Investigating the effect of residual and highway connections in speech enhancement models

João Felipe Santos **Tiago H. Falk**
Centre Énergie, Matériaux, Télécommunications
Institut National de la Recherche Scientifique
Montréal, QC, Canada
{jfsantos,falk}@emt.inrs.ca

Abstract

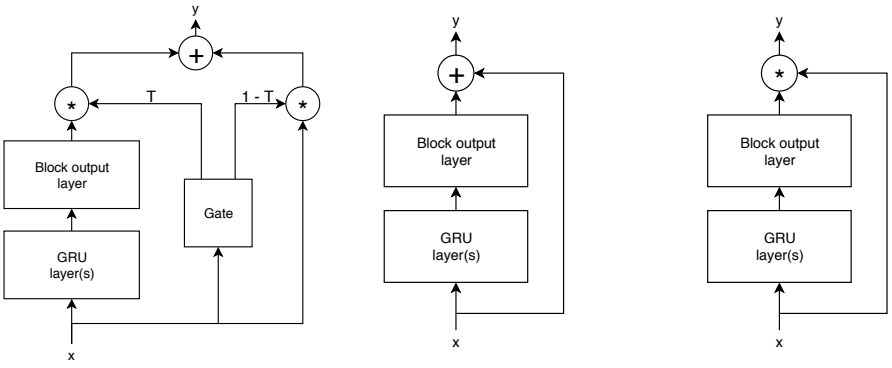
Residual and skip connections play an important role in many current generative models. Although their theoretical and numerical advantages are understood, their role in speech enhancement systems has not been investigated so far. When performing spectral speech enhancement, residual connections are very similar in nature to spectral subtraction, which is the one of the most commonly employed speech enhancement approaches. Highway networks, on the other hand, can be seen as a combination of spectral masking and spectral subtraction. However, when using deep neural networks, such operations would happen in a transformed spectral domain, as opposed to traditional speech enhancement where all operations are often done directly on the spectrum. In this paper, we aim to investigate the role of residual and highway connections in deep neural networks for speech enhancement, and verify whether or not they operate similarly to their traditional, digital signal processing counterparts. We visualize the outputs of such connections, projected back to the spectral domain, in models trained for speech denoising, and show that while skip connections do not necessarily improve performance with regards to the number of parameters, they make speech enhancement models more interpretable.

1 Introduction

Highway [Srivastava et al., 2015] and residual networks [He et al., 2015] have been proposed with the objective of improving activation and gradient flow in the training of deep neural networks. On the other hand, in tasks like image reconstruction or speech enhancement, the use of such skip connections serves a different purpose: if we model a corrupted signal $x = y + n$ as the addition of noise n to a clean signal y and x is the input to a neural network, we know that the task at hand is to predict n. In other words, to predict y, we have to alter the input x by subtracting n.

In speech enhancement, the two more commonly used approaches are spectral subtraction and spectral masking. In the first, a statistical model of n is used to predict its magnitude spectrum N, which is then subtracted from the input spectrum X to yield a clean magnitude spectrum estimate $\hat{Y}$. In spectral masking, instead of performing subtraction, we find a multiplicative mask M which aims at either blocking time-frequency cells dominated by noise (in the case of binary masks) or scaling down energies in such time-frequency cells to make them match that of the original clean signal. Recent work in speech enhancement has explored skip connections as a way of performing masking [Mimilakis et al., 2018] and spectral estimation [Santos and Falk, 2018]. Time domain approaches, such as SEGAN [Pascual et al., 2017], use a UNet-style network which employs multiple skip connections as well. Other works, such as Williamson and Wang [2017], perform spectral masking but

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) Highway (b) Residual (c) Masking

Figure 1: Diagrams for highway, residual, and masking blocks used in this paper

learn how to estimate an ideal mask instead of having the masking mechanism embedded in the neural network as a skip connection.

For better understanding of such models, we would like to understand whether there are any parallels between such connections and two traditional DSP approaches to speech enhancement, namely spectral subtraction and spectral masking. We also want to understand whether models using skip connections perform better for enhancement when such connections appear only once (resembling their DSP counterparts) or repeated as multiple blocks (like in highway and residual networks).

1.1 Residual and highway networks

Highway networks and residual networks are related, in the sense that residual networks can be seen as a special case of highway networks. A highway network block, as proposed in Srivastava et al. [2015], can be described by the following equation:

$$\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{\Theta_H}) \cdot T(\boldsymbol{x}, \boldsymbol{\Theta_T}) + \boldsymbol{x} \cdot (1 - T(\boldsymbol{x}, \boldsymbol{\Theta_T})) \tag{1}$$

Residual blocks [He et al., 2015], on the other hand, can be described by the following equation:

$$\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{\Theta_H}) + \boldsymbol{x} \tag{2}$$

which is equivalent (save for a multiplicative constant) to having a fixed gate T that outputs 0.5 for all inputs. In this paper, we additionally test another type of skip connection, which we call here a masking block. In this type of connection, the computation in the block serves only to compute a multiplicative gating function which is then applied to the input of the block, as follows:

$$\boldsymbol{y} = M(H(\boldsymbol{x}, \boldsymbol{\Theta_H}), \boldsymbol{\Theta_M}) \cdot \boldsymbol{x} \tag{3}$$

The reason for using this block is that we wanted to have a closer equivalent to spectral masking than highway networks.

In this paper, we used the above mentioned skip connections with stacked gated recurrent units (GRUs) followed by a feedforward output layer with linear activation for each block, as illustrated on figure 1.

1.2 Skip connections in speech enhancement models

Spectral subtraction Boll [1979] speech enhancement models assume distortion is additive and try to predict the magnitude spectrum of the distortion and subtract it from the magnitude spectrum of the input to yield an estimate of the clean signal:

$$\boldsymbol{y} = \boldsymbol{x} - N(\boldsymbol{x}, \boldsymbol{\Theta_N}) \tag{4}$$

$N(\boldsymbol{x}, \boldsymbol{\Theta_N})$ is a noise estimation model with parameters $\boldsymbol{\Theta_N}$. In speech enhancement, such models usually predict the noise magnitude spectrum based on several past input frames. This is very similar

to how the model based on residual blocks presented in this paper, and also the models proposed by Santos and Falk [2018] operate.

Spectral masking, on the other hand, is based on predicting time-frequency cells dominated by noise and creating a multiplicative mask that filters them out:

$$\boldsymbol{y} = \boldsymbol{x} \cdot M(\boldsymbol{x}, \boldsymbol{\Theta}_M) \tag{5}$$

Predicted masks can be either binary or ratio masks, although it has been shown that ratio masks potentially lead to better speech quality [Wang et al., 2014]. The masking model presented in the previous section, as well as the work done in [Mimilakis et al., 2018, Williamson et al., 2016] perform enhancement using masking. Highway networks, on the other hand, can be indirectly related to both spectral masking and subtraction: the output of each highway block is the sum between a masked input signal and a masked predicted signal.

2 Experiments

2.1 Datasets

For the experiments reported in this paper, we used a single speaker dataset which is publicly available[1]. It is a relatively small dataset that is comprised of the IEEE sentence list. This list contains 720 phonetically balanced sentences, separated into 72 lists with 10 sentences each, uttered by the same male speaker. The sampling rate is 16 kHz. We split the list into 680 sentences for training and validation (sentence lists 01 to 67, and 50 sentences (sentence lists 68 to 72) for testing. Both datasets can be reproduced by running the dataset generation scripts in our code repository[2].

For the denoising dataset, we mixed the training and validation sentences with noises from the DEMAND dataset [Thiemann et al., 2013] at SNRs of 12, 6, 3, 0, -3, and -6 dB. The testing sentences were mixed with four noises (babble, factory1, factory2, and volvo) from the NOISEX dataset [Varga and Steeneken, 1993], at SNRs of 13, 7, 4, 1, -2, and -5 dB. For each sentence, a random noise segment with the same length as the sentence was picked. Signal energy for speech signals were computed according to the P.56 standard (which aims at only considering energy from speech segments and discarding silence segments), while the energy for noise signals was computed by its overall RMS value. For each (noise type, SNR) pair, we mixed all sentences in the training + validation set or the test set, accordingly. The training, validation, and test sets have 64923, 3417, and 1200 sentences each.

The dereverberation dataset was constructed in a similar way, but using simulated room impulse responses (RIR) obtained using the fast Image-Source Method [Lehmann and Johansson, 2010]. We generated 20 RIRs for each reverberation time in the range 0.2 to 2.0 s, in 0.05 increments, and convolved 50 sentences to each RIR (given the large number of RIRs, we did not use the entire dataset for each). The training, validation, and test sets have 35150, 1850, and 3700 sentences each.

2.2 Model architectures, hyperparameters, and training

For training all models we used the same input representation: the log-magnitude spectrum of the signal's short-time Fourier transform with a window of 32 ms (512 samples) and 50% overlap, which corresponds to a dimensionality of 257 coefficients per frame. The inputs were standardized, but the outputs were not changed.

All models used 3 layers of gated recurrent units (GRUs) with 256 hidden units each. We tried both having a single skip connection wrapping the 3 GRU layers as a single residual or highway block, and having 3 blocks with one GRU layer each. All models were trained using Adam [Kingma and Ba, 2014] with standard hyper-parameters for 100 epochs, with a mini-batch size of 32.

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

[2] `https://github.com/jfsantos/irasl2018`

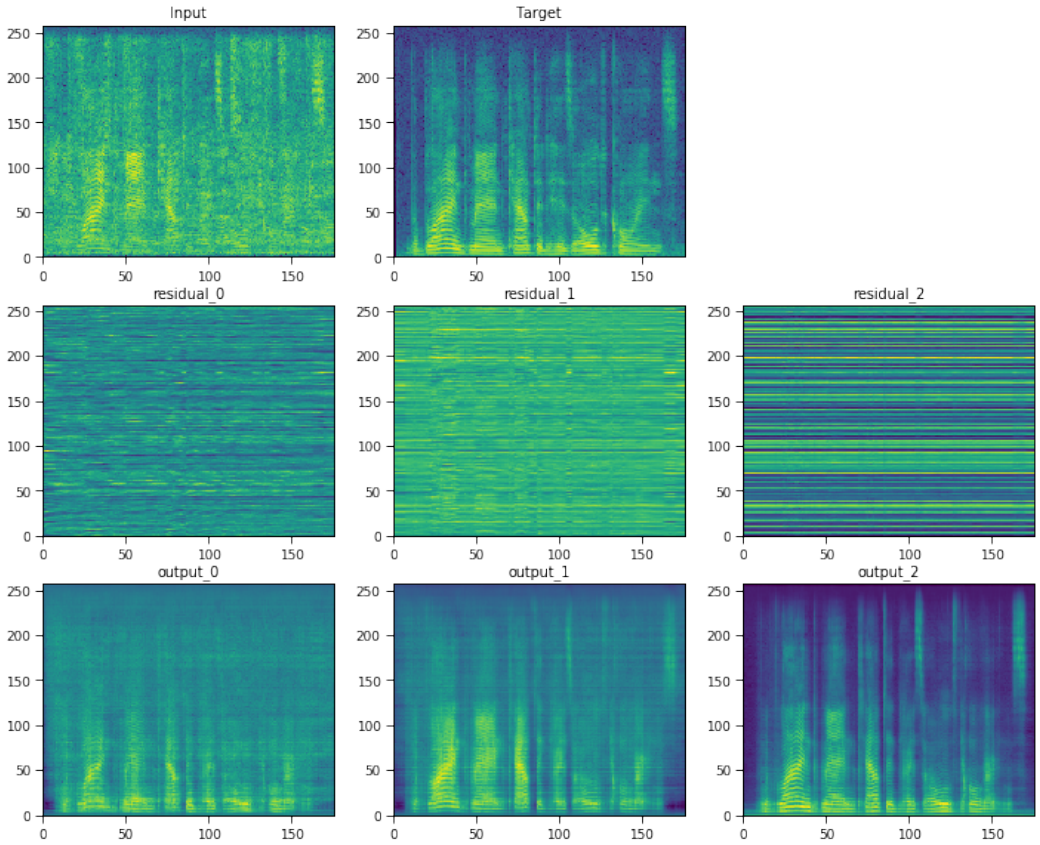


Figure 2: Outputs of the residual model

3 Analysis

3.1 Visualizing the role of skip connections

To understand how each type of skip connection and each block in a model works, we mainly used visualization of each output of a block. Since skip connections perform operations between the output of each block and its input, we can assume all operations happen in the domain of the transformed input, and can therefore be visualized by using the output layer of the model to bring them back to the STFT domain.

Figure 2 shows the intermediate and final outputs of the residual model with 3 blocks for a sentence corrupted by babble noise at 1 dB. The first row shows the input and respective target, the second row the residuals predicted in all blocks, and the last row the output of the residual block. The x and y axes represent the time and frequency bin respectively, with the color representing the log-magnitude (with blue representing low values, green intermediate values, and yellow high values). We can see that earlier blocks exhibit less frequency selectivity than later blocks, with the last block specializing strongly in certain frequencies. From block to block, we also see how the strongest components of the signal take shape, starting at the lower frequencies, followed by mid-bands and finally higher frequencies in the last stage.

Figure 4 shows a similar plot for the masking model, with the second row representing the predicted masks. Since we chose to use linear masks instead of bounded masks (such as the outputs of a tanh or sigmoid activation), these masks can flip the sign of time-frequency cells, which makes visualization a bit more complicated to follow, therefore we used a different colormap for plotting the mask, where values close to zero are shown in white, negative numbers in blue, and positive numbers in red. From the intermediate outputs, we can see in the first two plots that the model is

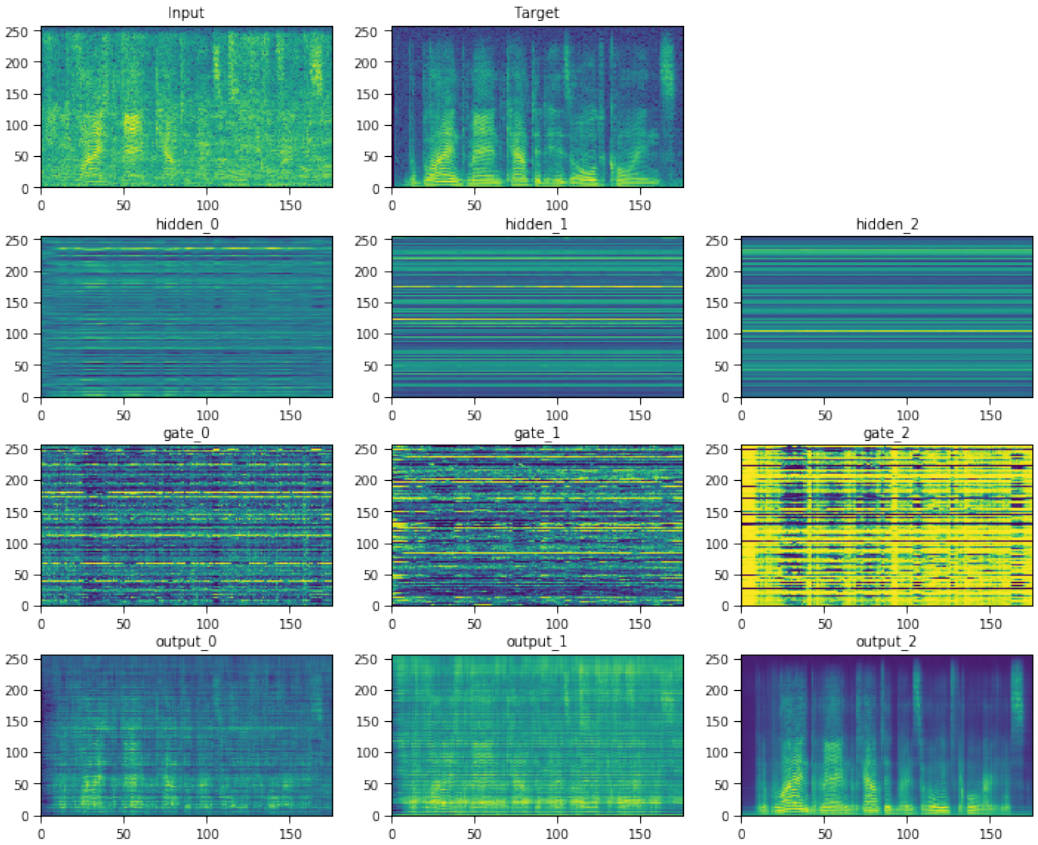


Figure 3: Outputs of the highway model

increasingly improving the predicted speech structure in all layers, but the signs of many frequency bands might be flipped before we reach the last layer.

Figure 3 shows a similar plot for highway models, with the second and third rows representing the outputs of $\boldsymbol{H}(\boldsymbol{x})$ and $\boldsymbol{T}(\boldsymbol{x})$, respectively. The interpretation of this model is more complicated since the output is a linear combination of the elementwise products of $\boldsymbol{T}(\boldsymbol{x})$ and $1 - \boldsymbol{T}(\boldsymbol{x})$ by $H(x)$ and x, respectively. We can interpret $T(x)$ in these plots as showing us which elements of $H(x)$ can we trust more than the current input, which those being represented by green and yellow (stronger confidence), while time-frequency cells marked in dark blue will be dominated by the input of the current block. We can see the first block selects regions where the speech signal is stronger than the noise as having a low value for $T(x)$, and decides to reject the input and use its own internal estimate for a number of bands. In the last layer, though, the model uses its internal estimate for a large portion of the signal, as evidenced by the predominance of yellow in the last output for $T(x)$. This seems counter-intuitive, especially as we notice that the output of the second blocks seems to have a larger amount of noise than that of the first block.

Although all results reported here take into account a single example for the denoising tests, these characteristics are common to different sentences and distortion types. The same observations are also valid for the reverberation models, although these use a different set of models so specific frequency ranges might not be the same but the observations in general still hold.

3.2 Objective quality and intelligibility assessment

We also evaluated the models using the objective speech quality and intelligibility metrics PESQ and STOI [Taal et al., 2010]. For both metrics, higher scores are better. As can be seen in figures 5-8, for most of the models there is not a large improvement over the baseline. We can argue that a positive aspect of the use of residual or masking models is that they are more interpretable than the

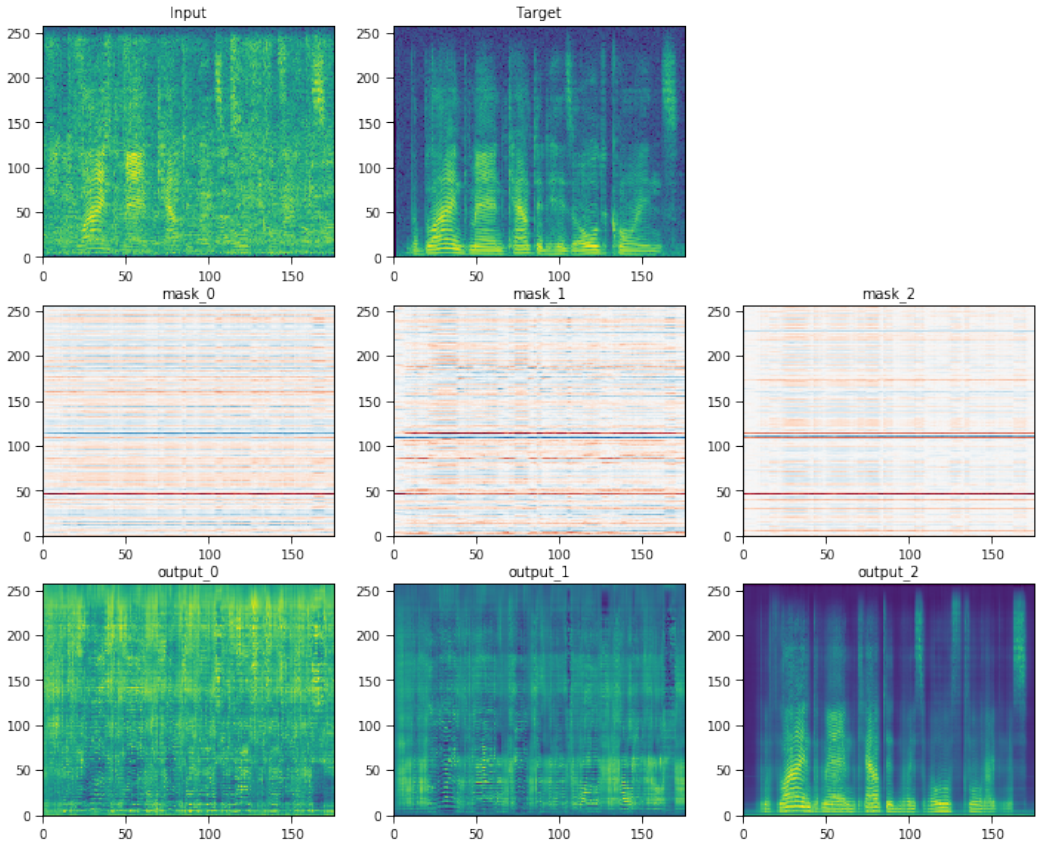


Figure 4: Outputs of the masking model

baseline. It should also be noted that direct comparison between these models is not completely fair since the models tested have different numbers of parameters, with the baseline model having the least parameters, followed by masking, residual, and highway, and models with 3 blocks have more parameters than their single block counterparts.

We also note that these models are not competitive with state-of-the-art models presented more recently in the literature. In this preliminary study, we looked into these models as building blocks for more advanced models who achieve higher performance in these tasks, such as those previously cited.

4 Conclusion

This paper shows early results of our investigation on the role of skip connections in speech enhancement models. Our preliminary experiments show that, although they have no significant impact in the performance of the models, such connections might help making the models more interpretable, as we can identify the contribution of each individual layer to the task. In the future, we intend to investigate more complex models, such as models based on the UNet architecture, as well as models that employ a temporal context window at the input instead of a single frame (such as the work in Santos and Falk [2018]), since those are more in line with state-of-the-art models in the literature.

Acknowledgments

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for their financial support. The authors also thank Kyle Kastner for the many discussions on the topic of speech enhancement and interpretability.

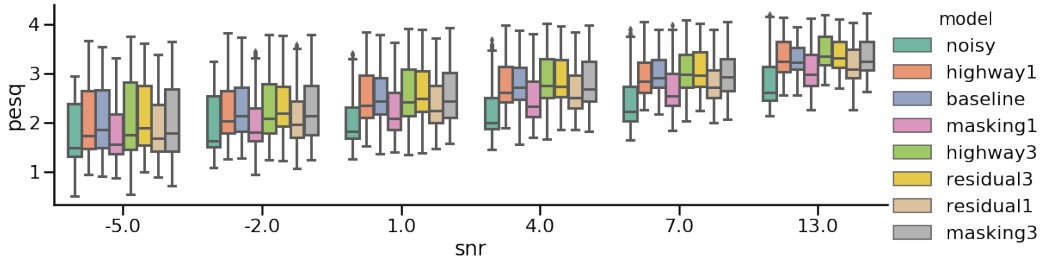


Figure 5: Denoising models - PESQ

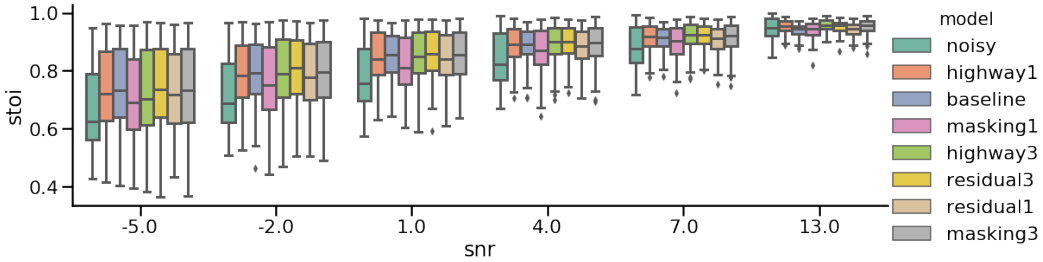


Figure 6: Denoising models - STOI

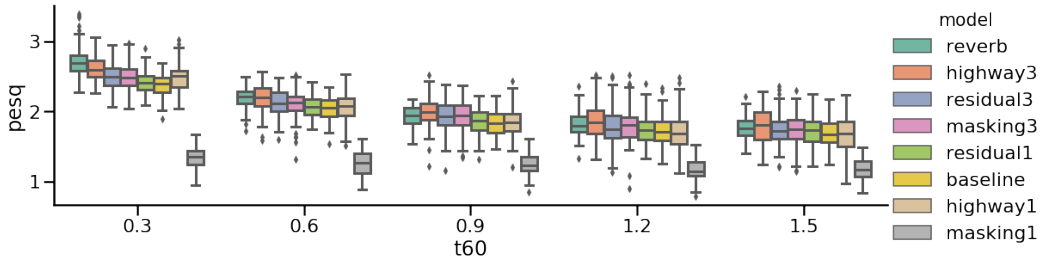


Figure 7: Dereverb models - PESQ

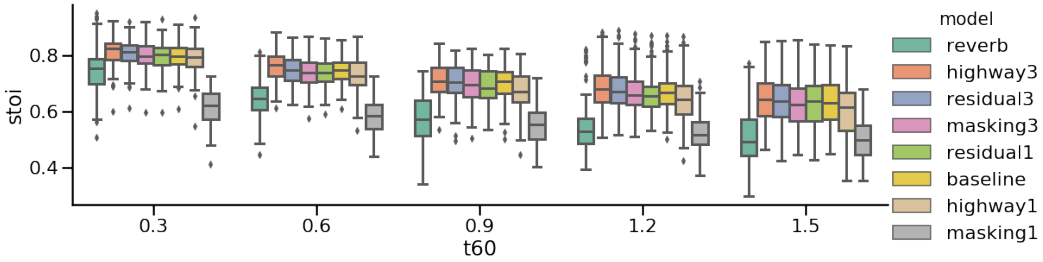


Figure 8: Dereverb models - STOI

References

S. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, volume 4, pages 200–203, April 1979. doi: 10.1109/ICASSP.1979.1170696. bibtex: boll_spectral_1979.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL `http://arxiv.org/abs/1512.03385`. arXiv: 1512.03385.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. URL `http://arxiv.org/abs/1412.6980`. arXiv: 1412.6980 bibtex: kingma_adam:_2014.

Eric A. Lehmann and Anders M. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1429–1439, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5299028`.

Stylianos Ioannis Mimilakis, Konstantinos Drossos, João Felipe Santos, Gerald Schuller, Tuomas Virtanen, and Yoshua Bengio. Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018. URL `http://arxiv.org/abs/1711.01437`.

Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech Enhancement Generative Adversarial Network. *arXiv:1703.09452 [cs]*, March 2017. URL `http://arxiv.org/abs/1703.09452`. arXiv: 1703.09452.

J. F. Santos and T. H. Falk. Speech Dereverberation With Context-Aware Recurrent Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1236–1246, July 2018. ISSN 2329-9290. doi: 10.1109/TASLP.2018.2821899.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training Very Deep Networks. *arXiv:1507.06228 [cs]*, July 2015. URL `http://arxiv.org/abs/1507.06228`. arXiv: 1507.06228.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4214–4217. IEEE, 2010. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5495701`.

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, May 2013. ISSN 0001-4966. doi: 10.1121/1.4806631. URL `http://scitation.aip.org/content/asa/journal/jasa/133/5/10.1121/1.4806631`.

Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, July 1993. ISSN 0167-6393. doi: 10.1016/0167-6393(93)90095-3. URL `http://www.sciencedirect.com/science/article/pii/0167639393900953`.

Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22 (12):1849–1858, 2014. URL `http://dl.acm.org/citation.cfm?id=2719964`.

D. S. Williamson and D. Wang. Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(7):1492–1501, July 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2017.2696307.

Donald S. Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3): 483–492, 2016. URL `http://ieeexplore.ieee.org/abstract/document/7364200/`.