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
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 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

$\mu$-law quantization: discrete softmax output distribution.

Unsupervised training and sequential inference.
Figure 2: Left – Residual layer. Right – Causal, dilated convolutions.

\[ \text{Gated Units} \quad \rightarrow \quad z_t' = \text{tanh}(W_f \ast x_t) \odot \sigma(W_g \ast x_t) \]
Wavenet: summary

- Causal, dilated convolutions
- $\mu$-law quantization – softmax output
- Skip connections
- Residual layers with gated units
- Time-complexity
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 → artifacts where introduced.
- $\mu$-law quantization → it amplified the background-noise.

It was key to remove the $\mu$-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 → no longer enforcing **temporal continuity**!

**Final 3x1 filters** and target field prediction
Autoregressive model → no longer enforcing *temporal continuity*!

Final 3x1 filters and target field prediction

**Figure 4:** Sample prediction
Wavenet for speech denoising: continuity?

Autoregressive model → no longer enforcing **temporal continuity**!

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

**Figure 5:** Target field prediction
Wavenet for speech denoising: continuity?

Autoregressive model → 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 → remove noise

Train: 28 speakers under 40 different noise conditions ≈ 10h
Test: 2 unseen speakers under 20 different noise conditions ≈ 40’

Architecture: optimized to adhere to our memory constraints.
30 residual layers ≈ 6.3 million parameters
Receptive field of 6,139 samples ≈ 384ms
Target field of 1601 samples ≈ 100ms

Parallel inference on 1601 samples at once, results in a \textbf{denoising time of} \approx 0.56 \textbf{seconds per second} of noisy audio on GPU.
### Results: objective measures

<table>
<thead>
<tr>
<th>Model</th>
<th>SIG</th>
<th>BAK</th>
<th>OVL</th>
<th>Target field length</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noise-only data augmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>2.74</td>
<td>2.98</td>
<td>2.30</td>
<td>1 sample*</td>
</tr>
<tr>
<td>10%</td>
<td>2.95</td>
<td>3.12</td>
<td>2.49</td>
<td>101 samples*</td>
</tr>
<tr>
<td>0 %</td>
<td>3.62</td>
<td>3.23</td>
<td>2.98</td>
<td>1601 samples</td>
</tr>
<tr>
<td><strong>Wiener filtering</strong></td>
<td>3.52</td>
<td>2.93</td>
<td>2.90</td>
<td></td>
</tr>
<tr>
<td><strong>Noisy signal</strong></td>
<td>3.51</td>
<td>2.66</td>
<td>2.79</td>
<td></td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Wiener filtering</th>
<th>Proposed Wavenet</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>2.92</td>
<td>3.60</td>
</tr>
</tbody>
</table>

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 → explore new costs!
- End-to-end approximation to **speech denoising**.

More audio examples:

[jordipons.me/apps/speech-denoising-wavenet](jordipons.me/apps/speech-denoising-wavenet)

Trained model and code:

[github.com/drethage/speech-denoising-wavenet](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! :)

Very nice speech denoising results with a Wavenet-based approach:

arxiv.org/abs/1706.07162
github.com/drethage/speech ...
jordipons.me/apps/speech-de ...

drethage/speech-denoising-wavenet
A neural network for end-to-end speech denoising. Contribute to speech-denoising-wavenet development by creating an account on GitHub.
github.com

5:49 PM - 22 Jun 2017

33 Retweets 92 Likes
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

Miles Brundage @Miles_Brundage · Jun 22
Original: vocaroo.com/sf1emmn9jK0kCqgk (ostensibly) denoised: vocaroo.com/sf1UQFbcjCLMq (inter-speech noise gone, but bad quality for voice proper)

Miles Brundage @Miles_Brundage · Jun 22
And the best part - (ostensibly) the noise: vocaroo.com/s12Z36PkJAAZx (aka Trump with lisp)

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 :

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?

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

Miles Brundage @Miles_Brundage · Jun 22
Another 1 - JFK didn’t work well, too low quality I think.
Original: vocaroo.com/s1EFOFqK0QXp0
Denoised: vocaroo.com/s1LLc6jH5CT
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