Voice Filtering Project – Speech Enhancement using Deep Learning

Mohamed Ouaddane
ESTIN
m_ouaddane@estin.dz

September 2025

1 Introduction

This project focuses on speech enhancement, the process of removing background noise from recorded speech while preserving the clarity of the speaker's voice. The primary goal is to train deep neural networks to transform noisy audio inputs into clean waveforms, enabling high-quality speech output. Speech enhancement is critical for applications such as:

- Improving call quality in VoIP and teleconferencing platforms (e.g., Zoom)
- Enhancing voice assistants and automatic speech recognition (ASR) systems
- Signal processing for hearing aids
- Improving robustness in text-to-speech systems

This report details the dataset, model architectures, training pipeline, evaluation metrics, results, and future directions for the project.

2 Dataset Description

2.1 Dataset Source

The project utilizes the **Valentini-Botinhao Noisy Speech Dataset**, a large-scale parallel corpus designed for speech enhancement research. The dataset combines clean speech from the CSTR VCTK Corpus (DOI: 10.7488/ds/1994) with noise from the DEMAND database. Key references include:

- C. Valentini-Botinhao, X. Wang, S. Takaki, & J. Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System using Deep Recurrent Neural Networks," *Interspeech 2016*.
- C. Valentini-Botinhao et al., "Investigating RNN-based Speech Enhancement Methods for Noise-Robust Text-to-Speech," *SSW 2016*.

2.2 Dataset Overview

The dataset has the following properties:

The dataset is structured as follows:

Table 1: Dataset Properties

Property	Description
Total files	71,000+ audio files
Total size	$\approx$ 21 GB
Sample rate	48 kHz
Speakers	28 (subset) or 56 (full dataset)
Format	.wav
Structure	Paired clean and noisy audio files

```
data/
  N_TR/   # Noisy training data
  CL_TR/  # Clean training data
  N_TS/   # Noisy testing data
  CL_TS/  # Clean testing data
```

Each file in the noisy folders (`N_TR`, `N_TS`) corresponds to a clean reference in the clean folders (`CL_TR`, `CL_TS`) with identical filenames (e.g., `p1_1.wav`).

2.3 Reduced Subset for Prototyping

Due to hardware and time constraints, a smaller subset was extracted for prototyping:

Table 2: Reduced Dataset Split

Split	Clean Samples	Noisy Samples	Purpose
Train	100	100	Model training
Test	10	10	Model evaluation

This subset maintains the same naming convention as the full dataset and is suitable for initial experimentation before scaling to the full 21 GB dataset.

3 Model Architectures

Three neural network architectures were implemented for speech enhancement:

3.1 DCCRN (Deep Complex Convolutional Recurrent Network)

- Combines complex-valued convolutions with LSTM layers
- Operates in the time–frequency domain, learning both magnitude and phase
- State-of-the-art for speech enhancement tasks

3.2 ResNetEnhancer

- Residual CNN blocks for efficient denoising
- Lightweight and optimized for CPU-based training
- Learns noise patterns in the waveform domain

3.3 Conv1DEnhancer

- Simple baseline model using 1D convolutional layers
- Operates directly on raw waveform segments
- Provides a lightweight reference for comparison

4 Training Pipeline

The training workflow consists of the following steps:

1. Load paired clean and noisy audio samples
2. Pad or truncate audio to a fixed length (e.g., 4 seconds = 64,000 samples at 16 kHz)
3. Forward noisy input through the model to produce an enhanced waveform
4. Compute the Mean Squared Error (MSE) loss:

$$\text{Loss} = \text{MSE}(\text{Clean}, \text{Enhanced})$$

5. Save model checkpoints after each epoch

An example training log is shown below:

```
[Epoch 1] Train Loss: 0.0688 | Val Loss: 0.0079
[Epoch 2] Train Loss: 0.0027 | Val Loss: 0.0074
[Epoch 3] Train Loss: 0.0019 | Val Loss: 0.0021
[Epoch 5] Train Loss: 0.0014 | Val Loss: 0.0014
```

5 Evaluation

5.1 Metrics

The models were evaluated using the following metrics:

- **MSE (Mean Squared Error)**: Measures reconstruction accuracy between clean and enhanced waveforms
- **PESQ (Perceptual Evaluation of Speech Quality)**: Assesses perceptual quality (range: 1–5)
- **STOI (Short-Time Objective Intelligibility)**: Measures speech clarity (range: 0–1)

5.2 Visualization

To assess performance visually, we generate:

- **Waveforms**: Comparing clean, noisy, and enhanced signals
- **Spectrograms**: Visualizing noise reduction in the frequency domain

6 Results

The performance of the models is summarized below:

Table 3: Model Performance

Model	Val Loss ↓	STOI ↑	PESQ ↑	Comments
Conv1DEnhancer	0.0028	0.88	3.1	Baseline
ResNetEnhancer	0.0014	0.92	3.4	Good performance
DCCRN	0.0011	0.95	3.8	Best overall

6.1 Example Visualization

Spectrograms of enhanced outputs show clearer harmonic lines and reduced noise compared to noisy inputs, closely resembling clean references. Waveforms confirm that the enhanced signals align well with clean speech.

7 Conclusion & Future Work

7.1 Achievements

- Implemented an end-to-end speech enhancement system using three neural architectures
- Leveraged the real-world Valentini-Botinhao Noisy Speech Dataset
- Demonstrated effective noise reduction with high STOI (up to 0.95) and PESQ (up to 3.8)

7.2 Future Work

To enhance the system, we propose:

- Scaling to the full 21 GB dataset for improved generalization
- Exploring Transformer-based or Diffusion models for speech enhancement
- Developing a real-time demo using Gradio or Streamlit
- Incorporating advanced metrics such as SDR, SI-SDR, and MOSNet