从时域到频域：基于多分支 CNN 网络的 AI 音频检测模型

NKUMMF2025138

October 24, 2025

摘要

- 问题一：
  - 从**时域、频域及声学统计量**等角度提取 11 类特征，多分支并行融合判别；
  - 基于**多分支卷积神经网络（Multi-Branch CNN）**的端到端 AI 音频检测与评分模型；
- 问题二：
  - 设计**分支探针机制**，依据各分支判别准确率加权，构建可解释的 AI 痕迹评分；
- 问题三：
  - 引入**频谱均衡、高频注入、环境噪声混入**等扰动测试模型鲁棒性；

- 模型在测试集上准确率达 **89.24% ± 0.43%**，在多种扰动下表现稳健，具**低计算开销与良好可扩展性**。

判别模型的建立与求解

特征说明

特征名称	维度	特征类型
rms	(1, 862)	时域能量
zcr	(1, 862)	时域变化
hjorth	(3, 1)	全局统计
log_mel	(128, 862)	频域谱图
mfcc	(13, 862)	声学特征
centroid	(1, 862)	频谱形状
contrast	(7, 862)	频谱对比度
flatness	(1, 862)	频谱形状
f0	(1440,)	音高曲线
hnr	(1998,)	声学质量
formant	(3, 2000)	共振峰

Table 1.各特征的名称、维度及类型

模型建立

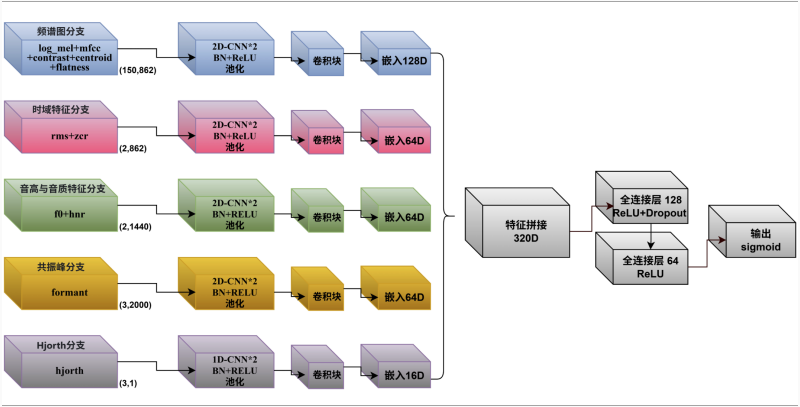


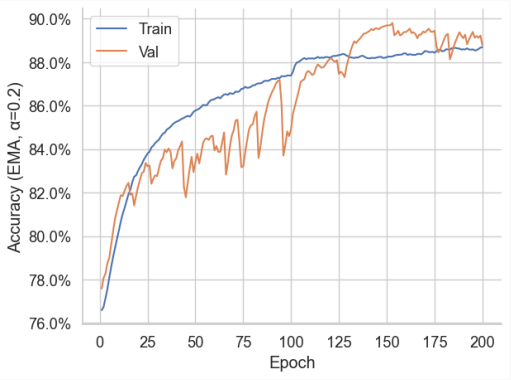
Figure 1.多分支 CNN 示意图

模型建立

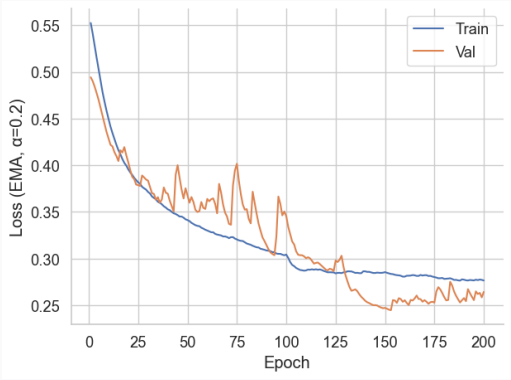
分支名称	输入特征	处理与输出
频谱图分支（Spectrogram）	log_mel, mfcc, contrast, centroid, flatness	2D CNN 提取时频模式 → GAP → 128 维表示
时域分支（Time-Domain）	rms, zcr	1D CNN 提取时间结构 → 64 维向量
音高与音质分支（Pitch & Quality）	f0, hnr	1D CNN 建模调型与音质 → 64 维向量
共振峰分支（Formant）	formant	1D CNN 处理时间序列 → 64 维向量
Hjorth 参数分支（Hjorth Param.）	hjorth (Activity, Mobility, Complexity)	两层全连接 → 16 维表示

Table 2.多分支特征处理结构概览

结果分析



(a) 准确率（Train vs Val）



(b) 损失（Train vs Val）

Figure 2.训练过程的准确率与损失对比

消融实验 i

为探究各输入分支在多分支 CNN 模型中的贡献度，本研究设计了消融实验，通过控制不同分支的启用/关闭状态，评估其对模型分类精度的影响。具体而言，我们分别在以下三种设置下进行测试：

- 全分支开启；
- 单独关闭某一分支，其余保持开启；
- 单独开启某一分支，其余全部关闭。

消融实验 i

Table 3.消融实验结果（准确率%）

频谱图分支	时域分支	音高与音质分支	共振峰分支	Hjorth 参数分支	准确率
1	1	1	1	1	89.24%
0	1	1	1	1	29.87%
1	0	1	1	1	79.20%
1	1	0	1	1	88.77%
1	1	1	0	1	88.40%
1	1	1	1	0	89.10%
1	0	0	0	0	77.87%
0	1	0	0	0	30.02%
0	0	1	0	0	47.76%
0	0	0	1	0	50.67%
0	0	0	0	1	25.20%

消融实验 i

虽然消融实验能够衡量各分支在整体判别任务中的贡献度，但这种方法仍存在一定局限性：

- **无法独立量化单分支判别能力**
- **分支间存在互补与冗余效应**
- **评估结果依赖当前模型权重分布**
- **未能刻画特征在不同类型样本上的贡献差异**

针对以上不足，我们在后续模型中引入了分支探针机制，使每个分支能够在保持原有多分支协作的同时，**单独**完成二分类任务，从而获得更客观、细粒度的特征贡献评估结果。

评分机制

评分公式的定义

受消融实验结果的启发，为了细化模型的评价指标，我们改进了多分支 CNN 网络，为每个分支添加了一个由线性层和 `Sigmoid` 函数构成的**探针头结构**。

对于一段音频，每个分支的探针都可以独立判断该音频是否为 AI 生成，并输出**分类概率**。

评分规则

记各分支探针头给出的“音频为 AI 生成”的概率为 $p_i \in [0,1]$，则作品的综合评分（AI 痕迹强度）定义为

$$S \;=\; \sum_{i} \omega_i \, p_i \;\in\; [0,1], \tag{1}$$

其中 ω_i 为分支的评分权重。S 越大表示该音频越可能为 AI 生成。

分支权重的确定

在综合各分支结果为音乐作品打分时，我们以各探针头单独决策时的**准确率**为依据，确定了各分支的评分权重。

设各分支探针在独立判别（AI/非 AI）下的准确率
$acc_i, i \in \{\text{spec,time,prosody,formant,hjorth}\}$。

为避免随机猜测带来的偏置，先做机会校正（chance-corrected）并归一化：

$$\omega_i = \frac{acc_i - 0.5}{\sum_j (acc_j - 0.5)}, \qquad \sum_i \omega_i = 1. \tag{2}$$

实验结果 i

Table 4.分支探针头独立预测准确率及由此推导出的归一化权重

分支（Branch）	准确率 a_i	权重 w_i
spec（频谱图）	0.7725	0.1987
time（时域）	0.7834	0.2067
prosody（音高音质）	0.7718	0.1982
formant（共振峰）	0.7717	0.1981
hjorth（Hjorth）	0.7718	0.1982
主分类器（参考）	0.8924	–

鲁棒性分析

鲁棒性分析 i

- 构造扰动
- 分析扰动实验结果
- 结合**数据流**进行归因分析

扰动构造 i

从模型和数据两个角度入手，构造扰动：

1. **模型角度：**修改音频，对模型特征进行直接干扰
   - **时域分支：**轻度动态压缩、滤波器、相位扰动；
   - **频域分支：** EQ 频段调节；
   - **类共振峰分支：**对常见人声频段进行增益/衰减；
   - **全局统计分支：**局部加速/减速，改变节奏特征；
   - **音高与谐噪比分支：**对音高作 ± 0.15 semitone 平移，加入低幅白噪声。
2. **数据角度：**音轨合并扰动
   将原始音频与自然环境噪声片段（如 *city park*, *forest*, *rain*）按不同响度比例混合，构造更具真实感的样本。

扰动实验结果及分析 i

对真实标签为"AI 生成" 的音频施加前述两类扰动，构造测试集，分别计算：

$$S = \text{加权融合评分}, \quad \texttt{prob\_main} = \text{主分类概率}$$

每类扰动设置 `mild`、`medium`、`strong` 三档强度，并统一归一化 RMS 以消除响度影响。

不同扰动类型下的鲁棒性表现 i

下表汇总了五组混淆效果较大的实验结果。

Table 5.不同扰动类型下主分类概率与综合评分均值对比

扰动方式	示例	样本数	prob_main 均值	S 均值
comboAll	多种扰动叠加版本	1	0.764	0.776
pitchHNR	音高调整 + 白噪音注入	1	0.659	0.757
specEQ	EQ	7	0.604	0.727
highFreqInject	高频注入（HI）	8	0.461	0.667
ambientNoise	环境噪声混入（雨声、森林、城市）	4	0.422	0.671

归因分析：分支数据流特征与扰动敏感性的关联 i

目的：探究检测器在不同分支上的数据流特征与其对扰动的敏感性之间的关系。

方法：

- 对各 CNN 分支的嵌入空间（即输入全连接层前的一维张量）进行降维可视化；
- 分析各分支在嵌入空间中的类别可分性；
- 采用 AUC、ACC、Fisher 比率与 logit 统计量量化每个分支对最终决策的贡献。

归因分析：分支数据流特征与扰动敏感性的关联

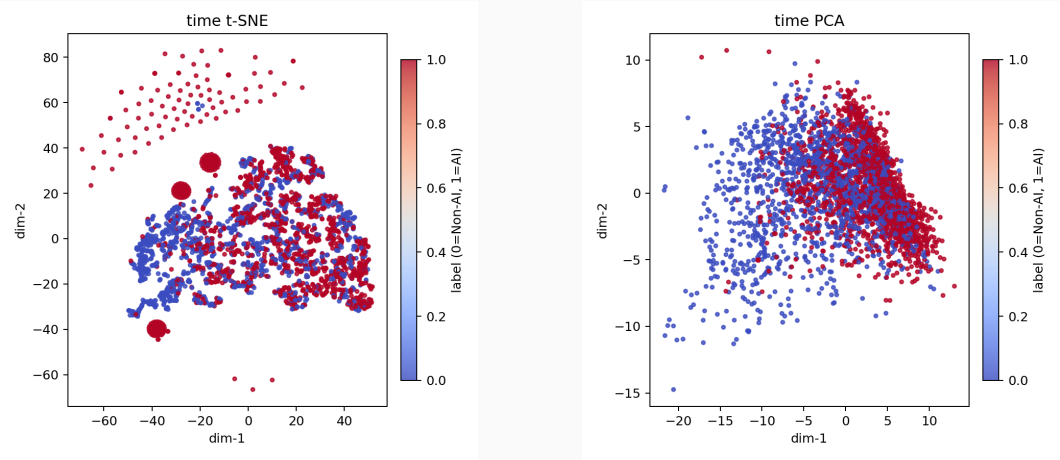


Figure 3.时域分支嵌入可视化：t-SNE（左）与 PCA（右）。

归因分析：共振峰分支嵌入特征

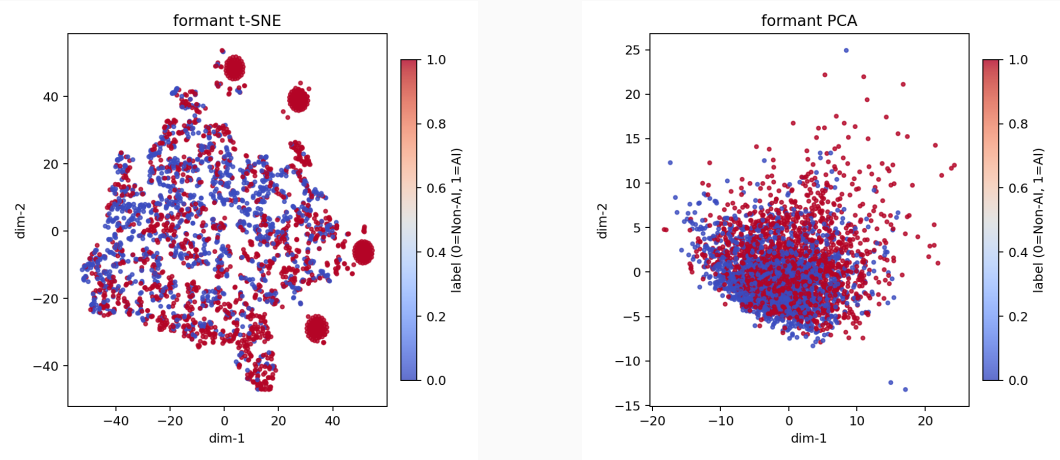


Figure 4.formant 分支嵌入可视化：t-SNE（左）与 PCA（右）。

量化分析：contrib 指标定义 i

Contribution 指标：

- **Grad×Input，**

$$\mathrm{contrib}_k \approx \sum_{j \in \mathcal{I}_k} \frac{\partial y}{\partial z_j} z_j,$$

  其中 $\mathcal{I}_k$ 表示第 k 个分支在拼接向量中的索引集合。

- $\mathrm{contrib}_k > 0$ 表示该分支对判决结果具有正向推动作用（倾向于判为 AI），$\mathrm{contrib}_k < 0$ 则表示负向作用（倾向于判为非 AI）。

量化分析：分支贡献的相关性与对冲关系

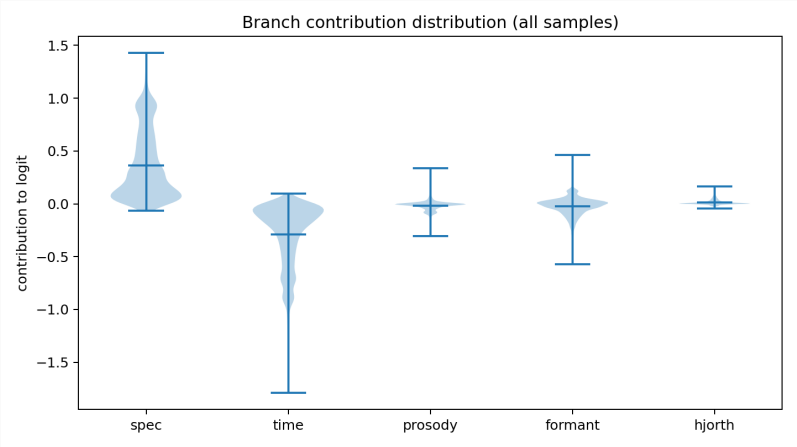


Figure 5.各分支 logit 贡献的分布（所有样本）。

量化分析结果：分支贡献对比

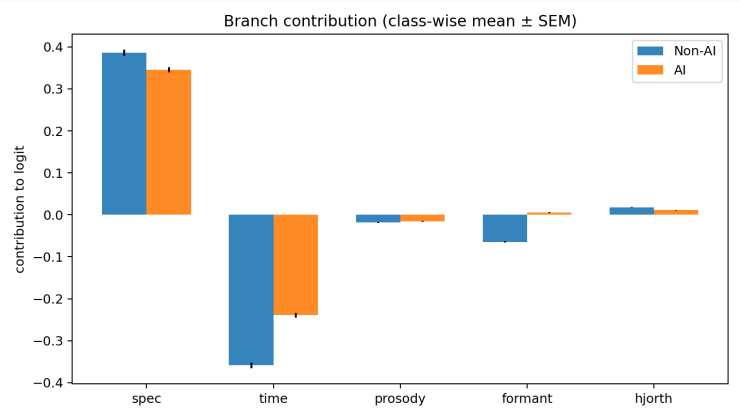


Figure 6.各分支在 AI 与非 AI 样本上的平均贡献值（均值 ± 标准误差）。

量化分析结果：分支贡献对比 i

结果概要：

- 图 6 展示了不同分支在 AI 与非 AI 样本上的平均贡献值（均值 ± 标准误差）；
- `spec` 分支在两类样本中均呈显著**正贡献**，推高 AI 判别概率；
- `time` 分支整体为**负贡献**，对 AI 判别起抑制作用，且幅度较大；
- `prosody`、`formant` 与 `hjorth` 分支的贡献幅度较小，表明其对最终判决的直接推动作用有限。

量化分析：分支贡献的相关性与对冲关系

相关性分析：

- 计算样本层面的皮尔逊相关系数矩阵；
- 结果如图 7 所示：
  - spec 与 time 分支呈显著**负相关**；
  - 表明它们在多数样本中存在“对冲”关系：一个分支推动判为 AI 时，另一个往往抑制；
  - prosody 与 time、formant 呈**正相关**，暗示在部分样本中可能协同作用。

量化分析：分支贡献的相关性与对冲关系

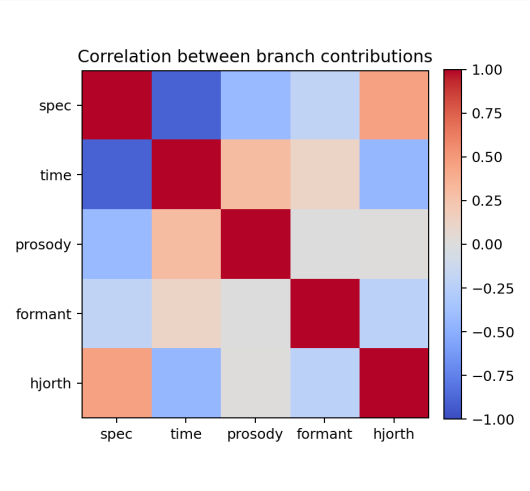


Figure 7.各分支 logit 贡献的相关性矩阵（皮尔逊相关系数）。

量化分析：分支判别能力指标定义 i

目的：评估各分支嵌入的判别能力，分析最终决策对不同分支的依赖性。

主要指标：

- **AUC（Area Under the ROC Curve）**
  - 表示在所有可能阈值下模型正确区分正负样本的概率；
  - ROC 曲线越靠近左上角，说明召回率高、误报率低；
  - AUC 值越接近 $1 \rightarrow$ 判别能力越强；AUC $= 0.5 \rightarrow$ 等价于随机猜测；
  - 图 8 展示各分支的 ROC 曲线。
- **ACC（Accuracy）**
  - 在固定分类阈值（本文取 0.5）下的分类准确率；
  - 反映分支在常规决策条件下的直接预测性能；

量化分析：分支判别能力指标定义 ii

- 强调“硬判别”能力。

- **Fisher 比率（Fisher_mean）**
  - 衡量类间分离度与类内紧凑度的比值：

$$\text{Fisher Ratio} = \frac{\sum_d (\mu_d^+ - \mu_d^-)^2}{\sum_d (\sigma_d^{+2} + \sigma_d^{-2})}$$

  - Fisher 比率越大，表示嵌入在不同类别间的分布差异越明显；
  - 说明该分支在特征空间中具有更好的可分性。

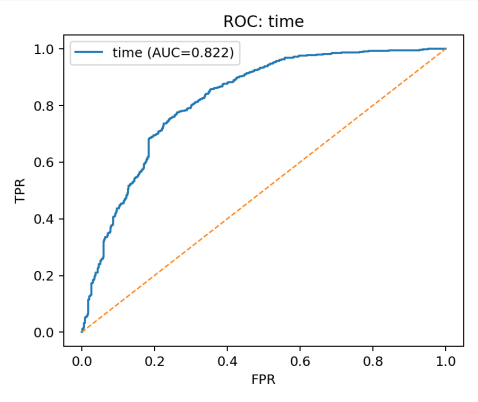
量化分析结果：主要分支的判别能力 i

结论概览：扰动敏感性与各分支的单独判别能力高度一致。

定量结果（5 折交叉验证）：

- **时域分支（time）：** AUC $= 0.8251 \pm 0.0076$，ACC $= 0.7627 \pm 0.0025$，Fisher 比率 $= 0.1324$；
- **formant 分支：** AUC $= 0.8161 \pm 0.0077$，ACC $= 0.7413 \pm 0.0101$，Fisher 比率 $= 0.0912$；
- 均显著高于 `prosody`（AUC 0.7191，Fisher 0.0071）与 `hjorth`（AUC 0.6963，Fisher 0.0054）；
- 表明模型决策更依赖：
  - **时域包络 / 动态结构**；
  - **人声共振峰布局特征**。

量化分析结果：主要分支的判别能力



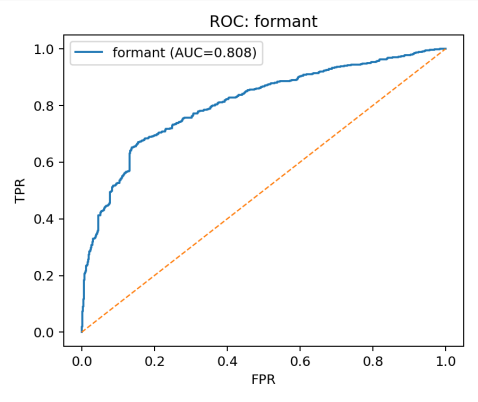


Figure 8.时域分支与 formant 分支 ROC 曲线。

量化分析结果：分支可分性指标对比

对比分析：

- 图 9 展示各分支的 AUC、ACC 与 Fisher 比率；
- 时域与 formant 分支在三个指标上均占优；
- prosody 与 hjorth 分支指标普遍偏低，表明其独立判别能力有限；
- 说明融合决策阶段模型主要依赖结构性强、时频特征稳定的分支。

量化分析结果：分支可分性指标对比

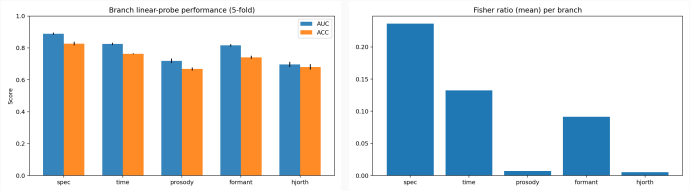


Figure 9.分支可分性定量指标：AUC/ACC（左）、Fisher 比率（右）。

量化分析结果：分支可分性指标对比

总结：当时域包络或共振峰结构被扰动时，模型主分类概率显著向“非 AI”区域偏移，验证了模型对这些关键声学特征的依赖性。

模型优缺点

模型优点

- 融合层充分利用多源信息的互补性；
- 具有较好的鲁棒性，即使在分类器受到干扰的情况下，基于多分枝独立判断的加权评分仍然稳健
- 特征提取部分与分类器解耦、各分支之间解耦，模型扩展性强
- 模型轻量、计算成本低、可实时检测并跨平台部署。

模型不足

- 特征融合层为简单拼接，缺乏对分支质量的动态自适应；
- 训练数据覆盖有限，对新风格或新算法生成的音乐适配性不足。

改进思路

1. **解耦式分支扩展（Plug-in）：**利用当前架构“特征提取与分类器解耦、分支间解耦”的优势，一方面可以设计可热插拔的分支接口，便于不断添加新的分支；另一方面，可以换用其他比 MLP 更强的分类器来进行的最终的分类
2. **数据层面优化：**通过扩充多样化的训练样本缓解数据覆盖不足问题，增加不同风格、不同生成算法的 AI 音频，并在训练中引入动态压缩、全通相位扰动、微小时间伸缩等数据增强手段，以提升模型的泛化与鲁棒性。
3. **结构层面优化：**在多分支融合层中引入一致性正则或分支级 dropout 机制，以抑制单一分支对判决的主导作用，促使模型在多个分支间学习到更加均衡的判别信息。
4. **自适应机制引入：**将静态拼接式融合改为基于注意力或门控机制的动态加权，使融合层能根据分支质量自适应分配权重，从而提升整体决策的灵活性与解释性。

谢谢大家！