A Study of Training Targets for Deep Neural Network-Based Speech Enhancement Using Noise Prediction

Babafemi O. Odelowo and David V. Anderson

Center for Signal & Image Processing

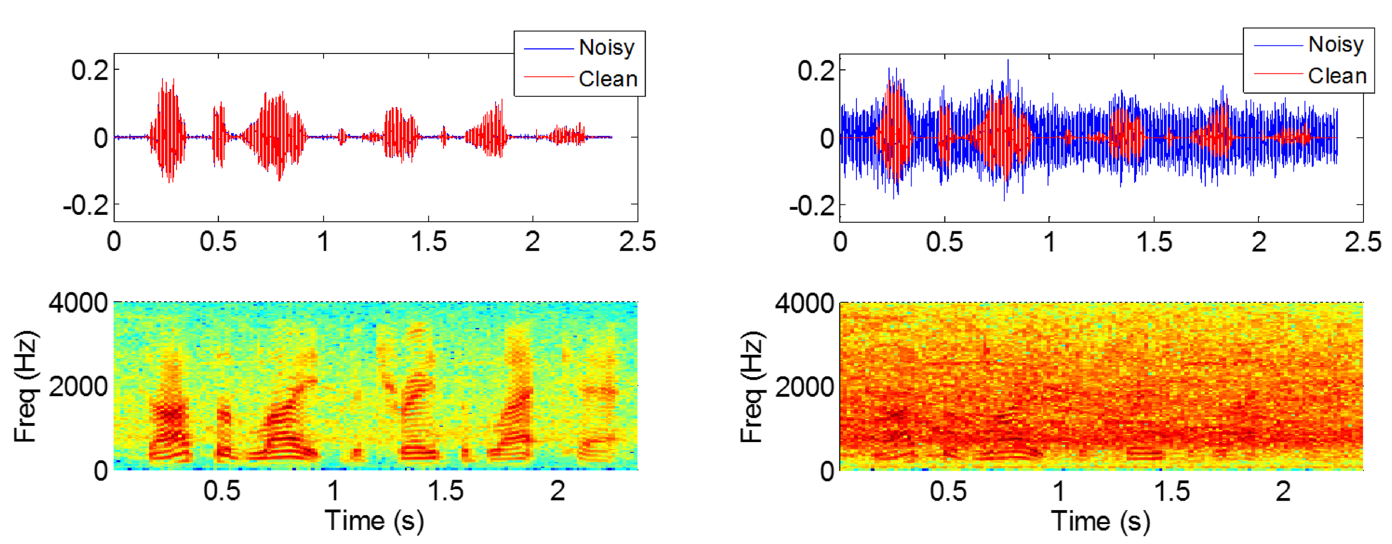
Objective

Evaluate the performance of different training targets for deep neural network (DNN) speech enhancement based on noise prediction

Compare the performance of the speech enhancement systems based on noise prediction to that of a conventional SE system based on prediction of clean speech.

Noise Prediction Rationale

Spectrograms of Noisy and Clean Speech Signals



(a) 20dB SNR (b) -5dB SNR

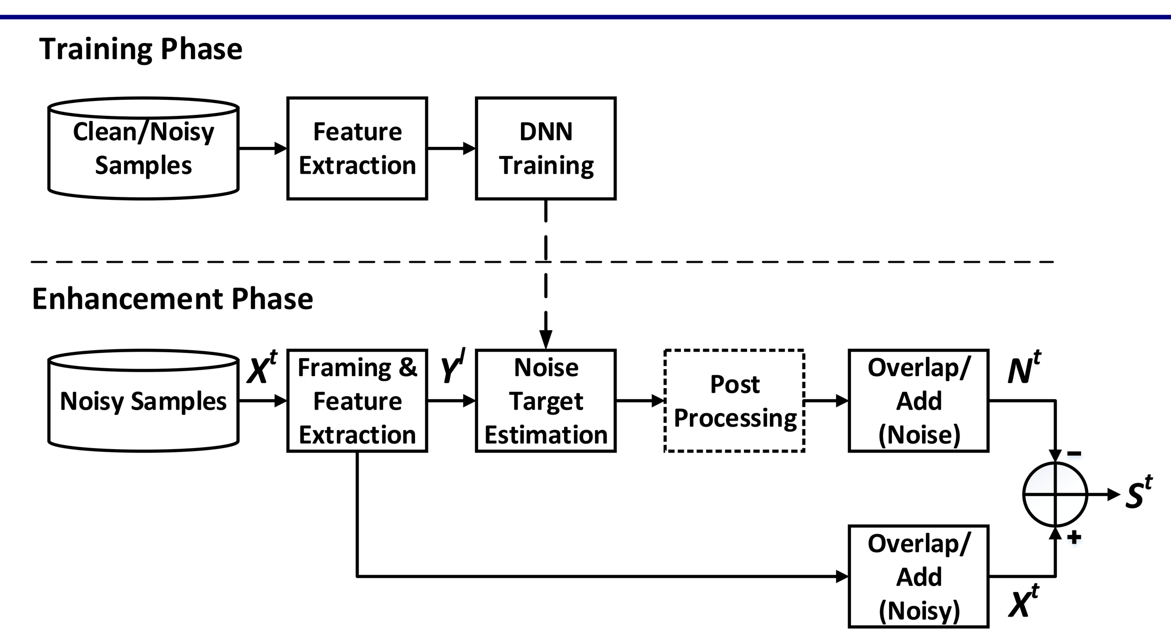
The speech signal is dominated by the noise at average SNR lower than 0dB

- Learn a mapping between the noisy signal input and the added noise [1]

The noisy signal phase is dominated by the phase of the noise at low SNR values

- Use the noisy signal phase to reconstruct the noise

System Block Diagram



In the training phase, the network learns a mapping between the noisy input spectra and a noise target

In the enhancement phase, an estimate of the added noise is reconstructed, and the noise-free speech is obtained by time domain subtraction

Training Targets

1. Log Magnitude Spectrum (LogFFT)
   - STFT magnitude spectrum is log compressed
2. Fourier Magnitude Spectrum Mask (FFT-MASK)
   - Ratio of the noise and noisy speech magnitude spectra

$$M_{FFT}(t,\omega) = \frac{N(t,\omega)}{X(t,\omega)}$$

where $N(t,\omega)$ and $X(t,\omega)$ are respectively the magnitude spectra of the added noise and noisy speech

3. Noise Ratio Mask (NRM)

$$NRM(t,\omega) = \left(\frac{N^2(t,\omega)}{S^2(t,\omega)+N^2(t,\omega)}\right)^{\frac{1}{2}}$$

where $N^2(t,\omega)$ and $S^2(t,\omega)$ are respectively the added noise and speech power spectral densities.

- Equivalent to a frequency domain square root Wiener filter

Experiments

Noise-free speech and noise data were obtained from the IEEE Corpus and non-speech sound database respectively

Training datasets of about 50 hours were made by adding 50 noise types to clean speech; testing was done with 10 noise types

Performance of noise prediction models compared to that of a conventional, noise-aware trained (NAT) [2], speech prediction model

Results

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	3.027	3.506	3.686	3.720	**3.765**
15	2.701	3.394	3.481	3.511	**3.590**
10	2.380	3.265	3.240	3.258	**3.380**
5	2.072	3.114	2.982	2.975	**3.134**
0	1.791	**2.932**	2.708	2.665	2.845
-5	1.503	**2.708**	2.409	2.327	2.513
AVG.	2.246	3.153	3.084	3.076	**3.205**

PESQ scores for proposed and NAT systems in seen noise

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	0.961	0.937	**0.981**	0.974	0.977
15	0.926	0.928	**0.968**	0.958	0.962
10	0.872	0.916	**0.947**	0.934	0.941
5	0.799	0.897	**0.917**	0.899	0.910
0	0.708	0.872	**0.874**	0.851	0.868
-5	0.608	**0.834**	0.817	0.787	0.808
AVG.	0.812	0.897	**0.917**	0.901	0.911

STOI scores for proposed and NAT systems in seen noise

Results

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	3.182	3.426	3.292	3.413	**3.530**
15	2.875	3.242	3.007	3.134	**3.266**
10	2.569	**3.020**	2.715	2.842	2.976
5	2.288	**2.760**	2.420	2.538	2.664
0	2.036	**2.475**	2.137	2.233	2.345
-5	1.779	**2.182**	1.865	1.942	2.032
AVG.	2.455	**2.851**	2.573	2.684	2.802

PESQ scores for proposed and NAT systems in unseen noise

SNR (dB)	Noisy	NAT	LogFFT	FFT-MASK	NRM
20	0.958	0.935	0.965	0.965	**0.970**
15	0.925	0.922	0.937	0.939	**0.949**
10	0.876	0.900	0.893	0.899	**0.915**
5	0.813	0.862	0.832	0.842	**0.863**
0	0.736	**0.804**	0.755	0.767	0.793
-5	0.650	**0.727**	0.666	0.677	0.706
AVG.	0.826	0.858	0.841	0.848	**0.866**

STOI scores for proposed and NAT systems in unseen noise

Observations

In seen noise:

- The noise prediction models, in general, perform well in enhancing the intelligibility of noisy speech
- The NRM model performs best among the noise models in enhancing speech quality
- The NRM outperforms the NAT model on average; however, it is worse than the latter at low SNR values

In unseen noise:

- The NRM is the best among the noise models in enhancing both the quality and intelligibility of noisy speech
- The NRM model performs better than the NAT model in enhancing quality at higher SNR values but is worse at lower SNR values and on average
- The NRM model performs better than the NAT model in enhancing intelligibility at higher SNR values and on average but is slightly worse at lower SNR values

Summary

The NRM was the best all-round noise target. It outperformed the NAT model in seen noise conditions and in improving intelligibility in unseen noise, but fell short at lower SNR values

REFERENCES

[1] B. O. Odelowo and D.V. Anderson, IEEE Int'l Conf. on Machine Learning and Applications, 2017

[2] Y. Xu et al., IEEE Tran. Audio, Speech, and Language Processing, 2015