

ESC-50 环境声音检索与分类系统报告

数字信号处理课程大作业

2025 年 12 月 25 日

摘要

本报告详细阐述了基于 ESC-50 数据集的环境声音检索与分类系统的设计与实现。项目采用“从零构建”的理念，手动实现了快速傅里叶变换 (FFT)、短时傅里叶变换 (STFT) 及梅尔频率倒谱系数 (MFCC) 等核心 DSP 算法，深入理解了频域分析的数学本质。在此基础上，构建了基于 MFCC 特征的无监督检索系统，并设计了 ResNet 风格的卷积神经网络 (CNN) 进行有监督分类，实现了 75.00% 的测试准确率。进一步地，项目引入了 PANNs、AST、CLAP 等前沿预训练模型进行迁移学习对比，其中 CLAP 模型达到了 97.25% 的分类精度。最后，本研究还探索了基于 Gemini 大语言模型和 CLAP 零样本学习的音频理解能力。实验结果不仅验证了经典信号处理方法的有效性，也展示了现代深度学习在音频领域的强大潜力，体现了从底层信号解析到高层语义理解的技术跨越。

目录

1 引言	3
2 数据集与实验环境	3
2.1 ESC-50 数据集概述	3
2.2 数据划分与预处理	3
3 核心与基础：DSP 算法实现及其物理意义	4
3.1 快速傅里叶变换 (FFT)	4
3.2 短时傅里叶变换 (STFT)	5
3.3 梅尔频率倒谱系数 (MFCC)	5
3.3.1 Log-Mel 频谱图	6
3.3.2 倒谱系数 (MFCC)	6
4 任务一：基于 MFCC 的声音检索系统	7
4.1 系统设计	7
4.2 实验结果：帧长与帧移的影响	7

目录	2
5 任务二：基于 CNN 的分类系统	8
5.1 模型架构：ResNet-Audio	8
5.2 训练策略	8
5.3 实验结果与分析	9
6 大模型时代的音频理解	9
6.1 模型介绍	10
6.2 对比实验结果	10
7 总结与思考	11
7.1 不变的数学基石	11
7.2 演进的特征范式	11
A 附录：代码与文件索引	12

1 引言

环境声音识别 (Environmental Sound Classification, ESC) 是机器听觉领域的核心任务之一，旨在让计算机系统能够感知并理解周围环境中的非语言声音。与语音识别和音乐检索不同，环境声音通常具有非平稳性、背景噪声复杂、声源多样（如自然界声音、动物叫声、城市噪音等）的特点，这对信号处理和特征提取提出了更高的要求。

ESC-50 (Environmental Sound Classification 50) 是该领域的各类算法得标准评估基准。本项目以此为基础，构建了一套完整的音频分析系统。本项目的核心目标不仅在于追求高识别率，更在于通过“从底层算法到顶层应用”的完整实现，深入探讨数字信号处理 (DSP) 技术与现代人工智能 (AI) 模型的内在联系。

2 数据集与实验环境

2.1 ESC-50 数据集概述

ESC-50 数据集由 Karol Piczak 于 2015 年发布，包含 2000 条带标注的环境音频片段。

表 1: ESC-50 数据集统计信息

参数	描述
样本总数	2000 条
类别总数	50 类 (分为 5 大组)
每类样本数	40 条
音频时长	5 秒
采样率	44.1 kHz
数据格式	单声道 WAV
划分	5-Fold Cross Validation

数据集的 50 个类别被划分为 5 个大类，涵盖了生活中常见的声音场景（见表2）。

2.2 数据划分与预处理

本项目严格遵循官方推荐的 5 折交叉验证：

- 训练集/数据库：Fold 1-4 (1600 条样本)。
- 测试集/查询集：Fold 5 (400 条样本)。

表 2: ESC-50 类别分布

大类	包含类别示例
动物声音 (Animals)	狗叫、猫叫、猪叫、牛叫、鸟鸣等
自然界声音 (Natural)	雨声、海浪、风声、雷声、水流等
人类非语音 (Human)	咳嗽、喷嚏、呼吸、脚步声、笑声等
室内声音 (Interior)	敲门、键盘声、闹钟、吸尘器、玻璃破碎等
城市/室外噪声 (Exterior)	直升机、电锯、警笛、汽车喇叭、发动机等

在 DSP/MFCC 与 CNN 训练/检索流程中，音频通过 `load_audio` 按配置采样率重采样（默认 44.1kHz，可在脚本参数中调整），并用 `normalize_audio` 做峰值归一化到 [-1,1]。CLAP/Gemini 等则使用各自模型采样率。

3 核心与基础：DSP 算法实现及其物理意义

傅里叶分析构成了音频信号处理的基石。在本项目中，我们没有调用现成的函数（如 ‘numpy.fft’ 或 ‘librosa.stft’），而是选择手工实现每一行核心代码，以求彻底掌握其数学原理与物理意义。

3.1 快速傅里叶变换 (FFT)

傅里叶变换揭示了信号的时频二象性：任何时域信号都可以看作是不同频率正弦波的叠加。离散傅里叶变换 (DFT) 将这一思想数字化：

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

直接计算 DFT 的时间复杂度为 $O(N^2)$ 。本项目实现了经典的 **Cooley-Tukey Radix-2 算法**，利用旋转因子 $W_N^{kn} = e^{-j2\pi kn/N}$ 的周期性和对称性，将 DFT 递归分解为偶数项和奇数项两部分：

$$X[k] = \text{DFT}_{N/2}\{x_{\text{even}}\}[k] + W_N^k \cdot \text{DFT}_{N/2}\{x_{\text{odd}}\}[k] \quad (2)$$

通过递归分治，复杂度降低至 $O(N \log N)$ 。

算法实现细节：

1. **位反转置换 (Bit-reversal Permutation)**: 为了实现原位运算 (In-place)，输入序列需要按索引的二进制位反转顺序重排。例如， $N = 8$ 时，索引 1(001₂) 变为 4(100₂)。

2. 蝶形运算 (Butterfly Operation)：核心计算单元，通过加减法结合旋转因子 W_N^k 完成两点 DFT。

我们通过与标准库 ‘numpy.fft’ 对比验证了实现的精度，复数域相对误差仅为 2.54×10^{-8} ，证明了底层实现的准确性。

3.2 短时傅里叶变换 (STFT)

由于环境声音往往是非平稳的（例如狗叫声是间歇的，雨声是持续的），单纯的 FFT 无法描述频率随时间的变化。STFT 引入了“时间窗”的概念，对信号分帧加窗处理：

$$STFT\{x[n]\}(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n - mH]e^{-j\omega n} \quad (3)$$

其中 $w[n]$ 为窗函数， H 为帧移 (Hop Length)， m 为帧索引。

本项目使用 **Hann** 窗以减少频谱泄露：

$$w[n] = 0.5 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (4)$$

物理意义： STFT 产生了一个二维的时频矩阵（频谱图）。这是一个包含丰富信息的图像数据，它打通了音频处理与计算机视觉 (CV) 的桥梁，使得我们后续能够使用卷积神经网络 (CNN) 来处理一维的声音信号。

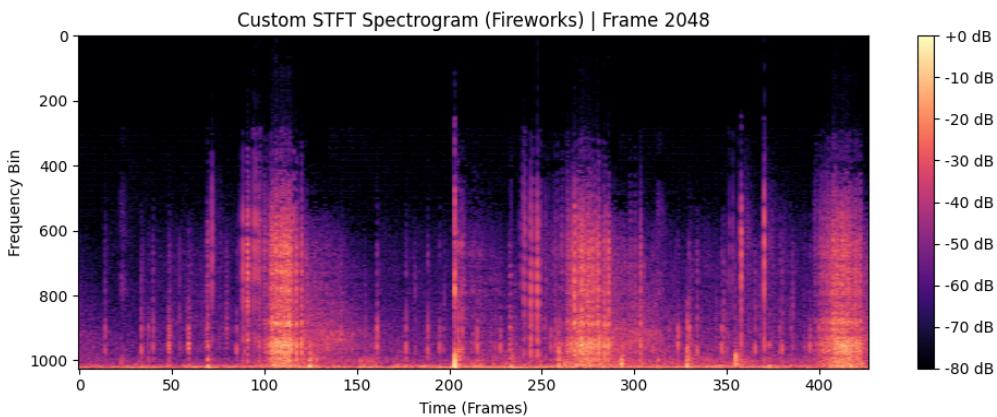


图 1：自定义 STFT 频谱图可视化（示例：Fireworks）

3.3 梅尔频率倒谱系数 (MFCC)

虽然 STFT 提供了完整的时频信息，但它并不符合人耳的听觉特性。人耳对频率的感知是非线性的（对低频更敏感）。MFCC 通过以下步骤模拟这一特性：

3.3.1 Log-Mel 频谱图

在进行倒谱变换之前，我们需要先计算 Log-Mel 频谱图，这是许多现代深度学习模型（如所有的 ResNet、AST、PANNs）的标准输入特征。

1. **预加重 (Pre-emphasis):** 使用滤波器 $y[n] = x[n] - \alpha x[n-1]$ ($\alpha = 0.97$) 提升高频分量，平衡频谱能量。2. **Mel 滤波器组:** 将线性频率 f 映射到 Mel 非线性尺度 m :

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (5)$$

在此尺度上设计三角滤波器组，计算每个滤波器的对数能量。

Log-Mel 频谱图有效地压缩了频率维度（从 1025 维线性频率压缩到 40-128 维 Mel 频率），同时保留了人耳感知的关键信息。

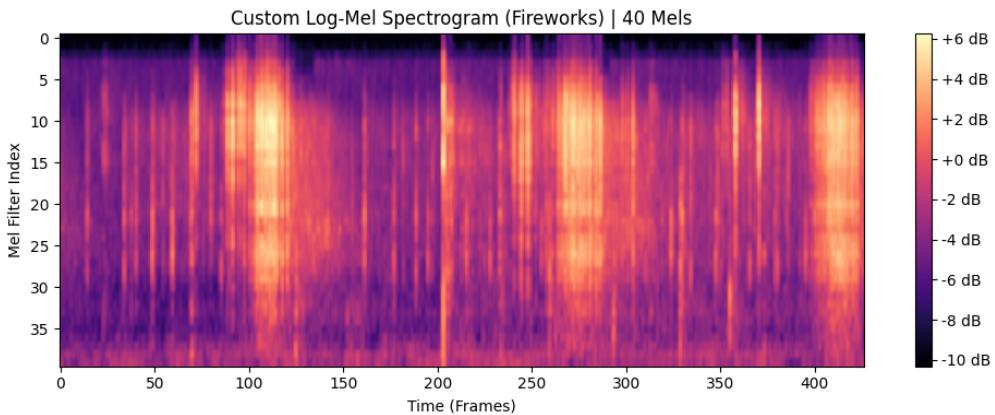


图 2: 自定义 Log-Mel 频谱图可视化（示例：Fireworks, 40 Mels）

3.3.2 倒谱系数 (MFCC)

为了去除频谱特征的相关性，我们对 Log-Mel 谱进行离散余弦变换 (DCT-II):

$$c[k] = 2 \sum_{n=0}^{N-1} x[n] \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right] \quad (6)$$

通常取前 13 维系数作为 MFCC 特征。

数学与感知的桥梁: MFCC 舍弃了详细的细微末节，保留了声音的“包络”特征（共振峰结构），这对于区分音色 (Timbre) 至关重要。从图3中可以看出，MFCC 的能量主要集中在低阶系数（底部），这反映了信号的主要频谱包络信息。

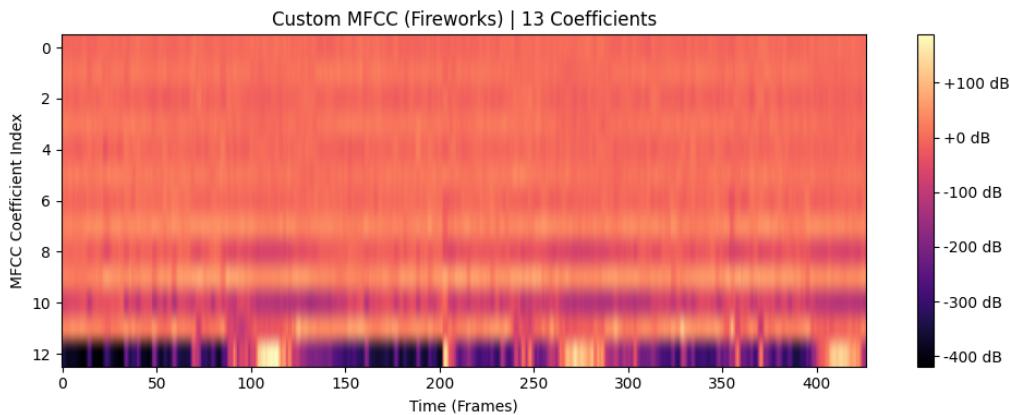


图 3: 自定义 MFCC 系数可视化（示例：Fireworks，13 系数）

4 任务一：基于 MFCC 的声音检索系统

4.1 系统设计

检索系统的核心在于如何度量两个音频片段的相似性。我们采用 MFCC 作为特征载体：

- **特征提取**: 计算每一帧的 13 维 MFCC 系数。
- **特征聚合**: 计算整个音频片段 MFCC 的均值向量 (Mean) 和标准差向量 (Std)，拼接得到 26 维的全局特征向量。这相当于捕捉了声音的平均音色和音色的变化范围。
- **相似度度量**: 使用余弦相似度 (Cosine Similarity)。

4.2 实验结果：帧长与帧移的影响

我们在不同的帧长 (512-4096) 和帧移 (256-2048) 下进行了全排列网格搜索实验。Top-10 和 Top-20 的检索精度如表3和表4所示。

表 3: MFCC 检索 Top-10 精度 (Frame Length vs Hop Length)

Frame Length	Hop Length			
	256	512	1024	2048
512	0.6500	0.6475	0.6450	0.6525
1024	0.6575	0.6525	0.6525	0.6400
2048	0.6775	0.6775	0.6625	0.6600
4096	0.6575	0.6575	0.6625	0.6600

表 4: MFCC 检索 Top-20 精度 (Frame Length vs Hop Length)

Frame Length	Hop Length			
	256	512	1024	2048
512	0.7775	0.7800	0.7775	0.7725
1024	0.7925	0.7900	0.7900	0.7875
2048	0.7950	0.7950	0.7950	0.7925
4096	0.7800	0.7775	0.7775	0.7750

关键发现：

- 最佳配置：**帧长 2048、帧移 256/512 时取得了最佳性能 (Top-10 67.75%)。这说明对于环境声音，较长的分析窗口 (约 46ms) 能提供更好的频率分辨率。
- 过度平滑：**当帧长增加到 4096 时，性能反而下降，可能是因为时间分辨率过低，平滑掉了短促声音的细节。
- 聚类效应：**Top-20 精度显著高于 Top-10，说明同类声音在特征空间中形成了较好的聚类。

5 任务二：基于 CNN 的分类系统

如果说 MFCC 是人工设计的特征工程，那么卷积神经网络 (CNN) 则是数据驱动的特征学习。

5.1 模型架构：ResNet-Audio

我们设计了一个 ResNet-18 变体 (参数量约 450K)，专门处理单通道的 Log-Mel 频谱图，模型架构见图4。

5.2 训练策略

- 特征输入：**40 维 Log-Mel 频谱图。
- 优化器：**Adam，学习率 0.001。
- 损失函数：**交叉熵损失 (CrossEntropyLoss)。
- 训练轮次：**50 Epochs。
- 批大小：**32。

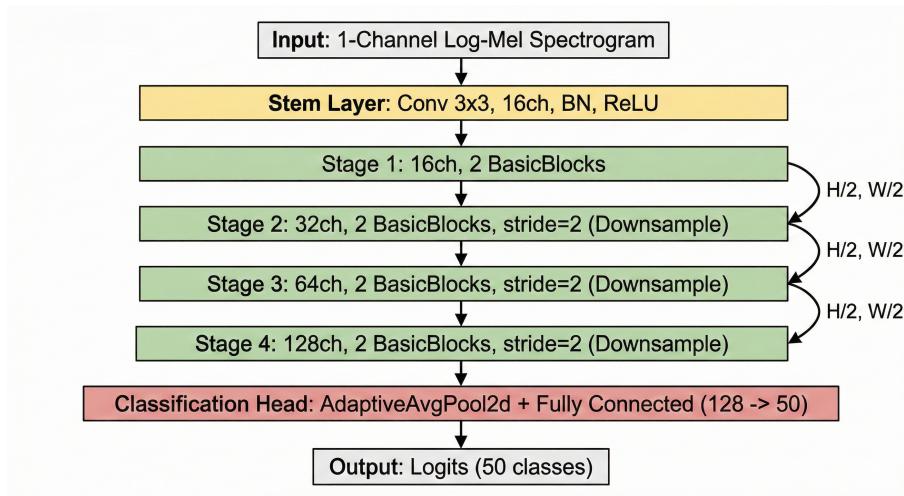


图 4: ResNet-Audio 模型架构

5.3 实验结果与分析

我们同样进行了帧长与帧移的网格搜索，结果如表5所示。

表 5: CNN 分类测试准确率 (Frame Length vs Hop Length)

Frame Length	Hop Length			
	256	512	1024	2048
512	0.6725	0.6975	0.7025	0.6450
1024	0.7025	0.6950	0.7225	0.6700
2048	0.6925	0.7150	0.7500	0.7200
4096	0.6650	0.7000	0.7100	0.7000

结果分析：

- 最优性能**: 在帧长 2048、帧移 1024 时达到最高准确率 **75.00%**。
- 2:1**: 表现较好的组合 (1024/512, 2048/1024) 的帧长与帧移比值通常为 2:1，这保证了 50% 的帧重叠率，既保留了信息又避免了过度冗余。
- 过拟合现象**: 从训练曲线 (图5) 可见，训练集准确率最终超过 90%，而测试集停滞在 75%，表明模型在小数据集上存在过拟合倾向。

6 大模型时代的音频理解

为了探索音频理解的上限，我们引入了基于大规模预训练数据的模型。

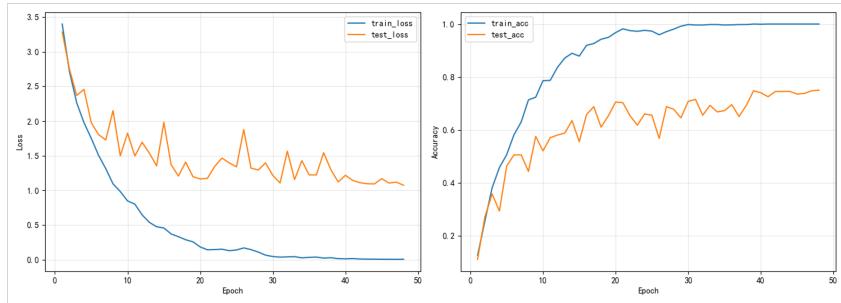


图 5: CNN 训练过程曲线 (训练集 Acc > 90%，验证集 Acc 停滞在 75%，显示过拟合)

6.1 模型介绍

- **PANNs (Cnn14)**: 在大规模音频数据集 AudioSet (200 万条) 上预训练的 CNN 模型 [3]。
- **AST (Audio Spectrogram Transformer)**: 基于 Vision Transformer 架构，同样在 AudioSet 上预训练 [2]。
- **CLAP (Contrastive Language-Audio Pretraining)**: 在 LAION-Audio-630K 图文对上进行对比学习预训练，支持零样本分类 [5]。
- **Gemini 3 Flash**: Google 目前 (2025 年 12 月) 的次旗舰多模态大语言模型，具备通用的多模态理解能力 [1]。

6.2 对比实验结果

我们对比了自研模型与各基线模型的性能 (见表6)。迁移学习采用 Linear Probe 方式 (冻结 backbone, 训练线性头)。

表 6: 不同模型性能对比总表

模型	方法	预训练数据	准确率
ResNet (Ours)	Supervised (Scratch)	None (仅 ESC-50)	75.00%
Gemini Flash	Zero-shot LLM	Multimodal (Web)	78.00%
Human Accuracy	Crowdsourcing	Biological Evolution	81.30% [4]
PANNs	Transfer Learning	AudioSet (2M)	90.50%
CLAP (Zero-shot)	Zero-shot	LAION-Audio (630K)	91.50%
AST	Transfer Learning	AudioSet (2M)	95.00%
CLAP (Transfer)	Transfer Learning	LAION-Audio (630K)	97.25%

深度洞察：

- **超人类的表现:** 现代音频基础模型 (CLAP, AST, PANNs) 的准确率 ($>90\%$) 已经大幅超越了人类听觉的平均水平 (81.30%)。这表明在特定领域的分类任务上,“机器听觉”已经实现了对“生物听觉”的超越。
- **人类水平的门槛:** 有趣的是, 我们的自研小模型 (75%) 和通用大模型 Gemini 3 Flash (78%) 恰好处于接近人类水平但略低的区间。这说明达到人类的“直觉”水平相对容易, 但要达到“专家”水平 ($>90\%$) 则需要大规模的领域专业训练数据。
- **多模态的威力:** CLAP 利用文本语义引导音频特征学习, 其 Zero-shot 能力 (91.50%) 甚至超过了 PANNs 的监督迁移结果, 这意味着模型真正理解了声音的语义。

7 总结与思考

本项目通过完整的系统实现与对比实验, 不仅验证了算法的有效性, 更揭示了音频处理技术发展的内在逻辑。

7.1 不变的数学基石

纵观从基础的 MFCC 到最前沿的 AST、CLAP 大模型, 我们发现一个有趣的现象: 尽管模型架构日新月异, 但它们的输入端无一例外地都指向了频谱图。这深刻说明, 频域分析依然是连接物理声学与与机器智能的“通用语言”。傅里叶变换 (FFT) 这一诞生于 19 世纪的数学工具, 跨越了算法的代际更迭, 依然是支撑现代人工智能大厦的底层基石。

7.2 演进的特征范式

通过对比不同模型的表现, 我们清晰地看到了表示学习 (Representation Learning) 的三次范式跃迁:

- **规则驱动 (DSP):** 如 MFCC, 依赖人工设计的物理和听觉规则, 可解释性强 (捕捉共振峰), 但难以描述复杂的语义。
- **数据驱动 (Deep Learning):** 如 ResNet, 通过 CNN 自动提取频域纹理特征, 性能大幅提升, 但本质上仍是模式识别。
- **语义驱动 (Foundation Models):** 如 CLAP, 将声音映射到文本语义空间, 实现了真正的“听音识意”, 使得机器听觉超越了单纯的分类, 具备了理解能力。

未来的音频技术, 必将是“经典信号处理”与“大规模预训练”的深度共生。底层的物理模型提供精确的时频描述, 上层的数据模型赋予其广义的语义理解, 共同推动机器听觉向着“超人类”的智能水平迈进。

参考文献

- [1] Google DeepMind. Gemini 3 flash. <https://deepmind.google/models/gemini/flash/>, 2025. Accessed: 2025-12-25.
- [2] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.
- [3] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition, 2020.
- [4] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.
- [5] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, 2024.