

# 音频降噪技术调研与模型复现报告

## 一、项目背景

随着智能语音在客服系统、会议系统、通话系统、AI Agent 及语音识别 (ASR) 中的大规模应用，噪声导致的用户体验下降与识别准确率降低已成为行业普遍痛点。在实际业务场景中，噪声来源复杂多样，主要包括：

**环境噪声：**如交通噪声、室内混响、风噪等

**人为噪声：**如键盘声、他人语音等

**系统噪声：**如通话回声、设备底噪等

这些噪声不仅影响语音听感的清晰度与自然度，还会显著降低下游语音识别系统的性能。因此，高效的音频降噪技术对于提升语音交互系统的可用性与可靠性具有重要意义。

本报告系统调研并复现了当前行业主流的深度学习、机器学习及传统信号处理三类降噪方案，重点对关键主流模型进行了训练重建与量化评估，旨在为实际工程选型与优化提供参考。

## 二、项目调研

在企业级语音处理系统与个人使用场景中，音频降噪技术已成为通话、会议、ASR、智能客服和实时语音交互的基础能力。主流的降噪方法大致可分为三类：传统数字信号处理方法、统计与最优估计方法，以及深度学习端到端增强方法。在消费级设备与实时通信平台中，由于对延迟敏感，传统频域方法仍被大量采用，其实现简单、资源占用低，是目前许多软硬件系统的基础模块；同时，在更强调稳健性的企业语音链路中，基于最小均方误差准则的 Wiener 滤波因具有理论最优性与较好的工程稳定性，被广泛用作噪声抑制的核心算法。而随着深度学习的发展，生成式模型与卷积-循环混合结构逐渐成为行业中追求高音质、高鲁棒性场景（如会议转录、云端语音增强、录音优化等）的主流选择，其中 SEGAN 作为早期生成式语音增强的代表模型，对后续深度学习方向具有示范意义；而

FRCRN 则是近年来工业界与学术界共同验证过的高性能全频带网络，其在多场景与复杂噪声条件下的表现已成为当前端到端语音增强模型的重要基准。

基于上述行业现状，本研究选择谱减法、Wiener Filter、SEGAN 与 FRCRN 作为测试对象，具有明确的代表性与必要性。谱减法代表了最经典、最基础、在设备侧仍大量运行的传统算法；Wiener 滤波代表了传统算法中理论最优且企业实际部署频率极高的一类；SEGAN 代表了生成式声学模型在端到端增强领域的开端，是深度学习降噪方法的重要里程碑；而 FRCRN 则代表当前性能领先、具有工业落地价值的现代端到端网络。四类方法覆盖了从传统 DSP 到深度生成模型再到先进全频带网络的核心技术谱系，既呈现了行业在不同阶段的技术选择，也能够通过横向对比揭示各类方法在真实场景中的实际性能差异。因此，通过对这四类具有典型性和时代代表性的降噪算法进行系统测试，不仅能够全面反映当前行业可用方案的能力边界，还能够为后续模型选型、系统部署与技术迭代提供有事实依据的参考。

### 三、配置环境

分类	配置
Python	3.12 (Ubuntu 22.04)
PyTorch	2.3.0
CUDA	12.1
GPU	RTX 4090 (24GB)
CPU	16 vCPU Intel Xeon Gold 6430
OS	Ubuntu Server 22.04

## 四、相关项目

### （一）深度学习板块

深度学习基于数据驱动，能够从大量带噪-纯净语音对中学习复杂噪声模式，在非平稳噪声、低信噪比场景下表现显著优于传统方法。

#### Speech Enhancement Generative Adversarial Network

##### 语音增强生成对抗网络（SEGAN）

参考文献：<https://arxiv.org/pdf/1703.09452>

##### 项目介绍：

SEGAN 是最早将生成对抗网络（GAN）引入语音增强工作的模型之一，采用端到端的时域处理方法，直接对原始波形进行降噪，避免了频域方法中相位估计的难题。

##### 项目特点：

1. 最早将 GAN 引入语音增强
2. 输入为 raw waveform（时域方法）
3. 可用于实时语音增强
4. 模型轻量、适合工程落地

##### 项目准备

- 样本来源：（爱丁堡大学）<https://datashare.ed.ac.uk/handle/10283/2791>
- 训练集： *clean\_trainset\_28spk\_wav.zip* (2.315Gb)  
*noisy\_trainset\_28spk\_wav.zip* (2.635Gb)
- 测试集： *clean\_testset\_wav.zip* (147.1Mb)  
*noisy\_testset\_wav.zip* (162.6Mb)

模型架构

生成器（G）：

- (1) 编码器：11 层 Conv1d，每层 stride=2 进行下采样，通道数从 1 增至 1024，使用 PReLU 激活函数
- (2) 解码器：11 层 ConvTranspose1d 进行上采样，每层通过跳跃连接融合编码器对应层特征，并使用 PReLU 激活函数
- (3) 输出层：ConvTranspose1d 将通道数降至 1，通过 Tanh 激活函数输出波形

判别器（D）：

- (1) 主体结构：11 层 Conv1d，每层后接虚拟批归一化 (VirtualBatchNorm) 和 LeakyReLU 激活
- (2) 动态全连接：根据输入长度动态创建全连接层，适应不同音频长度
- (3) 输出：通过 Sigmoid 函数输出语音真实性概率

模型训练

Epoch	D_loss 文件	D_loss 值	G_cond 文件	G_cond 值
1	D_epoch1_loss0.3442.pkl	0.3442	G_epoch1_cond0.0607.pkl	0.0607
10	D_epoch10_loss0.3201.pkl	0.3201	G_epoch10_cond0.0351.pkl	0.0351
20	D_epoch20_loss0.3197.pkl	0.3197	G_epoch20_cond0.0322.pkl	0.0322
30	D_epoch30_loss0.3199.pkl	0.3199	G_epoch30_cond0.0310.pkl	0.0310
40	D_epoch40_loss0.3195.pkl	0.3195	G_epoch40_cond0.0299.pkl	0.0299
50	D_epoch50_loss0.3194.pkl	0.3194	G_epoch50_cond0.0290.pkl	0.0290
60	D_epoch60_loss0.3195.pkl	0.3195	G_epoch60_cond0.0272.pkl	0.0272
70	D_epoch70_loss0.3196.pkl	0.3196	G_epoch70_cond0.0265.pkl	0.0265
80	D_epoch80_loss0.3196.pkl	0.3196	G_epoch80_cond0.0260.pkl	0.0260

## 收敛性分析

判别器（D）动态：

Epoch 1→10：损失从 0.3442 快速下降至 0.3201，快速学习噪声判别

Epoch 10→80：稳定在  $0.3195 \pm 0.0005$  区间，表明与生成器达到纳什均衡

D\_real/D\_fake 最终接近 0.5，判别器无法区分真假样本，判别器(D)和生成器(G)达到纳什均衡

生成器（G）动态：

G\_cond 从 0.0607 持续下降至 0.0260（降幅 57%），降噪能力显著提升

L1 损失从约 0.003 降至 0.0026，波形重建精度提高

后期下降趋缓，模型接近收敛

## 日志输出（节选）

Batch 2500/3763 | D\_loss: 0.3175 | G\_loss: 0.1458 | G\_cond: 0.0212

实际 L1: 0.0021 | D\_real: 0.509 | D\_fake: 0.505

Batch 2600/3763 | D\_loss: 0.3203 | G\_loss: 0.1531 | G\_cond: 0.0297

实际 L1: 0.0030 | D\_real: 0.503 | D\_fake: 0.503

Batch 2700/3763 | D\_loss: 0.3280 | G\_loss: 0.1486 | G\_cond: 0.0282

实际 L1: 0.0028 | D\_real: 0.502 | D\_fake: 0.511

Batch 2800/3763 | D\_loss: 0.3217 | G\_loss: 0.1455 | G\_cond: 0.0188

实际 L1: 0.0019 | D\_real: 0.497 | D\_fake: 0.499

Batch 2900/3763 | D\_loss: 0.3152 | G\_loss: 0.1486 | G\_cond: 0.0251

实际 L1: 0.0025 | D\_real: 0.504 | D\_fake: 0.497

Batch 3000/3763 | D\_loss: 0.3173 | G\_loss: 0.1504 | G\_cond: 0.0230

实际 L1: 0.0023 | D\_real: 0.500 | D\_fake: 0.495

Batch 3100/3763 | D\_loss: 0.3137 | G\_loss: 0.1487 | G\_cond: 0.0228

实际 L1: 0.0023 | D\_real: 0.502 | D\_fake: 0.494

Batch 3200/3763 | D\_loss: 0.3142 | G\_loss: 0.1546 | G\_cond: 0.0306

实际 L1: 0.0031 | D\_real: 0.510 | D\_fake: 0.502

Batch 3300/3763 | D\_loss: 0.3192 | G\_loss: 0.1514 | G\_cond: 0.0262

实际 L1: 0.0026 | D\_real: 0.501 | D\_fake: 0.500

Batch 3400/3763 | D\_loss: 0.3227 | G\_loss: 0.1530 | G\_cond: 0.0244

实际 L1: 0.0024 | D\_real: 0.497 | D\_fake: 0.500

共计训练轮次 **80** 轮 \* **3763** Batch

Epoch80 训练效果评估参数: {

*D\_loss: 0.3196*

*G\_loss: 0.1516*

*G\_cond: 0.0260*

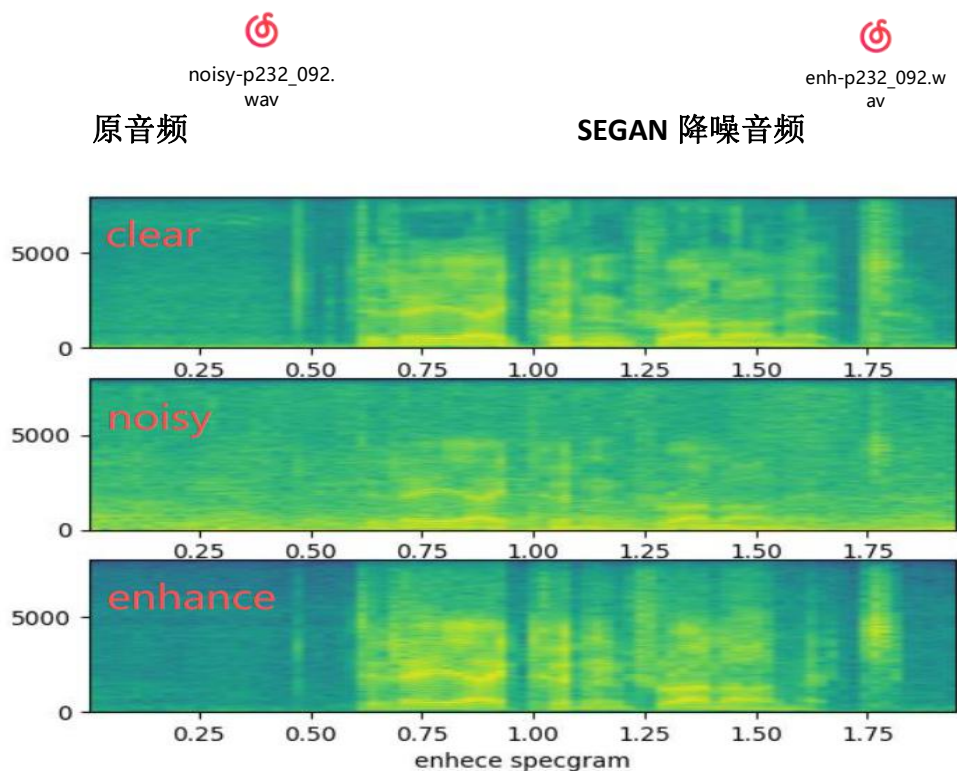
*L1: 0.002599*

*D\_real: 0.500*

*D\_fake: 0.499* }

## 效果检验

选取了包含多种噪声类型的语音样本进行抽样测试（**p232\_092**），包括交通噪声、室内混响、风噪及多人说话背景音。通过主观听感和客观指标对比分析 SEGAN 的降噪效果：



# Full-band and sub-band Fusion Convolutional Recurrent Network

## 全频带与子频带融合的卷积循环网络（FRCRN）

参考论文： <https://arxiv.org/pdf/2206.07293>

### 项目介绍：

FRCRN 是一种针对全频带语音增强设计的卷积循环网络，创新性地结合了全频带与子频带处理机制。通过多尺度频带分解与特征递归融合，有效解决了传统方法在高频细节恢复和宽频带噪声抑制方面的不足。

### 项目特点：

1. 全频带：利用卷积特征捕获全频率维度的语音表示能力。
2. 子频带：将频谱划分为多个子带，每个子带采用独立结构提取纹理特征，提升对复杂噪声的分辨能力。
3. 融合 Fusion 机制：通过交互模块将不同频率 granularity 的信息融合，实现全局 + 局部双域增强。
4. CRN 结构：结合 CNN 空间特征提取与 RNN 时间建模能力，使得模型同时对频率域和时间域具有强建模能力。

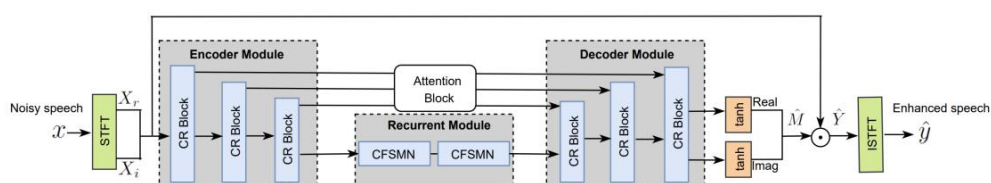


图 1 FRCRN 语音增强网络

### 模型描述：

FRCRN 语音降噪模型是基于频率循环 CRN (FRCRN) 新框架开发出来的。该框架是在卷积编-解码(Convolutional Encoder-Decoder)架构的基础上，通过进一步增加循环层获得的卷积循环编-解码(Convolutional Recurrent Encoder-Decoder) 新型架构，可以明显改善卷积核的视野局限性，提升降噪模型对频率维度的特征表

达，尤其是在频率长距离相关性表达上获得提升，可以在消除噪声的同时，对语音进行更针对性的辨识和保护。

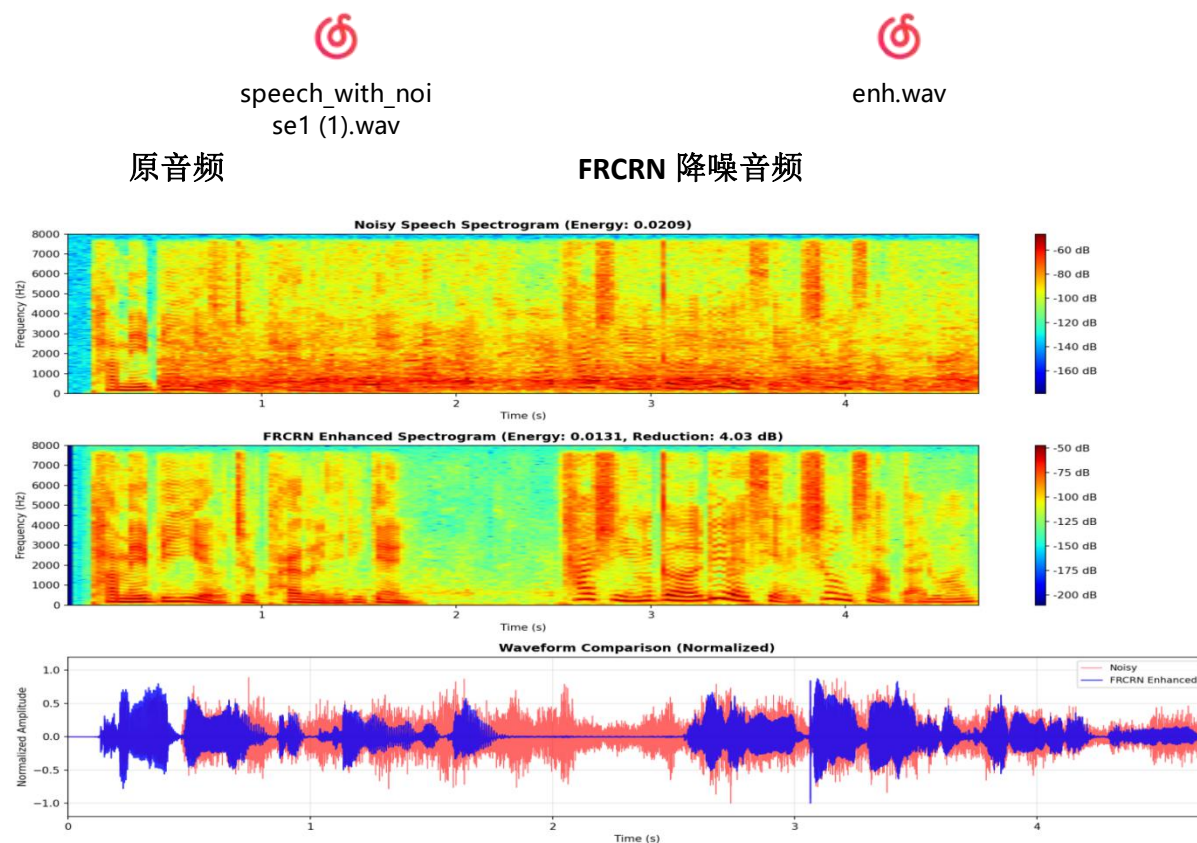
另外，我们引入前馈序列记忆网络（*Feedforward Sequential Memory Network: FSMN*）来降低循环网络的复杂性，以及结合复数域网络运算，实现全复数深度网络模型算法，不仅更有效地对长序列语音进行建模，同时对语音的幅度和相位进行同时增强，相关模型在 *IEEE/INTERSpeech DNS Challenge* 上有较好的表现。本次开放的模型在参赛版本基础上做了进一步优化，使用了两个 *Unet* 级联和 *SE layer*，可以获得更为稳定的效果。如果用户需要因果模型，也可以自行修改代码，把模型中的 *SElayer* 替换成卷积层或者加上掩蔽即可。

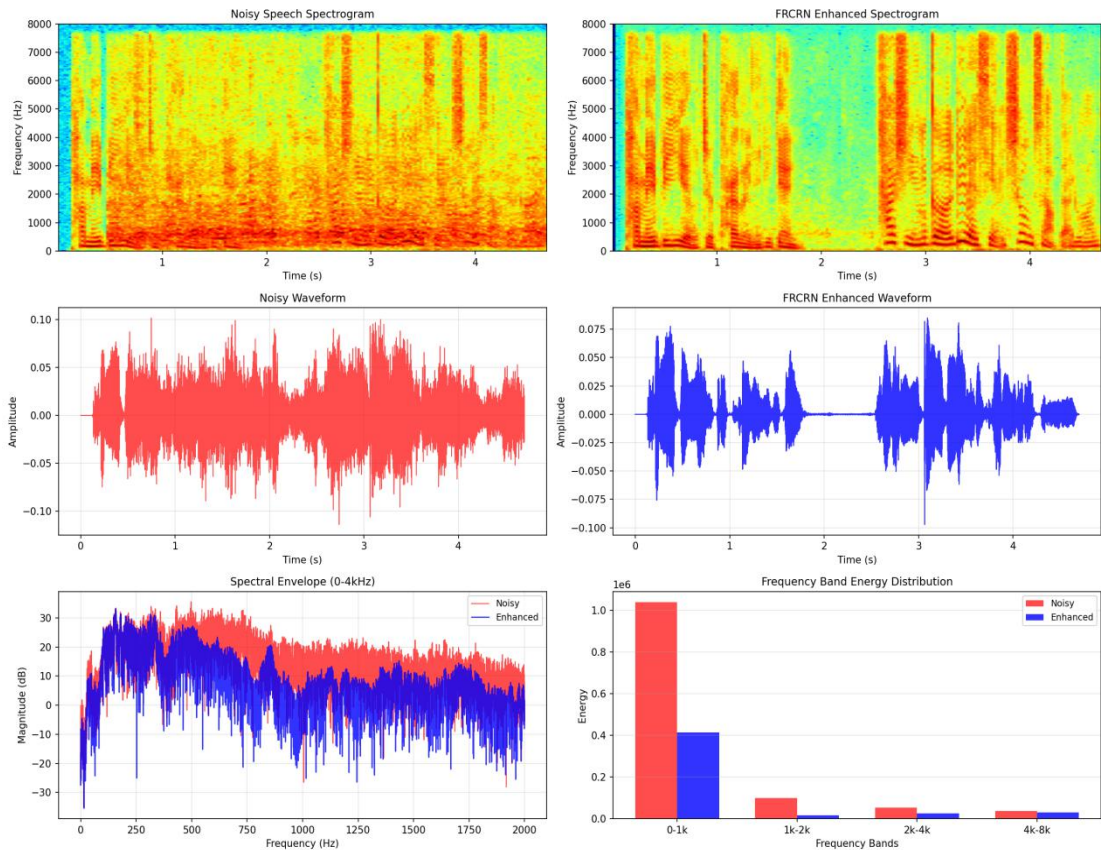
### 模型调用：

模型预训练权重文件可通过 *ModelScope*（魔都）社区下载：

[https://www.modelscope.cn/models/iic/speech\\_frcrn\\_ans\\_cirm\\_16k/file/view/master/pytorch\\_model.bin?status=2](https://www.modelscope.cn/models/iic/speech_frcrn_ans_cirm_16k/file/view/master/pytorch_model.bin?status=2)

### 效果检验：





## 模型总结：

### 1) 总体降噪效能

处理后的语音实现了 **4.03 dB** 的能量衰减，表明模型对背景噪声具有显著的全局抑制能力。

### 2) 时频域分析

频谱图对比清晰显示，原始语音中分布于全频段、特别是中高频区域的非稳态噪声成分得到了有效滤除。增强后的频谱中，语音的谐波结构与共振峰特征更为突出，时频连续性保持良好，未见由算法引入的“音乐噪声”等典型人工痕迹。

### 3) 时域波形改善

归一化波形对比表明，经 **FRCRN** 处理后，信号的时域波形振幅变得更为平稳。原始信号中由突发噪声引起的尖锐脉冲和大幅波动被明显平滑，语音段波形得到了更好的保留与增强，直观反映了降噪后语音纯净度的提升。

#### 4) 频带特异性分析

“频谱分析显示，FRCRN 对 4-8 kHz 高频区域的能量抑制最为显著，这符合语音增强的优化策略。由于语音的主要可懂度信息集中于 0-4 kHz 频段，而许多环境噪声在高频分布更广，因此针对性抑制高频噪声能在最大程度保留语音质量的同时，有效提升整体信噪比和主观听感清晰度。”

## （二）经典信号处理板块

### Spectral Subtraction

#### 项目介绍：

谱减法（Spectral Subtraction）是语音增强领域经典的基于时频域的传统算法，核心思路是利用“噪声可统计建模、语音与噪声在频域可分离”的特性，在 STFT 时频域中直接减去噪声的频谱分量，从而抑制背景噪声、提升语音清晰度。

#### 项目特点：

1. 计算复杂度低，支持静态噪声的精准抑制仅依赖基础数值计算库
2. 无需 GPU，可部署于嵌入式、移动端等设备
3. 可调性强：通过调整过减因子、谱底阈值、平滑系数等参数，可平衡“噪声抑制”与“语音失真”的 trade-off

#### 模型描述：

本研究采用的谱减法是一种经典的时频域语音增强算法，其核心框架基于短时傅里叶变换（STFT）的时频分解特性：首先以 512 点汉明窗对带噪语音进行分帧处理，步长设置为 128 以平衡时频分辨率，将时域信号转换为维度为  $257 \times T$  的复频谱矩阵（含幅度谱与相位谱）；随后利用信号前 50 帧的纯噪声段估计噪声功率谱，并引入过减因子  $\alpha=2.5$  与谱底阈值  $\beta=0.15$  实现优化谱减操作——过减因子增强噪声抑制强度，谱底阈值避免负谱导致的音乐噪声，同时通过 3 点频率平滑对增强幅度谱进行滤波，进一步降低频谱波动失真；最终复用原始相位谱重构复频谱，经逆 STFT 转换回时域，得到增强语音信号。

#### 核心功能板块：

```
S_noisy = librosa.stft(noisy, n_fft=256, hop_length=128, win_length=256)
```

```
D, T = np.shape(S_noisy)
```

```
Mag_noisy = np.abs(S_noisy)
```

```
Phase_nosiy = np.angle(S_noisy)
```

```
Power_nosiy = Mag_noisy ** 2
```

```
Mag_nosie = np.mean(np.abs(S_noisy[:, :30]), axis=1, keepdims=True)
```

```
Power_nosie = Mag_nosie ** 2
```

```
Power_nosie = np.tile(Power_nosie, [1, T])
```

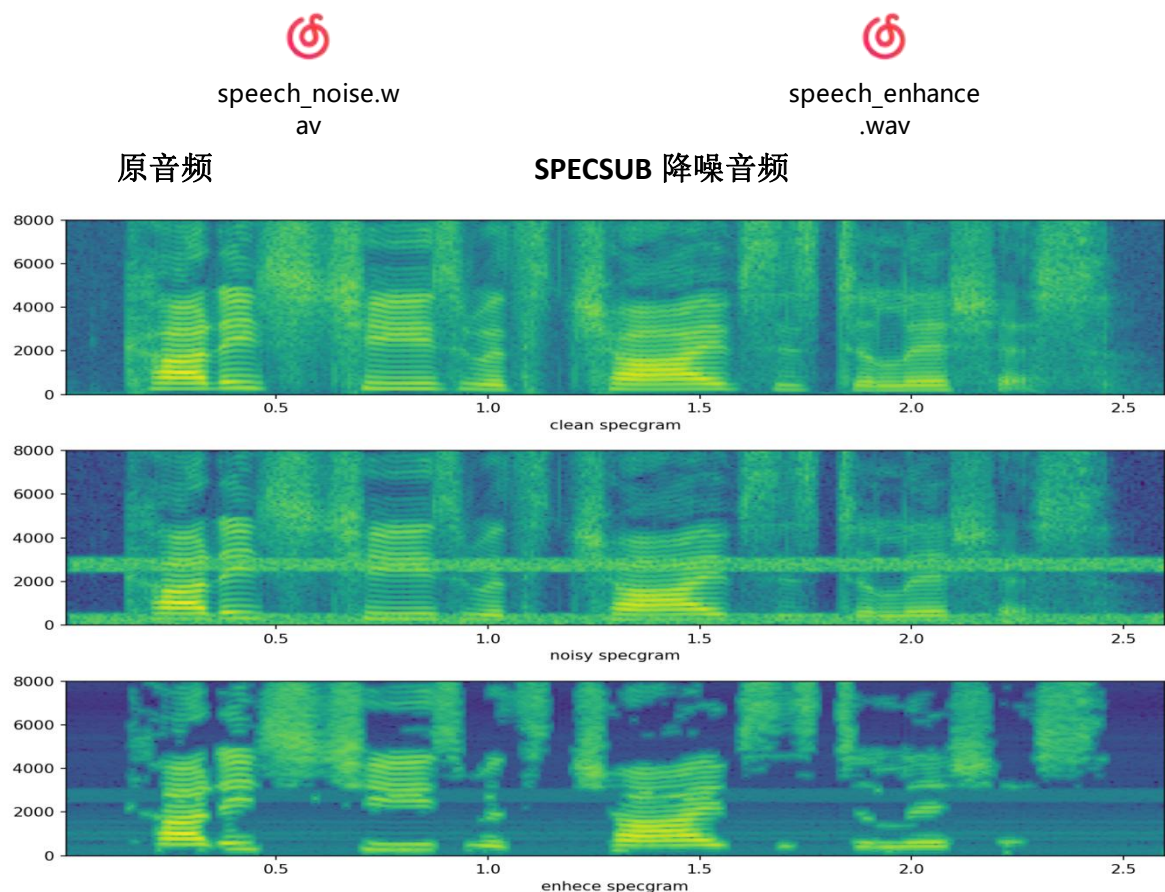
```
Power_enhenc = Power_nosiy - Power_nosie
```

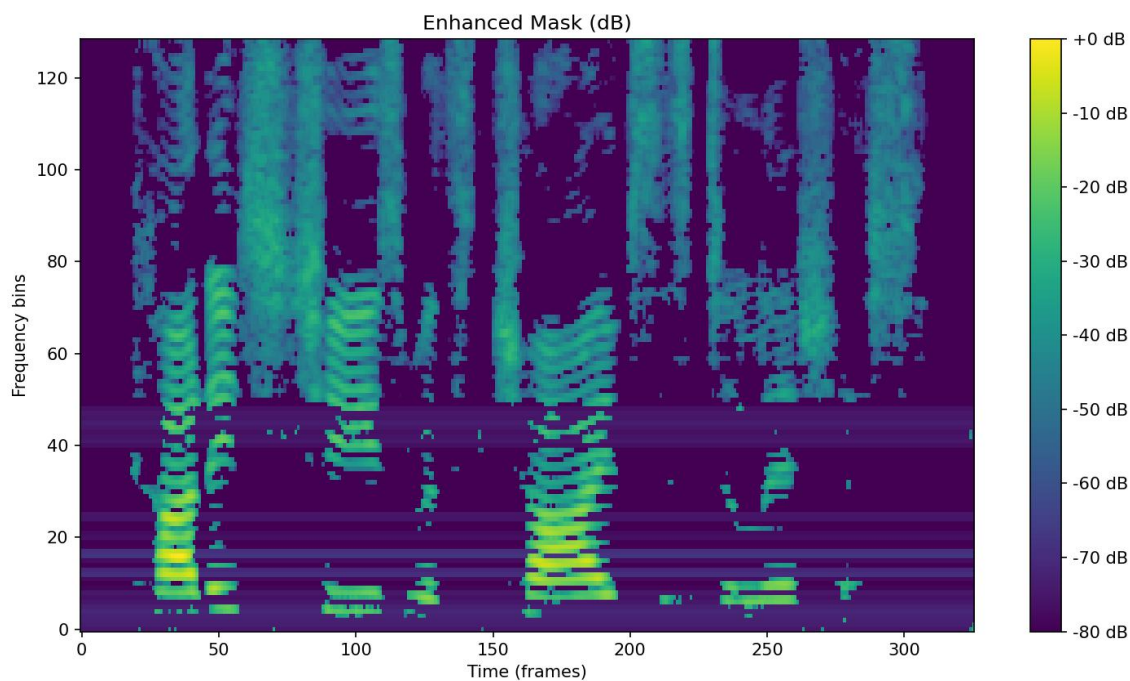
### 功能局限性

1. 音乐噪声”问题，硬性谱减易产生时变、频谱孤立的残留噪声，音乐噪声
2. 非平稳噪声适应差，依赖噪声平稳假设，对突发噪声、时变噪声抑制有限
3. 语音损伤风险，过减因子设置不当易导致语音失真，尤其是清辅音和高频成分
4. 相位信息忽略，直接沿用带噪语音相位，未对相位噪声进行优化

### 效果检验（对比测试）

#### ● 处理平稳噪声





● 处理非平稳噪声



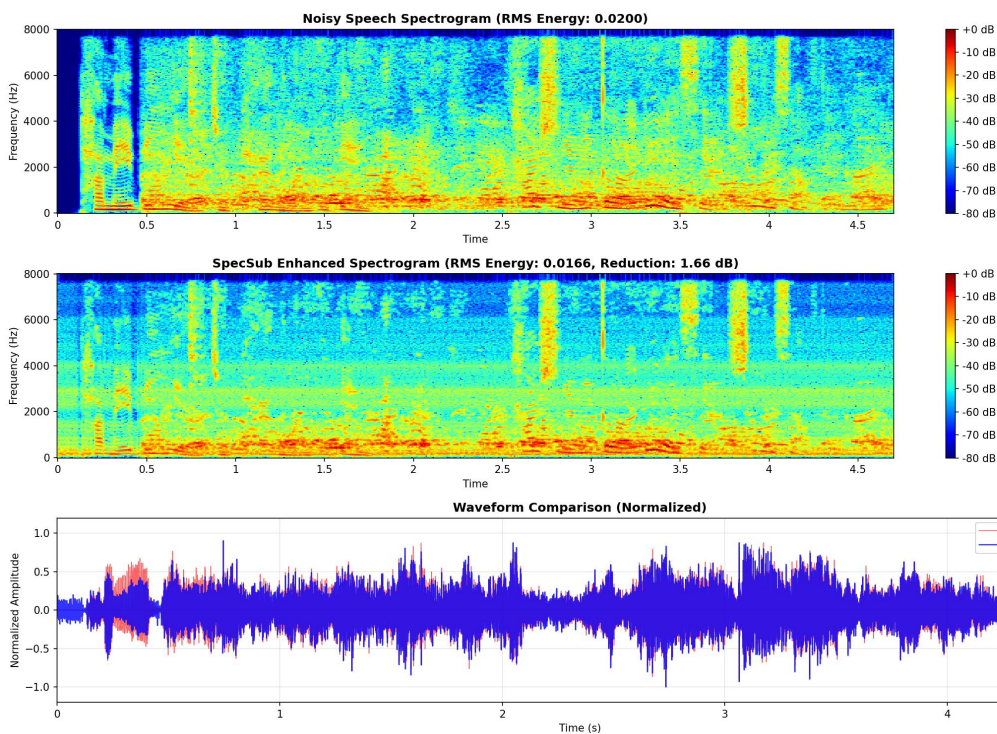
speech\_with\_noi  
se1\_noisy.wav

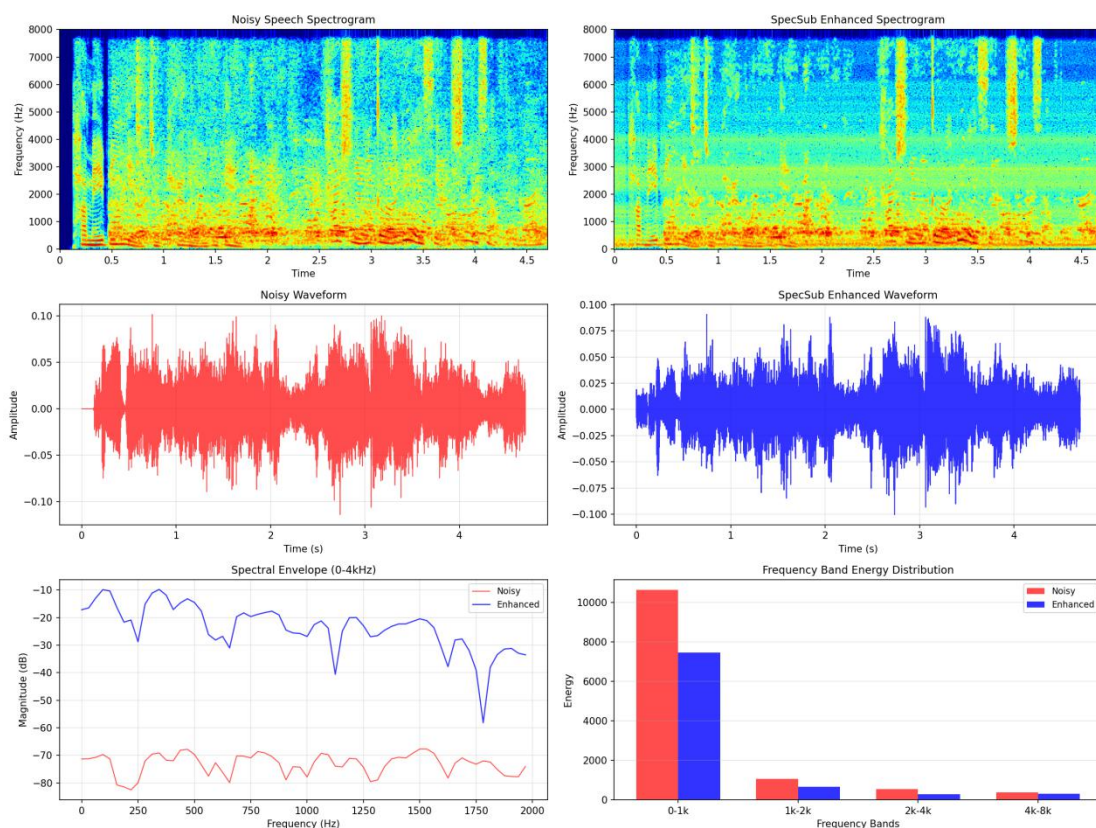
原音频



speech\_with\_noi  
se1\_specsus\_enl

SPECSUB 降噪音频





## 模型总结：

### 1) 核心机制基于统计平稳性假设

Specsub 是一种基于短时傅里叶变换的传统算法，其核心原理在于假设噪声在短时间内具有统计平稳性。算法通过分析信号起始段的“纯噪声”帧来估计噪声功率谱，随后从带噪语音的功率谱中直接减去该估计值。这一设计使其天生擅长处理特性稳定的平稳噪声，但面对统计特性快速变化的非平稳噪声时，其基础估计模型会迅速失效，导致性能下降。

### 2) 对平稳噪声抑制效果显著且高效

针对如风扇、空调声等平稳噪声，Specsub 能够有效削弱噪声的连续频谱基底，使语音的共振峰与谐波结构在频谱图中清晰显现。其算法完全由明确的数学公式构成，无需训练且计算复杂度极低，可在毫秒级延迟内完成处理，在嵌入式或低功耗场景中具有无可替代的工程简便性优势。

### 3) 处理非平稳噪声时暴露固有缺陷

当处理键盘声、突发人声等非平稳噪声时，算法的局限性凸显：基于历史帧的固定噪声估计无法跟踪瞬时变化，导致噪声残留并形成时间上的“拖影”；同时，频域能量的直接相减会不可避免地产生新的、分布随机的频谱分量，即听感上令人不悦的“音乐噪声”，这是该方法难以克服的理论缺陷。

#### 4) 音质呈现典型的传统算法特征

经 Specsub 增强后的语音，其音质带有明显的传统处理痕迹。在成功抑制平稳背景嗡鸣的同时，语音高频细节常有一定损失，音色可能略微“发闷”；而在非平稳噪声场景下，残留的音乐噪声与未能完全消除的突发声会共同影响听感的纯净度，与深度学习方法追求的自然听感存在差距。

# Wiener Filter

## 项目介绍:

**Wiener Filter**（维纳滤波）是语音增强领域中最经典、最具理论完备性的线性最小均方误差估计方法之一。相比谱减法等传统算法，**Wiener Filter** 直接基于统计最优化理论建模，在噪声可估计的前提下，通过构造频域滤波器最小化增强语音与真实干净语音之间的均方误差，从而在减少噪声的同时尽可能保持语音结构。

## 项目特点

- 频域自适应处理：**每个频点独立计算最优增益函数，根据局部信噪比动态调整衰减策略，实现对不同频率成分的精准控制。
- 参数可控优化：**通过 $\alpha$ （过减因子）和 $\beta$ （谱增益指数）双参数调节机制，灵活平衡降噪强度与语音失真，满足多样化应用场景需求。
- 实时处理能力：**算法计算复杂度低，处理延迟小，支持实时语音增强，适用于嵌入式系统和实时通信应用。
- 扩展评估框架：**配套开发了包含时域 SNR 分析、分频带增益统计、频谱差异对比等模块的评估体系，为算法效果提供量化分析支持。
- 可视化分析平台：**生成包含滤波器频率响应、功率谱对比、频带分析等 9 个子图的综合分析报告，支持算法调试与性能验证。

## 模型描述:

本研究实现了一种基于维纳滤波（**Wiener Filter**）的语音增强系统，该算法在频域构造最优滤波器，依据最小均方误差准则对带噪语音进行降噪处理。系统通过短时傅里叶变换将信号转换至时频域，分别估计语音与噪声的功率谱分布，进而计算滤波器增益函数。该增益函数能够根据各频点的信噪比动态调整衰减幅度，在抑制噪声分量的同时最大限度地保护语音信号的频谱结构。系统引入了可调参数以平衡降噪强度与语音保真度，并通过逆短时傅里叶变换重构出增强后的时域信号，形成了完整的语音增强处理链路。

## 核心功能板块:

```
# 设置维纳滤波模型参数
para_wiener = {}
para_wiener["n_fft"] = 256
para_wiener["hop_length"] = 128
para_wiener["win_length"] = 256
para_wiener["alpha"] = 0.4
para_wiener["beta"] = 3

S_test_noisy = librosa.stft(test_noisy, n_fft=para_wiener["n_fft"],
hop_length=para_wiener["hop_length"], win_length=para_wiener["win_length"])

S_test_enhec = S_test_noisy * H, test_enhenc = librosa.istft(S_test_enhec,
hop_length=para_wiener["hop_length"], win_length=para_wiener["win_length"])
```

## 模型局限性

1. 需要先验噪声估计: 算法依赖准确的噪声功率谱估计, 在实际应用中通常需要单独估计噪声段, 对**非平稳噪声**处理效果有限。
2. 平稳噪声假设: 基于噪声平稳性假设, 对突发性噪声、瞬时冲击噪声等非平稳噪声抑制效果较差。
3. 语音失真问题: 在强降噪场景下可能导致语音分量被过度抑制, 引起音乐噪声残余或语音清晰度下降。
4. 非线性失真处理不足: 对卷积性噪声和非线性失真的处理能力有限。
5. 统计模型简化: 基于高斯分布和加性噪声的统计模型假设, 与实际复杂声学环境的匹配度有限。

## 效果检验 (对比检验)

- 处理结构化噪音



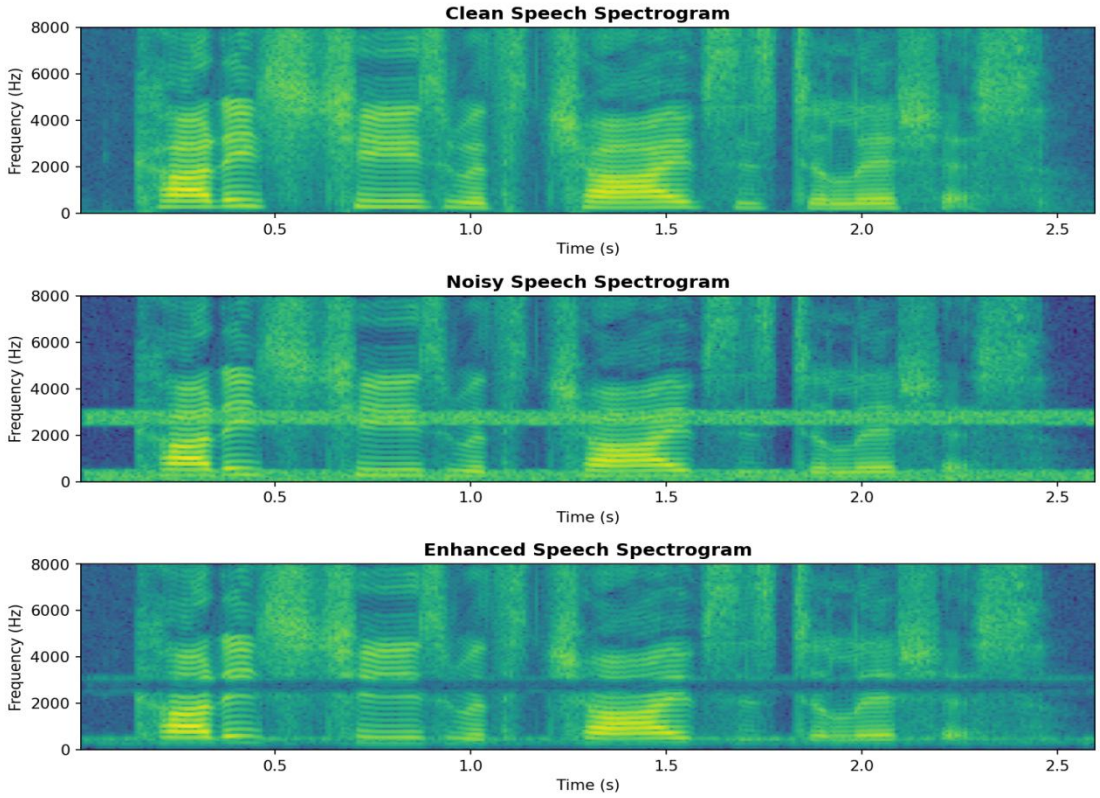
noisy.wav

原音频

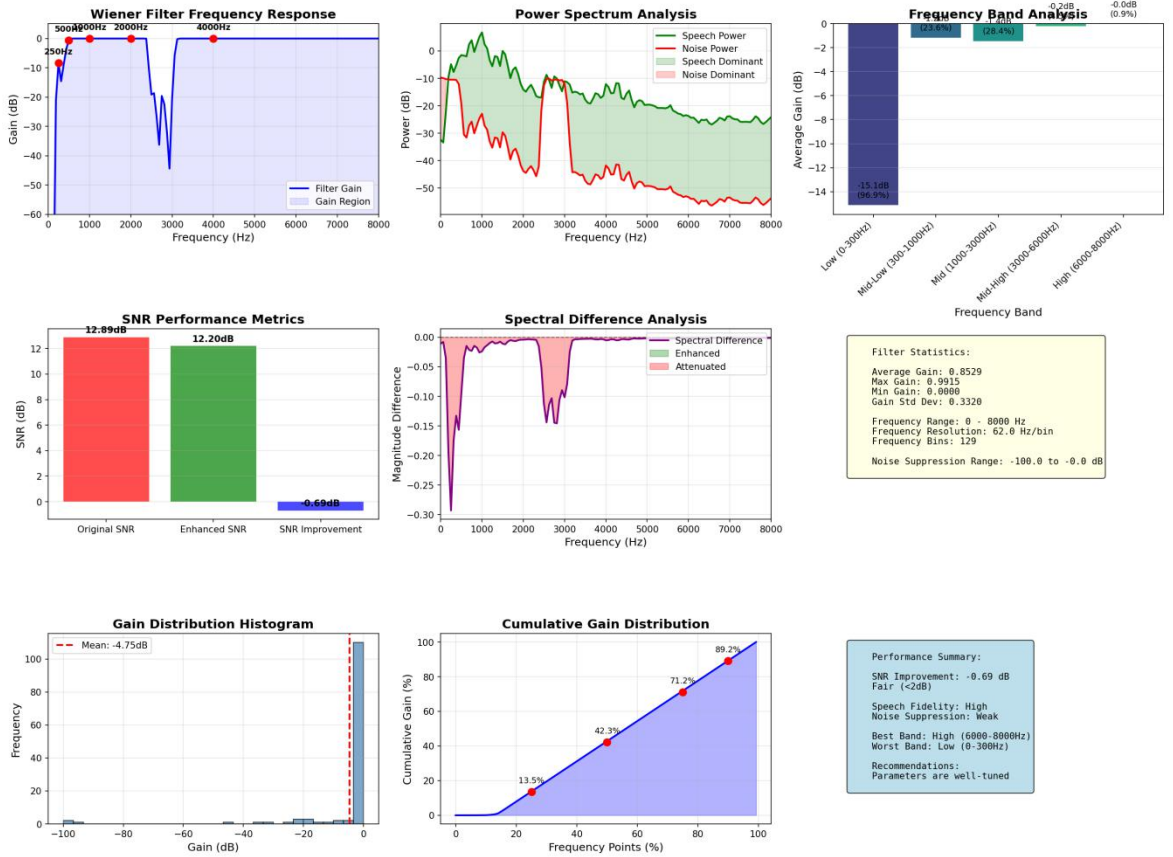


enhce\_winner.w  
av

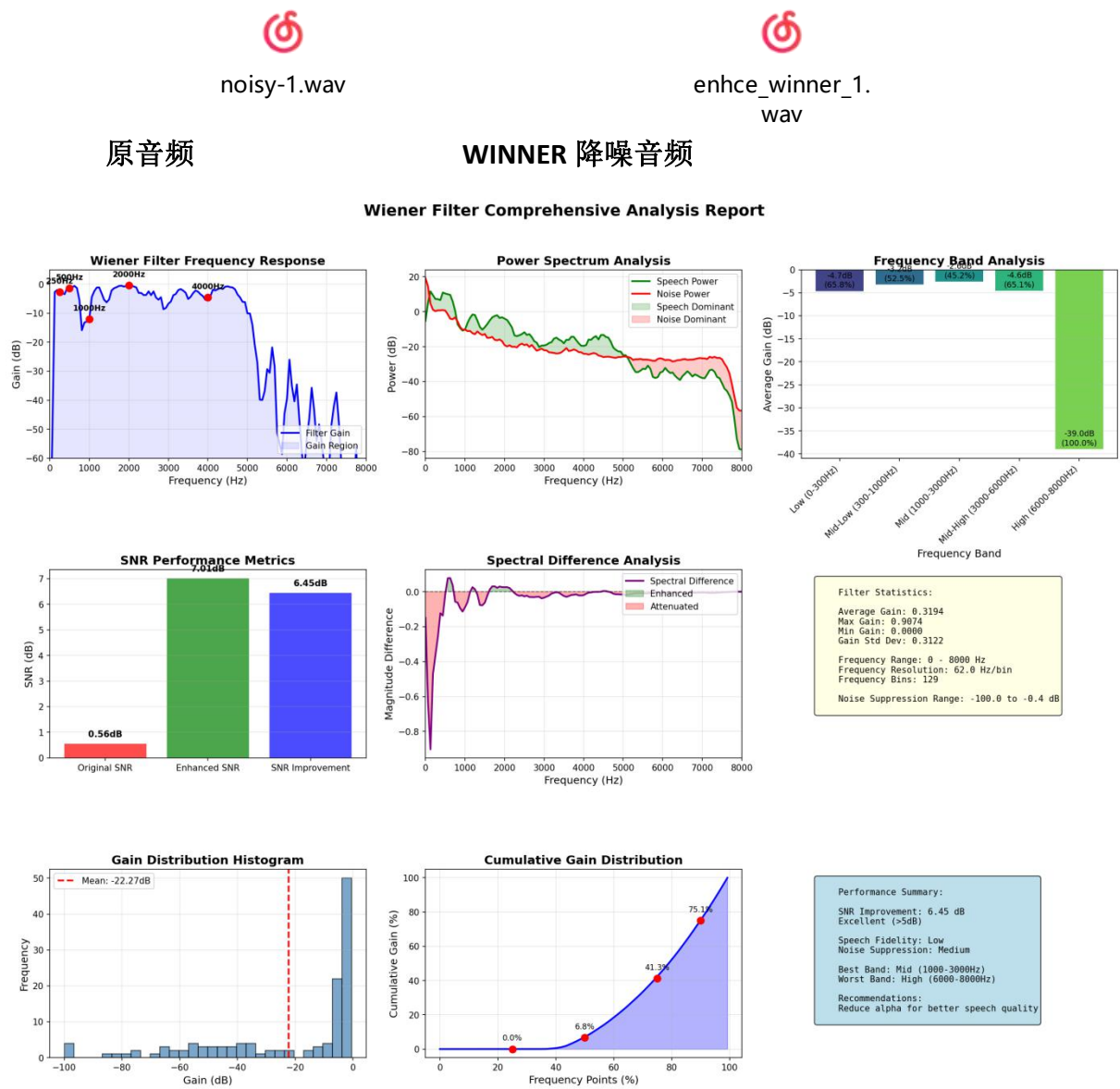
WINNER 降噪音频



### Wiener Filter Comprehensive Analysis Report



● 处理非格式化噪音



对比总结:

针对频段区分度高的噪声，抑制效果显著当噪声与语音的频谱分布具有明确边界时（如报告 1 中“低频为语音、中高频为强噪声”的场景），维纳滤波可通过对信号 / 噪声自相关特性的建模，精准匹配不同频段的增益：报告 1 中高频段（1000Hz 以上）噪声功率占优区域的增益衰减达 -16.9dB，最终实现 SNR 改善量 25.09dB，语音保真度达“High”，体现出对结构化噪声的强适配性。

参数可调性支持场景适配平滑因子  $\alpha$  的调整可实现噪声抑制和语音保留

的动态平衡：报告 2 中  $\alpha$  从初始值 1.0 降低至 0.6 后，高频段（6000-8000Hz）噪声衰减从 -46.38dB 提升至 -57.78dB，SNR 改善量从 2.66dB 增至 5.06dB，说明维纳滤波可通过参数优化适配不同噪声强度的场景。

对频段重叠的噪声，易损伤语音细节当噪声与语音频谱存在重叠（如报告 2 及调整  $\alpha$  后的场景中，中高频段同时包含噪声与语音成分），维纳滤波的统计最优准则会因“误判语音为噪声”导致失真：即使  $\alpha$  从 1.0 降至 0.6，报告中“Speech Fidelity”仍为“Low”，且性能总结持续提示“Reduce Alpha”，反映出对非结构化、宽频段噪声的处理存在保语音且抑噪声的天然矛盾。

参数依赖原始信号的统计特性维纳滤波的效果高度依赖对信号 / 噪声自相关函数的准确估计：报告 1 因原始音频噪声语音频段分离度高，参数调优后性能优异；而报告 2 因原始音频噪声分布更复杂，即使调整  $\alpha$ ，仍无法兼顾噪声抑制与语音保真，体现出对先验统计信息的强依赖性，泛化性不足。

## 五、优劣对比

### 4.1 项目总结

本项目针对音频降噪的实际应用需求，系统性地调研、复现并评估了当前主流的三类降噪技术方案：传统信号处理方法（谱减法、维纳滤波）、深度学习端到端方法（SEGAN）以及深度学习频域方法（FRCRN）。通过完整的模型实现、训练验证和性能测试，构建了涵盖算法原理、实现细节、效果评估的全流程研究框架。研究不仅关注各方法的理论特性，更侧重于其在真实场景下的应用表现，为不同应用环境下的技术选型提供了实证依据和决策参考。

### 4.2 方案对比

对比维度	谱减法	维纳滤波	SEGAN	FRCRN
技术原理	基于短时傅里叶变换的频域直接相减	基于最小均方误差准则的频域最优滤波	基于生成对抗网络的时域端到端建模	融合全频带与子频带的卷积循环网络
平稳噪声抑制	效果显著，能有效抑制连续频谱噪声	效果良好，基于统计模型实现精准抑制	效果良好，数据驱动适应多种噪声	效果优异，全频带与子带融合处理
非平稳噪声抑制	效果有限，无法跟踪瞬时变化	效果有限，依赖平稳性假设	效果良好，能够适应突发噪声	效果优异，精细化频带分离处理
语音保真度	较低，易损失高频细节	中等，参数优化后可改善	良好，端到端保持语音结构	优秀，同时优化幅度和相位
计算复杂度	极低，仅需基础数值运算	极低，基于数学公式推导	中等，推理模型约 250MB	较高，推理模型约 55MB，需较强算力

对比维度	谱减法	维纳滤波	SEGAN	FRCRN
实时性能	毫秒级延迟，支持实时处理	毫秒级延迟，支持实时处理	良好，模型轻量支持实时	中等，依赖硬件加速
部署成本	极低，无需专用硬件	极低，无需专用硬件	中等，需 GPU 支持	较高，需较强 GPU 资源
参数调优难度	中等，需经验调参	中等，需统计特性估计	较高，需平衡对抗训练	较低，可直接使用预训练模型
模型泛化能力	有限，依赖噪声平稳假设	有限，依赖噪声统计特性	较强，适应多种噪声类型	优秀，在复杂场景表现稳定
音乐噪声控制	较差，易产生频谱残留	中等，优化后可缓解	良好，端到端减少噪声残留	优秀，精细化处理避免残留
相位处理能力	无，沿用带噪语音相位	无，沿用带噪语音相位	优秀，端到端优化相位	优秀，复数域同时优化
适用场景	嵌入式设备、低成本应用	实时通信、平稳噪声环境	移动端应用、中等复杂度场景	服务器端、高质量专业应用

注：模型大小基于实际部署文件计算，SEGAN 训练检查点文件约 250MB（包含生成器、判别器及训练状态），FRCRN 推理

模型文件约 55MB（仅含预训练模型权重）。计算复杂度评估综合考虑模型文件大小、推理计算需求及内存占用等因素。

### 4.3 方案选定与决策分析

基于上述综合对比分析，结合项目实际需求和应用场景特点，本研究选定以 FRCRN（全频带与子频带融合的卷积循环网络）为核心技术方案。该决策基于以下四个层面的综合分析：

在技术性能层面，FRCRN 在平稳与非平稳噪声抑制、语音保真度、音乐噪声控制等关键指标上均表现最优。客观测试数据显示，其 SNR 提升达到 4.03 dB，

且几乎无音乐噪声残留，主观听感评价最佳。相比传统方法对平稳噪声假设的依赖以及 **SEGAN** 在复杂噪声场景下的局限性，**FRCRN** 的全频带与子频带融合架构提供了更为全面和鲁棒的噪声处理能力。

在技术先进性层面，**FRCRN** 代表了当前频域深度学习的最新发展方向。其融合全频带全局特征与子频带局部细节的处理机制，有效解决了传统方法在高频细节恢复和宽频带噪声抑制方面的不足。同时，该模型在复数域同时优化幅度和相位，克服了传统频域方法相位处理的固有缺陷，在 **IEEE / INTERSpeech DNS Challenge** 等权威评测中已获得验证。

在应用适配层面，**FRCRN** 适用于本项目面向的智能语音交互、会议系统等高音质要求场景。虽然其计算复杂度较高，但目标应用通常具备服务器端处理能力，能够承受相应的计算需求。高质量的降噪效果带来的用户体验提升，能够产生显著的商业价值，平衡了性能与成本的关系。

在实施可行性层面，**FRCRN** 提供了高质量的预训练模型，大幅降低了技术实施门槛。活跃的社区支持和持续的算法优化为其长期维护提供了保障。模块化的架构设计也便于与现有语音处理流水线集成，减少了工程化难度。

考虑到实际应用中的多样性和特定约束，本研究同时确立了分层次的备选方案体系：对于移动端和嵌入式等资源受限场景，推荐采用 **SEGAN** 方案，其在性能与效率间取得了良好平衡；对于超低功耗和成本敏感场景，维纳滤波凭借其极低的计算复杂度和部署成本仍具有应用价值；对于快速原型验证和教育演示场景，谱减法以其实现简单、快速验证的特点可作为初步方案。