Speech Enhancement: An investigation with raw waveform

Yujia Yan, and Ye He

Electrical And Computer Engineering
University of Rochester

Summary

This project serves as an investigation for building a system that operates on raw audio waveforms directly.

- We proposed a lattice-ladder structured neural networks with gated dilated convolutional layers as its basic building block.
- We performed training on the dataset we built, with a lot of operations for data augmentation. We evaluated this system with unseen speeches, unseen noises with unseen room impulse response.

Our results indicate that this approach is able to produce better speech for low input quality. Due to limited time and resources and high computational burden, many properties of this kind of systems are still remained for further investigation.

Introduction

Real world speeches are noisy. Increasing the overall quality, at least intelligibility has a vast demand nowadays, in areas such as communications, hearing aids, speech recognition and content production, etc. The goal of our project is to explore both traditional statistical spectrum domain methods and methods formulated with neural networks for speech enhancement.

With the increasing popularity of convolutional neural networks which is designed and restricted to learn time-invariant operators, and with the idea of building something from scratch (Tabura Rasa), some attempts [1][2] have been made trying to directly work on time domain with raw audio waveforms and without any notion of the well-established set of basis, namely, the Fourier Transform. Directly working on the time domain may have the potential to overcome the limits (time-frequency uncertainty, phase reconstruction, etc.) of using a time-frequency representation.

In this project, we made an investigation in this direction.

Baseline: Wiener Filter

We process with the above formula frame by frame. The estimated signal at time k and frequency bin m, $S_m(k)$ is given by

$$S_m(k) = H_m(k)Y_m(k) \tag{1}$$

where

$$H_m(k) = \frac{P_{xx,m}(k)}{P_{yy,m}(k)} \tag{2}$$

Equation 2 can be reformulated with SNR term [3]

$$H_m(k) = \frac{P_{xxm}(k)}{P_{xxm}(k) + P_{nnm}(k)} = \frac{\eta_m}{1+\eta_m} \tag{3}$$

where $\eta_m = \frac{P_{xx,m}(k)}{P_{nn,m}(k)}$, the Signal-to-Noise Ratio. Then we have a smoothing equation for η_m

$$\eta_m = \alpha_\eta \frac{|S_m(k-1)|^2}{P_{nn,m}(k)} + (1-\alpha_\eta)\max(0, \gamma_m(k) - 1) \tag{4}$$

where $\gamma_m(k) = \frac{P_{yy,m}(k)}{P_{nn,m}(k)}$ is a posteriori SNR and α_η is a smoothing parameter.

Gated Dilated Convolutional Layer

We incorporate similar idea as used by Wavenet[1], but the difference is that how we apply gating. The dilated convolution is defined as

$$(x\ [*]_k\ y)[n] = \textstyle\sum_m x[m]y[n-km] \tag{5}$$

where $[*]_k$ represents dilated convolution with dilating step k, which can be intuitively explained as convolving with skip step k. There are no downsampling operations after the convolution.

Denote the two inputs and the output of our layer as x_1, x_2, y respectively,

$$\begin{aligned} g &= \sigma(w_1^{gate}[*]_k x_1 + w_2^{gate}[*]_k x_2 + b^{gate}) \\ \tilde{y} &= c\tanh(w_1^{out}[*]_k x_1 + w_2^{out}[*]_k x_2 + b^{out}) \\ y &= g \circ \tilde{y} + (1-g) \circ (x_1 + x_2) \end{aligned} \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function,c is the scale parameter, and $\circ$ is the element-wise product. g can be interpreted as the gate to determine which portions of the input and the transformed input should pass the layer.

Results

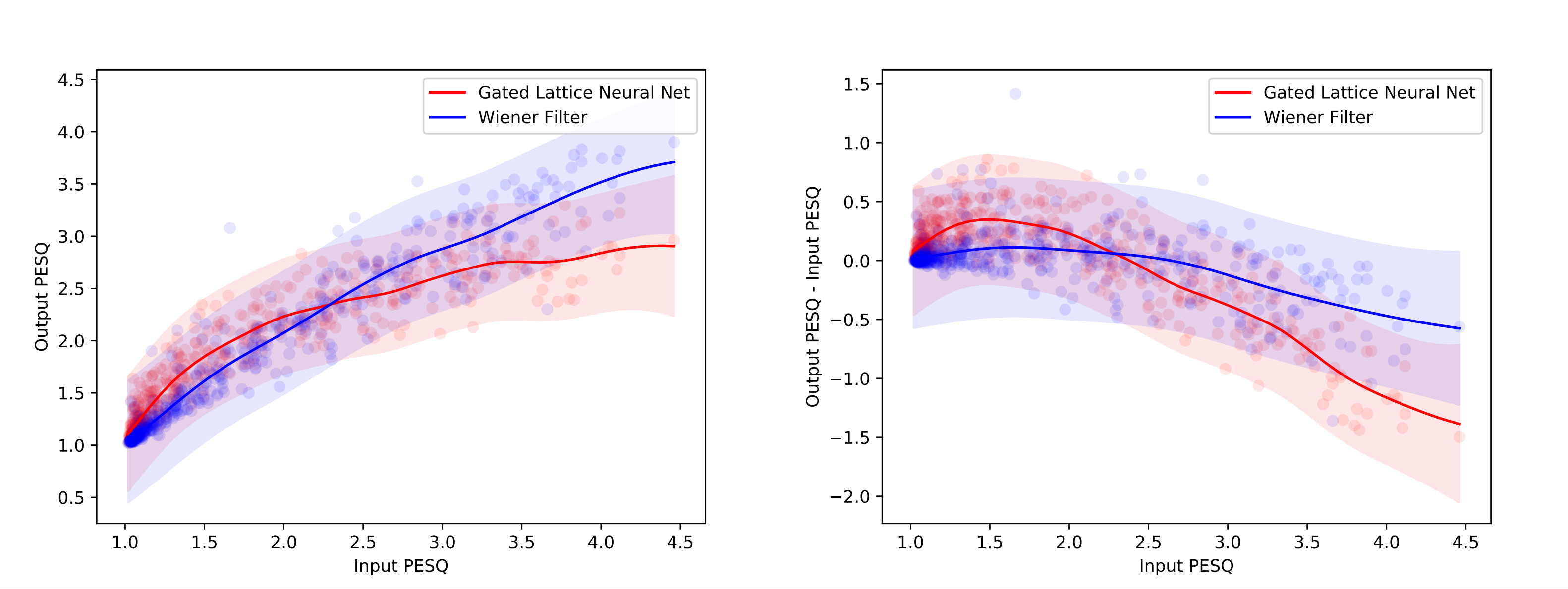

Figure 1: PESQ results: raw PESQ score versus input PESQ(above), PESQ improvement versus input PESQ(below)

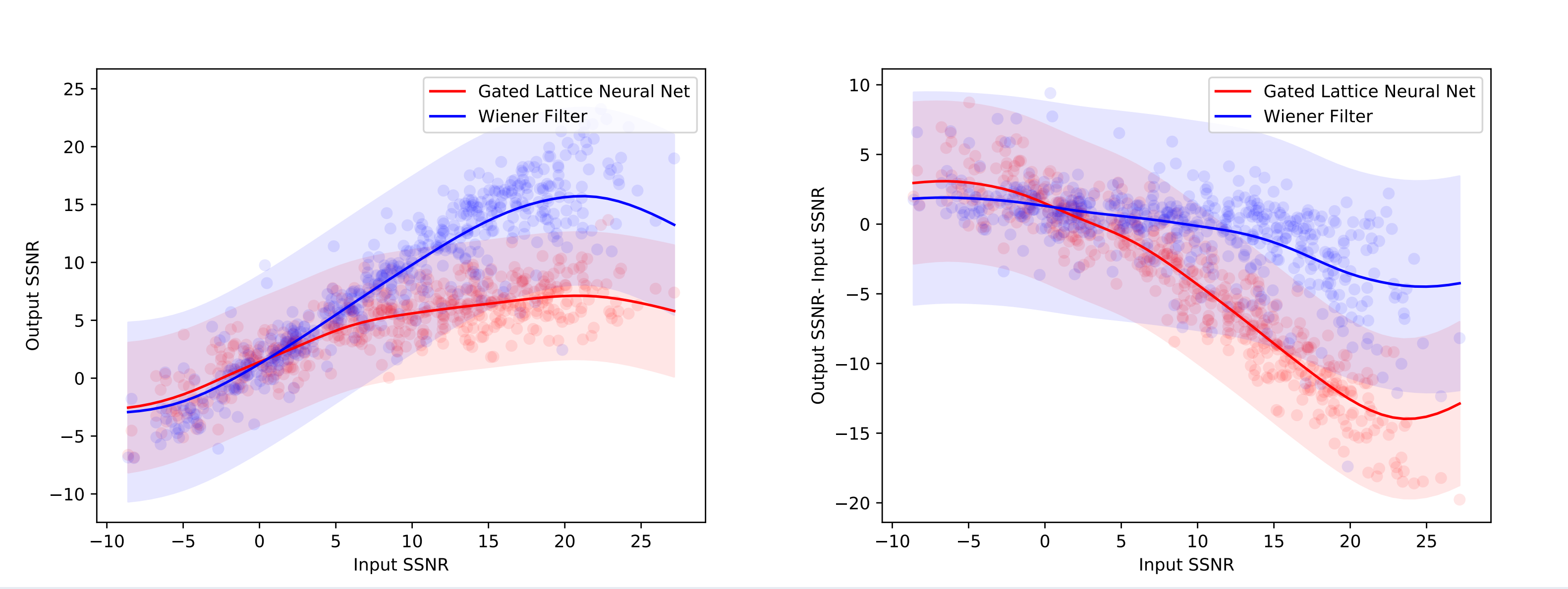

Figure 2: SSNR results: raw SSNR score versus input SSNR(above), SSNR improvement versus input SSNR(below)

Lattice-Ladder Structure

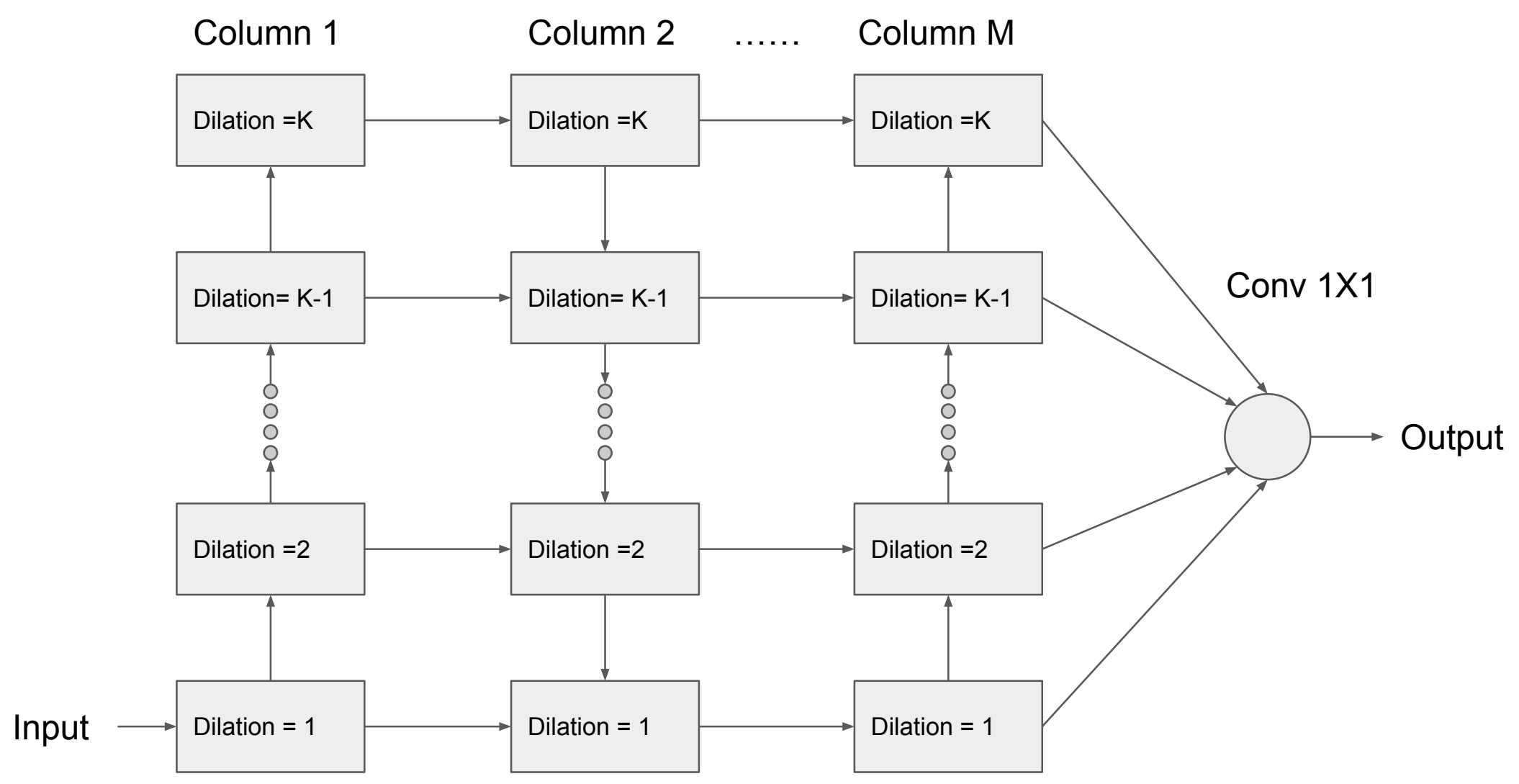

Figure 3: The lattice Architecture

The dilation step k for each dilating convolutional layers is calculated as follows

$$k = base^{dilation-1} \tag{7}$$

Conclusion

The result produced on unseen speeches, unseen noises with unseen room impulse responses suggests that our proposed model is able to outperform our baseline Wiener filter for inputs with low quality. Operating directly on raw audio waveforms is still remained for further investigation.

References

[1] Aaron 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.

[2] Santiago Pascual, Antonio Bonafonte, and Joan Serrà.
Segan: Speech enhancement generative adversarial network.
arXiv preprint arXiv:1703.09452, 2017.

[3] Yariv Ephraim and David Malah.
Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6):1109–1121, 1984.