Audioformer: A Transformer-Based Approach for Audio Denoising

Jimming He, Alex Kwon, Suhas Chundi
Department of Computer Science, Stanford University

Problem

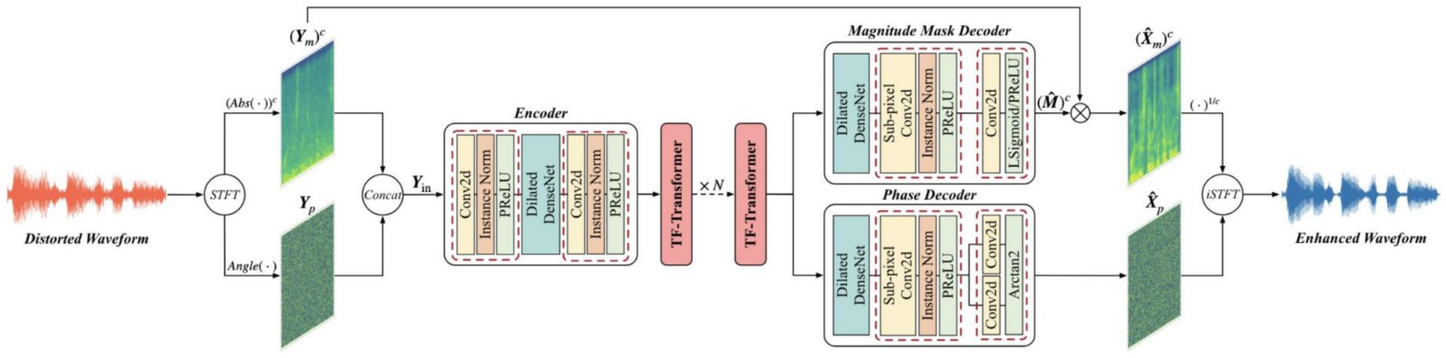
Monaural Speech Enhancement: identify a principal speaker's voice amongst general background audio

- Desired in real world settings where audio is often distorted by reverberations, background noise, and microphone quality
- Current DL models treat this as a supervised learning problem: given input noisy audio, produce matching clean audio
- Evaluation metric: *perceptual evaluation of speech quality (PESQ) score* ranging from 1.0 (poor) to 4.5 (perfect)

Methods

We apply a transformer architecture to the magnitude denoising problem:

- Input magnitude spectra image split into 8 chunks along time axis
- CNN embeds each chunk to pass to a transformer encoder block
- Outputs of the transformer encoder block are concatenated and upsampled by convolutional 2-D transpose layers to the original magnitude spectra shape
- Final layer is a learnable sigmoid transformation: $\sigma_{\beta,\alpha}(t) = \frac{\beta}{1 - e^{1-\alpha t}}$



- Output treated as a multiplicative mask for the noisy magnitude input. The power law compression technique defines the magnitude mask's prediction target:

$$\hat{X}_{predicted} = (X_{noisy} \odot M)^{1/0.3}$$

- X_{clean} and X_{noisy} represent the clean and noisy magnitude spectra, respectively. The elementwise product is taken to obtain the predicted clean magnitude

$$\hat{M} = (X_{clean}/X_{noisy})^{0.3}$$

Datasets

- VoiceBank + DEMAND dataset: a common benchmark for speech enhancement problems that combines clean audio with artificial noise.
- Audio segments are 1 − 10 seconds in length, and are cropped or padded to 2 seconds for a uniform input size.
- 90 − 10 train validation split of the 11572 examples in the training dataset.

Experiment 1: PESQ Hyperparameter Tuning

Attention Heads	Transpose Layers			
	2	4	6	8
2	1.85	1.87	1.92	1.88
4	1.98	2.01	**2.07**	2.02
8	1.87	1.91	1.95	1.92
12	1.80	1.84	1.88	1.86

Ran on a small subset of the data (50 audio clips)

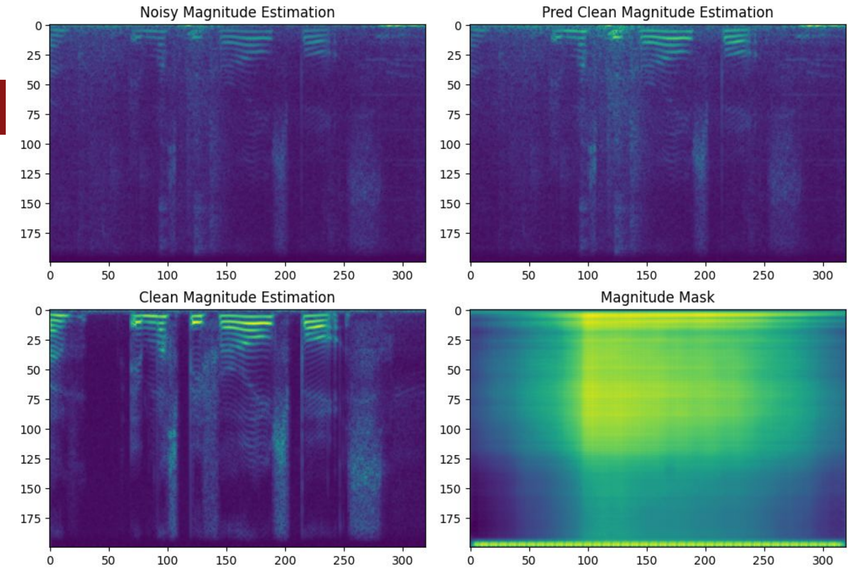
Experiment 2: Full Dataset Evaluation

Model	Input	PESQ
SEGAN	Time Domain	2.16
WaveUNet	Time Domain	2.41
MetricGAN+	Magnitude Spectra	3.15
MP-SENet	Magnitude + Phase	3.50
Audioformer	Magnitude Spectra	1.92

PESQ results on full dataset with Audioformer vs. other models

Analysis

- 4 attention heads and 6 transpose layers was the sweet spot
- But this didn't translate well to training on the entire model, actually worsening PESQ score significantly



Ground-truth noisy, predicted clean, and ground-truth clean magnitude for an audio clip

The predicted clean magnitude somewhat isolated the principal speaker amongst background noise, but to a significantly lower extent than the ground-truth clean magnitude

Conclusions

Our study found that the transformer-based Audioformer model underperformed in monaural speech enhancement compared to existing models. Key takeaways include:

Hyperparameter Optimization: Effective tuning requires a larger dataset. Initial tuning on a small subset did not scale well.

Model Complexity: The transformer architecture may overfit noise patterns.