# SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement

Tian Gao[1], Jun Du[1], Li-Rong Dai[1] and Chin-Hui Lee[2]

[1]NELSLIP, University of Science and Technology of China, Hefei, Anhui, China [2]Georgia Institute of Technology, Atlanta, Georgia, USA

## What's new in this study

**Speech enhancement in low SNR environments via**:

- SNR-based problem decomposition
- Progressive learning
- Compact DNN with less parameters

## Background

**Deep learning for speech enhancement**:

- Learning time-frequency mask (IBM or IRM) as classification (Wang *et, al.*, 2014)
- Learning target spectra as regression (Xu *et, al.*, 2014, 2015) with a classical DNN configuration (Figure 1): 1799(257*7)_2048_2048_2048_257, 12.6M parameters
- Learning soft mask as regression (Huang *et, al.*, 2014; Weninger *et, al.*, 2014)

**Challenge**:

- One challenge is the performance degradation in low SNR environments.

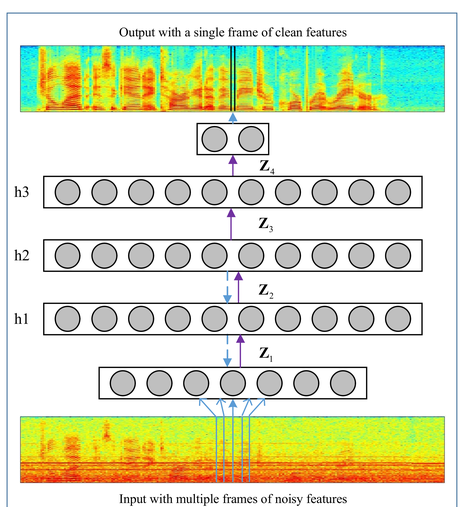



Figure 1: *Regression DNN-based speech enhancement*

## Progressive Learning

- **Method**: The direct mapping process is decomposed into multiple stages with an SNR gain achieved in each stage as shown in Figure 5.

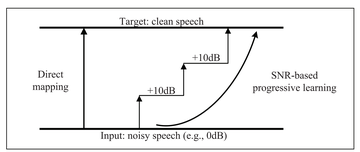



Figure 2: *Illustration of SNR-based progressive learning.*

## DNN implementation

We guide hidden layers to learn targets explicitly, as shown in Figure 3:

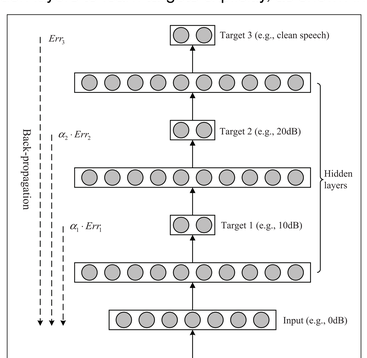



Figure 3: *DNN architecture for progressive learning.*

- **DNN**: 1799(257*7)_2048_257_2048_257_2048_257, 6.3M
- **forward pass**: linear active function in the target layers
- **backward pass**: objective function defined for the 3 targets ($Err_1$, $Err_2$, $Err_3$):

$$Err = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{X}_n^t - X_n^t\|_2^2) \quad (1)$$

back-propagated gradients in a weighted sum fashion as:

$$\boldsymbol{\epsilon} = \frac{\partial(Err_3)}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}_{1\le\ell\le L_3+1} + \alpha_2 \frac{\partial(Err_2)}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}_{1\le\ell\le L_2+1} + \alpha_1 \frac{\partial(Err_1)}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}_{1\le\ell\le L_1+1} \quad (2)$$

## Experiment configuration

**Training data**:

- clean speech: WSJ0 (about 12h)
- noise: 115 noise types
- SNR configuration: Table 1

Table 1: Target SNR configurations for progressive learning

| Input | Target 1 | Target 2 | Target 3 |
|---|---|---|---|
| -5dB | 5dB | 15dB | clean speech |
| 0dB | 10dB | 20dB | clean speech |
| 5dB | 15dB | 25dB | clean speech |

**Testing configuration**:

- three unseen noises from the NOISEX-92 corpus: babble, factory and destroyer engine

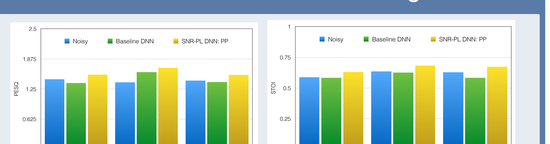
**Post-processing**: average multiple estimated features to further improve the overall performance

## Results: Single-SNR training

Table 2: A detailed PESQ and STOI comparison of different single-SNR training systems at 0dB SNR on the test set of three unseen noise environments (N1: Babble, N2: Factory, N3: Destroyer engine), among: Noisy, DNN baseline, estimations of different levels of SNR and SNR-based progressive learning combined with post-processing (denoted as SNR-PL DNN: PP).

| | N1 (0dB) | | N2 (0dB) | | N3 (0dB) | |
|---|---|---|---|---|---|---|
| System | PESQ | STOI | PESQ | STOI | PESQ | STOI |
| Noisy | 1.683 | 0.711 | 1.689 | 0.757 | 1.636 | 0.749 |
| **Baseline DNN (12.6M)** | **1.775** | **0.710** | **1.875** | **0.702** | **1.760** | **0.694** |
| SNR-PL DNN: Out1 | 1.828 | 0.730 | 1.850 | 0.764 | 1.693 | 0.763 |
| SNR-PL DNN: Out2 | 2.015 | 0.747 | 2.023 | 0.764 | 1.866 | 0.757 |
| SNR-PL DNN: Out3 | 1.789 | 0.731 | 1.894 | 0.722 | 1.760 | 0.710 |
| **SNR-PL DNN: PP (6.3M)** | **2.007** | **0.766** | **2.017** | **0.783** | **1.928** | **0.781** |

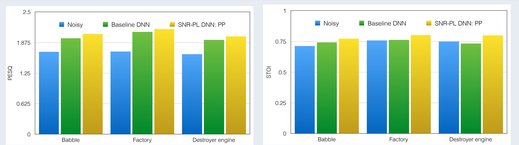
## -5dB Results: Multi-SNR training




(a) PESQ (b) STOI

Figure 4: *PESQ and STOI comparison for multi-SNR training system at -5dB.*

## 0dB Results: Multi-SNR training




(a) PESQ (b) STOI

Figure 5: *PESQ and STOI comparison for multi-SNR training system at 0dB.*

## Demo

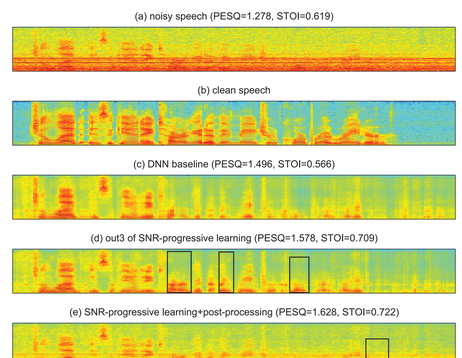



Figure 6: *Spectrograms of an utterance corrupted by Destroyer engine noise at -5dB SNR and enhanced by multi-SNR training: (a) noisy speech, (b) clean speech, (c) DNN baseline (PESQ=1.496, STOI=0.566); (d) out3 in the proposed DNN (PESQ=1.578, STOI=0.709); (e) further post-processing (PESQ=1.628, STOI=0.722).*

## Conclusion

- A novel SNR-based progressive learning framework was proposed for DNN based speech enhancement.
- It was implemented by guiding hidden layers in the DNN architecture to learn targets explicitly.
- It can improve performance in low SNR environments and reduce parameters by 50%.