A Wavenet for Speech Denoising

Jordi Pons

work done in collaboration with **Dario Rethage** *and* **Xavier Serra**

Music Technology Group (Universitat Pompeu Fabra, Barcelona)

Summer 2017 – Presented at **Pandora** and **Dolby** (Bay Area)

www.jordipons.me – @jordiponsme

Outline

Motivation

Wavenet

Wavenet for speech denoising

Motivation

Introduction: personal motivation

Until today it has been **standard practice** to use **time-frequency representations** as frontend – i.e. Automatic Speech Recognition.

Wiener filtering is commonly used for **source-separation** and **speech denoising** – how does it (typically) works?

1. Extract time-frequency representation – STFT.
2. Algorithm operates over the magnitude spectrogram.
3. Algorithm estimates a clean magnitude spectrogram.
4. Reconstruct audio using the phase of the mixture.

Can we do better? **End-to-end learning?**

Previous work: end-to-end learning for audio

- Discriminative models for **music audio classification** tasks.
  (Dieleman *et al.*, 2014) or (Lee *et al.*, 2017)
- Discriminative models for **speech audio classification** tasks.
  (Collobert *et al.*, 2016) or (Zhu *et al.*, 2016)
- **Generative models** for **music** audio signals.
  (Engel *et al.*, 2017) or (Mehri *et al.*, 2016)
- **Generative models** for **speech** audio signals.
  (van den Oord et al., 2016) or (SEGAN: Pascual *et al.*, 2017)

generative models are autoregressive – except for SEGAN!

..it looks like possible, specially with autoregressive models!

Previous work: end-to-end speech denoising

- Tamura *et al.* (1988) used a four-layered feed-forward network operating directly in the raw-audio domain.
- Pascual *et al.* (2017) used of an end-to-end generative adversarial network for speech denoising – *a.k.a.* SEGAN.
- Qian *et al.* (2017) proposed a Bayesian Wavenet.

In all three cases, they provide better results than their counterparts based on processing magnitude spectrograms!

Our study **adapts Wavenet**'s model **for speech denoising**.

Wavenet

Wavenet: an autoregressive generative model

Proposed by van den Oord *et al.* in 2016 – based on PixelCNN

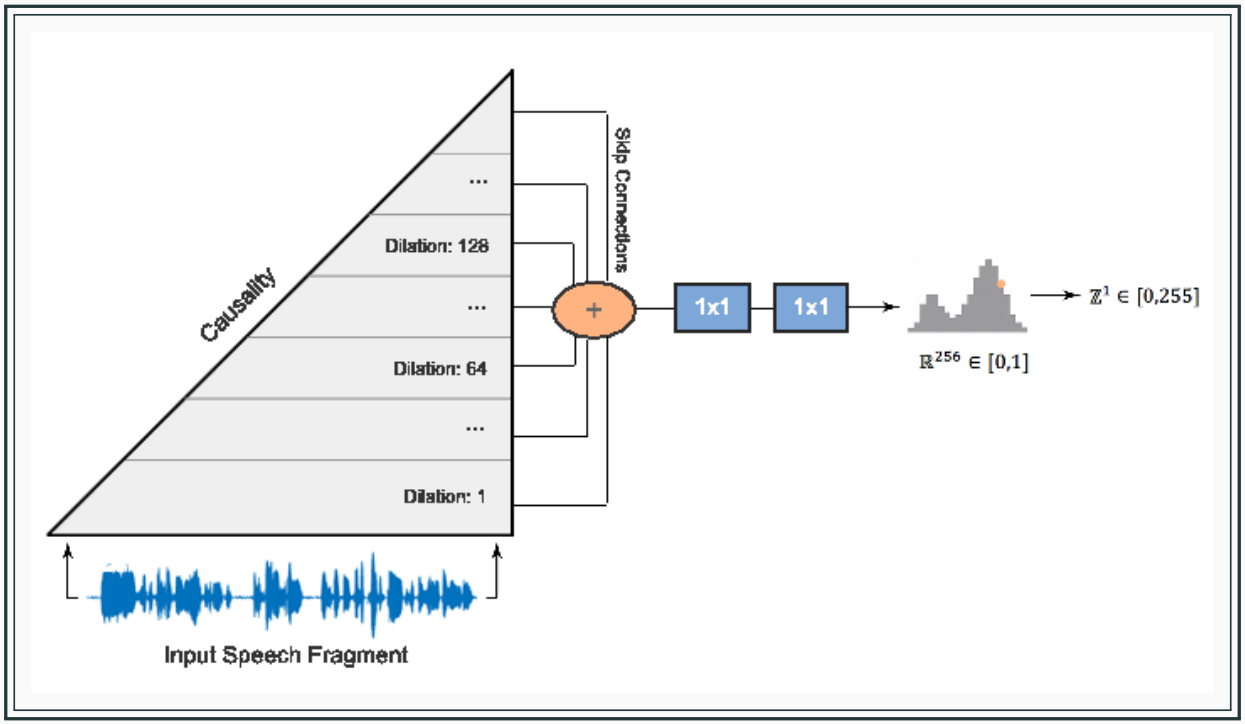
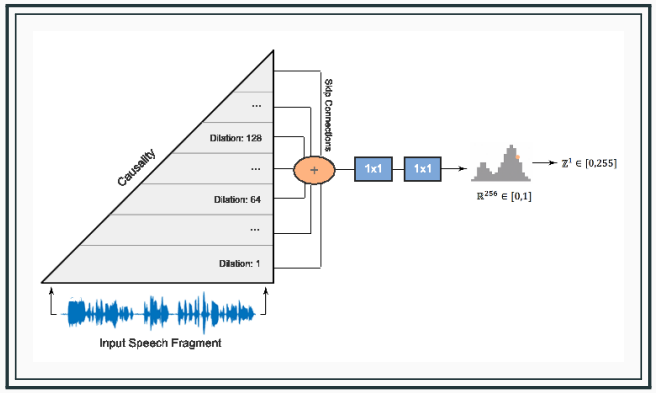


Figure 1: Wavenet architecture overview

μ-**law quantization**: discrete softmax output distribution.

Unsupervised training and **sequencial inference**.

Wavenet: building blocks



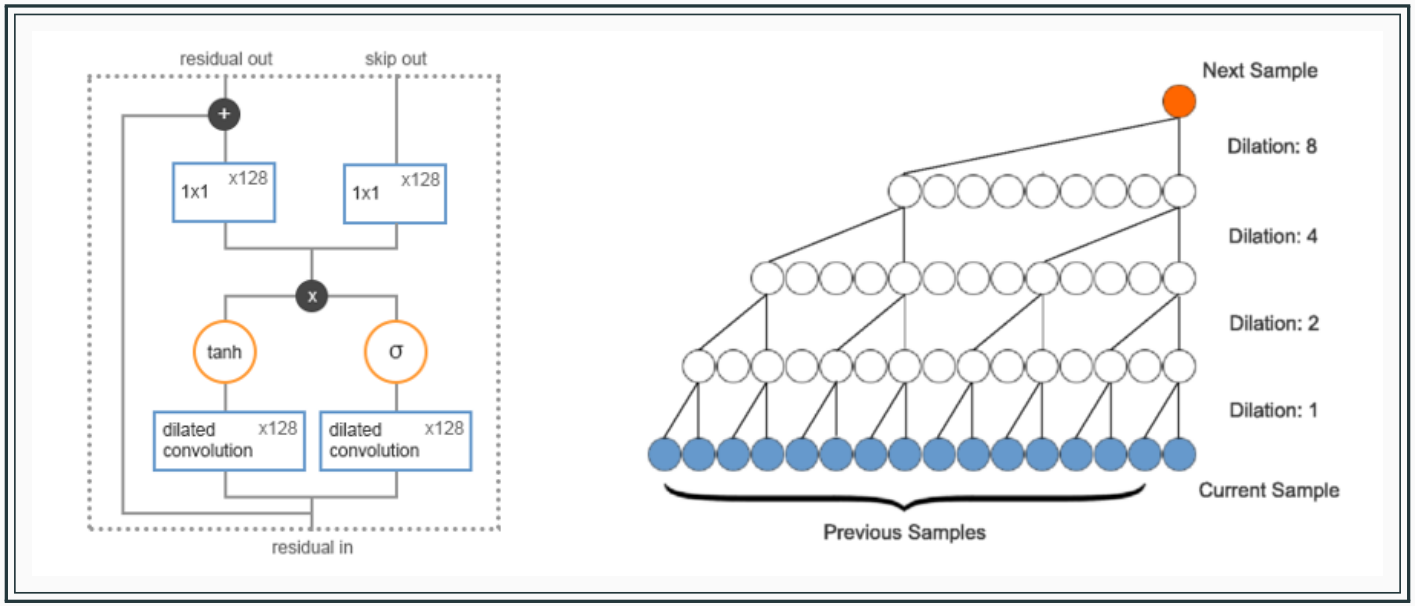
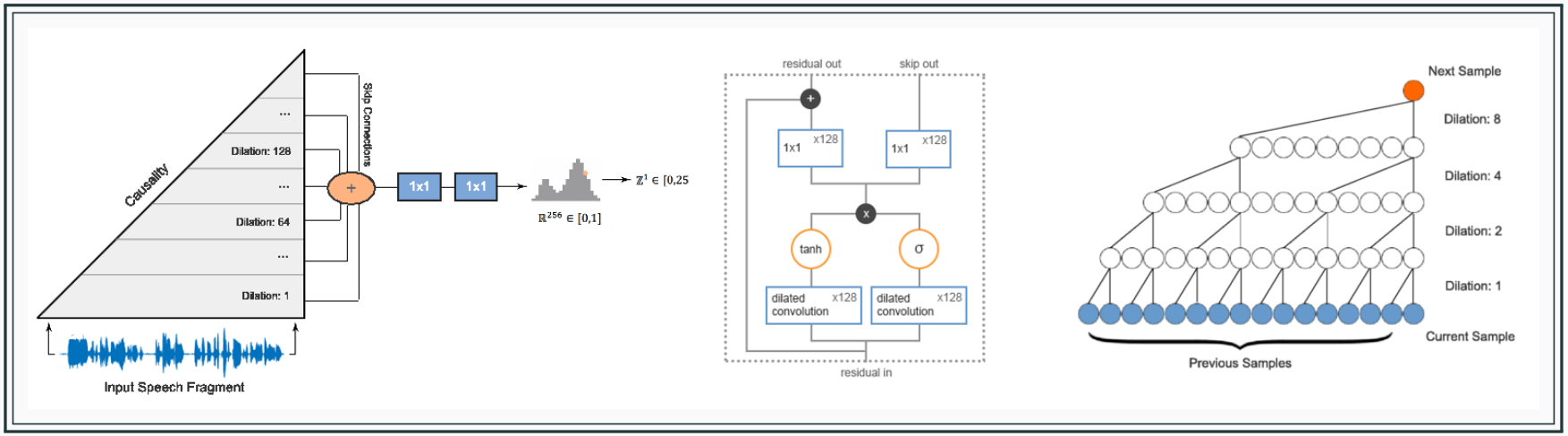


Figure 2: *Left* – Residual layer. *Right* – Causal, dilated convolutions.

$$\textit{Gated Units} \quad \rightarrow \quad z_{t'} = \tanh(W_f * x_t) \odot \sigma(W_g * x_t)$$

Wavenet: summary



- Causal, dilated convolutions
- μ-law quantization – softmax output
- Skip connections
- Residual layers with gated units
- Time-complexity

Wavenet for speech denoising

Wavenet for speech denoising

Main difference with respect to Wavenet:

non-causal: non-autoregressive and parallelize inference.

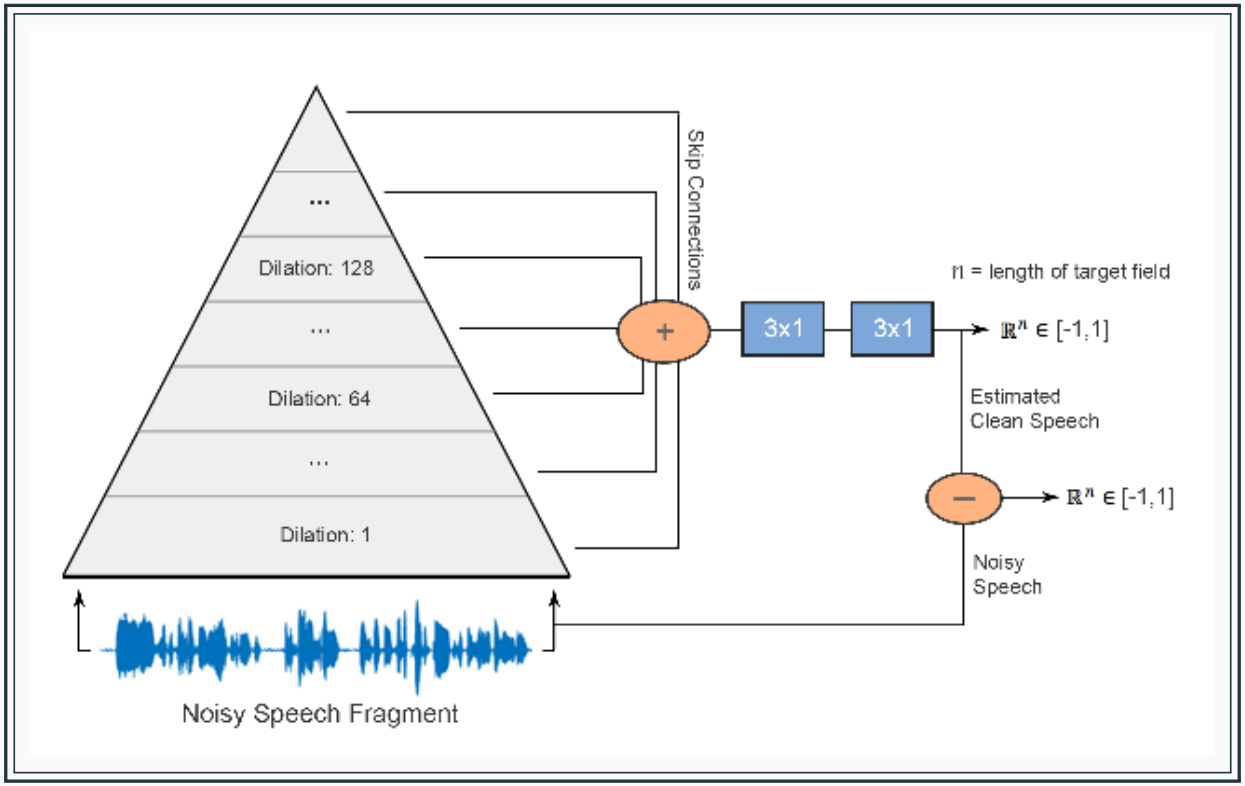


Figure 3: Wavenet for speech denoising: architecture overview

Wavenet for speech denoising: real-valued output

Original Wavenet:

- discrete softmax output $\rightarrow$ artifacts where introduced.
- μ-law quantization $\rightarrow$ it amplified the background-noise.

It was key to remove the μ-law quantization!

We predict raw audio – without any pre-processing.

Implication 1: **loss** function! $\rightarrow$ $\mathcal{L}(\hat{v}_t) = |v_t - \hat{v}_t| + |b_t - \hat{b}_t|$

Implication 2: **discriminative** Wavenet! $p(x) \rightarrow p(y|x)$

- Supervised learning – minimizing a regression loss function.
- Not using a probabilistic framework (sampling a distribution).

..new opportunities!

Wavenet for speech denoising: producing silence

Challenge: difficulties in **producing silence**.

Data driven solution: **data augmentation**.

In: noise – without speech

Out: zeros – silence

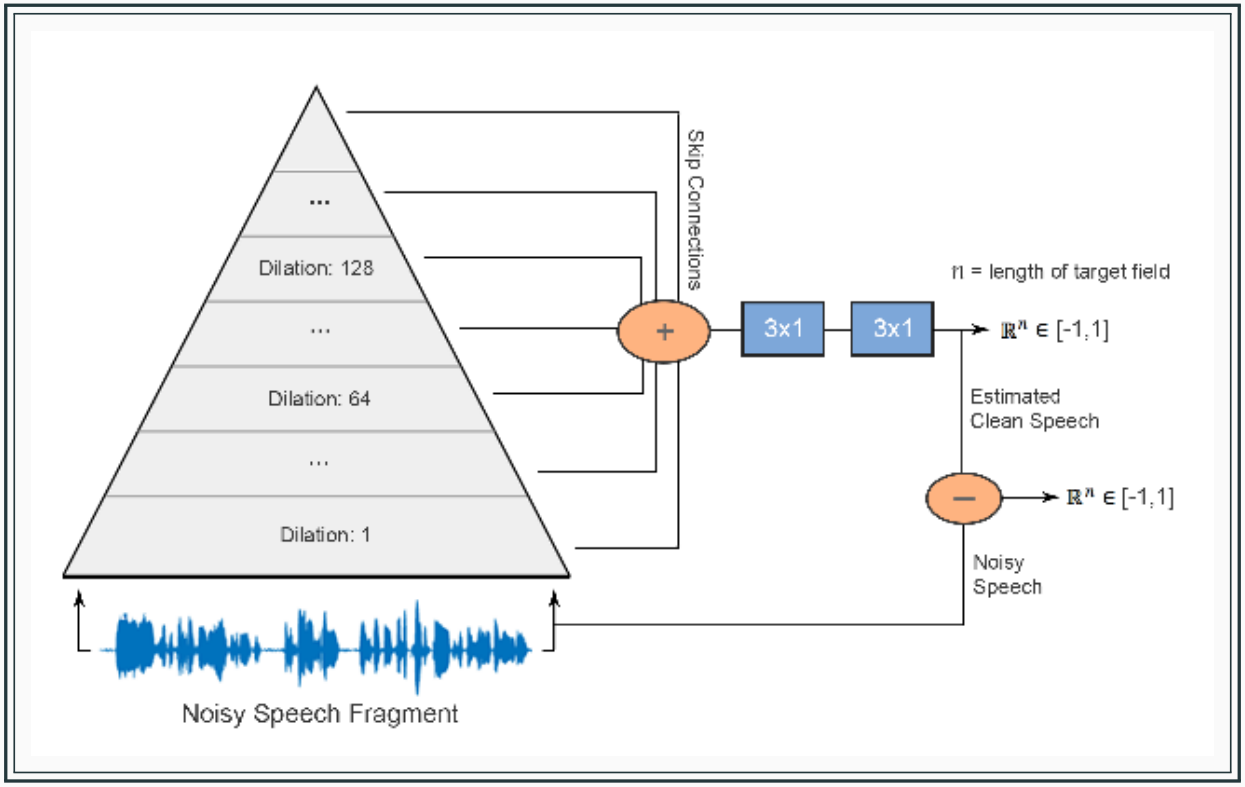
We trained models with 10–20% more examples having only noise.

Generated samples where more perceptually pleasant!

Wavenet for speech denoising: continuity?

~~Autoregressive model~~ $\rightarrow$ **no** longer enforcing **temporal continuity**!

Final 3x1 filters and target field prediction



Wavenet for speech denoising: continuity?

~~Autoregressive model~~ $\rightarrow$ **no** longer enforcing **temporal continuity**!

Final 3x1 filters and **target field prediction**

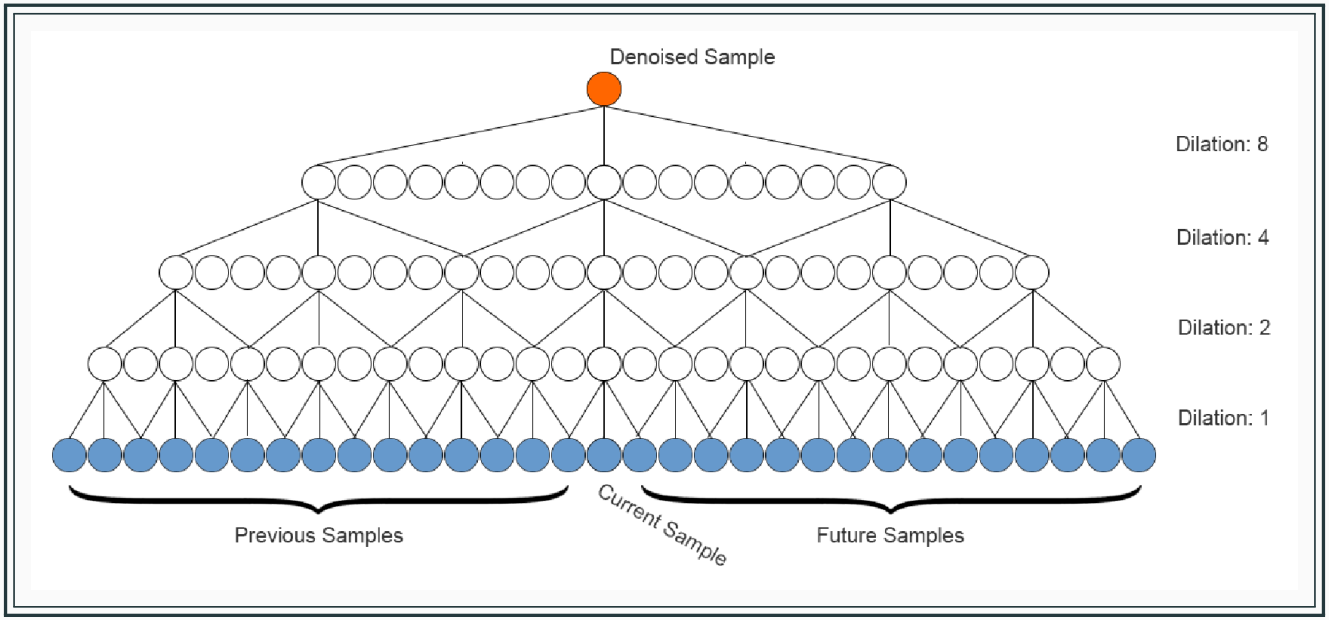


Figure 4: Sample prediction

Wavenet for speech denoising: continuity?

~~Autoregressive model~~ $\rightarrow$ **no** longer enforcing **temporal continuity**!

Final 3x1 filters and **target field prediction**

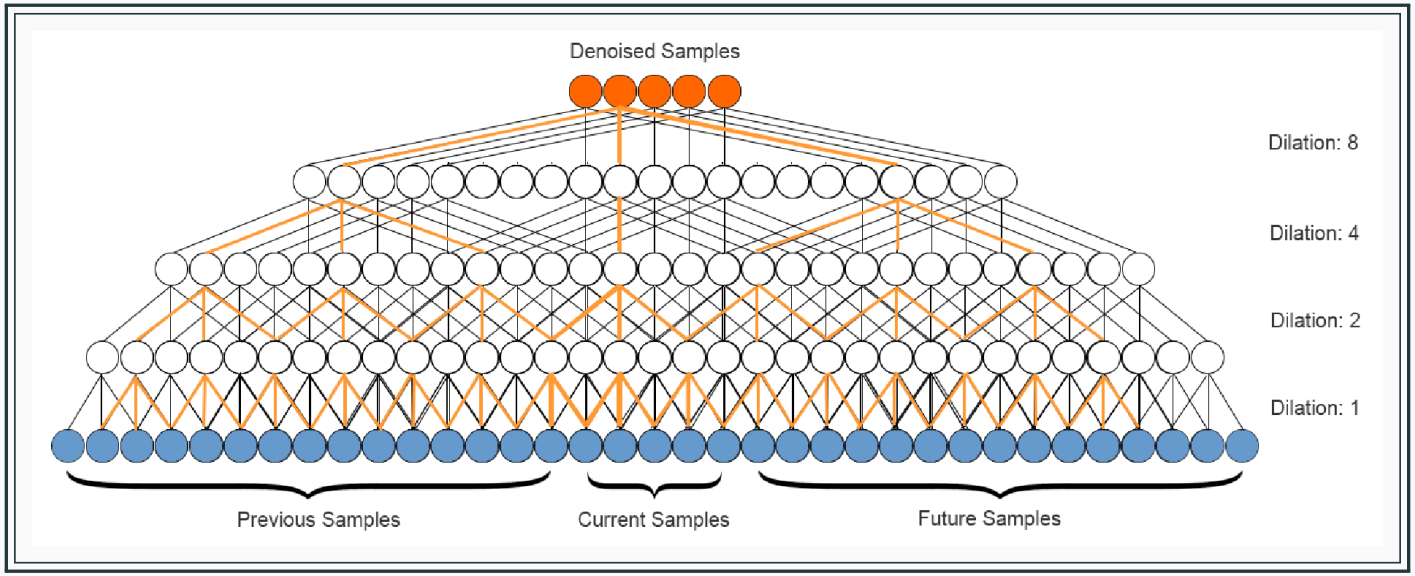


Figure 5: Target field prediction

Wavenet for speech denoising: continuity?

~~Autoregressive model~~ $\rightarrow$ **no** longer enforcing **temporal continuity**!

Final 3x1 filters and **target field prediction**

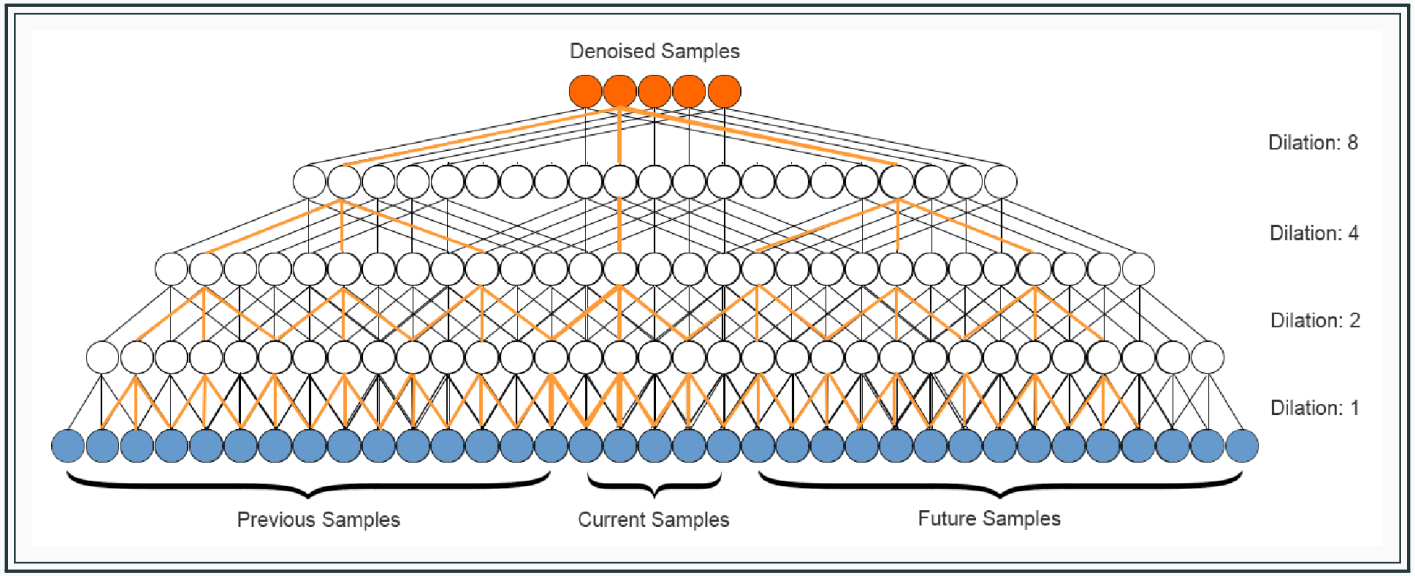


Figure 6: Target field prediction

target field + fully convolutional = **one shot denoising**!

Experimental setup

Dataset and problem:

Voice + (environmental) background noise $\rightarrow$ remove noise

Train: 28 speakers under 40 different noise conditions $\approx$ 10h

Test: 2 *unseen* speakers under 20 *different* noise conditions $\approx$ 40'

Architecture: optimized to adhere to our memory constraints.

30 residual layers $\approx$ 6.3 million parameters

Receptive field of 6,139 samples $\approx$ 384ms

Target field of 1601 samples $\approx$ 100ms

Parallel inference on 1601 samples at once, results in a **denoising time** *of* $\approx$ **0.56 seconds per second** *of noisy audio on GPU.*

Results: objective measures

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*
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.

computational approximations of quality ratings on test set

scores range: 1–5, higher scores are better

Wiener filtering method based on a priori SNR estimation

Perceptual test

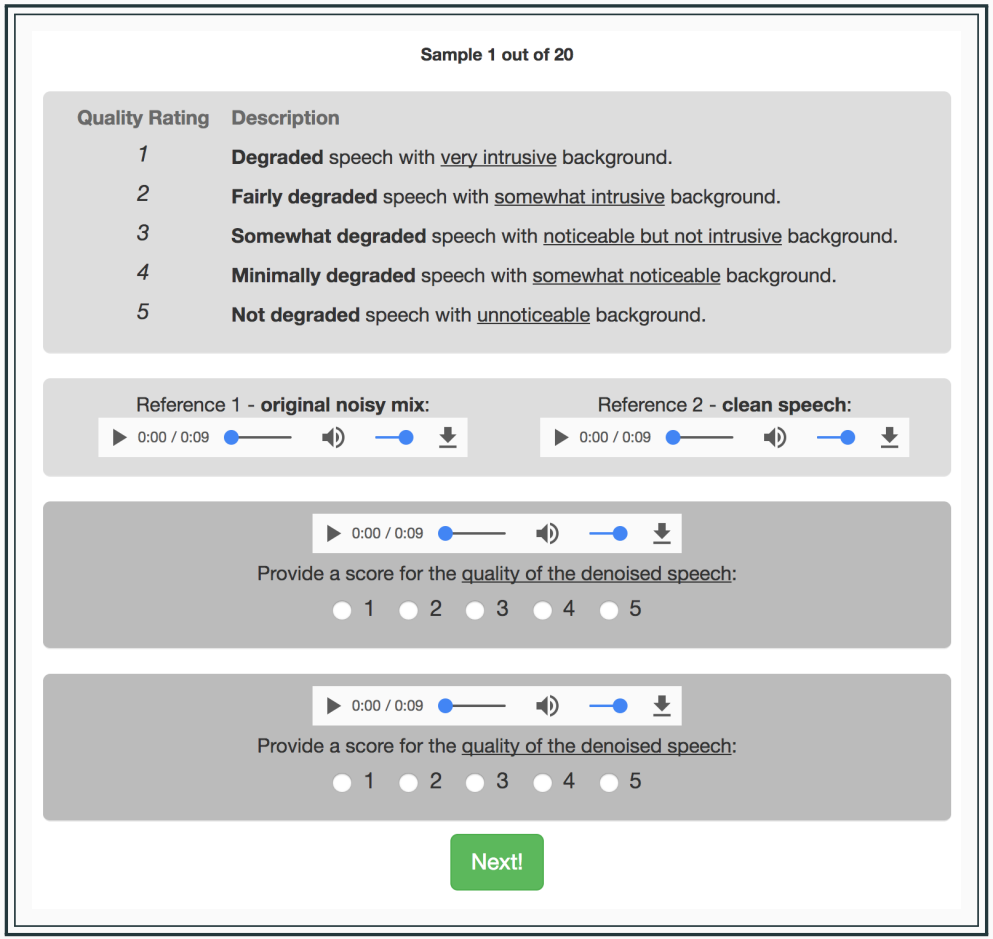


Figure 7: GUI used for the perceptual test

MOS scale

Results of the perceptual test

Measurement	Wiener filtering	Proposed Wavenet
MOS	2.92	3.60

subjective MOS measures on perceptual test set

quality ratings from 1–5, higher scores are better

(33 participants)

..statistically significant preference (t-test: p-value $\ll 0.001$)!

Wavenet for speech denoising: summary and conclusions

- **Non-causal** adaptation of Wavenet's architecture.
- Turned Wavenet into a **discriminative** model.
- **Minimize** the **time-complexity** of the model!
- Operating **directly** on the raw audio $\rightarrow$ explore new costs!
- End-to-end approximation to **speech denoising**.

More audio examples:

jordipons.me/apps/speech-denoising-wavenet

Trained model and code:

github.com/drethage/speech-denoising-wavenet

A Wavenet for Speech Denoising

Jordi Pons

work done in collaboration with **Dario Rethage** *and* **Xavier Serra**

Music Technology Group (Universitat Pompeu Fabra, Barcelona)

Summer 2017 – Presented at **Pandora** and **Dolby** (Bay Area)

www.jordipons.me – @jordiponsme

Does “Wavenet for Speech Denoising” generalize?

..this is a funny **twitter story**! :)

Does “Wavenet for Speech Denoising” generalize?

Miles Brundage @Miles_Brundage · Jun 22
Tried out WaveNet denoiser... not great, but entertaining, results... also ran it fast/didn't do any fine tuning. Trying to see how to share

1 2 8

Miles Brundage @Miles_Brundage · Jun 22
Original: vocaroo.com/i/s1dmmYKxkGgk
(ostensibly) denoised: vocaroo.com/i/s1UQFbcjCLMq (inter-speech noise gone, but bad quality for voice proper)

1

Miles Brundage @Miles_Brundage · Jun 22
And the best part - (ostensibly) the noise: vocaroo.com/i/s1236PkJAAZx (aka Trump with lisp)

1 2

Miles Brundage @Miles_Brundage · Jun 22
Suspect I could get good results with fine-tuning (that was 1st attempt) but main goal was just to see if I could run it, so done for now :)

1 1

Miles Brundage @Miles_Brundage · Jun 22
That was false, I want to do it again now that I have it installed. What should I denoise?

1

Miles Brundage @Miles_Brundage · Jun 22
crickets the earliest known recording of JFK it is!

2 1

Miles Brundage @Miles_Brundage · Jun 22
Another 1 - JFK didn't work well, too low quality I think...
Original: vocaroo.com/i/s1EOFgK6QXp0
Denoised: vocaroo.com/i/s1LLlc8jH5CT

A Wavenet for Speech Denoising

Jordi Pons

work done in collaboration with **Dario Rethage** *and* **Xavier Serra**

Music Technology Group (Universitat Pompeu Fabra, Barcelona)

Summer 2017 – Presented at **Pandora** and **Dolby** (Bay Area)

www.jordipons.me – @jordiponsme