ELEC5305 Project Proposal

Noise-Robust Speech Enhancement Using Classical and Machine Learning Methods

Wenkuan HU
Student ID: 510026747
GitHub: https://github.com/Therressia/ELEC5305-project–510026747-

September 7, 2025

Project Overview

This project aims to design and evaluate a speech enhancement system capable of improving intelligibility in noisy environments. Background noise such as traffic, crowds, or machinery significantly degrades speech quality in everyday applications such as telecommunication and hearing aids.

The proposed solution combines classical signal processing approaches (spectral subtraction and Wiener filtering) with a lightweight machine learning model trained on speech–noise mixtures. The goal is to deliver a working prototype that demonstrates measurable improvements in signal-to-noise ratio (SNR) and perceptual quality metrics such as PESQ.

Background and Motivation

Speech enhancement has been studied for decades, beginning with statistical and matrix decomposition approaches. For example, Fujimoto and Ariki (2007) proposed a combination of Gaussian Mixture Model (GMM)-based speech estimation and temporal-domain Singular Value Decomposition (SVD), which improved robustness under low signal-to-noise ratio conditions [**?**]. Although effective in controlled scenarios, these traditional methods were limited by assumptions on noise distribution and dataset scale.

More recently, deep learning methods have dominated the field. Lee et al. (2024) introduced a novel knowledge distillation (KD)-based framework in which an Automatic Speech Recognition (ASR) model serves as a teacher and a Speech Enhancement (SE) model acts as a student [**?**]. This approach successfully mitigated gradient conflicts in joint training and achieved state-of-the-art performance on noisy LibriSpeech and CHiME-4 datasets.

Beyond ASR, enhancement modules have also been shown to improve robustness in other speech-related tasks. Chen and Zhang (2024) proposed TRNet, a two-level refinement network that leverages speech enhancement for noise-robust speech emotion recognition (SER) [**?**]. Their results demonstrated that SE improves both low-level acoustic features and high-level emotion representations under noisy conditions.

These studies show the historical development and ongoing relevance of speech enhancement, motivating this project to combine interpretable DSP methods with modern lightweight neural models.

Proposed Methodology

The project will be implemented using MATLAB and Python.

Step 1 – Data Collection:

- Speech: TIMIT dataset (phonetically balanced speech).
- Noise: DEMAND or DNS-Challenge dataset.
- Mixtures at SNR = 0, 5, 10 dB.

Step 2 – Baseline DSP Methods:

- Compute short-time Fourier transform (STFT):

$$X(k,m) = \sum_{n=0}^{N-1} x[n+mH]w[n]e^{-j2\pi kn/N}$$

  where $w[n]$ is a Hann window and H is hop size.

- Apply spectral subtraction and Wiener filtering.
- Reconstruct using inverse STFT.

Step 3 – Machine Learning Enhancement:

- Train a small feedforward DNN on noisy-clean pairs.
- Features: log-mel spectrograms; Target: ideal ratio mask (IRM).

Step 4 – Evaluation:

- Objective metrics: SNR improvement, PESQ, STOI.
- Subjective: A/B listening tests.

Expected Outcomes

1. MATLAB/Python prototype for speech enhancement.
2. Comparative analysis: DSP vs machine learning methods.
3. Spectrogram visualizations before and after enhancement.
4. GitHub repository with code, audio samples, and documentation.

Timeline (Weeks 6–13)

- Weeks 6–7: Literature review; dataset preparation.
- Weeks 8–9: Implement DSP baselines.
- Weeks 10–11: Develop/train lightweight DNN.
- Weeks 12–13: Evaluation; final report; GitHub documentation.

References

1. Fujimoto, M., & Ariki, Y. (2007). Combination of GMM-based speech estimation and temporal domain SVD-based enhancement for noise robust speech recognition. *Speech Communication*, 49(5), 344–356.
2. Lee, G. W., Kim, H. K., & Kong, D. J. (2024). Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition. *IEEE Access*, 12, 72707–72720.
3. Chen, C., & Zhang, P. (2024). TRNet: Two-level refinement network leveraging speech enhancement for noise-robust speech emotion recognition. *Applied Acoustics*, 225, 110169.