A Wavenet for Speech Denoising

Dario Rethage[*]
dario@rethage.net
Music Technology Group
Universitat Pompeu Fabra

Jordi Pons[*]
jordi.pons@upf.edu
Music Technology Group
Universitat Pompeu Fabra

Xavier Serra
xavier.serra@upf.edu
Music Technology Group
Universitat Pompeu Fabra

Abstract

Currently, most speech processing techniques use magnitude spectrograms as front-end and are therefore by default discarding part of the signal: the phase. In order to overcome this limitation, we propose an end-to-end learning method for speech denoising based on Wavenet. The proposed model adaptation retains Wavenet's powerful acoustic modeling capabilities, while significantly reducing its time-complexity by eliminating its autoregressive nature. Specifically, the model makes use of non-causal, dilated convolutions and predicts target fields instead of a single target sample. The discriminative adaptation of the model we propose, learns in a supervised fashion via minimizing a regression loss. These modifications make the model highly parallelizable during both training and inference. Both computational and perceptual evaluations indicate that the proposed method is preferred to Wiener filtering, a common method based on processing the magnitude spectrogram.

1 Introduction

Over the last several decades, machine learning has produced solutions to complex problems that were previously unattainable with signal processing techniques [4, 12, 38]. Speech recognition is one such problem where machine learning has had a very strong impact. However, until today it has been standard practice not to work directly in the time-domain, but rather to explicitly use time-frequency representations as input [1, 34, 35] – for reducing the high-dimensionality of raw waveforms. Similarly, most techniques for speech denoising use magnitude spectrograms as front-end [13, 17, 21, 34, 36]. Nevertheless, this practice comes with its drawbacks of discarding potentially valuable information (phase) and utilizing general-purpose feature extractors (magnitude spectrogram analysis) instead of learning specific feature representations for a given data distribution.

Most recently, neural networks have shown to be effective in handling structured temporal dependencies between samples of a discretized audio signal. For example, consider the most local structure of a speech waveform ($\approx$ tens of milliseconds). In this range of context, many sonic characteristics of the speaker (timbre) can be captured and linguistic patterns in the speech become accessible in the form of phonemes. It is important to note that these levels of structure are not discrete, making techniques that explicitly focus on different levels of structure inherently suboptimal. This suggests that deep learning methods, capable of learning multi-scale structure directly from raw audio, may have great potential in learning such structures. To this end, discriminative models have been used in a end-to-end learning fashion for music [6, 15] or speech classification [5, 20, 39]. Raw audio waveforms have also been successfully employed for generative tasks [7, 18, 30, 22]. Interestingly, most of these generative models are autoregressive [7, 18, 30], with the exception of SEGAN – which is based on a generative adversarial network [22]. We are not aware of any generative model for raw audio based on variational autoencoders.

Previous discussion motivates our study in adapting Wavenet's model (an autoregressive generative model) for speech denoising. Our main hypothesis is that by learning multi-scale hierarchical representations from raw audio we can overcome the inherent limitations of using the magnitude

[*]Contributed equally.

spectrogram as a front-end for this task. Some work in this direction already exists. Back in the 80s, Tamura et al. [27] used a four-layered feed-forward network operating directly in the raw-audio domain to learn a noise-reduction mapping. And recently: Pascual et al. [22] proposed the use of an end-to-end generative adversarial network for speech denoising, and Qian et al. [24] used a Bayesian Wavenet for speech denoising. In all three cases, they provide better results than their counterparts based on processing magnitude spectrograms.

Section 2 describes the original Wavenet architecture, and section 3 describes the modifications we propose. In section 4 we experiment with and discuss some architectural parameters. And finally, section 5 concludes by highlighting the most relevant contributions.

2 Wavenet

Wavenet is capable of synthesizing natural sounding speech [30]. This autoregressive model shapes the probability distribution of the next sample given some fragment of previous samples. The next sample is produced by sampling from this distribution. An entire sequence of samples is produced by sequentially feeding previously generated samples back into the model – this enforces time continuity in the resulting audio waveforms. A high-level visual depiction of the model is presented in Figure 1. Wavenet is the audio domain adaptation of the PixelCNN generative model for images [19, 31]. Wavenet retains many PixelCNN features like: causality, gated convolutional units, discrete softmax output distributions and the possibility of conditioning the model – while introducing dilated convolutions and non-linear quantization [30]. Some of Wavenet's key features are presented below:

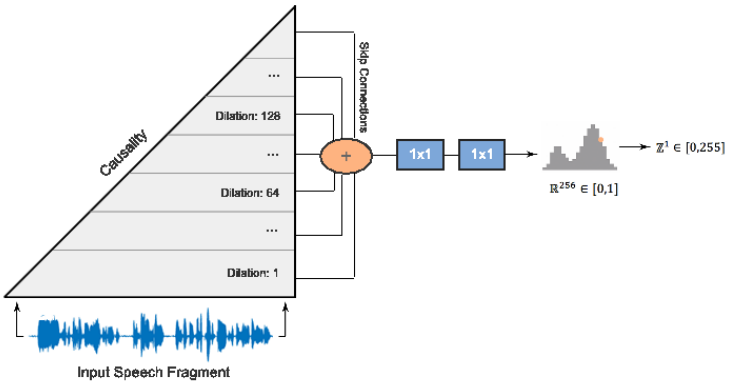

Figure 1: Overview of Wavenet.

Gated Units As in LSTMs [9], sigmoidal gates control the activations' contribution in every layer: $z_{t'} = \tanh(W_f * x_t) \odot \sigma(W_g * x_t)$, where $*$ and $\odot$ operators denote convolution and element-wise multiplication, respectively. f, t, t' and g stand for filter, input time, output time and gate indices. W_f and W_g are convolutional filters. Figure 2 (*Left*) depicts how sigmoidal gates are utilized.

Causal, dilated convolutions Wavenet makes use of causal, dilated convolutions [30, 37]. It uses a series of small (length = 2) convolutional filters with exponentially increasing dilation factors. This results in a exponential receptive field growth with depth. Causality is enforced by asymmetric padding proportional to the dilation factor, which prevents activations from propagating back in time – see Figure 2 (*Right*). Each dilated convolution is contained in a residual layer [8], controlled by a sigmoidal gate with an additional 1x1 convolution and a residual connection – see Figure 2 (*Left*).

μ-law quantization When using a discrete softmax output distribution, it is necessary to perform a more coarse 8-bit quantization to make the task computationally tractable. This is accomplished via a μ-law non-linear companding followed by an 8-bit quantization (256 possible values):

$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

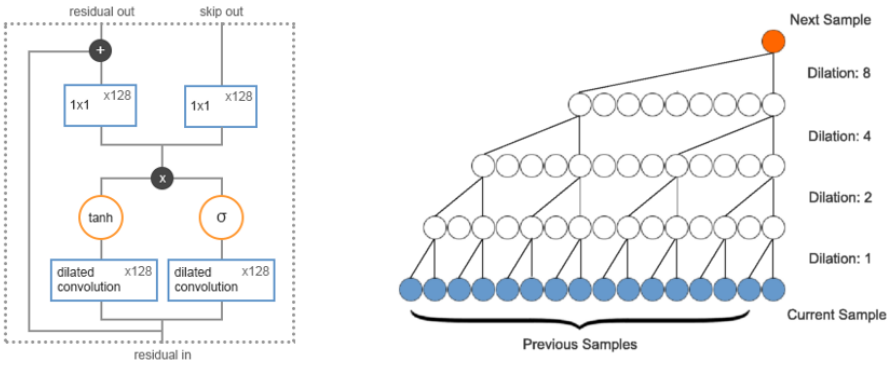

Figure 2: *Left* – Residual layer. *Right* – Causal, dilated convolutions with increasing dilation factors.

Skip Connections These offer two advantages. First, they facilitate training deep models [26]. And second, they enable information at each layer to be propagated directly to the final layers. This allows the network to explicitly incorporate features extracted at several hierarchical levels into its final prediction [14]. Figure 1 and 2 (*Left*) provide further details in how skip connections are used.

Context Stacks These deepen the network without increasing the receptive field length as drastically as increasing the dilation factor does. This is achieved by simply stacking a set of layers, dilated to some maximum dilation factor, onto each other – and can be done as many times as desired [30]. For example, Figure 2 (*Right*) is composed of a single stack.

Time-complexity A significant drawback of Wavenet is its sequential (non-parallelizable) generation of samples. This limitation is strongly considered in the speech-denoising Wavenet design.

3 Wavenet for Speech Denoising

Speech denoising techniques aim to improve the intelligibility and the overall perceptual quality of speech signals with intrusive background-noise. The problem is typically formulated as follows: $m_t = s_t + b_t$, where: $m_t \equiv$ mixed signal, $s_t \equiv$ speech signal, $b_t \equiv$ background-noise signal. The goal is to estimate s_t given m_t. Speech denoising, while sharing many properties with speech synthesis, also has several unique characteristics – these motivated the design of this Wavenet adaptation. A high-level visual depiction of the model is presented in Figure 3. Its key features are presented below:

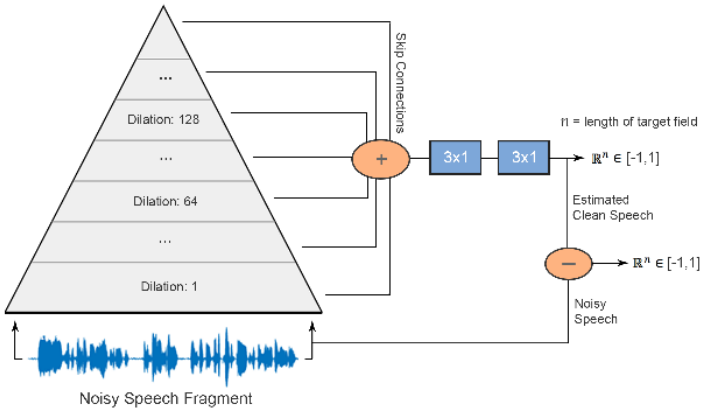

Figure 3: Overview of the speech-denoising Wavenet.

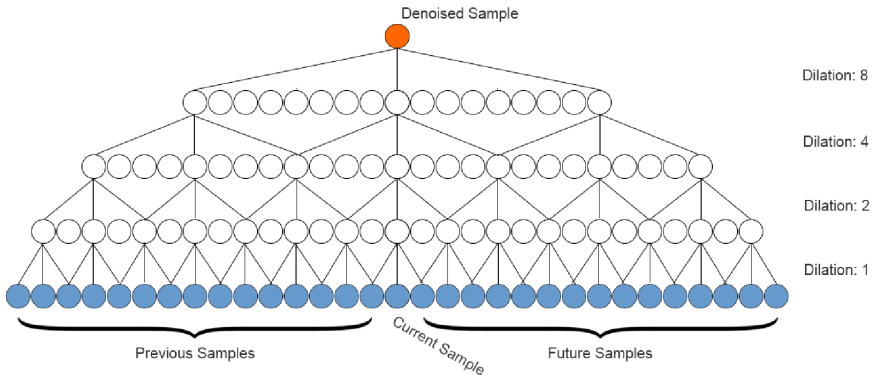

Figure 4: Non-causal, dilated convolutions with exponentially increasing dilation factors.

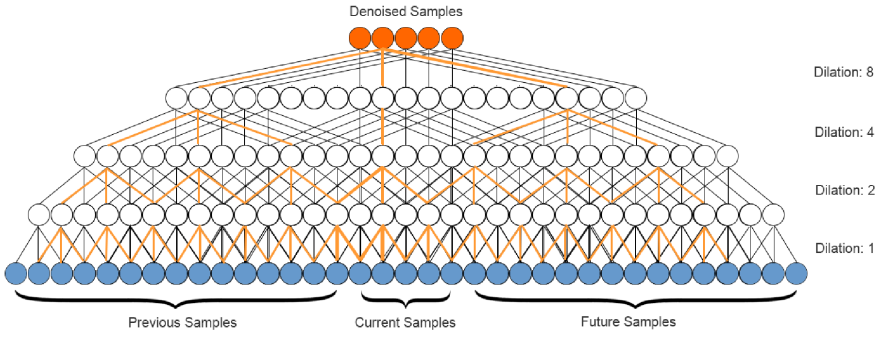

Figure 5: Predicting on a target field.

Non-causality Contrary to audio synthesis, in speech denoising, some future samples are generally available to help make more well informed predictions. Even in real time applications, when a few milliseconds of latency in model response can be afforded, the model has access to valuable information about samples occurring shortly after a particular sample of interest. As a result, and given that Wavenet's time-complexity was a major constraint, the autoregressive causal nature of it was removed in the proposed model. A logical extension to Wavenet's asymmetric dilated convolution pattern (shown in Figure 2) is to increase the filter length to 3 and perform symmetric padding at each dilated layer. If the sample we wish to enhance is now taken to be at the center of the receptive field, this has the effect of doubling the context around a sample of interest and eliminating causality. The model has access to the same amount of samples in the past as samples in the future to inform the prediction. This can be seen in Figure 4. Early experiments with larger filters of length 5, 11 and 21 showed inferior performance.

Real-valued predictions Wavenet uses a discrete softmax output to avoid making any assumption on the shape of the output's distribution – this is suitable for modeling multi-modal distributions. However, early experiments with discrete softmax outputs proved disadvantageous. Instead, the potentially multi-modal output distribution allowed for artifacts to be introduced into the denoised signal. This suggests that real-valued predictions (assuming uni-modal gaussian-shaped output distributions) seem to be more appropriate for our problem. Moreover, experiments with discrete softmax outputs resulted in output distributions with high variance – signifying low confidence with the value having highest probability. μ-law quantization was also disadvantageous because it disproportionately amplified the background-noise. For these reasons, the proposed model predicts raw audio – without any pre-processing.

Energy-conserving loss We propose utilizing a two term loss that enforces energy conservation:

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

The background-noise signal is estimated by subtracting the denoised speech from the mixed input, as shown in Figure 3 – and in the following formulation: $\hat{b}_t = m_t - \hat{s}_t$. As a result, the network predicts both components of the signal and, since the second output is produced by a parameterless operation on the speech estimate, the loss produced by this output is directly representative of the model's ability to perform the actual task of interest. Note that: $\hat{m}_t = \hat{s}_t + \hat{b}_t = \hat{s}_t + (m_t - \hat{s}_t) = m_t \rightarrow \hat{m}_t = m_t$, then: $\hat{s}_t + \hat{b}_t = s_t + b_t \rightarrow 0 = s_t - \hat{s}_t + b_t - \hat{b}_t$. Therefore, the proposed way of estimating $\hat{b}_t$ tailors the two term loss towards conserving the energy of the original mixture: $E_{\hat{m}_t} \equiv E_{m_t}$. Previous work considered energy-conservative pipelines: for source separation [23, 3], or for speech enhancement [34]. Finally, note that the first-term of the proposed loss corresponds to the standard L1 loss – that is used as baseline for comparison in Table 1.

Discriminative model Note that the proposed model is no longer autoregressive and its output is not explicitly modeling a probability distribution, but rather the output itself. Furthermore, the model is trained in a supervised fashion – by minimizing a regression loss function. As a result: the proposed model is no longer generative (like Wavenet), but discriminative.

Final 3x1 filters As mentioned, the proposed architecture is not autoregressive. That is, previously generated samples are not fed back into the model to inform future predictions – which enforces time continuity in the resulting signal. Early experiments produced waveforms with sporadic point discontinuities that sounded very disruptive. Replacing the kernels of the two final layers with 3x1 filters instead of 1x1 filters reimposed this constraint. Larger kernels were also considered, although these did not improve our results.

Target field prediction The proposed model does not predict just one, but a set of samples in a single forward propagation – see Figure 5. Parallelizing the inference process from 1 sample to on the order of 1000 samples offers significant memory and time savings. This is because overlapping data is used for predicting neighboring samples, and by predicting target fields these redundant computations are done just once.

The receptive field length (rf) of the model is the number of input samples that go into the prediction of a single denoised output sample. In order to maintain that every output sample in the target field (tf) has a full receptive field of context contributing to its prediction, the length of the fragment presented to the model must be equal to: $rf + (tf - 1)$. Finally, note that the cost is computed sample-wise – during training, individual sample costs of a target field are averaged.

Conditioning The model is conditioned on a binary-encoded scalar corresponding to the identity of the speaker. This condition value is the bias term in every convolution operation. Condition values 1-28 represent each of the 28 speakers comprising the training set. In addition, we add an auxiliary code (all zeros) denoting any speaker identity so that the trained model can be used for unknown speakers. The same training data is presented to the model either conditioned to its speaker ID or to zeros. 1/29% of training samples have their condition values set to zeros. Note that this conditioning procedure can also be interpreted as a data augmentation strategy.

Noise-only data augmentation A form of augmentation in which a proportion of training samples contain only background-noise was also employed after observing that our model had difficulties producing silence. In section 4.3 we experiment with 10% and 20% noise-only training samples.

Denoising step The network is presented with a noisy speech fragment and the condition value is set to zero. By default the network denoises the input in batches, iteratively appending each denoised fragment to the previous. Alternatively, the fully-convolutional architecture we use is flexible in the time domain. Therefore, it permits denoising on a different length of audio than the one used during training. This allows the model to denoise an entire audio sample in one-shot – given sufficient memory availability[2].

[2]On a Titan X Pascal (12GB-VRAM) it was possible to denoise up to 25s of audio using one-shot denoising.

4 Experiments

4.1 Dataset

The used dataset [22, 29] was generated from two sources: speech data was supplied by the Voice Bank corpus [32] while environmental sounds were provided by the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [28]. The subset of the Voice Bank corpus we used features 30 native english speakers from different parts of the world reading out ≈ 400 sentences – 28 speakers are used for training and 2 for testing. Recordings are of studio quality sampled at 48kHz – and subsampled to 16kHz for this study. The subset of DEMAND that we used provides recordings in 13 different environmental conditions such as in a park, in a bus or in a cafe – 8 background-noises are mixed with speech during training and 5 background-noises are used during testing. DEMAND was produced with a 16-channel array sampled at 48kHz, however for the purposes of this work all channels were merged and subsampled to 16kHz. During training, two artificial noise classes were added – in total 10 different noise classes are available during training. Training samples are synthetically mixed at one of the following four signal-to-noise ratios (SNRs): 0, 5, 10 and 15dB with one of the 10 noise types. This results in 11,572 training samples from 28 speakers under 40 different noise conditions. Test samples are also synthetically mixed at one of the following four different SNRs: 2.5, 7.5, 12.5 and 17.5dB with one of the 5 test-noise types – resulting in 20 noise conditions for 2 speakers. As a result, the test set features 824 samples from unseen speakers and noise conditions. For both sets, the samples are on average 3 seconds long with a standard deviation of 1 second. No preprocessing to the audio (such as pre-emphasis filtering [22] or μ-law quantization[30]) is used, allowing the pipeline to be end-to-end in the strictest sense.

4.2 Basic experimental setup

The proposed model features 30 residual layers – as in Figure 2 (*Left*). The dilation factor in each layer increases in the range 1, 2, ..., 256, 512 by powers of 2. This pattern is repeated 3 times (3 stacks). Prior to the first dilated convolution, the 1-channel input is linearly projected to 128 channels by a standard 3x1 convolution to comply with the number of filters in each residual layer. The skip connections are 1x1 convolutions also featuring 128 filters – a RELU is applied after summing all skip connections. The final two 3x1 convolutional layers are not dilated, contain 2048 and 256 filters respectively, and are separated by a RELU. The output layer linearly projects the feature map into a single-channel temporal signal by using a 1x1 filter. This parameterization results in a receptive field of 6,139 samples ($\approx$ 384ms). The target field is comprised of 1601 samples ($\approx$ 100ms) – optimized to adhere our memory constraints. The relatively small size of the model (6.3 million parameters) together with its parallel inference on 1601 samples at once, results in a denoising time of ≈ 0.56 seconds per second of noisy audio on GPU. Unless explicitly stated, we assume using the following basic experimental setup: training was done with the energy-conserving loss, conditioning to speaker ID and without data augmentation. Code and trained models are available online[3].

We set as baselines for comparison: *i)* the noisy signal, and *ii)* a signal processing method based on Wiener filtering – a widely used technique for speech-denoising [34, 25] or source-separation [2, 3, 23, 11]. The baseline algorithm uses a Wiener filtering method based on a priori SNR estimation [25], as implemented here[4].

4.3 Evaluation based on computational measurements

The goal is to measure the quality of the denoised speech along three dimensions: signal distortion, background-noise interference and overall quality. To this end, we consider three measures [10]: **SIG** - predictor of signal distortion; **BAK** - background-noise intrusiveness predictor; and **OVL** - predictor of overall quality. These measures operate in a 1–5 range, aiming to computationally approximate the Mean Opinion Score (MOS) that would be produced from human perceptual trials.

Experiments were conducted to study how noise-only data augmentation, target field length, energy-conserving loss and conditioning influence the model's performance. Computed MOS measures relating to these experiments are presented together in Table 1. Italicized rows correspond to the basic experimental setup introduced above. Non-italicized rows represent a single parameter modification of this basic setup.

[3] https://github.com/drethage/speech-denoising-wavenet

[4] `https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip`

Table 1: Computed MOS measures on test set. Ranging from 1–5, higher scores are better.

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.

The basic experimental setup with no data augmentation (0 %) achieves better results across all metrics. However, informal listening clearly shows that training with 10% noise-only augmentation allows the model to produce silence in moments where no speech is present without degrading the signal, which is perceptually pleasant when aurally evaluating denoised samples. 20 % noise-only augmentation achieves lower computed MOS ratings and further listening reveals that it does not improve the quality of the denoised samples.

Moreover, we observe that training with longer target fields is crucial in achieving any significant denoising. Models trained with small target field lengths produce audio which not only fails to denoise, but also sounds fuzzier than the input. In addition, we observe that models with a small target field length require impractically long inference times (as a result of many redundant computations). Due to this, the results for smaller target field lengths in Table 1 are computed with the 20-sample perceptual test set.

Although the improvement is barely noticeable aurally, the proposed energy-conserving loss also achieves slightly better results than the standard L1 loss. This might be occurring because the energy conserving loss is simply two times the L1 loss – what is caused by the way the proposed model computes the background-noise estimate[5]. We leave for future work studying the impact of the proposed loss in cases where the background-noise is estimated through a more powerful model.

Results suggest that conditioning on speaker ID is beneficial for achieving a better speech denoising. Improvements are consistent throughout measures, although these are marginal. Informal listening confirms the results depicted by the computational measurements: the background-noise and algorithmic artifacts in conditioned samples are marginally, but noticeably reduced.

When comparing the proposed model with the baseline Wiener filtering method, one observes that OVL and SIG results are comparable, showing that Wiener filtering similarly preserves the quality of the speech signal. However, the proposed method removes the background-noise more effectively than Wiener filtering. In line with this, informal listening confirms that Wiener filtering takes small risks: no strong speech signal distortion is measured, but likewise background-noise is not heavily removed. When comparing the Wiener filtering MOS measures with the ones measured on the noisy signal, one observes that the baseline algorithm denoises without introducing algorithmic artifacts.

As seen, computational measures alone do not reveal a clear best performing configuration. Based on previously discussed informal listening and computed MOS measures, we consider the following model to achieve the best sounding results: conditioned, energy-conserving loss, with a target field of 1601 samples and 10% noise-only data augmentation.

4.4 Perceptual evaluation

Perceptual tests were conducted with 33 participants to get subjective feedback on the effectiveness of the speech-denoising Wavenet. 20 audio samples were chosen to compose the perceptual test set: 5 samples from each of the four SNRs, with an equal number of samples coming from each of the 2

[5]Since the background-noise estimate is computed by a parameterless subtraction ($\hat{b}_t = m_t - \hat{s}_t$), then: $\mathcal{L}(\hat{s}_t) = |s_t - \hat{s}_t| + |b_t - \hat{b}_t| = L_1(\hat{s}_t) + |(m_t - s_t) - (m_t - \hat{s}_t)| = L_1(\hat{s}_t) + |-s_t + \hat{s}_t| = 2 \cdot L_1(\hat{s}_t)$

speakers in the test set. Aside from these constraints, the samples were chosen randomly. Participants were presented with 4 variants of each sample: *i)* the original mix with speech and background-noise, *ii)* clean speech, *iii)* speech denoised by Wiener filtering, and *iv)* speech denoised with the best performing Wavenet – defined at the end of section 4.3. Participants were asked to *"give an overall quality score, taking into consideration both: speech quality and background-noise suppression"* for each of the last two variants. The first two variants were presented as references. Participants were able to give a score between 1–5, with a 1 being described as *"degraded speech with very intrusive background"* and a 5 being *"not degraded speech with unnoticeable background"* [10, 16]. MOS quality measurement is obtained by averaging the scores from all participants. We also compute the t-test (H_0 being that means are equal) to study whether obtained results are statistically significant or not. Table 2 presents the results of the perceptual evaluation, showing that participants significantly preferred (t-test: p-value < 0.001) the proposed method over the one based on Wiener filtering.

Table 2: Subjective MOS measures on perceptual test set. Ranging from 1–5, higher scores are better.

Measurement	**Wiener filtering**	**Proposed Wavenet**
MOS	2.92	3.60

5 Conclusion

We have presented a discriminative adaptation of Wavenet's model for speech denoising that features a non-causal and non-autoregressive architecture. This allows us to reduce the time-complexity of the model, one of the main drawbacks of Wavenet. However, removing autoregression also means that temporal continuity in the resulting signal is no longer enforced. In order to overcome this limitation, we observe that adjacent continuity is maintained by using 3x1 filters in the final layers of the model.

The proposed model is able to predict target fields instead of single samples – which further reduces its time-complexity while also significantly improving the performance of the model. In addition, the convolutional nature of the model makes it flexible in the time-dimension. As a result, it supports denoising variable-length audio – independently of how the model was trained. Therefore, the proposed model allows for one-shot denoising. This flexibility can be valuable since training and inference may be done on different hardware with varying memory availability.

Algorithmic artifacts appeared early on in our research. Switching to real-valued outputs was key in removing these artifacts introduced by the softmax loss used in Wavenet. Interestingly, note that directly operating in the raw audio domain enables considering alternative costs that can be motivated from a domain knowledge perspective.

Initially, it was also challenging for the model to generate silence. To mitigate this, we use the proposed noise-only data augmentation which has shown to be effective for this. However, the model still has its limitations, *i.e.:* its inability to deal with sudden interferences like honks in city traffic.

Although our focus is speech denoising, it is worth noting that the proposed model inherently estimates two sources: speech and background-noise – since the background-noise can be computed via subtracting the speech estimate to the input. These results are in line with recent work [33] showing that end-to-end pipelines are starting to prove effective for source separation tasks, as well.

It is important to note that no speech specific constraints are incorporated into our pipeline to overcome the aforementioned challenges. Instead, we either propose architectural improvements to our model or we propose new forms of data augmentation. The proposed model effectively denoises speech signals under noise conditions and speakers that it has never been exposed to – audio samples are available online for listening[6]. This implies that the proposed discriminative reformulation of Wavenet does not sacrifice its modeling capabilities while significantly reducing its time-complexity.

Further, our adaptation of Wavenet for speech denoising differs from the Bayesian variant [24] in many ways, particularly that: *i)* we do not explicitly manipulate probability distributions, and *ii)* that our model does not infer sequentially.

Finally, perceptual tests show that our model's estimates are preferred over the ones based on Wiener filtering. This confirms that it is possible to learn multi-scale hierarchical representations from raw audio instead of using magnitude spectrograms as front-end for the task of speech denoising.

[6]Speech and background-noise estimates: `http://jordipons.me/apps/speech-denoising-wavenet`

6 Acknowledgments

This work is partially supported by the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). We are grateful for the GPUs donated by NVidia, and also thanks to `http://foxnice.com` for hosting our demos.

References

[1] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[2] Julio J Carabias-Orti, Máximo Cobos, Pedro Vera-Candeas, and Francisco J Rodríguez-Serrano. Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings. *EURASIP Journal on Advances in Signal Processing*, 2013(1):184, 2013.

[3] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[4] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. *arXiv preprint arXiv:1704.08292*, 2017.

[5] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[6] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, 2014.

[7] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Yi Hu and Philipos C Loizou. Evaluation of objective measures for speech enhancement. In *Interspeech*, 2006.

[11] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(12):2136–2147, 2015.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[13] Anurag Kumar and Dinei Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. *arXiv preprint arXiv:1605.02427*, 2016.

[14] Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. *arXiv preprint arXiv:1703.01793*, 2017.

[15] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *arXiv preprint arXiv:1703.01789*, 2017.

[16] Philipos C Loizou. Speech quality assessment. In *Multimedia analysis, processing and communications*, pages 623–654. Springer, 2011.

[17] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[18] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[19] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[20] Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert. Convolutional neural networks-based continuous speech recognition using raw speech signal. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4295–4299, 2015.

[21] Shahla Parveen and Phil Green. Speech enhancement with missing data techniques using recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–733, 2004.

[22] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *Interspeech*, 2017.

[23] Jordi Pons, Jordi Janer, Thilo Rode, and Waldo Nogueira. Remixing music using source separation algorithms to improve the musical experience of cochlear implant users. *The Journal of the Acoustical Society of America*, 140(6):4338–4349, 2016.

[24] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Florencio Dinei, and Mark Hasegawa-Johnson. Speech enhancement using bayesian wavenet. In *Interspeech*, 2017.

[25] Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 629–632, 1996.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[27] Shin'ichi Tamura and Alex Waibel. Noise reduction using connectionist models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 553–556, 1988.

[28] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013.

[29] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA Speech Synthesis Workshop*, pages 146–152.

[30] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[31] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[32] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013.

[33] Shrikant Venkataramani and Paris Smaragdis. End-to-end source separation with adaptive front-ends. *arXiv preprint arXiv:1705.02514*, 2017.

[34] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.

[35] W Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. The microsoft 2016 conversational speech recognition system. *arXiv preprint arXiv:1609.03528*, 2016.

[36] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(1):7–19, 2015.

[37] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016.

[39] Zhenyao Zhu, Jesse H Engel, and Awni Hannun. Learning multiscale features directly from waveforms. *arXiv preprint arXiv:1603.09509*, 2016.