arXiv:2601.11768v1 [eess.AS] 16 Jan 2026

单音音乐中基频与准确浊音概率的轻量级自监督检测

Venkat Suprabath Bitra[1]，以及 Homayoon Beigi[2] b[1]哥伦比亚大学计算机科学系，美国纽约[2]哥伦比亚大学机械工程系和电气工程系，以及 Recognition Technologies, Inc.，美国纽约
venkat.s.bitra@columbia.edu, homayoon.beigi@columbia.edu

已被接收，将于 ICPRAM 2026 发表，2026年3月2日至4日，西班牙马贝拉

关键词:
自监督音高检测，无监督音高检测，基频，音高估计，共振，音乐音色迁移，浊音概率，音乐合成，音乐分析，常数Q变换，常数Q变换，DDSP，移位交叉熵损失，乐器建模，ResNeXt神经网络，音乐信息检索，音乐信息检索

摘要:
可靠的基频（F_0）和浊音估计对于神经合成至关重要，但许多音高提取器依赖大量标注语料库，在真实录音伪影下性能会下降。我们提出一种轻量、完全自监督的框架，用于联合估计F_0和进行有声性推断，旨在利用有限音频快速完成单乐器训练。通过在常数Q变换特征上进行移调等变学习，我们引入一种期望最大化式迭代重加权方案，利用移位交叉熵（SCE）一致性作为可靠性信号，抑制信息量不足的噪声帧／清音帧。所得权重提供置信度分数，可在无需人工标注的情况下，为独立的轻量浊音分类器生成伪标签。在 MedleyDB 上训练，并使用 MDB-stem-synth 基准真值进行评估后，我们的方法取得了具有竞争力的跨语料库性能（RPA 95.84，RCA 96.24），并展现出跨乐器泛化能力。

1 引言

基频（F_0）估计以及判断信号是否有声，是语音和音乐分析的核心基础，也是音乐信息检索（MIR）、音频编辑和神经合成等众多下游系统的基础（Beigi，2011；Yost，2009）。YIN（de Cheveigné 和 Kawahara，2002）、RAPT（Talkin，1995）和 pYIN（Mauch 和 Dixon，2014）等经典音高跟踪器利用周期性和启发式方法，在受控条件下效果良好，但在真实声学伪影和领域偏移下容易失效。

深度学习方法（例如 CREPE（Kim 等，2018））大幅提升了在基准数据集上的准确率，但有两个实际障碍限制了它们在快速构建定制音频处理流水线时的实用性。首先，大多数高性能模型都采用有监督学习，因此依赖规模庞大且经过细致标注的语料库，而这类语料库构建成本高，也难以扩展到不同乐器、录音条件和演奏技巧。其次，在精心整理的数据上训练的模型，部署到包含串音、混响和背景噪声的真实录音中时，往往泛化能力较差（Morrison 等，2024；Saxena 等，2024；Geirhos 等，2020）。这些障碍对现代合成工作流程尤其棘手，因为特征提取器必须迅速适配目标乐器或录音设置。

一个具有启发性的例子是可微分数字信号处理（DDSP）及相关的神经合成系统（Engel 等人，2020a；Hayes 等人，2023）。这些框架能够生成高质量、个性化的乐器，但前提是具备可靠的 F_0 和浊音特征。实践

a https://orcid.org/0000-0002-9254-2274b
https://orcid.org/0000-0003-0127-2385

者通常希望使用几分钟的音频，而不是几小时的音频来构建定制合成器，然而，当前许多音高提取器，无论是有监督的还是自监督的，都是在大型语料库上训练的，并未针对低数据量单乐器场景进行稳健收敛的设计。这种不匹配造成了瓶颈：即使仅用 10 到 15 分钟的音频就能实现合成个性化，特征提取器也未必能够做到。

本文将音高估计重新定义为一个低数据量、单乐器、快速训练的问题。我们的目标是学习一个<风格 id='1'>F</风格>$_0$/发声前端，使其能够利用少量目标音频高效训练，同时在真实录音中保持可靠。为突出真实的声学特性，我们聚焦于 MedleyDB（Bittner等人，2014），该数据库包含真实的多轨录音，因此存在串音、房间效应和非理想隔离。此设置与数字化重新合成或完美分离的数据集（例如 MDB-stem-synth（Salamon等人，2017））有实质区别；在这类数据集中，清音片段可能几乎无声，干扰也很少。若要使提取的特征成为合成的可靠信号，那么在 MedleyDB 上的鲁棒性至关重要。

为免除人工音高标签的需求并支持快速适配，我们采用自监督学习（SSL），并使用一种受 SPICE（Gfeller 等人，2020）和 PESTO（Riou 等人，2023）启发的移调等变目标。我们使用常数Q变换（CQT）表示（Brown，1991）训练模型，使其满足一个简单的一致性约束：如果输入音频按已知音程移调，模型的内部音高表示也必须相应移动相同的幅度。这一等变性“前置任务”促使网络直接编码音高结构，无需监督，并为从小规模、乐器专属集合中学习提供了一种自然机制。

除此之外，F_0, 我们还致力于在干扰下实现稳健的浊音检测。在真实录音中，“清音”并不意味着静音；其中往往含有串音和噪声带来的残余能量。我们通过移位交叉熵（SCE）损失利用这一特性：只有当音高表示稳定且明确时，等变目标才会达到最小值，而这一条件与真实的浊音内容高度相关。我们利用这一稳定性信号构建可靠的伪标签，并训练专用的浊音分类器，从而打造出一种兼具无需标签和能够抵御真实伪影能力的 F_0/浊音处理流程。

2 贡献

- 仅使用 MedleyDB（Bittner等人，2014）中声学条件极具挑战性的多轨录音进行训练，以确保模型能够抵御真实场景中的噪声、串音和伪影。
- 两阶段伪标签方法：无监督移位交叉熵损失生成置信度分数，无需人工标签即可训练轻量级二元浊音分类器。
- 通过在单一乐器上训练、在未见过的乐器上测试，进行严格的跨乐器评估，量化与音色无关的音高泛化能力。

3 方法

我们基于 PESTO 的自监督、移调等变学习范式，从分离的单音乐器分轨中估计基频（F_0）和发声状态。我们的方法专为单乐器、低数据量训练量身定制，采用：(i) 使数据增强与音乐移调对齐的常数Q变换前端；(ii) 配备托普利兹结构音高头的高效 ResNeXt1D 编码器；以及 (iii) 两阶段发声处理流程，将自监督稳定性信号转化为伪标签，以训练轻量级二元分类器。下一小节将详细介绍数据、特征提取、数据增强、架构和目标函数；再下一小节将介绍我们的基于EM的重加权策略。

3.1 数据集与特征提取

3.1.1 数据集

我们使用 MedleyDB（Bittner等人，2014）中针对特定乐器的分轨进行训练。该数据集包含真实的多轨录音，因此呈现出逼真的干扰，例如串音、房间混响和录音噪声。评估时，我们使用 MDB-stem-synth（Salamon等人，2017），它提供由 MIDI 控制合成生成的精确帧级真值F_0 标签。这两个数据集都提供了预先分离的分轨，因此训练和测试期间无需进行声源分离。第 4.1 节对 MedleyDB 和 MDB-stem-synth 进行了详细比较。

3.1.2 常数 Q 变换（CQT）输入

每个音频片段都会转换为对数幅度 CQT 表示。设单个时间帧为$x \in \mathbb{R}^B$，训练批次为$X \in \mathbb{R}^{N\times B}$，其中$B$为 CQT 频率箱的数量，$N$ 为采样帧的数量。我们使用$f_{\min} = \mathrm{A0} \approx 27.5\mathrm{Hz}$ 和每八度音程 36 个频率箱（即每半音$B_{\mathrm{st}} = 3$ 个频率箱）。该变换使用$B = 269$ 个频率箱，覆盖约 7.5 个八度音程，最高至$f_{\max} \approx 4.9\mathrm{kHz}$，涵盖有音高范围，最高至 C8 ≈4.2 kHz。

尽管在我们的实验中，目标声部主要占据大约 200-1200Hz 的F_0 范围，但我们仍保留较宽的 CQT 频带，为基于移调的自监督学习留出余量。当$f_s = 16\mathrm{kHz}$（奈奎斯特频率为= 8kHz）时，$\sim 4\mathrm{kHz}$ 以上的内容对这一F_0 范围提供的谐波证据有限。纳入这些内容主要是为了减少边缘效应，即降低音高移位将预测众数推至边界频率箱并导致表示坍塌的概率。

3.1.3 通过 CQT 移位实现移调等变性

自监督是在从同一底层音频中提取的移调配对（x，x'）上定义的。移调 τ 个半音对应于 $\delta = \tau B_{\mathrm{st}}$个频率箱的移位。我们通过沿 CQT 频率轴进行离散移位来近似移调后的视图，

$$x'(b) \approx x(b - \delta),$$

其中b 表示 CQT 频率箱的索引。学习目标通过施加等变性约束，使x' 的预测音高分布与x 移位 δ 个频率箱后的预测结果相匹配。实际操作中，移位采用边界安全填充（而非循环移位），以避免在频率边界处引入不符合物理规律的能量。

3.1.4 音高不变增强

为增强模型对响度和背景扰动的不变性，我们对每个视图分别进行独立的数据增强。令 x 表示对数幅度 CQT 帧，令 m =exp(x）表示其线性幅度。定义 $\mathcal{U}(a, b)$ 为 $[a, b]$,上的均匀分布，定义 $\mathcal{N}(\mu\ \sigma^2)$ 为均值为 μ 、方差为σ^2的高斯（正态）分布。

随机增益。我们采样 $g \sim \mathcal{U}(0。5, 1。5)$，并在对数空间中施加加性偏移：

$$x \leftarrow x + \log(g).$$

信噪比定向噪声。我们对 SNR 进行采样，令其服从 $_{\mathrm{dB}} \sim \mathcal{U}(15,50)$，并计算各频率箱上的信号功率 $P_s = \mathbb{E}[m^2]$。给定 SNR = $10^{\mathrm{SNR_{dB}}/10}$，我们设定$P_n = P_s/\mathrm{SNR}$，并采样 $n \sim \mathcal{N}(0,P_n)$:

$$m \leftarrow m + n, \qquad x \leftarrow \log(\max(m, \varepsilon)),$$

其中， ε 可避免对非正值取对数。

3.1.5 模型架构

我们的网络 $f_\theta(\cdot)$ 将一个 CQT 帧映射为音高 logits。它由一个 ResNeXt1D编码器和一个托普利茨结构线性分类器组成（图 1）。编码器包括一个 stem 卷积，以及三个由 ResNeXtBlock1D 模块构成的阶段，这些模块采用分组卷积，从而能够以更高的参数效率实现多尺度频谱聚合。音高头采用沿频率轴的一维卷积实现，施加托普利茨结构，相较于全连接分类器可减少参数量。该结构还支持高效的事后校准：无需重新训练，只需对分类器卷积核进行循环移位，即可校正系统性的音高分箱偏移。

3.1.6 训练目标

给定一批配对帧 $\{(x_i, x_i')\}_{i=1}^N$ 及其对应的逻辑值 $y_i = f_\theta(x_i)$ 和 $y_i' = f_\theta(x_i')$，我们定义 p_i = Softmax(y_i) 和 p_i' =Softmax(y_i')。我们使用三个损失项的加权和：

$$\mathcal{L}_{\mathrm{total}} = \frac{1}{N}\sum_{i=1}^{N} w_i \Big(\mathcal{L}_{\mathrm{equiv},i} + \mathcal{L}_{\mathrm{invar},i} + \mathcal{L}_{\mathrm{sce},i} \Big)$$

其中，w_i 是由我们的迭代重加权方案（将在下一小节介绍）生成的样本权重。

3.2 使用期望最大化进行迭代重加权

3.2.1 直觉

SCE 在音高移调下施加自一致性约束：如果某一帧包含连贯的谐波结构，那么移调视图上的预测音高分布应与原始视图上预测结果的移位版本相匹配。有声帧大致满足

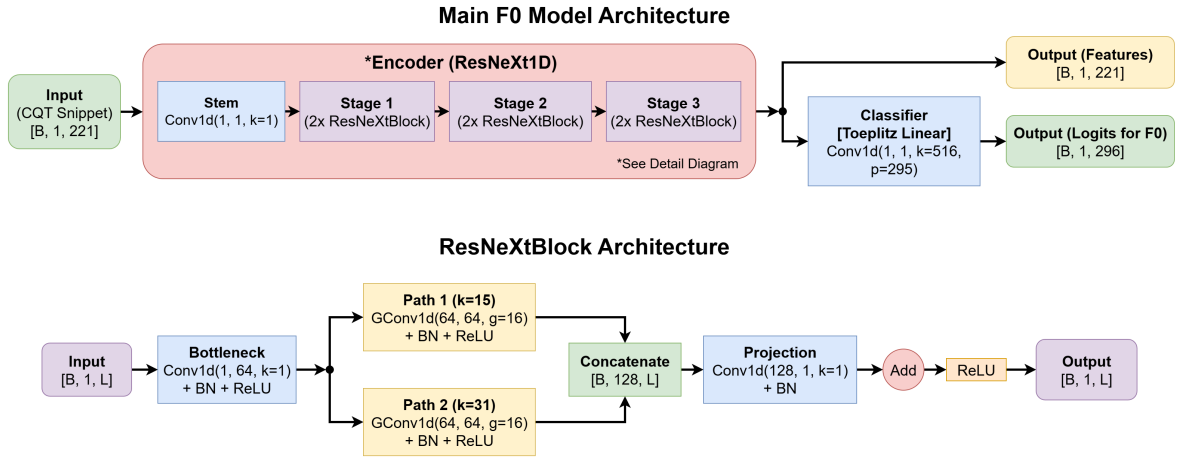


图 1：模型架构。ResNeXt1D编码器处理 CQT 帧，随后由托普利兹线性分类器（以一维卷积实现）生成音高 logits。

这一约束，从而产生对齐的目标和低方差梯度，使音高峰值更加突出。清音帧（噪声、串音、混响尾音）在移调后不会变成一致的音高模式，因此 SCE 会产生不一致的目标和高方差梯度。在低数据量条件下，一旦有声帧拟合完成，其 SCE 就会变小，优化过程可能因此主要受这些无信息的清音梯度支配。我们的 EM 式重新加权通过降低那些 SCE 表明其移调结构不可靠的帧的权重，来避免这一问题。

3.2.2 算法（EM 风格的迭代重加权）

训练过程交替进行权重估计（E 步）和加权优化（M 步）。令$\mathcal{D} = \{X_i\}_{i=1}^{M}$ 为训练集，并初始化$w_i \leftarrow 1$。我们每隔$K = 5$ 个训练轮次更新一次权重。

初始化。 使用$w_i = 1$训练f_θ，共K 个训练轮次。

E 步（权重更新）。在训练轮次索引e 处（即K的倍数），使用当前模型计算所有帧的$\mathcal{L}_{\text{sce},\,i}$。在本轮处理中进行归一化：

$$\hat{\mathcal{L}}_{\text{sce},i} = \frac{\mathcal{L}_{\text{sce},i} - \min_j \mathcal{L}_{\text{sce},j}}{\max_j \mathcal{L}_{\text{sce},j} - \min_j \mathcal{L}_{\text{sce},j} + \varepsilon} \in [0,1].$$

定义依赖于训练轮次的退火因子

$$\lambda(e) = \exp\left(\frac{e^{1.25}}{1000}\right) - 1,$$

并通过以下方式更新权重

$$\Delta w_i = \lambda(e)\,\hat{\mathcal{L}}_{\text{sce},i} \qquad w_i \leftarrow \max(0,\, w_i - \Delta w_i).$$

因此，随着训练的推进，SCE 较大的帧会逐渐被降低权重。

M 步（加权优化）。在接下来的 K 个训练轮次中，通过最小化加权目标来更新 θ。重复 E 步和 M 步，直至收敛。

3.2.3 为什么重新加权能够抑制有害梯度

设 $X \in \mathbb{R}^B$ 为一个 CQT 帧，$y = f_\theta(X) \in \mathbb{R}^K$为模型在 K 个音高分箱上的输出 logits。设 X' 为其移调视图，其对应的 $y' = f_\theta(X')$。定义 $p =$ 为 softmax(y)，$p' =$为 softmax(y')。令 roll(y δ) 表示按 δ 个分箱进行循环移位（即在对数频率空间中进行音高移调）。移位交叉熵（SCE）损失比较原始帧 X 及其音高移调版本 X'上的预测：

$$\mathcal{L}_{\text{sce}} = \frac{1}{2}\left(\text{CE}(y', \text{roll}(p, \delta)) + \text{CE}(y, \text{roll}(p', -\delta))\right)$$

其中 $y = f_\theta(X)$ 和 $y' = f_\theta(X')$。交叉熵相对于逻辑值的梯度遵循标准形式：

$$\frac{\partial \text{CE}(p,q)}{\partial y} = p - q$$

其中 $p = \text{Softmax}(y)$，而 q 是目标分布（视为固定值）。

干净的静音会助长平凡解 在合成数据集（例如 MDB-stem-synth）中，静音帧的值恰好为零：$X = \mathbf{0}$。输入为零时，模型输出会随时间保持恒定：$y = f_\theta(\mathbf{0})$ 会为所有静音帧生成相同的逻辑值。在采用常见初始化方式、权重衰减且训练数据具有对称性的情况下，网络倾向于对恒定输入给出近似均匀的预测，因此：

$$p = \mathrm{softmax}(f_\theta(\mathbf{0})) \approx \mathbf{u}$$

其中$\mathbf{u}$是均匀分布：

$$u_k = \frac{1}{K}$$

由于均匀分布在循环移位（$\mathrm{roll}(\mathbf{u}\ \delta) = \mathbf{u}$）下保持不变，因此预测和目标都会变成均匀分布：

$$\frac{\partial \mathcal{L}_{\mathrm{sce}}}{\partial y} \approx \mathbf{u} - \mathbf{u} = \mathbf{0}$$

因此，网络无法从静音帧中学到任何信息，且这些帧产生的梯度会消失。这是无害的，也不提供有用信息。

真实静音存在噪声，并会产生有害梯度 在真实录音（如 MedleyDB）中，“静音”帧并非真正静音。它们包含噪声、串音和混响：$X = \varepsilon$ （随机且非零）。深度网络f_θ 具有很强的非线性，并且对输入扰动十分敏感。即使是微小的噪声 ε ，也可能导致输出显著不同：

$$y = f_\theta(\varepsilon), \quad p = \mathrm{softmax}(y)$$

由于噪声会随帧和时间随机变化，p 会不可预测地波动。

关键问题 对于移调版本$X' = \mathrm{roll}(X\ \delta)$，$p(X')$与$\mathrm{roll}(p(X)\ \delta)$之间不存在一致的音高关系，原因如下：(i) 噪声没有可供移调的谐波音高结构；(ii) 不同的噪声实现（例如，混响尾音中的不同位置、不同的麦克风串音模式）在移调时无法保持任何音高对应关系；(iii) 深度网络f_θ 可能会通过其非线性层放大这些不一致性。

这使得 SCE目标 $q = \mathrm{softmax}(\mathrm{roll}(y\ \delta))$ 本身变得随机，且与预测值 $p' = \mathrm{softmax}(y')$ 不一致。梯度方差变为：

$$\mathrm{Var}\left[\frac{\partial \mathcal{L}_{\mathrm{sce}}}{\partial y_k}\right] = \mathrm{Var}[p_k - q_k] \approx \mathrm{Var}[p_k] + \mathrm{Var}[q_k]$$

这会在清音帧上产生高方差的噪声梯度，即“坏梯度”。模型在不同小批次中接收到相互矛盾的信号，导致收敛效果不佳、在低信噪比区域出现伪峰，以及训练不稳定。

方案：使用 SCE 进行 EM 式重新加权。我们将发声视为潜变量，并将 SCE 损失用作可靠性信号。SCE 损失较低的帧具有一致的音高结构（有声），而 SCE 损失较高的帧则受到噪声影响/为清音。该加权损失的计算梯度如下：

$$\frac{\partial}{\partial \theta}\left[w_i \cdot \mathrm{CE}(p_i, q_i)\right] = w_i \cdot (p_i - q_i) \cdot \frac{\partial y_i}{\partial \theta}$$

当$\mathcal{L}_{\mathrm{sce},i}$ 较大时（噪声较大的帧），$w_i \approx 0$，因此梯度贡献趋于消失。这：

1. 保留对真正有声帧的强学习能力（结构稳定 $\rightarrow$ SCE 较低 $\rightarrow w \approx 1$） 2. 抑制来自噪声/清音帧的高方差梯度（结构不稳定 $\rightarrow$SCE 较高 $\rightarrow w \approx 0$）

EM方法将一个劣势（现实世界中受噪声影响的静音）转化为优势：模型会自动学会聚焦于具有真实音高内容的帧，同时忽略主要由噪声主导的帧。

3.3 监督式发声分类

自监督训练后，网络$f(\cdot)$被冻结，并对最终样本权重$w_i \in [0,1]$进行阈值处理（阈值为 θ ），以生成二元伪标签$v_i^* \in \{0,1\}$。随后，使用二元交叉熵，将轻量级线性分类器在 CQT 帧上针对这些伪标签进行训练，从而无需任何人工标注即可生成稳健的发声预测。架构细节见附录。

4 结果与讨论

4.1 录音与重新合成

图2对比了 MedleyDB（真实多轨录音）与 MDB-stem-synth（算法重合成）在相近音乐片段上的 CQT 幅度频谱图。这两个语料库在多个方面存在差异，并直接影响 F_0 跟踪和发声。

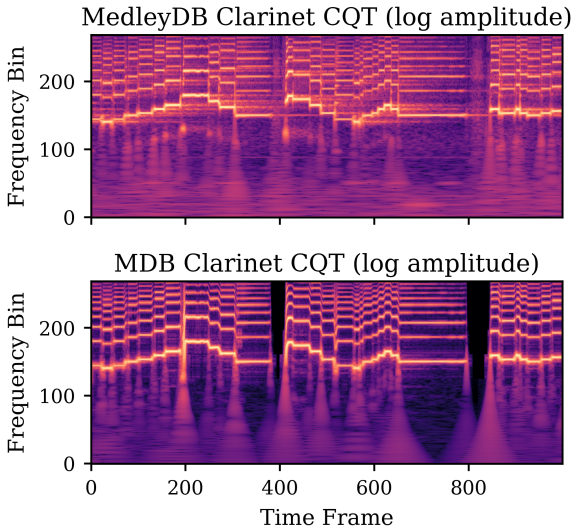


图2：声学比较：MedleyDB 与 MDB-stem-synth。（上）MedleyDB 录音受室内声学影响，呈现出更宽的泛音分量、调制边带以及指数衰减的尾部。（下）MDB-stem-synth 呈现出狭窄的谐波轨迹、突然而止的终止点以及完全静音的区域（黑色区域）。真实录音会因串音和混响保留残余能量，而合成音轨则完全为零。

MDB-stem-synth 呈现出清晰的谐波轨迹，分音带宽较窄，音符偏移 abrupt，音符之间往往会出现完全静音的间隙（精确为零）。相比之下，MedleyDB 包含由乐器动态、麦克风摆放位置和室内声学环境塑造的物理上合理的衰减尾部。真实演奏还包含频率调制（颤音）和幅度调制（震音），这些调制会拓宽谐波，并引入调制边带，在谐波列周围表现为较宽的脊线。

最后，MedleyDB 包含残余的房间反射和音轨间串音，这些会抬高频谱底噪并引入微弱的交叉谐波伪影。这些影响在 MDB-stem-synth 中并不存在，因为其中的“无发声”区域确实没有混响。这种差异有助于解释，为何在重新合成语料库上优化的模型在真实音频上的表现会下降：它们从未被迫处理串音导致的低信噪比帧、真实的衰减或调制扩展。因此，仅使用 MedleyDB 进行训练，能够针对真实世界合成和个性化流程所需的运行条件。

4.2 EM重加权对 MedleyDB 的影响

图 3 展示了 MedleyDB 上的典型表现：CQT 及预测的 $\hat{}F_0$（顶部）、学习得到的样本权重 w（中部），以及由此得出的浊音代理 $\hat{}v$（底部）。在各种录音条件下，我们都观察到一致的模式。在持续谐波区域，权重保持较高（$w \approx 1$）且稳定，预测的 $\hat{}F_0$ 会跟随占主导地位的谐波脊线。在音符释音、混响间隙和模糊低信噪比区域，权重会急剧下降，从而防止优化器对音高不可靠帧过拟合。

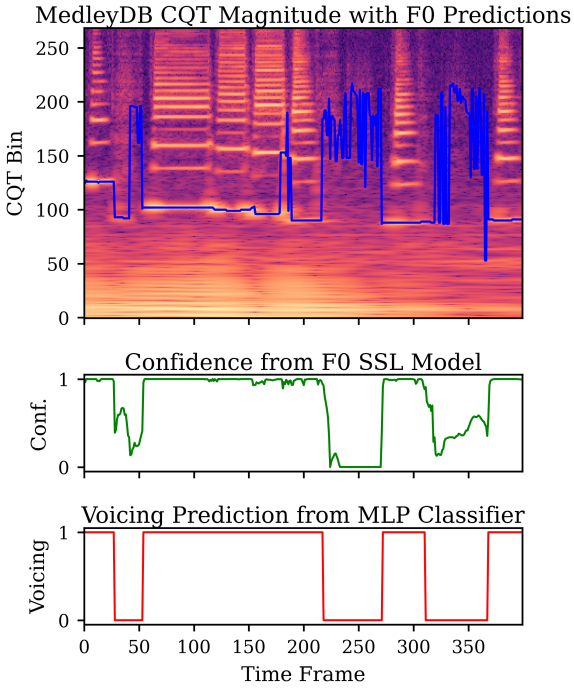


图3：MedleyDB 上学习到的权重行为。（上）带有预测 F_0（黑色）的常数Q变换频谱图。（中）在持续谐波上，EM样本权重w（绿色）保持较高，并在释放段和瞬态期间下降。（下）推导出的发声 $\hat{}v$（红色）能够有效门控浊音片段，自动区分稳定音高与含混内容。

图 4 进一步表明，学习到的权重呈现明显的双峰分布（质量集中在 0 和 1 附近），这说明重加权能够自然地区分包含移调一致的谐波结构的帧与以噪声/串音为主的帧。这种双峰性也使从 w 到发声估计值 $\hat{}v$ 的映射保持稳定，无需进行大量调优。

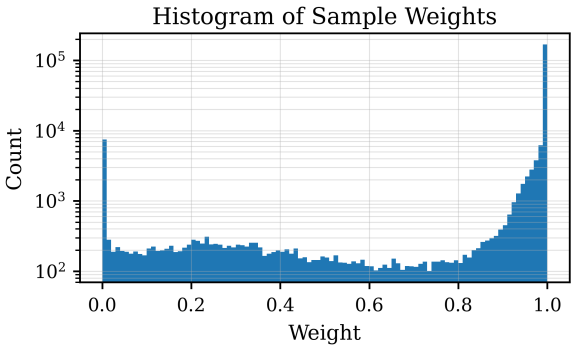


图4：Med-leyDB 上 EM样本权重的直方图（对数计数刻度）。呈现明显的双峰分布：大多数帧要么被确信地保留（$w \approx 1$），要么被降低权重（$w \approx 0$）。

4.3 跨语料库与跨乐器评估

为量化真实声学环境下的迁移效果，我们使用 MedleyDB 分轨进行训练，并在 MDB-stem-synth 上进行评估，后者提供准确的帧级 F_0 标签。我们报告：(i) 一项简洁的基线对比，用于说明我们的设置；(ii) 一种跨乐器方案，即使用单一 MedleyDB 乐器进行训练，并在 MDB-stem-synth 中未见过的乐器上进行测试。该方案能够在保持评估标签固定且精确的同时，单独考察音色泛化能力。这也反映了预期的使用场景：在带有串音和混响的真实录音上快速进行针对特定乐器的训练，然后将模型部署到更干净或录制条件不同的音频上。

基线背景。表1总结了在 MDB-stem-synth 上报告的有监督方法（CREPE、DeepF0）和自监督方法（SPICE、DDSP-inv、PESTO）的代表性结果。相比之下，我们的主要设置是在 MedleyDB 录音上训练，并使用 MDB-stem-synth 标签进行评估，从而引入有意的领域偏移。尽管存在这种失配，我们的方法仍具有竞争力：在 MedleyDB 上训练并在 MDB-stem-synth 上测试时，RPA = 95.84 和 RCA = 96.24；而在 MDB-stem-synth 上进行域内训练和测试时，RPA = 96.54 和 RCA = 97.55 。小幅下降（0。70 RPA，1。31 RCA）表明，在真实伪影条件下进行学习，既能在干净的评估数据上保持准确性，又能提升模型在实际运行条件下的鲁棒性。对于 PESTOv2，括号中的数值表示我们对以 VQT 报告的结果进行的基于CQT的插值，以便在统一的时频表示下进行更公平的比较。

表1：MDB-stem-synth 上的基线比较（RPA/RCA，以 % 计）。PESTOv2 括号中的数值表示使用常数Q变换而非 VQT 得到的结果。

模型	RPA	RCA
CREPE（Kim 等，2018）	96.7	97.0
DeepF0（Singh 等，2021）	98.3	98.4
SPICE（Gfeller 等人，2020）	89.1	–
DDSP-inv（Engel 等人，2020b）	88.5	89.6
PESTOv1（Riou 等人，2023）	94.6	95.0
PESTOv2（Riou 等人，2025年）	94.4	95.0
我们的方法（MDB →MDB）	96.54	97.55
我们的方法（Medley →MDB）	95.84	96.24

跨乐器迁移。表 2 报告了跨乐器泛化结果：每列表示在单种 MedleyDB 乐器上训练，每行表示在未见过的 MDB-stem-synth乐器上测试。在干净的合成语料库上，若干迁移的 RPA 超过 90%，这与其声学复杂度较低（部分频带更窄、静音准确）的特点一致，使谐波结构更易区分。在弦乐器家族内部，迁移效果最强（例如，小提琴↔大提琴），这反映了它们具有相似的激励机制和重叠的音域。低音乐器是最难处理的目标：对所有训练乐器取均值后，电贝司的 RPA 最低（≈ 68.7 %），低音提琴次之（≈ 77.2%），这表明低频主导和瞬态突出的起音会加剧跨音色失配。在所有条件下，RCA 始终高于 RPA，说明残余错误主要源于八度混淆，而非任意音高漂移；对于合成和条件控制而言，这是一种较为理想的失效模式。

4.4 短片段加性重合成

为进行定性验证，我们使用模型逐帧的 $\hat{F}_0$和有声度权重 $\hat{v}$ 驱动一个简单的加性合成器（详见附录）。图 5 比较了以下音频的常数Q变换：(a) 目标 MedleyDB 音频；(b) 合成输出；(c) MDB-stem-synth 音频。

合成谐波列紧密跟随目标的阶梯式 $\hat{F}_0$ 轮廓，且转换发生在正确的帧上，表明音高跟踪准确。有声门控抑制了模糊的音符间区域，使合成的常数Q变换中出现清晰的静音区段，并消除了真实录音中持续存在的残余振铃（例如释放尾音）。细小的三角形——

表 2：在 MedleyDB 上训练（列），在 MDB-stem-synth 上测试（行）。指标为 % RPA / RCA。

训练所用乐器 →	单簧管		大提琴		低音提琴		电贝司		小提琴	
测试乐器 ↓	RPA	RCA	RPA	RCA	RPA	RCA	RPA	RCA	RPA	RCA
单簧管	94.29	95.37	94.31	95.96	80.61	92.68	96.88	96.98	94.69	95.30
大提琴	83.61	86.62	83.29	91.15	96.34	96.65	94.80	95.23	87.96	90.26
低音提琴	54.33	60.45	74.97	82.62	92.88	93.44	88.48	89.68	75.35	81.22
电贝司	47.22	60.44	60.68	82.23	89.59	91.36	86.22	88.56	59.69	80.02
小提琴	91.50	95.37	91.37	96.65	86.05	91.39	93.48	93.85	93.76	96.07

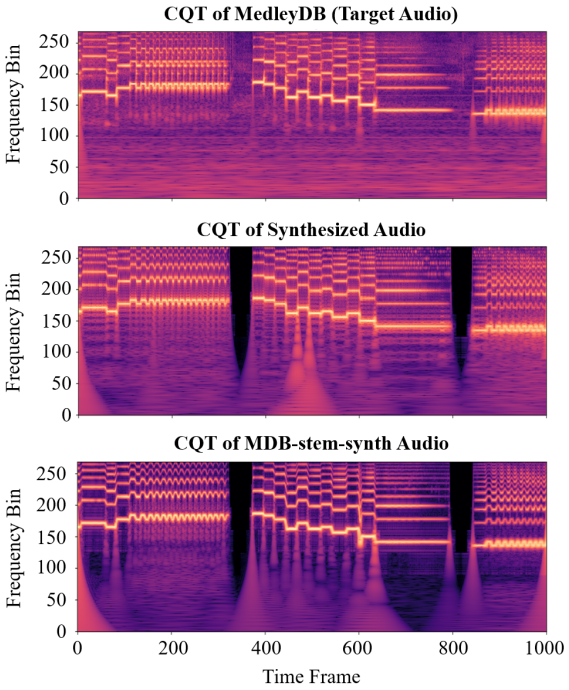


图5：（a）目标音频（常数Q变换）。来自MedleyDB单簧管的轮廓呈现清晰的阶梯状，分音狭窄，音符间隔干净。（b）由$\hat{}F_0$ 和$\hat{}v$合成的音频（常数Q变换）。谐波列与目标保持一致。黑色间隙表示在预测的清音区间内通过发声控制进行静音。最左侧和最右侧的低频楔形区域源于片段淡入/淡出。（c）MDB-stem-synth合成（常数Q变换）。与真实录音相比，重新合成的音频呈现出狭窄、高对比度的谐波分音，起止边界清晰，静音准确（黑色间隙），且调制边带极少。

楔形区域出现在片段边界附近，这是由重叠相加窗函数造成的；延长淡入淡出时间可以减少这种现象，但此处保留这些区域，以明确标示门控位置。

5 结论

我们提出了一种轻量级、完全自监督的框架，用于从真实声学环境下的单音乐器分轨中联合估计F_0 和发声推理。该模型完全使用 MedleyDB 进行训练，因此能够接触到串音、混响以及低信噪比音符间区域；这些情况通常会导致在重合成数据上训练的音高跟踪器失效。

我们的核心贡献是一种EM式迭代重加权策略，该策略利用基于SCE的转置一致性作为可靠性代理指标。稳定的谐波帧保留较高权重并推动学习，而不可靠的帧则逐步降低权重，从而抑制高方差梯度。收敛后的权重可作为置信度分数，用于伪标签以及训练轻量级二元发声分类器，无需人工标注。

从定量结果来看，该方法在跨语料库迁移中仍具有竞争力（Medley→MDB：RPA 95.84，RCA 96.24），与域内训练的结果差距很小（MDB→MDB：RPA 96.54，RCA 97.55）；跨乐器矩阵则显示出较强的同族迁移能力，其中低音乐器仍是最具挑战性的目标（RCA ＞ RPA，错误主要由八度错误构成）。这些结果表明，所提出的提取器可作为稳健的前端，用于 DDSP 风格神经合成与个性化中的快速乐器特定条件化。

参考文献

Beigi, H. (2011).《说话人识别基础》。Springer，纽约。ISBN：978-0-387-77591-3。Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C. 和 Bello, J. P. (2014). MedleyDB：面向高强度标注音乐信息检索研究的多轨数据集。第15届国际音乐信息检索学会会议，台北，台湾。Brown, J. C. (1991). 常数Q谱变换的计算。《美国声学学会会刊》，89(1):425–434。de Cheveigné, A. 和 Kawahara, H. (2002). YIN，一种

用于语音和音乐的基频估计器。《美国声学学会会刊》，111(4):1917–1930。Engel, J., Hantrakul, L., Gu, C.-Z. 和 Roberts, A. (2020a). DDSP：可微分数字信号处理。载于国际学习表征会议（ICLR）。Engel, J., Hantrakul, L. H., Swavely, R. J., Roberts, A. 和 Hawthorne, C. G.-M. (2020b). 通过逆向音频合成进行自监督音高检测。Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. 和 Wichmann, F. A. (2020). 深度神经网络中的捷径学习。《自然·机器智能》，2(11):665–673。Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M. 和 Velimirovic, M. (2020). SPICE：自监督音高估计。IEEE/ACM 音频、语音与语言处理汇刊，28:1118–1128。Hayes, B., Shier, J., Fazekas, G., McPherson, A. 和 Saitis, C. (2023). 可微分数字信号处理在音乐与语音合成中的综述。载于《信号处理前沿》，第3卷。Huber, P. J. (1964). 位置参数的稳健估计。《数理统计年刊》，35(1):73–101。Kim, J. W., Salamon, J., Li, P. 和 Bello, J. P. (2018). CREPE：用于音高估计的卷积表征。ICASSP，IEEE 声学、语音与信号处理国际会议论文集，2018年4月：161–165。Mauch, M. 和 Dixon, S. (2014). PYIN：一种利用概率阈值分布的基频估计器。载于2014年IEEE声学、语音与信号处理国际会议（ICASSP），第659–663页。Morrison, M., Hsieh, C., Pruyne, N. 和 Pardo, B. (2024). 跨领域神经音高与周期性估计。arXiv预印本 arXiv:2301.12258。Riou, A., Lattner, S., Hadjeres, G. 和 Peeters, G. (2023). PESTO：采用自监督移调等变目标的音高估计。载于第24届国际音乐信息检索学会会议（ISMIR）论文集，第535–544页。Riou, A., Torres, B., Hayes, B., Lattner, S., Hadjeres, G., Richard, G. 和 Peeters, G. (2025). Pesto：采用自监督移调等变目标的实时音高估计。《国际音乐信息检索学会汇刊》，8(1):334–352。Salamon, J., Bittner, R. M., Bonada, J., Bosch, J. J., Gómez, E. 和 Bello, J. P. (2017). 用于多轨数据集基频自动标注的分析／合成框架。第18届国际音乐信息检索学会会议（ISMIR）论文集，第71–78页。

Saxena, K., Kumar, A., Srivastava, P., 和 Gupta, A.（2024）。基于主动元学习的交互式歌唱旋律提取，用于领域适配。arXiv 预印本 arXiv:2402.07599。Singh, S., Wang, R., 和 Qiu, Y.（2021）。Deepf0：面向音乐和语音信号的端到端基频估计。发表于 ICASSP 2021——2021 年 IEEE 国际声学、语音与信号处理会议（ICASSP），第 61–65 页。IEEE。Talkin, D.（1995）。一种稳健的音高跟踪算法（RAPT）。《语音产生与语音建模》，第 495–518 页。Yost, W. A.（2009）。音高感知。《注意、感知与心理物理学》，71(8):1701–1715。

附录

数据集详情与统计

本文研究使用了两个不同的数据集：MedleyDB，由真实场景中的多轨录音组成；以及 MDB-stem-synth，一个由干净的重合成音轨构成的数据集。正如论文正文所述，MedleyDB 用于提供声学复杂训练数据（包含串音、噪声和混响），而 MDB-stem-synth 则用于评估，以干净且精确标注的基准真值作为参照。

表3详细比较了本研究所用两个数据集中具体乐器子集的统计数据。比较内容包括每种乐器对应的片段总数、累计音频时长和平均片段长度。

对于所选乐器，MedleyDB 子集包含更多片段（129 对 105），且音频总时长显著更长（约 2.45 万秒，对比 1.79 万秒）。这为自监督训练提供了规模更大、声学多样性更丰富的语料库。值得注意的是，部分乐器（如小提琴、电贝司）的平均片段时长相近，而其他乐器（如单簧管、低音提琴）则差异较大，这反映了两个数据集在构成上的不同。

训练目标

一个批次<各N 帧的总计损失是每个样本损失的平均值，其中各个组成部分定义如下：

$$\mathcal{L}_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}\left(w_i(\mathcal{L}_{\text{equiv},i} + \mathcal{L}_{\text{invar},i} + \mathcal{L}_{\text{sce},i})\right)$$

表3：MedleyDB（真实录音）和 MDB-stem-synth（合成分轨）中所用乐器音频的统计比较。时长以秒为单位。

乐器	MedleyDB（录音）			MDB-stem-synth（合成）		
	片段	总计（秒）	平均（秒）	片段	总计（秒）	平均（秒）
单簧管	9	885.39	98.38	8	829.21	103.65
大提琴	13	2344.35	180.34	11	1816.80	165.16
低音提琴	17	2919.01	171.71	14	1722.81	123.06
电贝司	63	11,398.80	180.93	58	9908.25	170.83
小提琴	27	6986.28	258.75	14	3619.81	258.56
总计	129	24533.83	–	105	17896.88	–

该公式使用软权重w_i，以确保F_0特定的等变性和不变性损失主要由被识别为强有声的样本驱动。

用于F_0 估计（$\mathcal{L}_{\text{equiv}}$）的等变性损失。为了学习$F_0$，，我们强制施加音高等变性。我们为$k \in \{p_{\min},\dots,p_{\max}\}$定义一组权重$w_k = \alpha^k$，其中 α为常数。如果输入X 经过 δ 个频率箱的移调后生成X'，则相应的输出z 和z' 应满足$z' \approx v^{\delta} z$。损失采用对称Huber损失（Huber，1964）进行定义：

$$\mathcal{L}_{\text{equiv}} = \frac{1}{2}\left(\mathcal{H}\left(\frac{z'}{z} - \alpha^{\delta}\right) + \mathcal{H}\left(\frac{z}{z'} - \frac{1}{\alpha^{\delta}}\right)\right)$$

该损失确保将输入移调 δ 个频率箱后，会在中间表示z中产生相应的乘法偏移：具体而言，$z' \approx \alpha^{\delta} z$，其中 α 定义了与对数CQT结构相匹配的指数音高尺度。对称Huber损失能够抵抗瞬态和噪声帧的影响，同时确保模型将音程学习为一致的乘法因子，而无需标注数据。

音色表示的不变性损失（$\mathcal{L}_{\text{invar}}$）。虽然$F_0$ 对音高移位具有等变性，但乐器音色应保持不变。给定原始输出y以及移调版本的输出y'，音色表示应相似。这可以通过对称交叉熵损失（Beigi，2011）实现：

$$\mathcal{L}_{\text{invar}} = \frac{1}{2}\left(\text{CE}(y, p') + \text{CE}(y', p)\right)$$

该损失通过约束原始输入与移调输入的输出 logits y 和 y' 之间的相似性，促使模型学习不受音高影响的音色特征。由于音色特征例如频谱形状、谐波衰减率和共振峰在不同音高下保持一致，因此对称交叉熵损失会惩罚预测分布之间的差异。将音高与音色解耦对于下游任务至关重要，包括乐器识别、合成控制和音色迁移，从而确保学到的特征能够泛化到同一乐器演奏的不同音符。

移位交叉熵损失（$\mathcal{L}_{\text{sce}}$）。该损失强制满足循环等变性，是 E 步的关键信号。如果输入的音高移位了 δ 个频率箱，输出分布也应循环移位相同的数量。设y为原始输出，y' 为变换后的输出。损失为（Riou 等人，2023）：

$$\mathcal{L}_{\text{sce}} = \frac{1}{2}\left(\text{CE}(y', \text{roll}(p, \delta)) + \text{CE}(y, \text{roll}(p', -\delta))\right)$$

这种核心自监督损失施加了循环等变性约束：如果输入平移了 δ 个频率箱，输出概率分布也应循环平移相同的幅度。它使用对称交叉熵，将平移后输入的预测分布与原始分布经过相应循环移位后的版本进行比较。这一约束通过确保整个预测音高分布能够正确且一致地平移，来优化F_0 预测，并促使概率分布尖锐、定位准确且集中于正确的分箱。值得注意的是，$\mathcal{L}_{\text{sce}}$ 还可作为隐式的有声检测器：具有清晰谐波结构的帧会产生尖锐且平移规律明确的分布，因此 SCE 损失较低；而含噪或清音帧则会产生较平坦、不稳定的分布，因此 SCE 损失较高。这一特性使迭代重加权算法的 E 步能够自动识别并降低不可靠帧的权重，从而使F_0 学习能够抵御真实环境中的噪声和伪影，详见。

浊音分类器

自监督表征学习阶段完成后，网络$f(\cdot)$的权重将被冻结。迭代重加权过程中针对整个训练数据集计算出的最终样本权重w_i，将用于生成硬伪标签。这些权重的取值范围为 0 到 1，能够有效反映模型收敛后对某帧是否包含稳定有声音信号的置信度。通过对这些最终权重进行阈值处理，即可为$v_i^* \in \{0,1\}$ 分配硬伪标签：

$$v_i^* = \begin{cases} 1 & \text{if } w_i > \theta \\ 0 & \text{otherwise} \end{cases}$$

其中 θ 是预先设定的置信度阈值。这些伪标签将作为后续监督训练的目标。

直接在 CQT 频谱图上训练一个轻量级的单层线性分类头。该层将每个 CQT 输入帧X 映射为一个逻辑值，用于二分类（有声/无声）。分类器使用生成的伪标签v^*作为目标，通过标准二元交叉熵损失进行优化。我们特意选择小型单层架构，而非更深的多层网络，因为在初期迭代中，当F_0 估计仍不准确时，EM 重加权过程会不可避免地降低一些实际有声帧的权重。容量更大的模型往往会对这些不完善的伪标签过拟合，错误地将初期权重被降低、但实际上有声的帧学习为无声。单个线性层的容量有限，能够提供天然的正则化，促使分类器只捕捉最稳健的浊音模式，同时更好地泛化到迭代训练过程中暂时被误分类的帧。尽管此阶段的推理不会使用特征提取器$f(\cdot)$，但它通过 EM 过程生成高质量伪标签的作用至关重要。这种两阶段方法利用自监督过程中学到的音高结构来训练最终的高性能浊音检测器，无需任何外部真实标签。

谐波合成器

为定性验证我们预测的基频（$\hat{F}_0$）和发声（v）轮廓的准确性，我们采用分析合成流程。我们使用一个可微分的谐波合成器（如图 6 所示），仅使用预测的 $\hat{F}_0$ 和 v 作为输入来重建音频片段。随后优化合成器的参数，使其匹配目标音频，从而稳健地检验预测的轮廓是否足以实现高保真重建。

模型架构 合成器是一种专用的加性模型，通过对 H 个谐波求和来生成音频，其中每个谐波的振幅都是可学习参数，并随时间变化。模型的输入是逐帧的 $\hat{F}_0$个分箱索引和二元浊音标记v，两者的帧率均为 CQT 帧率（跳步长度为 160 个采样点）。

1. 输入预处理：首先使用 CQT 频率查找表，将 $\hat{F}_0$ 个分箱索引转换为赫兹。随后，将 $\hat{F}_0$ (Hz) 轮廓和浊音轮廓v 从帧率上采样至音频采样率（$f_s = 16000$ Hz）。我们对 $\hat{F}_0$ 轮廓使用线性插值，以确保频率平滑过渡；对浊音轮廓使用最近邻插值，以保持起音和消音边界清晰。由此得到采样率信号$f_0(t)$ 和$v(t)$。

2. 瞬时相位：第h次谐波的瞬时频率为$f_h(t) = h \cdot f_0(t)$。瞬时相位 $\phi_h(t)$通过对频率积分计算，具体以相位增量的累积和近似，并加上每个谐波可学习的全局相位偏移 ψ_h：

$$\phi_h(t) = \left(2\pi \sum_{i=1}^{t} \frac{h \cdot f_0(i)}{f_s} \right) + \psi_h$$

其中，f_s 是采样率（16000 Hz），而$\psi \in \mathbb{R}^H$ 是一个可学习参数向量。

3. 可学习振幅包络：合成器的核心是随时间变化的振幅控制。可学习参数矩阵$G_{\text{pre}} \in \mathbb{R}^{H \times T}$ 为T帧中的每一帧的H 个谐波分别存储一个预激活增益。这些增益经过 Softplus 函数处理以确保为正，G_{f} =Softplus(G^{pre})，然后线性插值至采样率，得到$g_h(t)$。

4. 合成：最终的音频信号$s(t)$由各个谐波相加生成，每个谐波都由其特定的增益包络、主发声门控$v(t)$以及可学习的对数主增益g_{master}进行调制。频率高于奈奎斯特极限（$f_s/2$）的谐波会被屏蔽

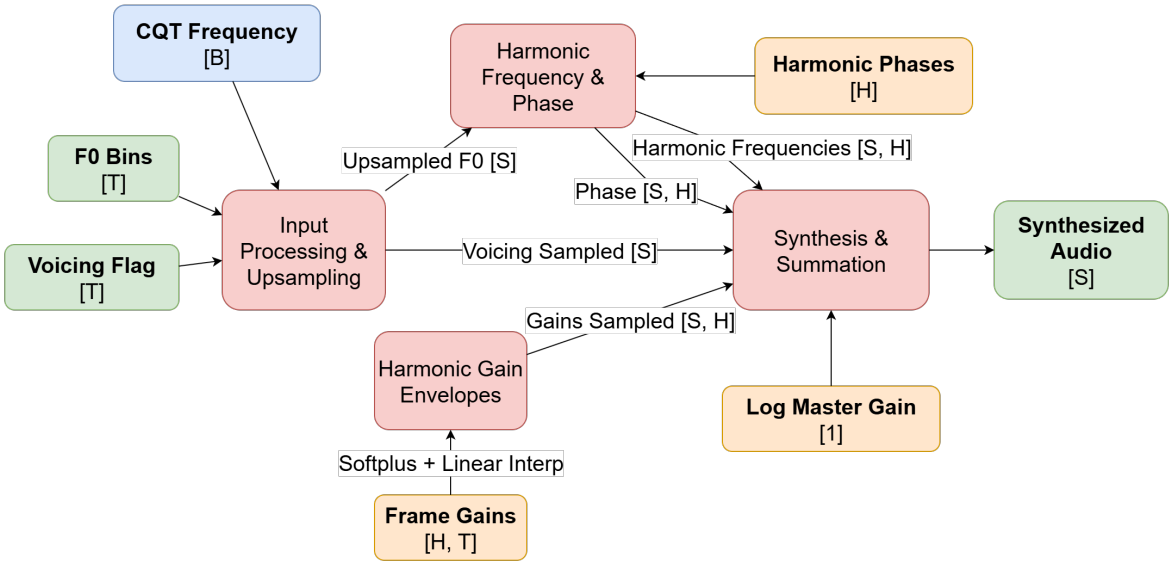


图 6：用于分析合成的谐波合成器架构。输入（“F_0 频率箱”、“浊音标志”）经过处理和上采样。可学习参数（“帧增益”、“谐波相位”、“对数主增益”）控制谐波频率、相位和振幅。合成阶段将这些组件组合起来，并应用浊音和奈奎斯特掩蔽，从而生成最终的音频输出。

out.

$$s(t) = 10^{g_{\text{master}}} \cdot v(t) \cdot \sum_{h=1}^{H} g_h(t) \cdot \sin(\phi_h(t))$$

优化 对于给定的目标音频片段$y(t)$及其对应的$\hat{F}_0$和v 轮廓，我们优化模型的可学习参数：逐帧增益G^{pre}、谐波相位 ψ以及主增益g_{master}。

目标是最小化合成音频 $s(t)$ 与目标信号 $y(t)$ 之间的感知差异。我们使用 auraloss 库中的多分辨率短时傅里叶变换损失（MR-STFT），该损失会在多个分辨率（快速傅里叶变换尺寸、跳步长度和窗长）下计算短时傅里叶变换的幅度的 L1 损失和对数幅度的 L2 损失。

为了使振幅包络更加平滑、逼真，我们添加了平滑正则项。该惩罚项是逐帧增益矩阵 G_{f} 的时间全变差（TV）均值，用于惩罚相邻帧之间大幅且突兀的变化：

$$\mathcal{L}_{\text{smooth}} = \frac{1}{H(T-1)} \sum_{h=1}^{H} \sum_{t=1}^{T-1} |G_{\text{f}}[h,t+1] - G_{\text{f}}[h,t]|$$

最终损失是一个加权和：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MR-STFT}}(s,y) + \lambda \cdot \mathcal{L}_{\text{smooth}}$$

使用 Adam优化器对模型进行固定轮数的训练。此过程实际上执行分析合成，在给定模型的 $\hat{F}_0$ 和发声预测的情况下，找出重构目标音频所需的最佳谐波振幅。