Whitepaper: Comparison of Noise Suppression Models for Ale

Introduction

Noise suppression is a critical technology in speech-based applications, especially for real-time virtual assistants like Ale. By removing unwanted background noise, noise suppression enhances the clarity and intelligibility of user commands, ensuring seamless communication.

This paper compares two noise suppression models: **DeepFilterNet3** and **SepFormer**, evaluating their performance, latency, ease of integration, and resource requirements. These evaluations guide the recommendation for their potential integration into Ale's real-time audio pipeline.

Models Overview

Model 1: DeepFilterNet3

- **Description**: A lightweight neural network optimized for real-time noise suppression.
- **Framework**: Implemented in PyTorch with pre-trained weights.
- **Features**: Designed for low-latency applications.
- **Repository**: DeepFilterNet3 GitHub

Model 2: SepFormer

- **Description**: A speech separation model adapted for noise suppression.
- **Framework**: Available via SpeechBrain on Hugging Face.
- **Features**: High-quality noise reduction at the cost of higher latency.
- **Repository**: SepFormer on Hugging Face

Comparison Criteria

Performance

Metric	DeepFilterNet3	SepFormer
SNR Improvement	11.12 dB	-3.60 dB
PESQ Score	1.33	2.26
STOI Score	0.95	0.95

- **Analysis**: While SepFormer excels in perceptual metrics like PESQ, DeepFilterNet3 demonstrates better signal-to-noise ratio improvement.

Latency

Metric	DeepFilterNet3	SepFormer
Inference Time	~2.12 seconds	~3.52 seconds (GPU)

- **Analysis**: DeepFilterNet3 is faster, making it more suitable for real-time applications on standard hardware. Note: Server-grade GPUs could significantly reduce inference times.

Ease of Integration

Aspect	DeepFilterNet3	SepFormer
Framework	PyTorch	SpeechBrain
Pre-trained Models	Yes	Yes
Custom Training	Optional	Limited

- **Analysis**: Both models offer pre-trained weights, but DeepFilterNet3 is simpler to integrate due to its lightweight architecture.

Resource Requirements

Upon reviewing the available information, specific details regarding the GPU, CPU, and memory consumption for both DeepFilterNet3 and SepFormer are limited. However, based on general practices and available discussions, we can provide some insights:

DeepFilterNet3:

- **CPU Usage**: There is an indication that DeepFilterNet can run on a standard CPU and RAM setup. A user inquired about running DeepFilterNet on a normal CPU and RAM combination instead of GPU memory.

GitHub

- **Memory Consumption**: Specific memory usage details are not provided. However, given its design for low-latency applications, it is likely optimized for efficient memory usage.

SepFormer:

- **GPU Usage**: Inference can be performed on a GPU by specifying the device as "cuda" during model initialization.
- **Memory Consumption**: While exact figures aren't provided, transformer-based models like SepFormer can be resource-intensive. For context, training NLP transformer models typically requires a minimum of 12 GB of VRAM and 32 GB of RAM, with 64 GB being more comfortable.

Reddit

In summary, while exact resource requirements are not specified, DeepFilterNet3 appears to be optimized for lower resource consumption, making it more suitable for deployment on standard hardware configurations. SepFormer, being a transformer-based model, may demand more substantial computational resources, particularly when processing longer audio files or during training phases.

Use Case in Ale

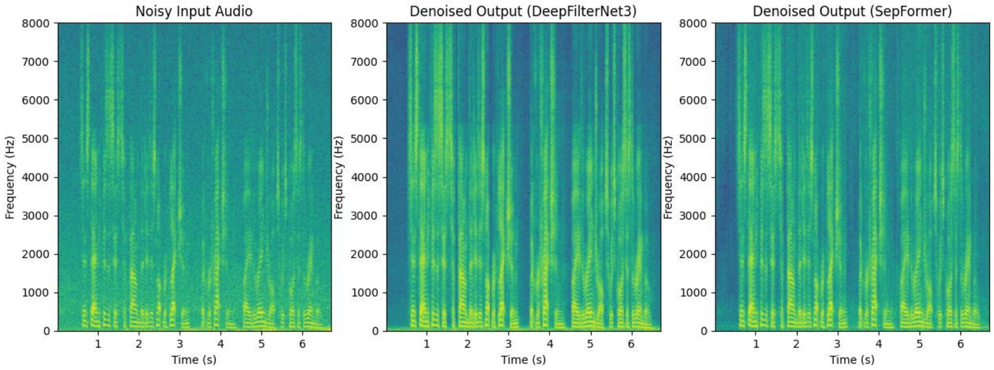
Ale, as a real-time virtual assistant, relies on efficient noise suppression for clear communication. Here’s how the models align with Ale’s requirements:

- **DeepFilterNet3**: Its low latency and lightweight nature make it ideal for real-time audio processing on standard devices. Its simplicity ensures seamless integration into Ale’s pipeline.
- **SepFormer**: While offering superior perceptual quality, its higher latency might hinder real-time responsiveness.

Recommendation: DeepFilterNet3 is the preferred choice for Ale due to its balance of speed, performance, and integration ease.

Benchmarks and Visualizations

Spectrogram Comparison



- **Left**: Noisy input audio
- **Center**: Denoised output (DeepFilterNet3)
- **Right**: Denoised output (SepFormer)

Observations

- DeepFilterNet3 effectively reduces noise, preserving speech intelligibility.
- SepFormer achieves higher perceptual quality but at the cost of increased computational demand.

References and Benchmarks

1. DeepFilterNet GitHub Repository
2. SepFormer on Hugging Face
3. PESQ and STOI Metrics: PESQ ITU-T P.862 and STOI paper.

Conclusion

This paper evaluated two state-of-the-art noise suppression models for potential integration into Ale. While both models have their strengths, DeepFilterNet3 is recommended for its low latency and ease of integration, aligning with Ale’s real-time processing needs.

Post Scriptum

In hindsight, this analysis and documentation could have been better presented and explored using platforms like Google Colab or Jupyter Notebook. These tools would have allowed for more interactive visualizations, code execution, and seamless sharing. Additionally, deploying the solution using Docker would enhance portability, ensuring consistent performance across different environments.