MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Speech Enhancement Using U-Net Architecture

Mini-Project Synopsis

submitted to

Manipal School of Information Sciences, MAHE, Manipal

Reg. Number	Name	Branch
241058020	Nikhil M	Big Data Analytics
241058024	Nikhil S G	Big Data Analytics
241058030	Vinayashree M Shet	Big Data Analytics

Under The Guidance Of

Mr. Raghudathesh G P

14/08/2024

MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

Table of Contents

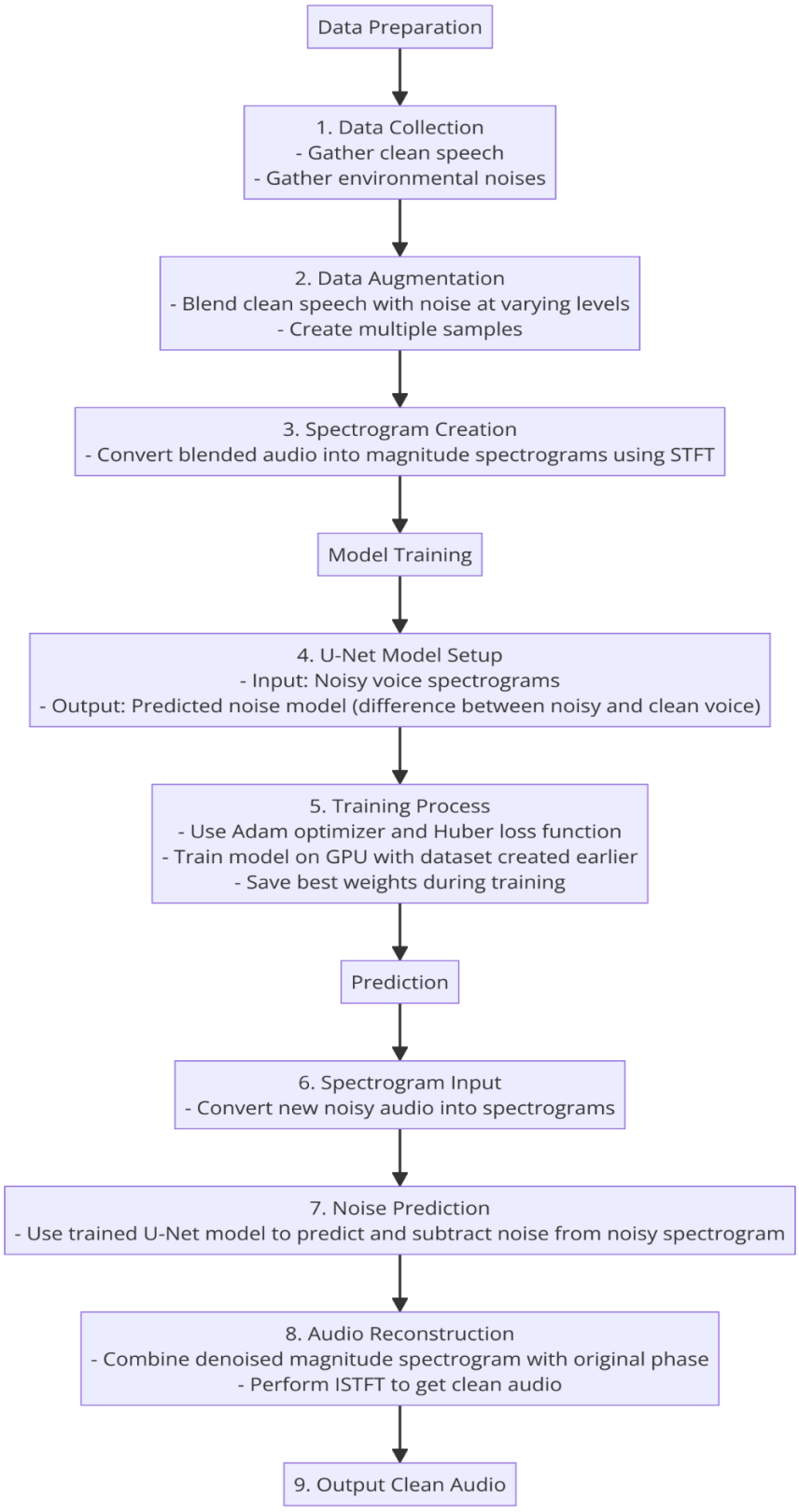
1. Introduction

This project is dedicated to developing a cutting-edge speech enhancement system that enhances audio clarity by efficiently reducing environmental noise. It utilizes magnitude spectrograms, which convert audio signals into 2D images showing time and frequency, and processes them through a U-Net deep learning model. The model is trained to identify and remove noise from voice spectrograms, resulting in much clearer audio. Training involves clean speech from the LibriSpeech dataset and diverse noise samples from the ESC-50 and SiSec datasets, using advanced data augmentation techniques and GPU optimization to manage various noise environments. The enhanced audio is then reconstructed by combining the cleaned spectrogram with the original phase. This system has versatile applications, including improving communication quality in telecommunications, enhancing audio for assistive devices, and refining sound in media production, making it a valuable tool for various professional and consumer uses.

2. Objective

- Data Preparation
- Data modelling/Modeling the data
- Benchmarking the results with other work

3. Block Diagram/Flowchart



4. Applications

4.1. Noise Reduction in Voice Recordings

Application: The system can be used to enhance speech clarity by removing background noise from voice recordings, making it easier to understand spoken words in environments where there is significant noise, such as crowded public spaces or noisy workplaces.

Examples:

Call Centers: Reducing background chatter to improve customer service interactions.

Lecture Recordings: Cleaning up recordings of lectures or presentations held in noisy environments to ensure the speaker's voice is clear.

Public Safety: Enhancing audio from police or emergency responder recordings, where clear communication is critical.

4.2. Improved Audio Quality in Telecommunication and Assistive Devices

Application: This technology can be integrated into telecommunication systems, such as VoIP services, mobile networks, or video conferencing tools, to ensure that voice communication remains clear, even in environments with high levels of ambient noise. It can also be used in assistive hearing devices to filter out unwanted background noise.

Examples:

Video Conferencing: Enhancing voice clarity during video calls, especially in remote work settings where participants might be in less-than-ideal acoustic environments.

Hearing Aids: Implementing real-time noise reduction in hearing aids to help users better focus on conversations in noisy settings like restaurants or busy streets.

Smartphones: Integrating noise reduction features in mobile phones to improve call quality in noisy environments like subways or city streets.

4.3. Audio Processing in Media Production

Application: In the media and entertainment industry, the system can be used during the post-production phase to improve the quality of audio recordings. This is particularly useful when the original audio is recorded in less-than-ideal conditions, where noise cannot be completely avoided.

Examples:

Film and Television Production: Enhancing dialogue clarity in scenes shot in noisy outdoor locations or crowded places.

Podcasting: Cleaning up interviews or discussions recorded in public spaces or with imperfect equipment, ensuring the final product sounds professional.

Music Production: Reducing noise in vocal tracks recorded in home studios or during live performances to produce cleaner audio tracks.

5. Software & Hardware Requirements

Software:

- Python: Main programming language.
- TensorFlow/Keras: For implementing and training the U-Net model.
- LibriSpeech, ESC-50, SiSec datasets: For obtaining clean speech and noise data.
- Google Colab: For using free GPU resources.

Hardware:

- GPU (Graphics Processing Unit): For faster training of deep learning models.
- High-Performance CPU: Supports general data processing and model training tasks.
- Storage: Enough space for storing and processing audio datasets, preferably over 5GB.

References

[1] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM Int. Conf. on Multimedia*, Orlando, FL, USA, Nov. 2015, pp. 1015-1018.

[2] O.Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234-241.

[3] Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Suzhou, China, Oct. 2017, pp. 745-751.

[4] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031-82057, 2021.

[5] Y. Deng, Y. Hou, J. Yan, and D. Zeng, "ELU-Net: An efficient and lightweight U-Net for medical image segmentation," *IEEE Access*, vol. 8, pp. 123045-123053, 2020.