Deep Residual-Dense Lattice Network for Speech Enhancement

Paper ID: 6844
Supplementary Material

Network Details

The configurations of the ResNets, ResLSTMs, DenseNets, and DenseRNets are given in Table 1.

Speech enhancement

Further enhanced speech spectrograms produced by the RDL-Nets and their counterparts are provided in Figures 1, 2, 3, and 4.

References

Germain, F. G.; Chen, Q.; and Koltun, V. 2018. Speech denoising with deep feature losses. *arXiv preprint arXiv:1806.10522*.

Nicolson, A., and Paliwal, K. K. 2019. Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*.

Pascual, S.; Bonafonte, A.; and Serra, J. 2017. SEGAN: Speech enhancement generative adversarial network. In *Interspeech*.

Rethage, D.; Pons, J.; and Serra, X. 2018. A Wavenet for speech denoising. In *ICASSP*, 5069–5073. IEEE.

Scalart, P., et al. 1996. Speech enhancement based on a priori signal to noise estimation. In *ICASSP*, volume 2, 629–632. IEEE.

Table 1: Configurations of the networks used within the Deep Xi (Nicolson and Paliwal 2019) framework (RDL-Net configuration described in main text).

	# convs units / LSTM size per block	# residual blocks	# dense blocks	# denseR blocks	# params.
ResNet (conv output size=64)	2/-	20	-	-	0.53M
		40	-	-	1.03M
		60	-	-	1.53M
		80	-	-	2.03M
ResLSTM	-/170	4	-	-	1.02M
	-/188	5	-	-	1.51M
	-/200	6	-	-	2.03M
DenseNet (*growth rate*=24)	4/-	-	5	-	0.57M
		-	7	-	0.97M
		-	9	-	1.48M
		-	11	-	2.10M
DenseRNet (*growth rate*=24)	8/-	8	-	2	0.60M
		12	-	3	1.05M
		16	-	4	1.44M
		24	-	6	2.02M

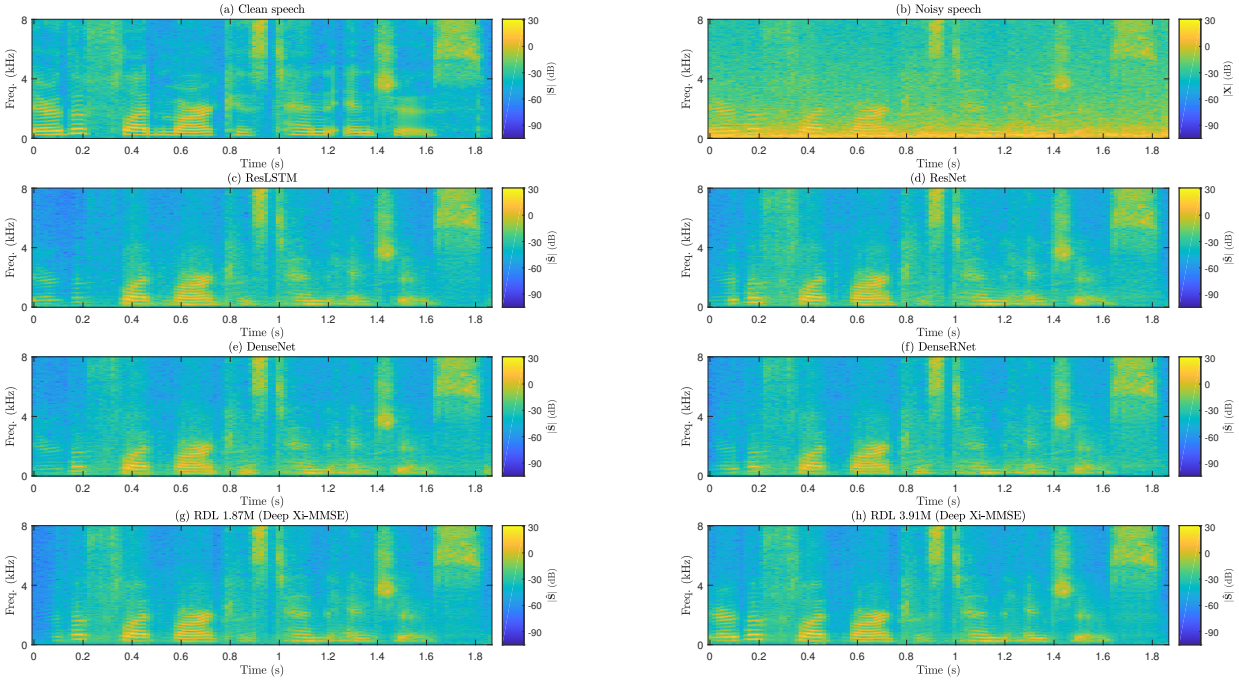

Figure 1: (a) Clean speech magnitude spectrogram ($|\mathbf{S}|$). (b) A recording of *street-music* mixed with (a) at an SNR level of -5 dB ($|\mathbf{X}|$). Enhanced speech ($|\hat{\mathbf{S}}|$) produced by (c) ResLSTM 2.03M, (d) ResNet 2.03M, (e) DenseNet 1.94M, (f) DenseRNet 2.02M, (g) RDL-Net 1.87M (Deep Xi-MMSE), and (h) RDL-Net 3.91M (Deep Xi-MMSE).

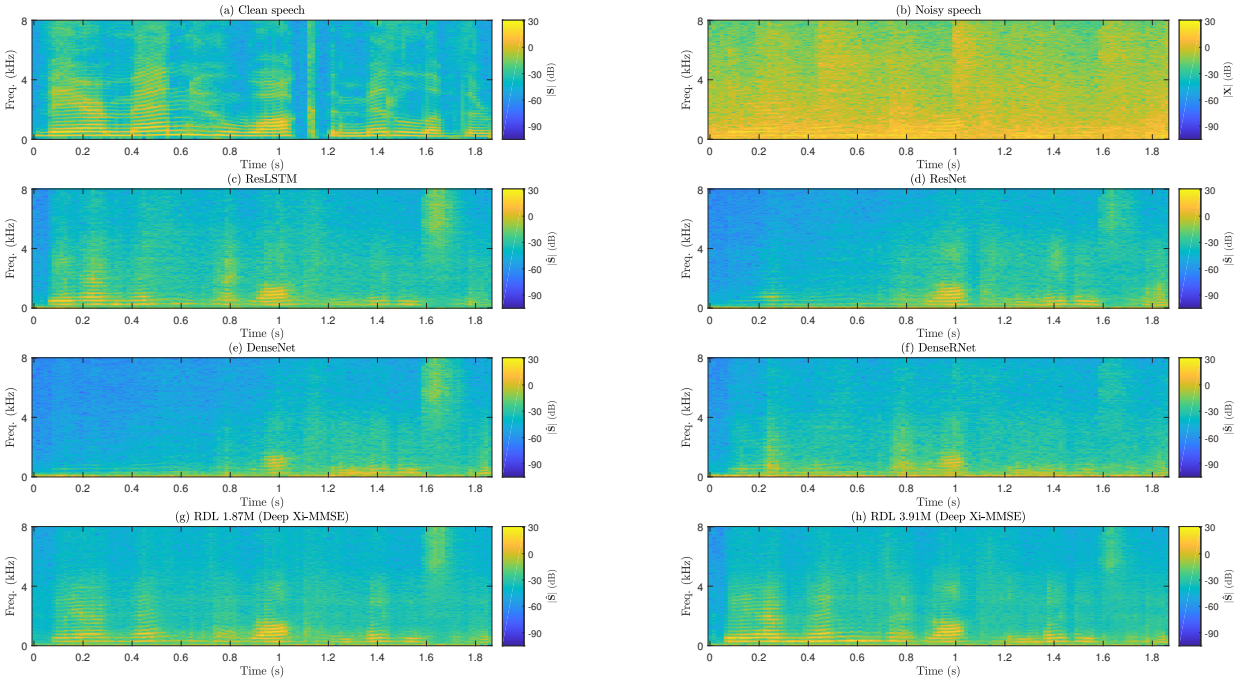

Figure 2: (a) Clean speech magnitude spectrogram ($|\mathbf{S}|$). (b) A recording of *factory* noise mixed with (a) at an SNR level of -5 dB ($|\mathbf{X}|$). Enhanced speech ($|\hat{\mathbf{S}}|$) produced by (c) ResLSTM 2.03M, (d) ResNet 2.03M, (e) DenseNet 1.94M, (f) DenseRNet 2.02M, (g) RDL-Net 1.87M (Deep Xi-MMSE), and (h) RDL-Net 3.91M (Deep Xi-MMSE).

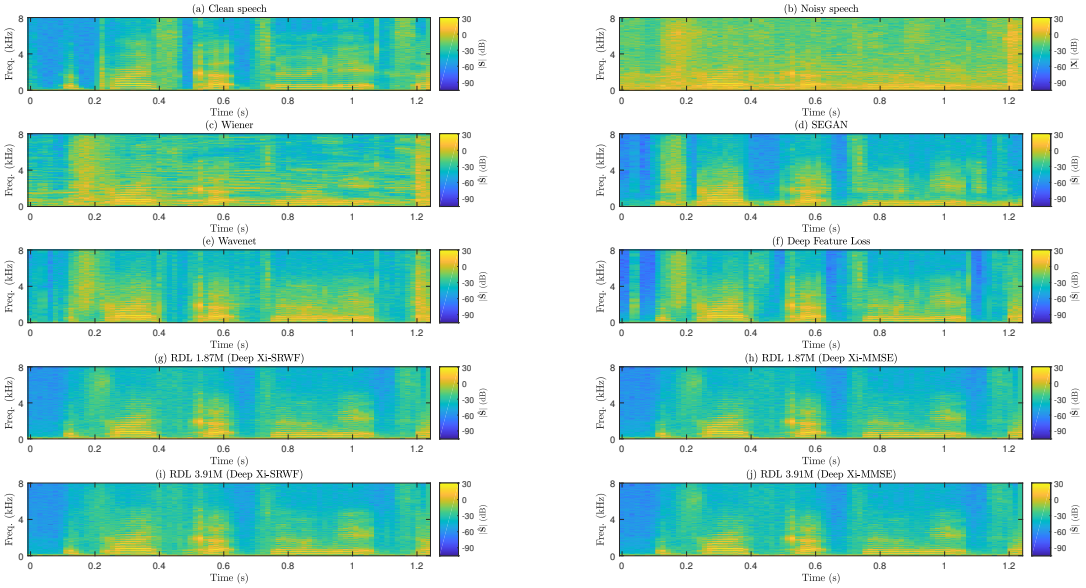

Figure 3: (a) Clean speech magnitude spectrogram ($|\mathbf{S}|$) of female *p*257 uttering sentence 70, "The price cuts are really exciting". (b) A recording of *crowd noise* mixed with (a) at an SNR level of 2.5 dB ($|\mathbf{X}|$). Enhanced speech ($|\hat{\mathbf{S}}|$) produced by (c) Wiener (Scalart and others 1996), (d) SEGAN (Pascual, Bonafonte, and Serra 2017), (e) Wavenet (Rethage, Pons, and Serra 2018), (f) Deep Feature Loss (Germain, Chen, and Koltun 2018), (g) RDL-Net 1.87M (Deep Xi-SRWF), (h) RDL-Net 1.87M (Deep Xi-MMSE), (i) RDL-Net 3.91M (Deep Xi-SRWF), (j) RDL-Net 3.91 (Deep Xi-MMSE).

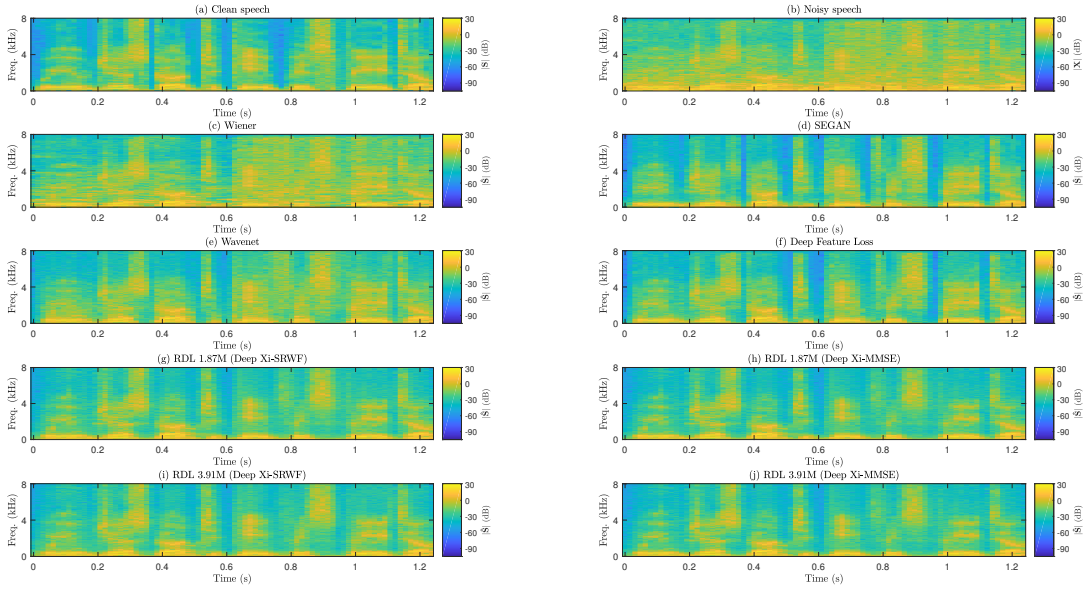

Figure 4: (a) Clean speech magnitude spectrogram ($|\mathbf{S}|$) of male *p*232 uttering sentence 415, "We have just got to keep speed on the ground". (b) A recording of *crowd noise* mixed with (a) at an SNR level of 2.5 dB ($|\mathbf{X}|$). Enhanced speech ($|\hat{\mathbf{S}}|$) produced by (c) Wiener (Scalart and others 1996), (d) SEGAN (Pascual, Bonafonte, and Serra 2017), (e) Wavenet (Rethage, Pons, and Serra 2018), (f) Deep Feature Loss (Germain, Chen, and Koltun 2018), (g) RDL-Net 1.87M (Deep Xi-SRWF), (h) RDL-Net 1.87M (Deep Xi-MMSE), (i) RDL-Net 3.91M (Deep Xi-SRWF), (j) RDL-Net 3.91 (Deep Xi-MMSE).