Physical Sensors and Systems for Environmental Signals
A Comparative Study of Denoising Techniques for Speech Audio Signals

Mirko Morello
m.morello11@campus.unimib.it
Mat. 920601

February, 2025

Università degli studi di Milano Bicocca / Artificial Intelligence for Science and Technology

Outline

Introduction

Introduction

- Environmental acoustic recordings are crucial for applications in ecology, urban planning, and environmental monitoring.
- However, these recordings are often contaminated with noise from various sources.
- In this study, we compare classical denoising techniques with modern neural network approaches.
- **Datasets:** Clean speech from LibriSpeech and noise from UrbanSound8K.

Related Work

Related Work

- **Classical Methods:**
  - Spectral Subtraction [1]
  - Wiener Filtering [2]
- **Neural Network Methods:**
  - Residual Autoencoder
  - U-Net (UNetSpec)
  - Hybrid Denoiser
  - Transformer Autoencoder

Methods

Spectral Subtraction

- **Process:**
  1. **STFT:** Compute the Short-Time Fourier Transform (STFT) of the noisy signal to obtain magnitude and phase.
  2. **Noise Estimation:** Estimate the noise spectrum (often using initial frames assumed to be noise-dominant).
  3. **Subtraction:** Subtract the estimated noise magnitude from the noisy magnitude. Use a max operation to avoid negative values.
  4. **iSTFT:** Reconstruct the time-domain signal by applying the inverse STFT (iSTFT) using the original phase.
- **Pros:** Simple and computationally efficient.
- **Cons:** May introduce “musical noise” artifacts due to imperfect noise estimation.

Wiener Filtering

- **Principle:** Minimizes the mean squared error (MSE) between the estimated clean signal and the true clean signal.
- **Process:**
  1. Estimate the power spectral density (PSD) of both the clean signal and the noise.
  2. Calculate the Wiener filter, which balances noise reduction and signal preservation.
  3. Apply the filter in the time domain to the noisy signal.
- **Pros:** Statistically optimal under assumptions of stationarity.
- **Cons:** Performance decreases when noise is non-stationary.

Overview of Neural Network Denoisers

- **Residual Autoencoder:** Processes raw waveforms in the time domain using residual learning.
- **U-Net (UNetSpec):** Enhances the magnitude spectrogram (frequency domain) with skip connections.
- **Hybrid Denoiser:** Combines both time-domain and frequency-domain processing for improved denoising.
- **Transformer Autoencoder:** Uses a simplified attention mechanism to weigh spectrogram features.

Residual Autoencoder (Details)

- **Architecture:**
  - **Encoder:** Series of 1D convolutional layers that extract temporal features.
  - **Decoder:** Transposed convolutions to reconstruct the signal.
  - **Residual Connection:** The network predicts the noise component; subtracting it from the input yields the denoised signal.
- **Advantage:** Direct processing of the raw waveform without domain conversion.

U-Net (UNetSpec) (Details)

- **Architecture:**
  - Operates on the magnitude spectrogram obtained from the STFT.
  - Uses an encoder-decoder structure with skip connections to preserve fine details.
  - Reconstructed magnitude is combined with the original phase for the final signal.
- **Advantage:** Effective at preserving and enhancing spectral details.

Hybrid Denoiser (Details)

- **Dual-Branch Architecture:**
  - **Time-Domain Branch:** Similar to the Residual Autoencoder.
  - **Frequency-Domain Branch:** Processes the magnitude spectrogram using a U-Net-like structure.
- **Fusion:** The outputs of both branches are concatenated and merged to produce the final denoised waveform.
- **Advantage:** Leverages complementary information from both the time and frequency domains.

Transformer Autoencoder (Details)

- **Architecture:**
  - Converts the time-domain signal to a spectrogram via STFT.
  - Incorporates a simplified attention block (channel-wise attention) in the bottleneck.
  - The decoder reconstructs the enhanced spectrogram, which is then used with the original phase to recover the waveform.
- **Advantage:** Provides global feature weighting with lower computational overhead compared to full transformer models.

Loss Functions Overview

- Two training loss variants are used:
  1. **Simple Loss (v1):** A combination of L1 loss and Mean Squared Error (MSE) computed in the time domain.
  2. **Hybrid Loss (v2):** Combines time-domain loss, frequency-domain loss, and a negative SI-SDR term.

Simple Loss (v1)

$$L_{\text{simple}} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \hat{x}_i| + \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

- **L1 Loss:** Penalizes absolute differences; robust to outliers.
- **MSE Loss:** Emphasizes larger errors.
- The sum of both encourages both overall fidelity and the preservation of fine details.

Hybrid Loss (v2)

- **Components:**
  - **Time-domain L1 Loss:**

$$L_{time} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{x}_i|$$

  - **Frequency-domain L1 Loss:**

$$L_{freq} = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \Big| |X(f,t)| - |\hat{X}(f,t)| \Big|$$

  - **Negative SI-SDR:** Optimizes the Scale-Invariant Signal-to-Distortion Ratio.
- **Overall Hybrid Loss:**

$$L_{\text{hybrid}} = \frac{1}{3} \left(L_{time} + L_{freq} + (-\text{SI-SDR}) \right)$$

- This loss encourages accurate reconstruction in both the time and frequency domains while directly minimizing signal distortion.

Metrics

Evaluation Metrics

- **PESQ (Perceptual Evaluation of Speech Quality):**
  - Measures the perceived quality of speech.
  - Scale: Approximately -0.5 to 4.5 (higher scores indicate better quality).
- **STOI (Short-Time Objective Intelligibility):**
  - Assesses the intelligibility of speech.
  - Scale: 0 to 1 (values closer to 1 indicate higher intelligibility).
- **SI-SDR (Scale-Invariant Signal-to-Distortion Ratio):**
  - Evaluates the overall distortion introduced by the denoising process.
  - Higher values denote less distortion.
- **MOS (Mean Opinion Score):**
  - A subjective measure of audio quality, typically rated from 1 to 5.

Experimental Setup

Experimental Setup

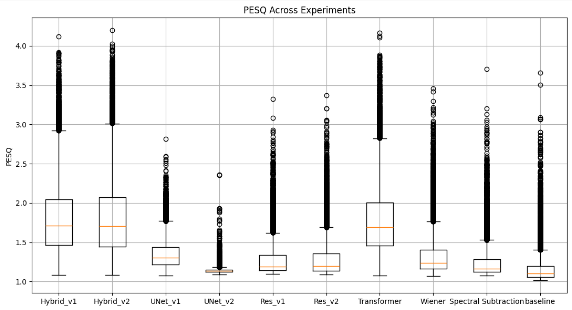
- **Optimizer:** AdamW with a learning rate of 3e-4 and weight decay of 1e-5.
- **Scheduler:** ReduceLROnPlateau to adjust learning rate when the validation loss plateaus.
- **Training:** Batch size of 24 over 10 epochs (demonstration setting; longer training is recommended).
- **Dataset:** Synthetic noisy data generated by mixing LibriSpeech with UrbanSound8K at various SNR levels.

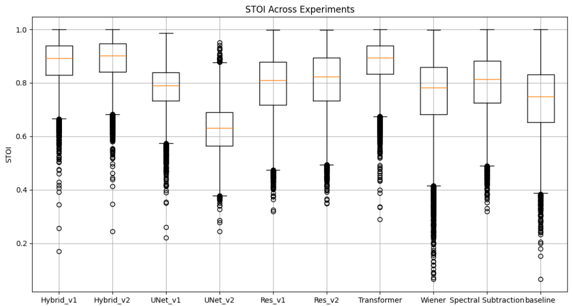
Results

Quantitative Results

Method	PESQ	STOI	SI-SDR (dB)	MOS
Baseline (No Denoising)	1.16 $\pm$ 0.19	0.74 $\pm$ 0.12	0.10 $\pm$ 4.22	2.84 $\pm$ 0.82
Spectral Subtraction	1.25 $\pm$ 0.21	0.80 $\pm$ 0.11	3.76 $\pm$ 5.72	3.83 $\pm$ 0.54
Wiener Filtering	1.33 $\pm$ 0.26	0.76 $\pm$ 0.14	-0.10 $\pm$ 6.51	3.82 $\pm$ 0.47
ResAutoencoder (v2)	1.29 $\pm$ 0.24	0.81 $\pm$ 0.11	3.18 $\pm$ 5.69	2.52 $\pm$ 0.63
U-Net (v1)	1.35 $\pm$ 0.18	0.78 $\pm$ 0.08	3.95 $\pm$ 3.63	3.54 $\pm$ 0.78
Hybrid (v2)	1.81 $\pm$ 0.50	0.89 $\pm$ 0.08	11.69 $\pm$ 5.27	3.79 $\pm$ 0.74
Transformer	1.78 $\pm$ 0.45	0.88 $\pm$ 0.08	11.65 $\pm$ 4.90	2.87 $\pm$ 0.50

Boxplots of Evaluation Metrics





Comparative Conclusions

- **Classical vs. Neural Methods:** Neural approaches (especially Hybrid and Transformer models) outperform classical methods in reducing distortion (SI-SDR) and improving intelligibility (STOI).
- **Architecture Insights:**
  - **Hybrid Denoiser:** Achieves the highest SI-SDR, indicating minimal distortion.
  - **U-Net:** Excels in preserving spectral details (PESQ and MOS).
  - **Transformer:** Provides competitive SI-SDR with slightly lower perceptual quality (MOS), suggesting room for further tuning.
- **Consistency:** Boxplots show that deep learning models not only improve mean performance but also reduce variability.
- **Trade-Offs:** High SI-SDR values must be balanced with perceptual quality, as indicated by MOS.

Future Work

Future Work

- Validate the models on real-world environmental recordings.
- Explore advanced architectures (e.g., full transformer models and GAN-based approaches).
- Develop adaptive and semi-supervised denoising methods.
- Investigate additional evaluation metrics that better capture perceptual quality.

Conclusion

Conclusion

- Deep learning approaches (Hybrid and Transformer) substantially outperform classical methods.
- Combining time- and frequency-domain information is key for effective denoising.
- Neural models achieve improved signal fidelity and intelligibility, though further tuning is needed for optimal perceptual quality.

Thank you for your attention!

Steven F Boll.
Suppression of acoustic noise in speech using spectral subtraction.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2):113–120, 1979.

Norbert Wiener.
Extrapolation, interpolation, and smoothing of stationary time series.
Wiley, 1949.