Experimental analysis of ablation studies

1. Stage1 vs Stage2 vs Two-stages

- Complexity

Methods	Para	Flops
Stage1	16K	2.1M
Stage2	62K	7.8M
Two-stages	78K	10M

- PESQ

Methods	-5db	0db	5db	10db	15db	Avg.
Noisy	1.43	1.74	2.16	2.42	2.81	2.12
Stage1(MSE)	1.74	2.00	2.45	2.61	2.97	2.35
Stage2(MSE)	1.53	1.76	2.31	2.53	2.95	2.26
Stage2(-SISNR)	1.40	1.72	2.30	2.51	2.99	2.23
Stage2(-SISNR+0.02*MSE)	1.73	1.99	2.49	2.66	2.98	2.37
Two-stages(MSE)	1.80	2.10	2.57	2.80	3.11	2.50
Two-stages(-SISNR)	1.75	2.07	2.55	**2.83**	**3.2**	2.48
Two-stages(-SISNR+0.02*MSE)	**1.85**	**2.27**	**2.63**	**2.83**	3.12	**2.56**

From the above table, we can see that:

① The model complexity (memory, amount of operations) of stage 2 is much higher than that of stage 1.

② If the two stages of the model are implemented separately for speech enhancement, stage 1 performs better overall, while stage 2 is more dependent on the setting of the loss function.

③ In terms of speech perception quality, we trained -SISNR and MSE, respectively, on the neural network of stage 2. From the results, the -SISNR loss function did not work well, and it scored lower than MSE. The reason is not clear.

④ The whole system (two stages)shows a similar pattern to the 3th point, which is that there is no advantage to using -SISNR alone.

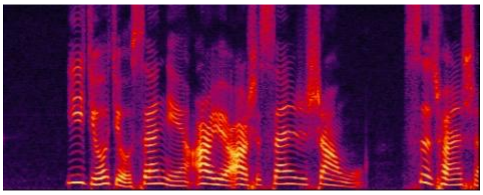
⑤ But by combining -SISNR (time domain loss) and MSE (frequency domain loss) to train either stage 2 or the whole system, both stage 2 and the whole system perform better than either loss function alone. This is why in the paper we chose to combine the two loss functions.

Additional notes:

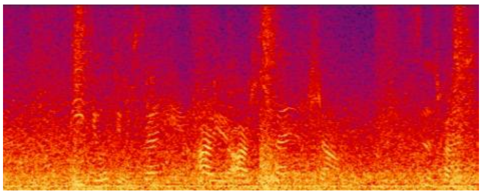
Exploring the specific effects of the time-domain and frequency-domain loss functions on the proposed system is not really the focus of this paper, which aims to provide researchers in speech enhancement with a scheme that can effectively reduce the complexity of the system. In the future, we will further analyze the specific reasons for this in the journal (more pages,more statistical charts and tables).

2. Proposed loss

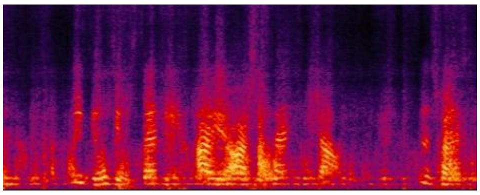
Since the proposed loss function is always used in stage 1, we only use stage 1 for the experiment.



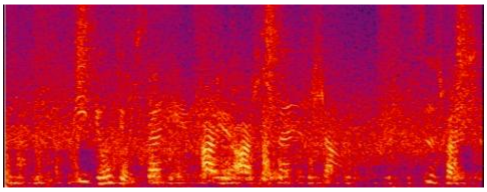
Clean



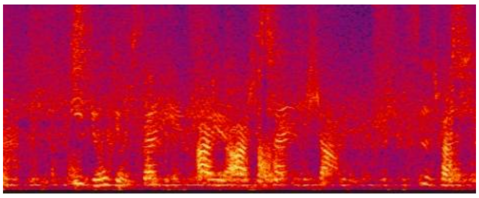
Noisy(-5db factory noise)



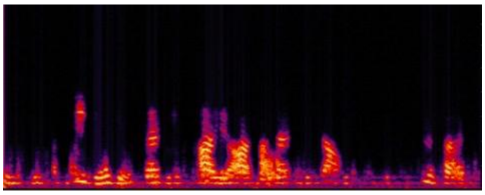
MSE



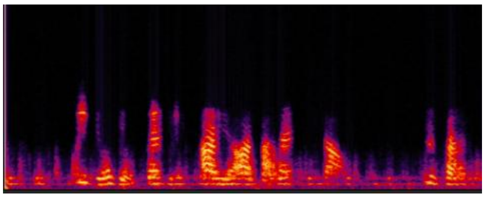
$\mathcal{L}_C$ **weight=0.4**



$\mathcal{L}_C$ **weight=0.7**



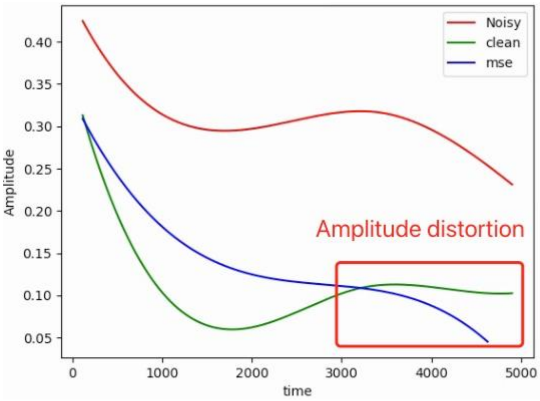
$\mathcal{L}_N$ **weight=0.4**



$\mathcal{L}_N$ **weight=0.7**

From the above figure, we can see that:

① The use of MSE acting on stage 1 alone may lead to problems of speech distortion, with some speech harmonic structures having disappeared.For a multi-stage systems, the speech distortion problem caused by stage 1 is fatal.



② The proposed loss function can control the training direction of stage 1 (retain more or suppress more).

③ The $\mathcal{L}_C$ loss function is more "preserving" and the noise is less suppressed, while $\mathcal{L}_N$ is the opposite.

④ The $\mathcal{L}_C$ loss function or $\mathcal{L}_N$ loss function is only suitable for use in multi-stage noise reduction systems, and Mse is still a reliable loss function in single-stage systems.

One-stage model loss function comparison(proposed vs MSE）

Methods	PESQ(Avg.)
Noisy	2.12
MSE	**2.35**
$\mathcal{L}_N$ **weight =0.4**	2.01
$\mathcal{L}_N$ **weight =0.7**	2.17
$\mathcal{L}_C$ **weight =0.4**	2.20
$\mathcal{L}_C$ **weight =0.7**	2.26