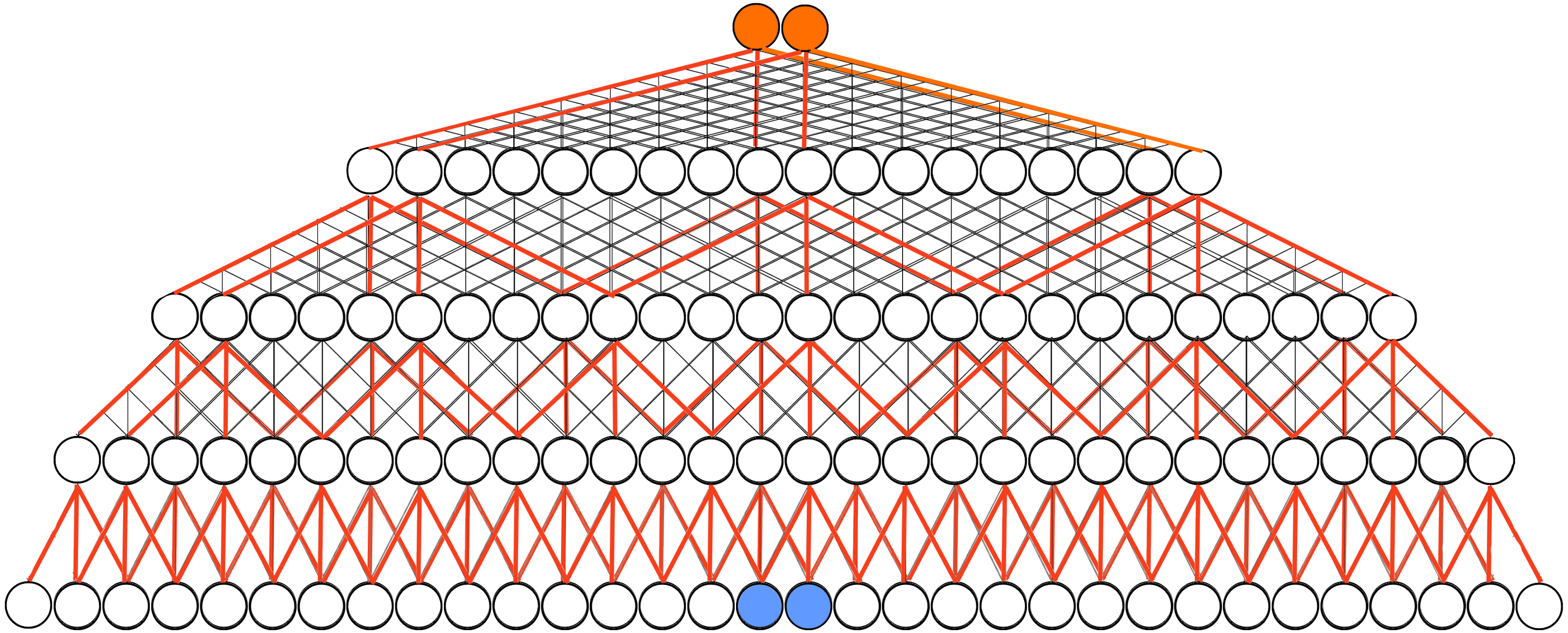


Proposed Model

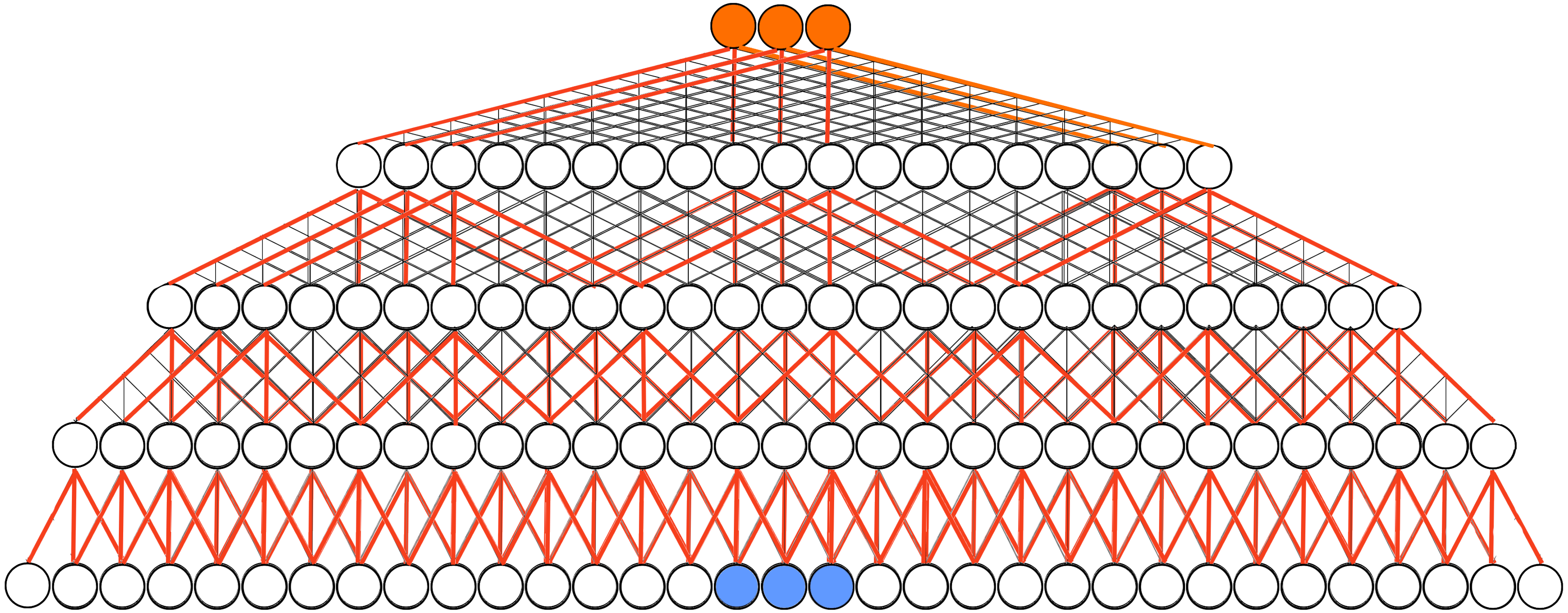
Target Fields



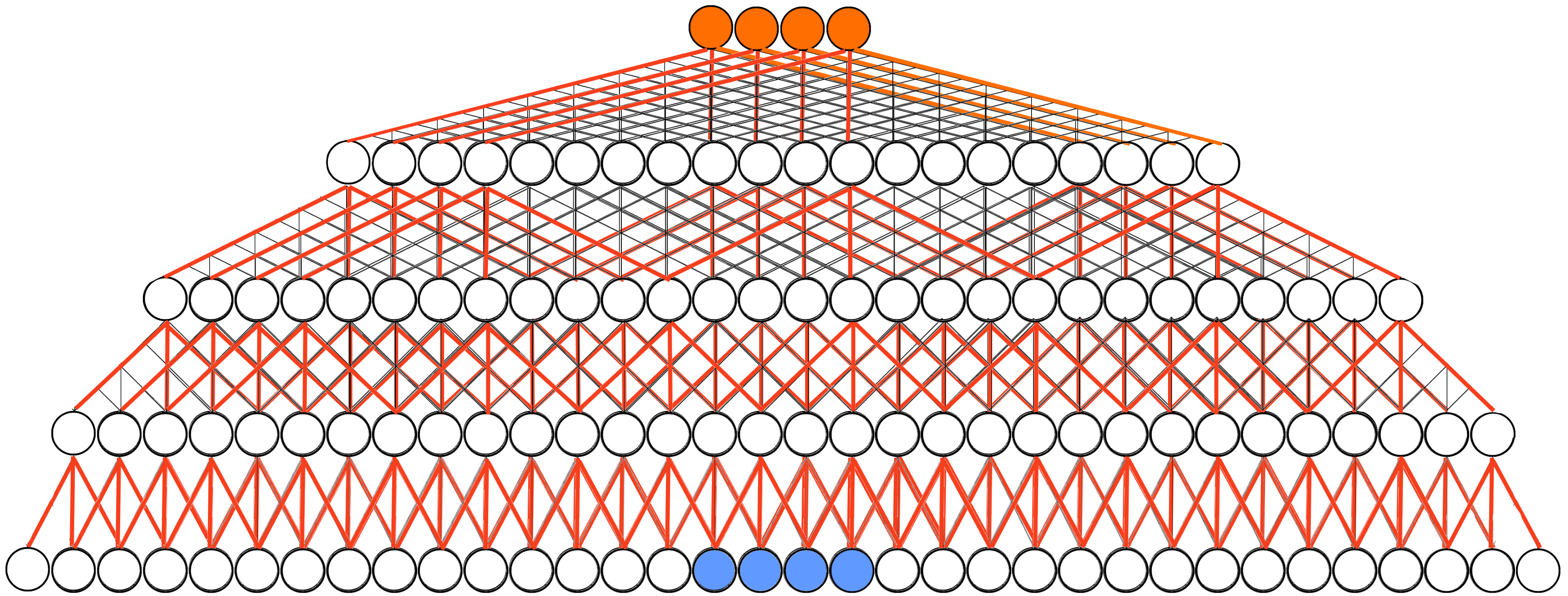
Target Fields



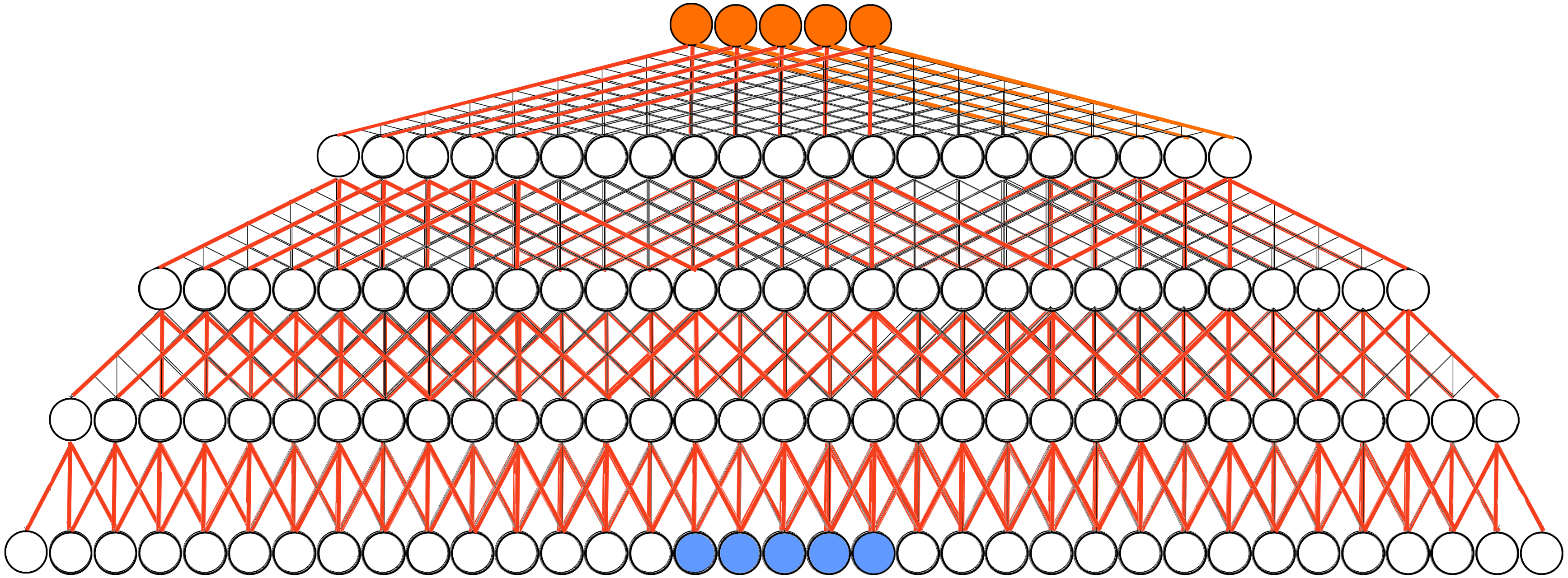
Target Fields



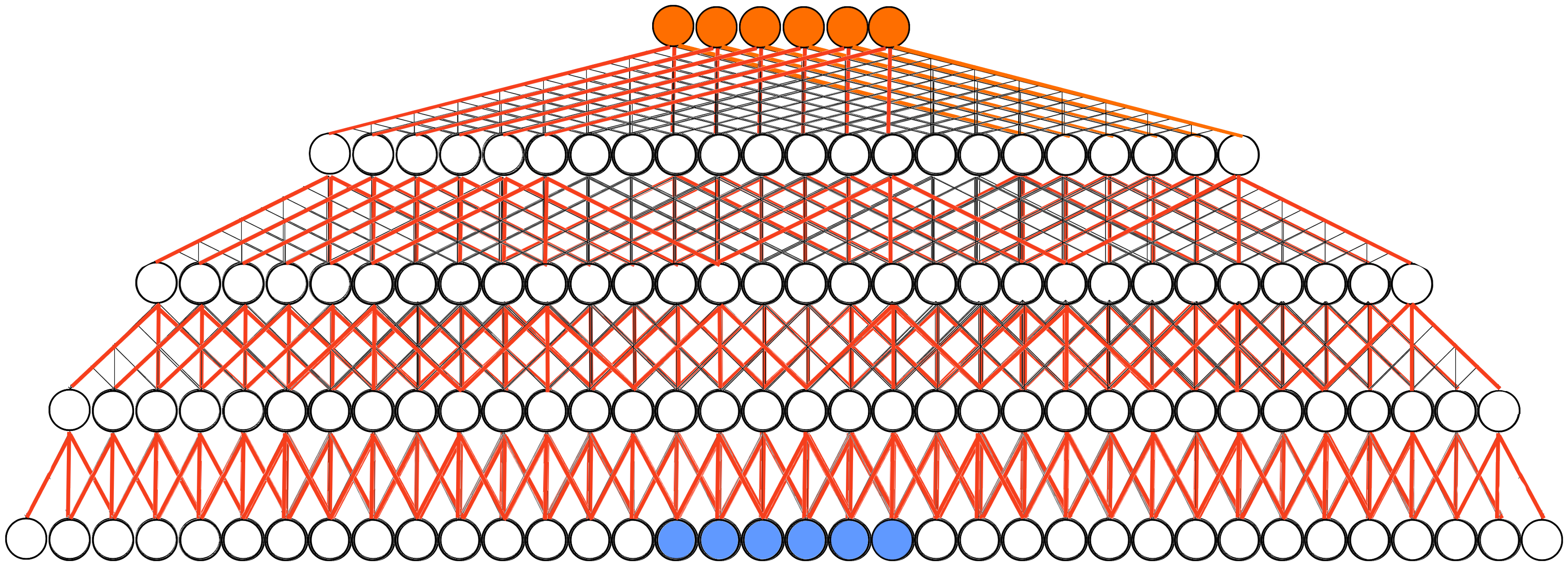
Target Fields



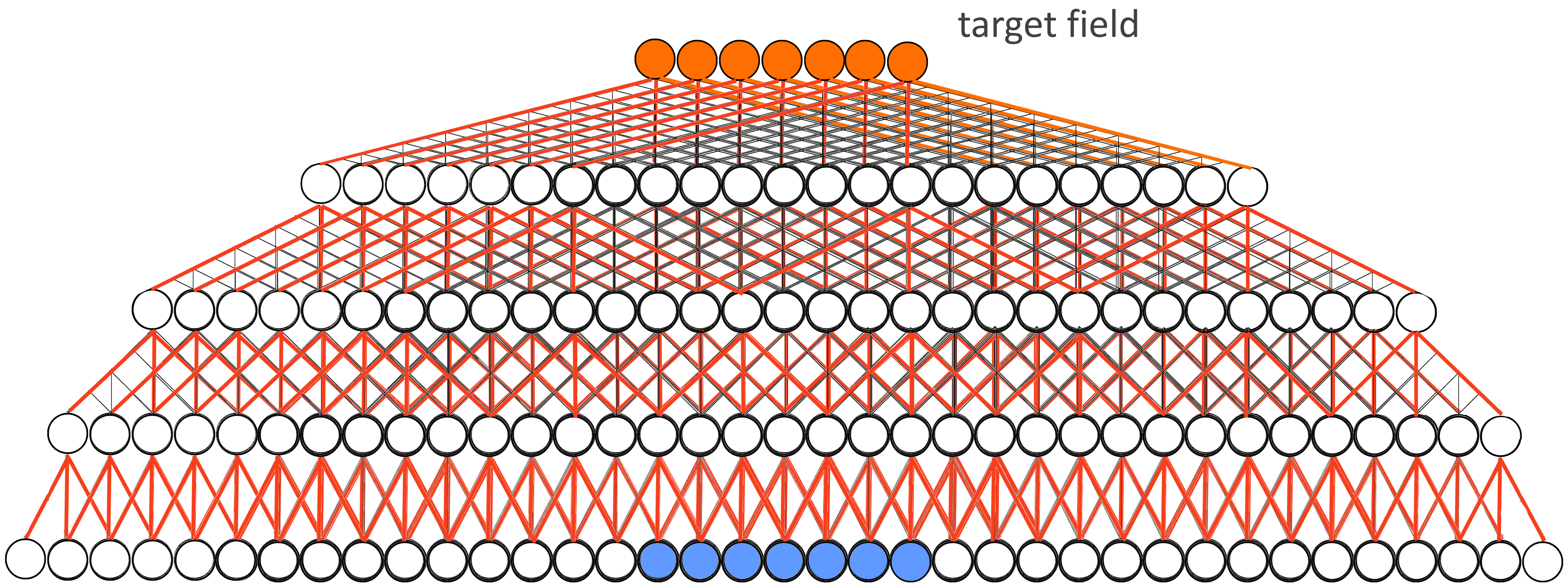
Target Fields



Target Fields

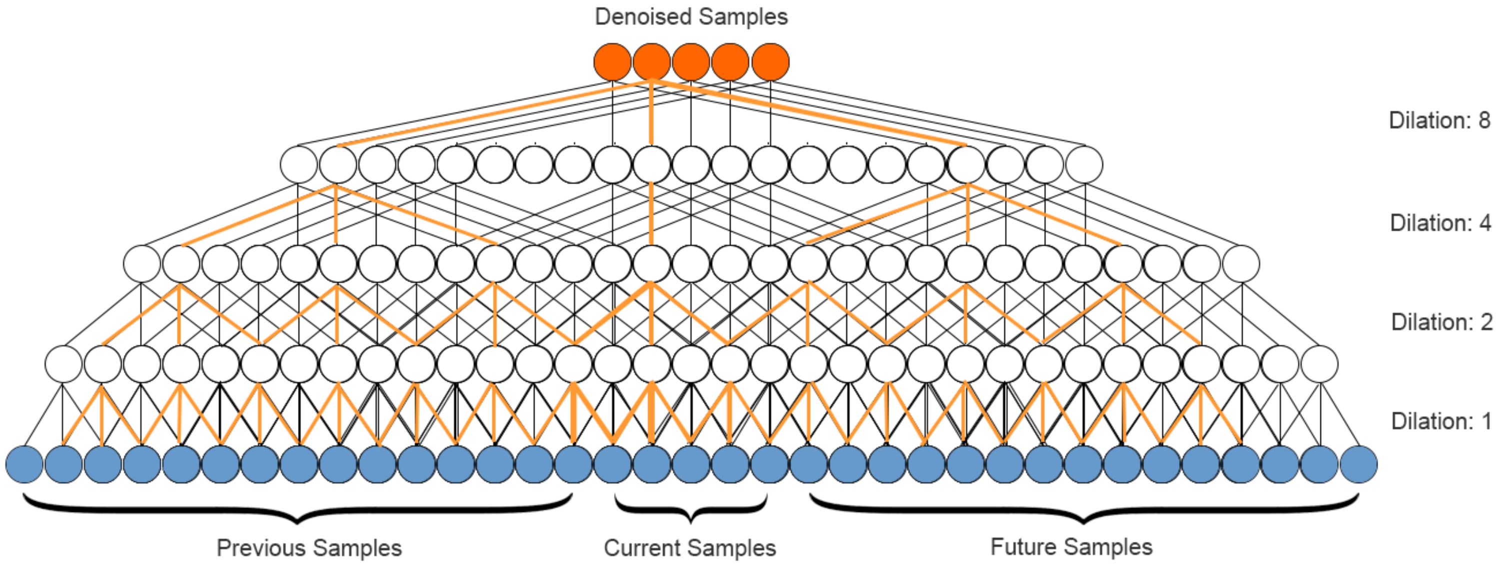


Target Fields



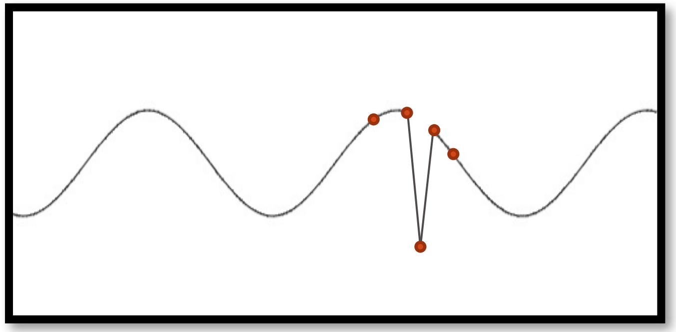
Target Fields

- Autoregression requires sequential, sample by sample, inference → slow
- Parallel prediction of target field benefits inference AND training

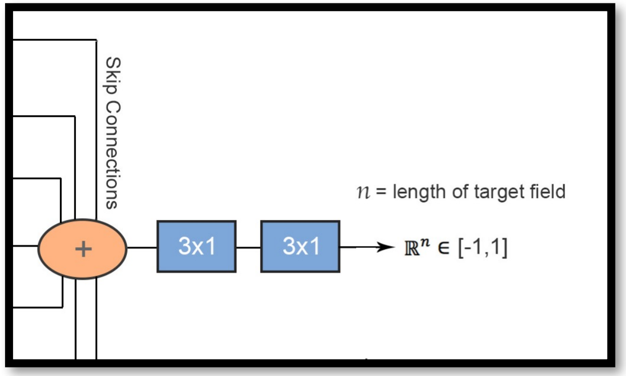


Enforcing Time Continuity

- Without auroregression, original Wavenet produces point discontinuities
- Very unpleasant sound
- 3x1 filters in final (non-dilated) layers allow time continuity to be reflected in the loss



Point discontinuity

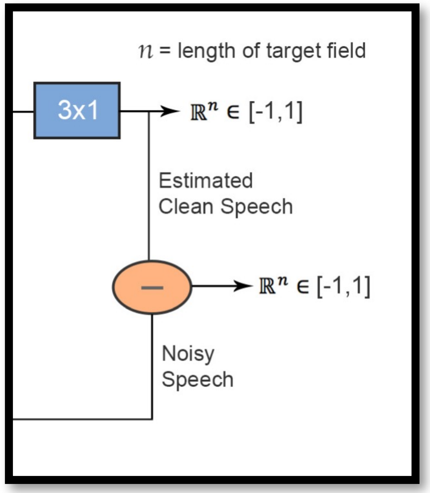


3x1 filters

Energy-Conserving Loss

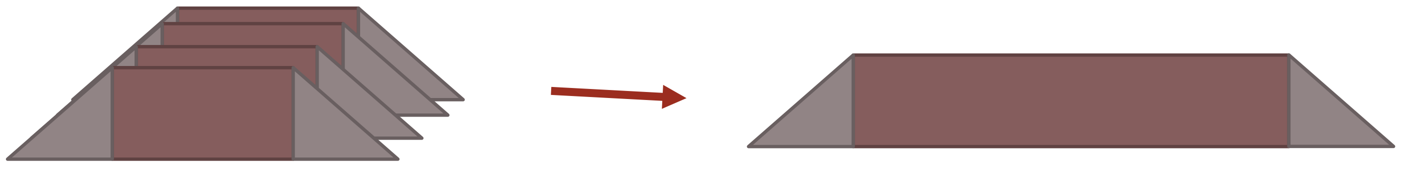
$$\mathcal{L}(\hat{s}_t) = |s_t - \hat{s}_t| + |b_t - \hat{b}_t|$$

- Goal: $E_{m_t} \equiv E_{\hat{m}_t}$
- Inspired by dissimilarity losses
- Empirically, reduces algorithmic artifacts



Flexibility in Temporal Dimension

- Same model can be deployed on reduced computational resources
- Audio input of arbitrary length → one-shot denoising
- Reduces redundant computations
- 25s of audio in single forward pass (Titan X Pascal)
- ~0.56s per 1 second of noisy audio
- Fully convolutional



Experiments

Setup

- 33 Layers
  - Dilations: 1, 2, ..., 256, 512
  - Stacks: 3
- 384ms Receptive Field
- 6.3m parameters

Data

- VCTK for voice
- DEMAND for environmental sounds

Unseen speakers in unseen noise conditions

Training SNR: 0dB – 18dB

Test SNR: 2.5dB – 17.5dB

Evaluation Metrics

- Should be perceptually meaningful
- MOS = mean opinion score (predicted) in range [1,5]
- Weighted combination of objective speech quality measures

- **SIG**: MOS rating of the signal distortion attending only to the speech signal
- **BAK**: MOS rating of the intrusiveness of background noise
- **OVL**: MOS rating of the overall effect

Results

Model	SIG	BAK	OVL	Model	SIG	BAK	OVL
Noise-only data augmentation				**Target field length**			
20%	2.74	2.98	2.30	1 sample*	1.37	1.79	1.28
10%	2.95	3.12	2.49	101 samples*	1.67	2.07	1.50
0 %	*3.62*	*3.23*	*2.98*	*1601 samples*	*3.62*	*3.23*	*2.98*
Loss				**Conditioning**			
L1	3.54	3.22	2.93	Unconditioned	3.48	3.12	2.88
Energy-Conserving	*3.62*	*3.23*	*2.98*	*Conditioned*	*3.62*	*3.23*	*2.98*
Wiener filtering	3.52	2.93	2.90	**Noisy signal**	3.51	2.66	2.79

*Computed on perceptual test set due to computational (time) constraints.

Best Configuration

- Energy-conserving loss
- 10% noise-only augmentation
- 100ms target field
- Conditioning

Mixed	Speech	Background	Wiener	12.5dB
Mixed	Speech	Background	Wiener	7.5dB
Mixed	Speech	Background	Wiener	2.5dB

Perceptual Evaluation

"give an overall quality score, taking into consideration both: speech quality and background-noise suppression"

- 33 participants
- 20 samples, 5 at each SNR
- 1-5 quality rating

Wiener Filtering	Proposed Model
2.92	3.60

Take away

- A discriminative adaptation of Wavenet for speech enhancement
- Reduction in time complexity, without sacrificing expressive capability
- Noise-only augmentation necessary for generating silence
- No speech-specific constraints
- Energy-conservation
- Perceptual trials: Preferred over Wiener Filtering
- Possible to learn multi-scale hierarchical representations from raw audio
- Audio samples online, source on GitHub

Future Work

- Continue exploring the idea of energy-conserving losses in neural audio processing models
- Better handling of short-time high energy events, e.g. honk in city traffic
- Apply to other audio domains
  - Music, multi-track separation

Thank you