

第 6 章 基于 Harmonic-CRNN 的主旋律提取方法

6.1 简介及相关工作

在多音音乐中，不同声部分别有着各自的旋律进行。其中，主旋律被定义为与主唱或者主乐器声音对应的基频序列。在听觉上，主旋律往往占据主导地位。在频域上，与主旋律对应的基频位置及其若干个泛音位置的能量都会相对显著。基于此，有研究人员便提出了一种谐波加权求和的基频估计方法^[113]。该方法首先进行短时傅里叶变换得到短时傅里叶频谱，并在频谱的每一帧上筛选出若干个峰值作为候选基频，然后通过谐波求和公式6.1计算出当前帧每个候选基频对应的显著性值。每个候选基频的显著性值表示它是当前帧基频的可能性，显著性值越大，可能性越大。

$$S = \sum_{n=1}^N a(n) |Amp(nf_0)| \quad (6.1)$$

其中， $Amp(nf_0)$ 表示在当前帧上候选基频 f_0 的第 n 个谐波所对应的幅值； $a(n)$ 是人为设定的权重函数，在求和之前，赋予 f_0 第 n 个谐波权重。

通过上述步骤，方法由短时傅里叶频谱计算得到一张显著性函数图，横轴为时间，纵轴为频率。最后，从显著性函数图的每一帧中选出最大值，如果值大于 0.52，那其位置对应候选基频被认为是当前帧的基频，如果值小于 0.52，则表示当前帧不包含基频。

自该方法提出以后，公式6.1和类似的规则常常以不同的名字和变化被研究人员用于主旋律提取任务^[10, 62, 73, 77]。其中，经典的 melodia 主旋律提取方法^[77]便是对公式6.1进行了改进。melodia 方法将基频候选值的范围限定在 32.5 赫兹到 1760 赫兹，并把该频率范围量化为 600 离散值 b ， $b \in [1, 600]$ ，每两个离散值之间的频率间隔为 10 音分。该方法同样先进行短时傅里叶变换，并在频谱的每一帧上筛选出若干个峰值作为候选基频。这些候选基频通过公式6.2转换为 600 个离散值中的一个。然后，根据公式6.3计算每帧的每个候选基频值的显著性值，得到显著性函数图。之后，方法基于旋律线连续、平滑等特性从显著性函数中提取出基频序列。

$$B(\hat{f}) = \lfloor \frac{1200 \times \log_2(\hat{f}/55)}{10} + 1 \rfloor. \quad (6.2)$$

其中， \hat{f} 表示候选基频。

$$S(b) = \sum_{n=1}^N \sum_{i=1}^I e(\hat{a}_i) \times g(b, h, \hat{f}_i) \times (\hat{a}_i)^\beta, \quad (6.3)$$

其中， N 表示所考虑的谐波个数， I 为筛选出的峰值数量， \hat{a}_i 和 \hat{f}_i 分别为第 i 个峰值的瞬时频率和幅值， $e(\hat{a}_i)$ 表示的是一个幅值门阀函数， $g(b, h, \hat{f}_i)$ 是个基于高斯分布的权重函数，以及 β 是一个幅值压缩参数，具体细节可参见原文。

随着神经网络模型在各个领域展示出其强大的性能，Rachel 等人在 2017 年提出了一种基于 HCQT（Harmonic Constant-Q Transform）和卷积神经网络的主旋律提取方法。由于在 CQT 频谱的频率维度，第 k 个频带频率 f_k 与最小频率 f_{min} 的关系是 $f_k = f_{min} \cdot 2^{k/M}$ ，其中， M 表示每个八度的频带数量，以及第 k 个频带频率的第 h 次谐波频率 $h \cdot f_k = h \cdot f_{min} \cdot 2^{k/M}$ 。因此，分别以 f_{min} 的 h 倍作为最小频率计算得到的 CQT 谱和以 f_{min} 为最小频率得到的 CQT 谱，在同一频带上，前者频率是后者的 h 倍，也就是说两频谱对应位置之间具有 h 倍的谐波关系。作者依次以 32.7 赫兹（即 C1）的 0.5, 1, 2, 3, 4, 和 5 倍为最小频率进行 CQT 变换，得到 6 个 CQT 频谱图，并将 6 个频谱图输入一个 5 层的 CNN 模型中，通过有监督的方式间接学习 6 个谐波之间的权重关系。该方法在测试集上取得了不错的效果。然而，该模型只能计算至 5 次谐波，而无法捕获更高次的谐波。此外，对于旋律帧的定位，方法也仅采用简单的阈值判断，这导致模型预测时产生了很高的旋律帧误报率。

Yuzhi Hang 等人在 2018 年，提出一种 HNN（Harmonic Neural Network）模型来提取主旋律。方法在 CQT 频谱中为每帧的每一个候选基频找到与其对应的 3 个泛音并拼接为一个 12 维的向量，以此作为当前帧当前候选基频的谐波结构特征。然后将每个候选基频的 12 维的谐波结构特征单独输入 Resnet 或 LSTM 中预测该基频候选值是基频的概率。对于旋律帧的定位，模型设计了另外一个以 CQT 频谱为输入的并行的 LSTM 网络。该方法最大的问题是虽然考虑了谐波结构特征，但方法将特征输入 Resnet 或 LSTM 网络，很难解释模型到底学到了什么。此外，方法只能计算至 4 次谐波，也未考虑候选基频以及基频序列的上下文关系。

受 HNN 模型的启发，本章提出一个基于谐波和卷积的 CRNN 模型。方法首先提取谐波结构特征。针对 HNN 方法中仅能考虑 4 个谐波的问题，我们采用 STFT 频谱代替 CQT 频谱。尽管通过增加八度包含的频带数量可以使得在 CQT 频谱上获得高次谐波，但是由于 CQT 在高频的频率分辨率比较低，而我们希望

在定位高次谐波频率时能够定位到更准确的位置，因此，STFT 频谱最为合适。进行短时傅里叶变换时，为了使得频谱的频率分辨率尽可能地高，我们将采样率和窗长分别设置为 44100 赫兹和 8192。方法考虑的候选基频范围是 44.16 到 1760 赫兹，因此，在 STFT 频谱上最高能捕获至 12 次谐波。

在提取谐波结构特征后，我们不像 HNN 模型将特征输入无法解释的 Resnet 或 LSTM 网络，而是受公式6.3启发，搭建一层卷积层来计算每个候选基频对应的谐波和。为了方便，本文将这层卷积称为谐波和卷积层。其中卷积核表示每个谐波位置的权重，对应公式6.3中的权重函数 $g(b, h, \hat{f}_i)$ 。而且由于我们设置的卷积核是 2 维的，即在时间帧上也具有感受野，因此在计算每帧候选基频位置对应的谐波和时可以考虑一定范围内的上下文信息。

经过这次谐波和卷积层，我们从 STFT 频谱得到一个显著性函数图。又考虑到旋律走向是基于一定的乐理，基频序列在时序上的上下文关系也很重要，因此，我们将显著性函数输入一个具有残差结构的 CRNN，希望能够得到更精确地显著性函数图。

对于旋律定位子任务，我们在谐波和卷积层之后，设计了一层全连接层对显著性函数中每一帧依次判定。原因是我们在显著性函数图中，基频及其泛音位置的能量会比较强，形成比较清晰的谐波结构，而非基频的谐波结构会比较弱，我们希望全连接层能够学习到这种差异，从而判别旋律是否存在。我们在 3 个公开测试集上对本章提出的模型进行评测。结果表明，本章提出的模型在三个测试集上均超过了目前性能最好的算法。

6.2 算法描述

为了能够清晰地描述本章节提出的算法，我们分为两个部分进行介绍：第一部分介绍基于 STFT 频谱提取的谐波结构特征，第二部分介绍本章设计的模型结构。

6.2.1 基于 STFT 频谱的谐波结构特征

首先进行短时傅里叶变换。为了能够捕获更高次的谐波，以及能够将高次谐波定位到更准确的频谱位置，短时傅里叶变换频谱应具备较高的奈奎斯特频率和频率分辨率。为此，我们设置采样率为 44100 赫兹，傅里叶点数为 8192，窗长为 2048，以及滑动窗口为 512，计算得到奈奎斯特频率为 22050 赫兹，频率分辨率为 5.38 赫兹的 STFT 频谱。

得到 STFT 频谱后，我们根据候选基频提取相应的谐波结构特征。本章算法考虑的候选基频范围是 44.16 到 1760 赫兹，因此可以考虑到的最高次谐波为 12 ($22050/1760 \approx 12.5$)。我们将候选基频范围以每两个音高类别间隔 20 音分进行

量化，共得到 320 个音高类。

特征提取过程如图6-1所示，对于每一帧，我们首先根据公式6.4将 320 个候选基频及各自的 N 个倍频映射到 STFT 频率的频率轴上。

$$Bin(f) = \frac{f \times 8192}{44100}, \quad (6.4)$$

然后，提取每个候选基频及其 N 个倍频所在位置的频谱并在频率维度拼接，作为相应候选基频的谐波结构特征。考虑到在频谱上倍频可能不会正好出现在基频的整数倍位置（可能会上下浮动），因此我们在提取基频和其泛音对应的频谱时，同时提取该位置上下 2 个位置的频谱作补充。这样，对于每个帧数为 L 的片段，都会产生一个 3 维的谐波结构特征，维度大小为 $(320, 5N, L)$ 。

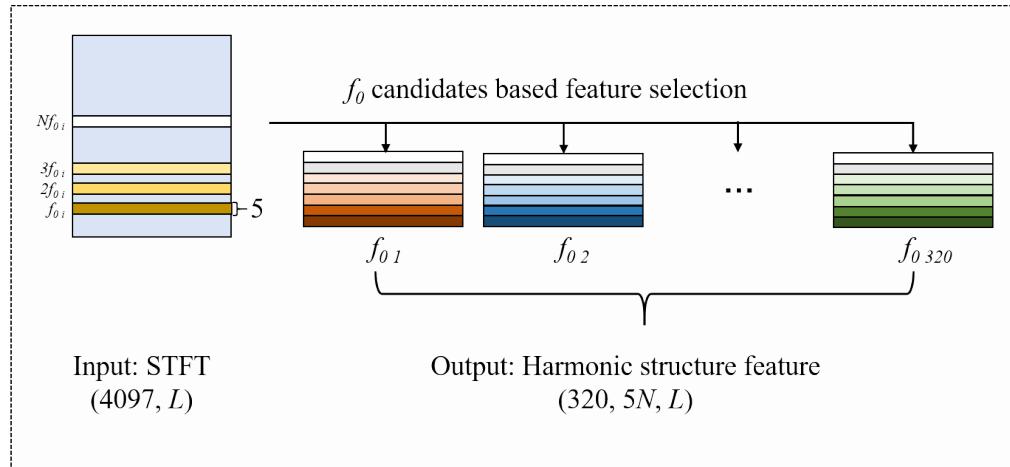


图 6-1 基于 STFT 频谱的谐波结构特征的提取。

在图??中，我们以 ADC2004 数据集中文件名为“pop2.wav”的音频片段为例，显示该音频片段某一帧的基频对应的谐波结构特征。首先进行 STFT 得到该音频片段的 STFT 频谱，如图左上角，这里我们仅显示了第 700 帧到 900 帧的频谱。为了更清楚地看到 STFT 频谱中基频的谐波结构，我们将 2000 赫兹以下的频谱放大，显示于图的右上角。从右上角的频谱中，我们可以看到在基频及其泛音位置能量比较大，呈现了比较清晰的谐波结构。以第 820 帧为例，该帧人工标注的基频为 340.95 赫兹，在 320 个音高类中，最接近候选基频值为 341.25 赫兹的第 63 类。我们设 $N = 9$ ，根据图6-1中的方法，提取该帧候选基频的谐波结构特征，并显示于左下角。同时，我们也随机提取了候选基频值为 322.09 赫兹对应的谐波结构特征，显示在图的右下角。对比两个不同候选基频对应的谐波结构特征，可以明显地看出，候选基频值为 341.24 赫兹的谐波结构特征更有意义。

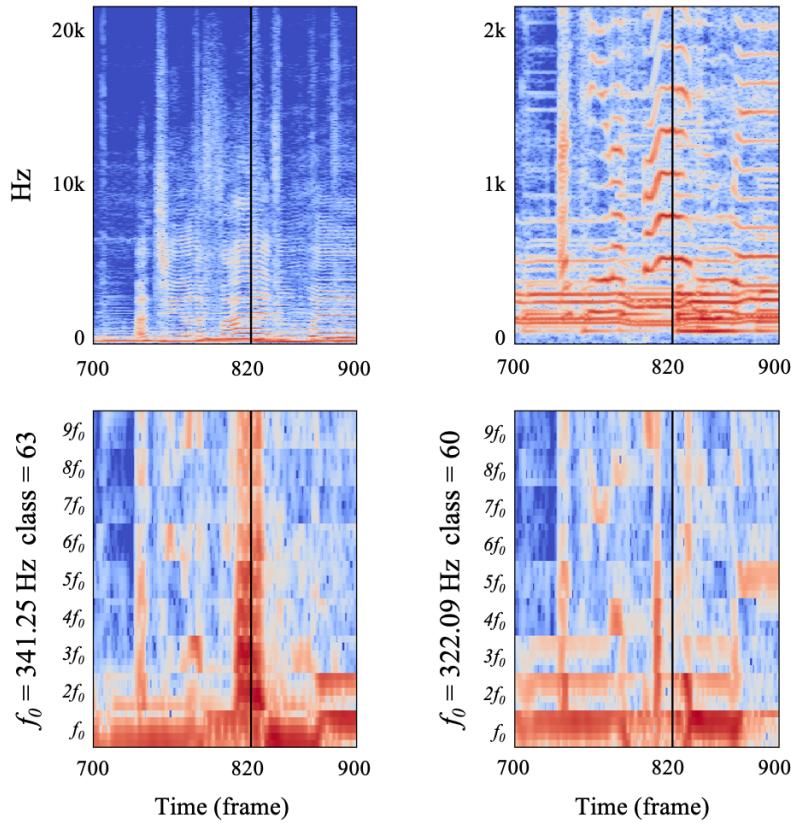


图 6-2 以 ADC2004 数据集中文件名为“pop2.wav”的音频片段为例，显示该音频片段第 820 帧基频所对应的谐波结构特征。

6.2.2 模型结构

给定一段音频片段，根据上节提出的特征提取方法，我们可以得到维度为 $(320, 5N, L)$ 的谐波结构特征。其中，每个候选基频对应着一个维度为 $(5N, L)$ 的二维特征。众所周知，卷积层的卷积操作是将卷积核中的各参数与输入矩阵相应位置的数值相乘后再求和。那么，我们将每个候选基频对应的维度为 $(5N, L)$ 的特征输入一层卷积核大小为 $(5N, W)$ 的卷积层，正好可以实现谐波加权求和操作。其中的 W 是卷积核在时间帧上的感受野，当 W 为1时，可以近似实现公式6.3。

如图??所示，320个候选基频共享这层谐波和卷积，都生成一个维度为 $(1, L)$ 的向量。将这320个维度为 $(1, L)$ 的向量在第一维上进行拼接，可以得到一张维度为 $(320, L)$ 的矩阵。该矩阵我们称为显著性函数图，图中的每个值表示与该值对应候选基频是基频的可能性。值越大，属于当前帧基频的可能性就越大。

考虑到仅用1层卷积得到显著性函数图很容易出现噪声干扰，我们将显

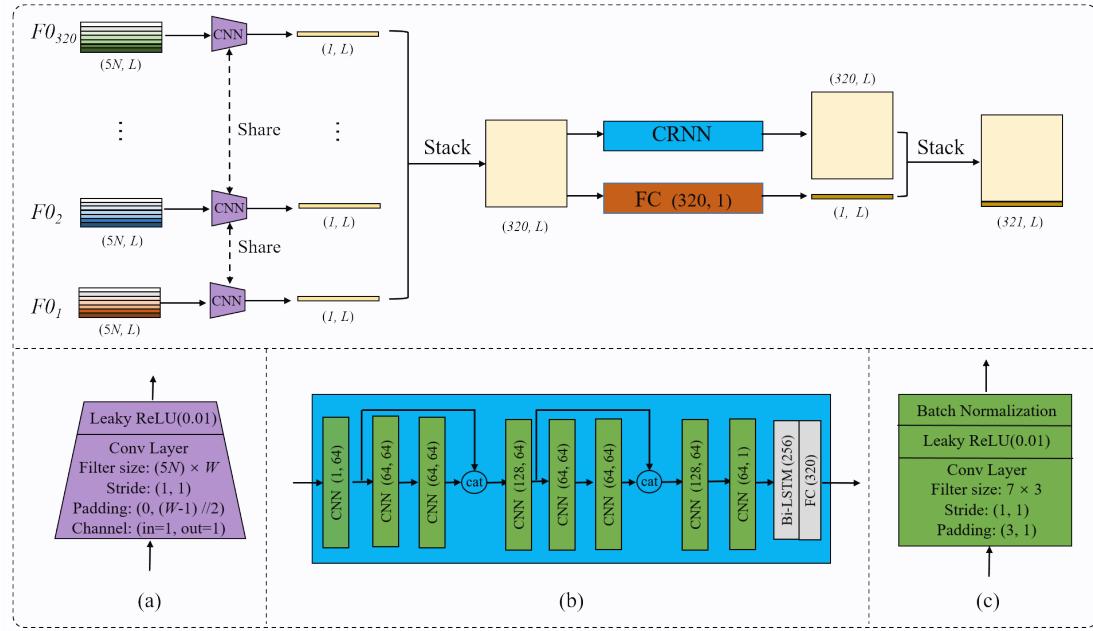


图 6-3 模型的整体框架

著性图继续输入一个卷积循环神经网络 CRNN (Convolutional Recurrent Neural Network) 中，希望可以学到更清晰的显著性函数图。CRNN 的结构如图??-(b) 所示，共包含了 8 个卷积层以及 1 个 Bi-LSTM 层。表??列出了 CRNN 中各层具体的参数设置，其中每个卷积层的卷积核都设为 $(7,3)$ ，目的是希望网络能够更多地捕获频率维度的信息。在每层卷积之后，都跟随着一层 LeakyRelu 激活函数 ($\alpha = 0.02$)，以及批归一化 BN 层。此外，我们还设置了两条跳跃连接 (skip connection)，这样不仅能够融合不同层次的特征，还可以提升收敛速度。由于旋律走向是基于一定乐理的，基频序列的上下文关系也很重要，为此我们增加了一个 Bi-LSTM 层。

除了判断音高类别，我们还需要判别帧内是否包含旋律。在谐波和卷积层得到的显著性函数图中，当帧内包含旋律时，基频位置及其泛音位置的能量会比较明显，形成比较明显的谐波结构，而如果帧中不包含旋律，能量分布会比较没有规律且能量都相对较弱。基于这个特点，我们搭建一个全连接层直接从显著性函数图中判别帧内是否包含旋律。这个全连接层会生成一个维度为 $(1, L)$ 的向量。该向量经过 LeakyReLU 之后，与 CRNN 的输出在频率维度进行拼接，作为模型最后的输出。

表 6-1 CRNN 的参数设置

	filter size, padding size	(inchannel, outchannel)
conv_layer_1	(7, 3),(3, 1)	(1, 64)
conv_layer_2	(7, 3),(3, 1)	(64, 64)
conv_layer_3	(7, 3),(3, 1)	(64, 64)
cat_layer	cat (conv_layer_1, conv_layer_3)	
conv_layer_4	(7, 3),(3, 1)	(128, 64)
conv_layer_5	(7, 3),(3, 1)	(64, 64)
conv_layer_6	(7, 3),(3, 1)	(64, 64)
cat_layer	cat (conv_layer_4, conv_layer_6)	
conv_layer_7	(7, 3),(3, 1)	(128, 64)
conv_layer_8	(7, 3),(3, 1)	(64, 1)
Bi-LSTM		256
fc_layer		320

6.2.3 损失函数

本章算法采用本文第五章提出的损失函数。该损失函数的提出是为了缓解训练集中大量无歌声片段造成的类别不平衡问题。损失函数 L 由两部分损失加权求和得到：

$$L = \alpha L_1 + \beta L_2, \quad (6.5)$$

第一部分损失 L_1 仅考虑模型在旋律帧上对音高类估计的损失，不考虑任何非旋律帧，它的定义为

$$L_1 = \begin{cases} CE(\mathbf{y}, \hat{\mathbf{y}}) & \text{如果当前帧包含旋律} \\ 0 & \text{否则} \end{cases}, \quad (6.6)$$

其中， CE 表示交叉熵损失函数， \mathbf{y} 是标注的正确基频， $\hat{\mathbf{y}}$ 是模型预测的概率。在 L_1 的计算中，由于没有考虑任何非旋律帧，因此，能够很大程度上缓解大量非旋律帧带来的负面影响。

第二部分损失 L_2 计算在所有帧上语音激活检测的损失，它的定义为

$$L_2 = BCE(\mathbf{v}, \hat{\mathbf{v}}), \quad (6.7)$$

其中， BCE 表示二进制交叉熵损失函数， \mathbf{v} 和 $\hat{\mathbf{v}}$ 分别是语音激活检测的真实值和估计值。这里的估计值 $\hat{\mathbf{v}}$ 通过计算所有 320 个音高类的输出概率之和得到。本章算法设置 $\alpha = 3$ 和 $\beta = 1$ 。

6.3 算法评测

6.3.1 数据集

本章算法按采用的训练集来自 MedleyDB 数据集。它是由美国纽约大学音频音乐研究实验室 (Music and Audio Research Lab) 的 Rachel Bittner 等人制作完成的，共包含 122 首多轨道音乐，其中大部分音乐的时长为 3 到 5 分钟。在 122 首多轨道音乐中，只有 108 首对主旋律基频进行了标注，其中以歌声为主旋律的有 61 首，剩余 47 首均以乐器声为主旋律。该数据集中所有音频的采样率为 44100Hz，并且连续两个音高标注的间隔约 5.8 毫秒。本章算法将 61 首以歌声为主旋律的歌曲分为两部分，其中 49 首用来作训练集，剩余的 12 首歌曲作为验证集。为了扩增训练样本，我们对训练集中的每首歌曲升高/降低 1 个半音和 2 个半音，这样训练样本数量被扩增为原来的 5 倍。我们采用三个常用的公开测试集，分别为 ADC2004 数据集，MIREX05 数据集，以及 ikala 数据集。

6.3.2 评测标准

按照惯例，我们采用常见的五个指标来评估算法性能。这五个指标分别为语音召回率 (VR)、语音虚警率 (VFA)、原始基音准确度 (RPA)、原始色度准确度 (RCA)，以及总体准确度 (OA)。mir_eval 工具箱被用来计算这五个指标，并且当估计的基音值和真实值之间的差值在 ± 50 音分范围内时，估计的基音值被认为是正确的。

由于 5 个评测指标中最重要的两个指标是 RPA 和 OA，因此在参数选择的实验中，我们通过计算两者的平均值来衡量模型的性能。为了方便，我们称这个平均值为 MRO。

6.3.3 训练参数

在训练阶段，我们设置的批大小为 16，以及原始学习率为 0.0001。在训练集上每完成一次迭代，学习率会被降低为先前值的 99%。当经过 10 次训练集的迭代，模型在验证集上的损失仍未下降，训练停止。我们采用 ADAM 优化算法^[1]最小化损失函数。

6.3.4 参数选择

本章提出的算法共包含 3 个超参数，第一个是提取谐波结构特征时，要考虑的谐波数量 N 。它决定了谐波和卷积层的卷积核在频率维度上的感受野。第二个是谐波和卷积层的卷积核在时间帧上的感受野 W 。第三个是每个训练样本的帧数 L ，也就是一次输入模型的帧数。在这小节，我们在验证集上进行一系列实验来确定这些超参。

第一个实验是确定在提取谐波结构特征时，要考虑的最佳谐波数量 N 。由于 STFT 频谱的奈奎斯特频率为 22050 赫兹，算法设置的最大候选基频为 1760 赫兹，因此，算法能够考虑到的最高次谐波为 12。我们分别以 1 到 12 个谐波提取谐波结构特征，并进行 12 次实验。每个实验我们都固定其他两个超参数不变，谐波和卷积层的卷积核在时间帧上的感受野 $W = 11$ ，以及每个训练样本的帧数为 64。实验结果如图??所示，左侧图为各模型在训练过程中，每经 500 次迭代在验证集上计算的 MRO 值。右侧图是训练完成后，筛选出每次实验的最佳模型，并在验证集上计算的 MRO 值。从结果图中很容易看出，当考虑谐波数量为 9 时，模型在验证集上性能最好。

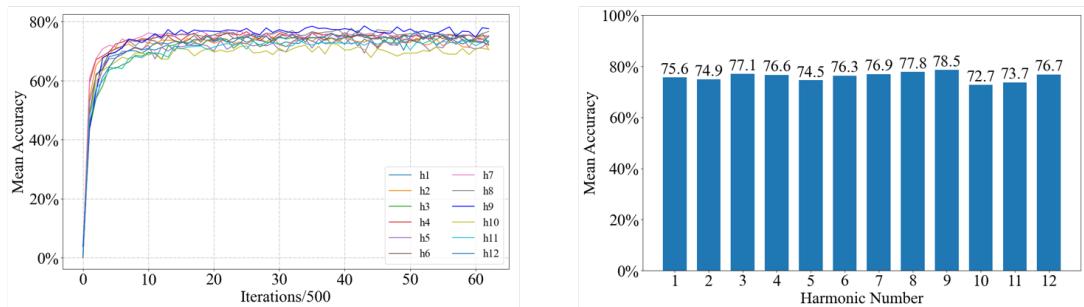


图 6-4 Mean metric with different harmonic number

在图??，我们将每次实验的最佳模型学到的权重卷积核显示出来。可以发现，当考虑的谐波数量 N 为 9 时，卷积核在当前帧的基频位置权重最大，其他 8 个谐波位置的权重也相对明显，这样能够使得基频能量在计算显著性值时仍然起着主导作用，同时各泛音能量对基频的显著性也有贡献。此外，我们发现当谐波数量为 3 和 8 时，模型性能也相对较好，但观察图??中与谐波数量为 3 和 8 对应的两个卷积核，可以看到这两个卷积核在基频位置的权重要低于其某个泛音位置的权重。这说明在计算显著性值时基频能量没有起到主导作用，但模型性能并没有很差。为了分析原因，以 ADC2004 数据集中的“daisy2.wav”为例，在图??中，我们将谐波数量为 3, 8, 9 的模型学到的显著性函数图显示出来。可以看到，谐波数量为 9 的模型，在基频和其泛音的位置能量比较强，而谐波数量为 3 和 8 的模型，由于学到的卷积核在泛音位置的权重高于基频位置的权重，因此得到的显著性图在基频及其子谐波位置的能量比较强。但是三者有个共同之处就是，在各自的旋律帧上都有很明显的谐波结构。我们初步分析的原因可能是模型最后的 Bi-LSTM 层在基于特征判别基频类别时，不仅基于基频位置的能量，而且还基于显著性函数图中基频的谐波结构。清晰的谐波结构特征更有

利于Bi-LSTM层做出正确的判别。对比这三个显著性函数图，可以看出，谐波数量为9的模型学到的权重卷积核不仅能够使得基频位置能量最显著，而且能够形成了清晰的谐波结构。

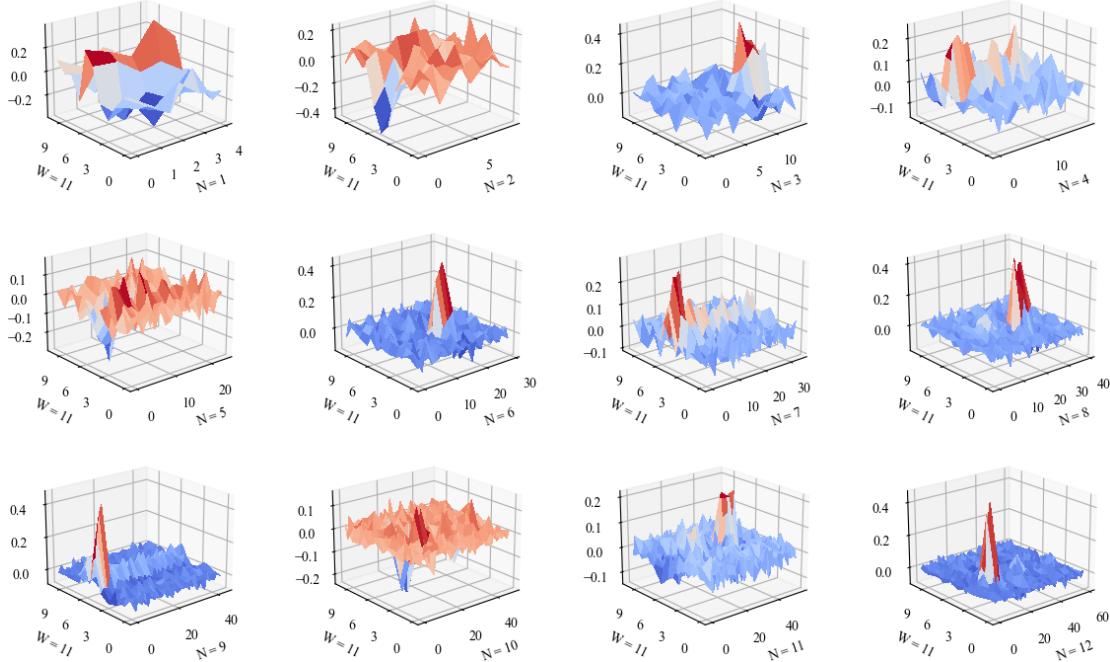


图 6-5 weights for choosing harmonic number

第二个实验是确定卷积核在时间维度的感受野 W 。我们固定谐波数量 N 为 9，分别采用 1, 3, 5, 7, 9, 11, 13, 17，进行 9 次实验。训练结果如图??所示，可以明显看出，当卷积核在时间维度的感受野为 11 帧时，模型的性能最好。同样地，我们将每次实验的最佳模型学到的卷积核显示出来，如??所示。可以发现，刚开始随着卷积核时间维度感受野的增加，性能逐步上升，而当增加到 11 帧的时候，性能达到最佳，之后便开始下降。这是符合预期的，因为在 STFT 频谱中，每一帧覆盖的时间非常短，仅 11.6 毫秒左右，因此，在计算显著性函数值的时候，距离当前帧比较近的前后帧信息可以用来补充当前帧的信息。但是如果考虑了太多前后帧信息，便会给当前帧带来很多干扰。

由上面两个实验可知，基于谐波和卷积层提取谐波结构特征时，考虑的最佳谐波数量为 9，时间维度上的感受野为 11 帧，因此，权重卷积核大小确定为 45×11 。第三个实验就是确定每个训练样本包含的帧数 L 。直观地认为，一次输入模型的帧数越长，输入模型的信息越丰富，模型性能越好。但在实验中发现以不同长度对训练集的每首音乐进行切分，会明显影响模型的性能。主要原因是在非常有限的带标注的训练集上，这个参数影响着训练样本的多样性。

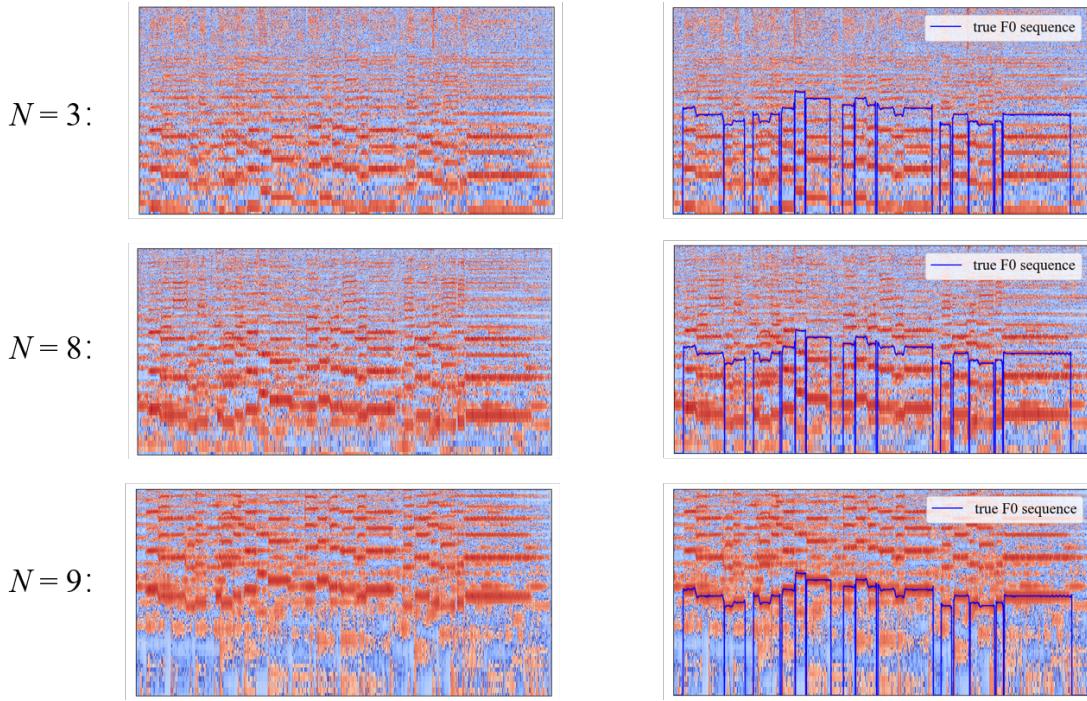


图 6-6 分别基于 3 个, 8 个, 和 9 个谐波的模型在第一层卷积之后得到的特征图

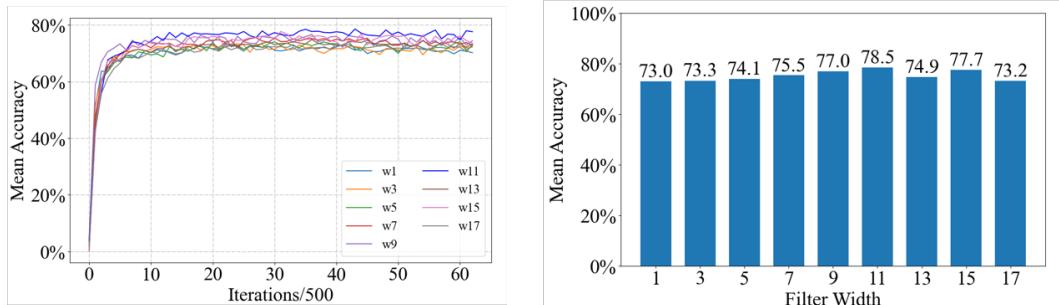


图 6-7 Mean metric with different filter width

我们分别选取 32, 64, 和 128 来切分训练集，在得到的三个训练集上分别训练 Harmonic-CRNN 模型。为了保证训练样本数量保持一致，每切分一个片段之后，均向后滑动 32 帧来切分下一个片段。实验结果显示以 32 帧, 64 帧, 和 128 帧切分训练集，模型在验证集上得到的 MRO 指标分别为 73.0%, 78.5%, 和 78.9%。虽然以 128 帧切分数据集比以 64 帧切分的结果高出 0.4%，但是一次输入模型 128 帧的计算量远远大于了 64 帧，每秒所执行的浮点运算次数 (Floating-point Operations per Second, FLOPS) 由原来的 228.36G 增加到了 456.71G。此外，实验发现在以 128 帧切分的训练集上训练，模型容易过拟合，这很可能是由于以

128 帧切分训练集得到的样本多样性低于以 64 帧切分的。因此，我们选定以 64 帧来切分样本。

6.3.5 实验结果

我们首先进行 1 组消融实验，来分析 CRNN 模块在整个模型中的作用。实验 1：将 CRNN 模块整体去除后在训练集上进行训练，即模型在经过谐波和卷积层之后，将结果直接和用于旋律定位的全连接层的结果进行拼接作为输出；实验 2：本章提出的算法。

图??显示了两次实验在验证集上得到的各项指标。我们发现实验 1 的结果非常差，特别是在旋律帧误报率指标 VFA 上，结果接近 1。这意味着模型基本没有能力判别旋律帧是否包含旋律。但是对于 RCA 指标，实验 1 的结果比较接近实验 2 的结果（64.3% vs. 76.6%），这说明实验 1 的模型能够在一定程度上估计出正确的基频值。

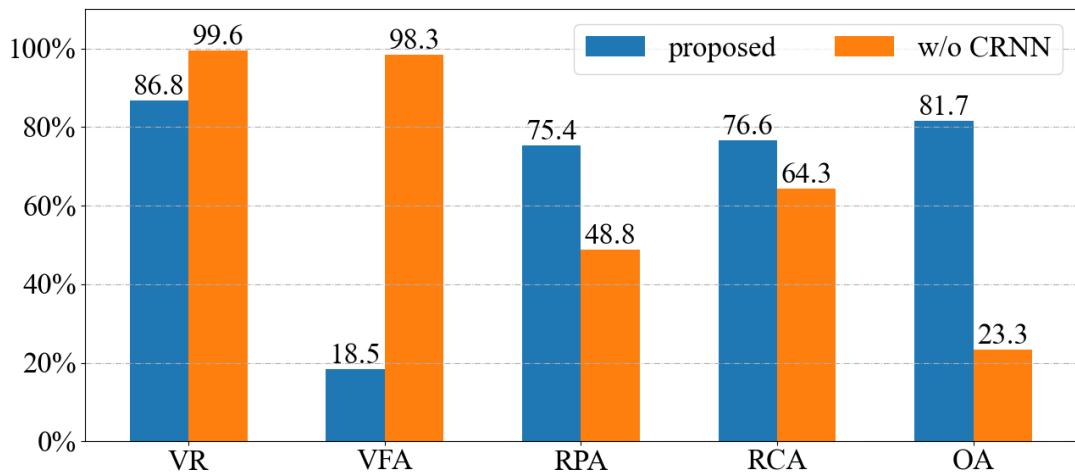


图 6-8 CRNN 的作用

为了更直观地分析，我们利用这两个模型分别对 ADC2004 中的“daisy2.wav”进行预测，预测结果如图??所示。图的左侧为不包含 CRNN 的模型预测的结果，右侧是本章算法的预测结果。我们首先在第一行，列出两个模型经过谐波和卷积层之后得到的显著性函数图。可以看出，在左侧显著性函数图中不仅在基频位置和其泛音位置比较显著，而且在其他位置也出现了有很多能量强的干扰，使得谐波结构非常模糊。在右侧显著性函数图中，基频位置及其谐波位置能量都比较强，没有太多的噪声干扰，形成了比较清晰的谐波结构。这应该归功于 CRNN 的存在。我们将第 8 层卷积之后的结果也显示出来，即图右侧的第 2 行。可以看

出经过多层 CNN 之后，基频位置及其泛音位置附近能量更加集中，谐波结构更加清晰。以此作为特征再输入 Bi-LSTM 层和全连接层进行预测，结果如图右侧的 3, 4 行所示。我们可以看到全连接层能够比较准确地定位出基频位置。这正符合我们在参数选择部分的第一个实验中的猜测，最后的 Bi-LSTM 层和全连接层在基于特征判别基频类别时，不仅是基于基频位置的能量，更多的是基于显著性函数图中基频的谐波结构。清晰的谐波结构特征更有利于 Bi-LSTM 和全连接层做出正确的判别。

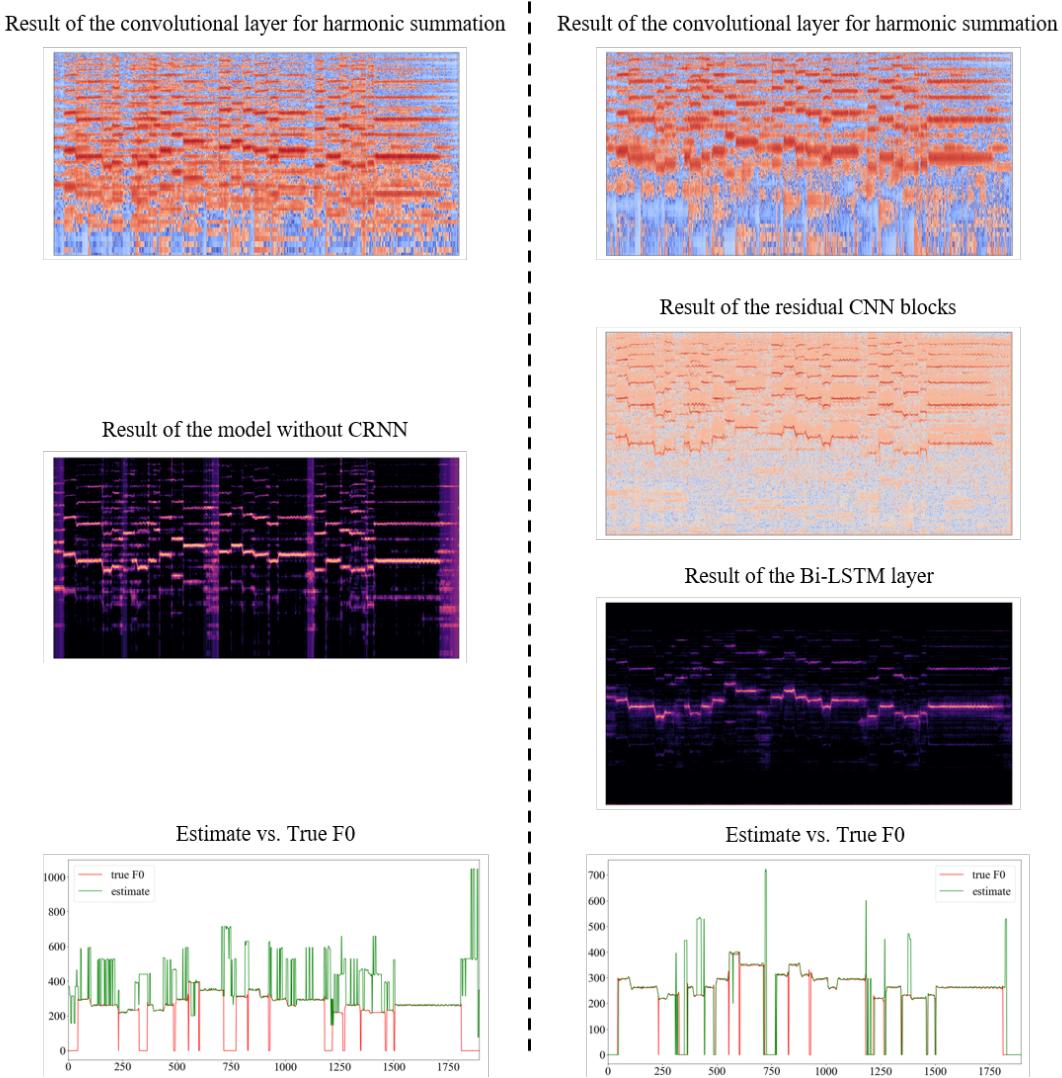


图 6-9 以“daisy2.wav”结果为例验证 CRNN 的作用。左侧为去掉 CRNN 部分的模型的结果；右侧是本章算法的结果

此外，对于旋律帧和非旋律帧的判别，我们模型是利用 1 层以谐波和卷积层得到的显著性函数图为输入的全连接层实现的。该全连接层利用显著性函数图中在旋律帧上谐波结构明显，而非旋律帧上能量分布不规律且都比较弱的特

性，对帧内是否存在旋律进行判别。基于上述两个对谐波结构的需求，在训练过程中，模型的谐波和卷积层以及 8 个卷积层都会尽力学习更清晰的谐波结构。因此，加入了 CRNN 能够使得显著性函数图中基频的谐波结构更加清晰，这不仅能够提高旋律帧定位的准确性 (VFA 由 98.3% 降低至 18.5%)，而且能够提高最后的全连接层对基频估计的准确性 (RPA 由 48.8% 提高至 75.4%，RCA 由 64.3% 提高至 76.6%)。

最后，我们在三个公开测试集上与现存算法中最具代表性的 8 种算法 Melodia^[23], DS^[85], JDC^[82], SED^[31], SSL^[?], HNN^[?], HRNet, 以及 HRNet_V2 进行对比。其中，HRNet 和 HRNet_V2 是本文第五章提出的算法，也是目前性能最好的算法，HRNet 表示仅在带正确标签的数据集上训练得到模型，HRNet_V2 表示首先经过音高细化后的 2000 首音频上进行预训练，然后在带正确标签的数据集上进行二次训练得到模型。算法 DS, JDC, SED, 和 SSL 算法均已开源，因此我们直接在测试集上采用他们的默认参数设置进行评测。而算法 HNN 没有开源，文章也没有详细给出网络结构及其参数且只在 MIREX05 集和部分 MIR1k 数据集上测试，因此，我们直接采用文章给出的 MIREX05 数据集上的结果。

表 6-2 Experiment results on ADC2004 dataset

Algorithms	VR	VFA	RPA	RCA	OA
ADC2004 (vocal)					
Melodia	81.6	12.0	71.8	74.8	73.9
DS	92.9	50.5	77.1	78.8	70.8
JDC	88.9	11.1	83.2	84.3	83.1
SED	91.1	19.3	84.7	86.3	83.7
SSL	87.4	10.0	81.5	81.7	82.1
HRNet	86.1	16.4	81.7	83.1	81.0
HRNet_V2	86.5	11.9	83.3	83.6	82.8
proposed	91.9	17.8	88.4	88.7	86.8

表??分别列出各个算法在 ADC2004, MIREX05, 以及 ikala 数据集上的结果。我们可以看出，本章提出的算法在大部分指标上超过了目前最好的结果。具体地，在 ADC2004 数据集上，算法在指标 RPA 上，RCA 和 OA 上分别超过在该数据集上表现最好的 SED 算法 3.7%, 2.4%, 和 3.1%。在 MIREX 数据集上，算法在指标 RPA 和 RCA 和经过预训练的 HRNet_V2 模型结果相当，其指标 OA 略低于 HRNet_V2 的 OA (87.8% vs. 89.1%)。由于本章提出的算法未经任何预训练，因此，我们与未经预训练的 HRNet 进行对比，结果表明在 MIREX 数据集上，算

表 6-3 Experiment results on MIREX05 dataset

	MIREX05 (vocal)				
Melodia	87.0	22.7	76.7	77.9	76.6
DS	93.6	42.8	76.3	77.3	69.6
JDC	88.2	4.2	82.6	83.2	87.6
SED	84.8	13.2	75.2	76.5	79.4
SSL	86.3	4.5	80.5	80.5	86.2
HNN	91.4	17.5	78.7	79.3	74.6
HRNet	86.8	7.5	81.4	81.9	85.6
HRNet_V2	90.2	5.9	85.8	85.9	89.1
proposed	90.1	10.0	85.9	85.9	87.8

表 6-4 Experiment results on Ikala dataset

	Ikala				
Melodia	86.0	19.0	77.3	80.4	78.0
DS	81.5	41.4	74.4	76.7	68.4
SED	80.5	9.7	76.3	76.6	80.9
HRNet_V2	85.5	10.2	83.4	83.5	85.5
Proposed	86.8	10.9	83.8	84.0	85.8

法在指标 VR, RPA, RCA 和 OA 上分别超过了 HRNet 算法 3.3%, 4.5%, 4.0%, 和 2.2%。与 JDC 算法相比，我们在指标 VR, RPA, RCA 和 OA 上分别超过了 JDC 算法 1.9%, 3.3%, 2.7%, 和 0.2%。对于 ikala 数据集，由于 JDC 和 SSL 算法的训练集包含了该数据集，因此在这个数据集上我们不与这两种算法进行对比。在 ikala 数据集上，本章提出算法和 HRNet_V2 在四个指标 VR, RPA, RCA 和 OA 上均远远超过了其他两个算法，并且本章提出的算法略高于经过预训练的算法 HRNet_V2。

特别地，本文提出的算法均大幅度超过基于谐波和理论的 *Melodia* 算法，*DS* 算法和 *HNN* 算法。从 3 个数据集上的结果可以看出，我们的结构不仅可以更准确地估计出音高值，而且可以相对准确地定位旋律帧。尽管本章算法的旋律帧误报率 VFA 指标还是比较高，但是和这三个基于谐波理论的算法对比，我们能够在保证相对高的旋律帧召回率 VR 的同时，大幅度地降低旋律帧误报率 VFA。

以上结果都表明了本文提出的 HCRNN 模型能够在各个数据集上有不错的

表，尤其在 ADC2004 数据集中，音高准确性的两个度量指标 RPA 和 RCA 均超过了 88%。我们通过聆听音频、分析频谱与对应结果发现，我们算法比较依赖基频和其泛音在 STFT 频谱上的表现，泛音结构越明显，预测结果会越准确。如果伴奏的能量比歌声强，泛音结构比较模糊，频谱看起来很杂乱，那我们算法预测的结果就会比较差。在 ADC2004 数据集中，歌声的能量都相对较强，因此结果都比较高。而在 MIREX05 数据集和 Ikala 数据集中，存在一些伴奏能量比较强的歌曲片段，因此，性能上会略低于 ADC2004 数据集。

6.4 本章总结

本章提出一个基于谐波和卷积的 CRNN 模型。方法首先提取谐波结构特征。我们采用 STFT 频谱代替 CQT 频谱。尽管通过增加八度包含的频带数量可以使得在 CQT 频谱上获得高次谐波，但是由于 CQT 在高频的频率分辨率比较低，而我们希望在定位高次谐波频率时能够定位到更准确的位置，因此，STFT 频谱最为合适。进行短时傅里叶变换时，为了使得频谱的频率分辨率尽可能地高，我们将采样率和窗长分别设置为 44100 赫兹和 8192。方法考虑的候选基频范围是 44.16 到 1760 赫兹，因此，在 STFT 频谱上最高能捕获至 12 次谐波。

受多种基于谐波和理论的主旋律提取方法的启发，在提取谐波结构特征后，我们搭建一个谐波和卷积层来计算每个候选基频对应的谐波和。其中卷积核表示每个谐波位置的权重。由于我们设置的卷积核是 2 维的，即在时间帧上也具有感受野，因此在计算每帧候选基频位置对应的谐波和时可以考虑一定范围内的上下文信息。经过这次谐波和卷积层，我们从 STFT 频谱得到一个显著性函数图。又考虑到旋律走向是基于一定的乐理，基频序列在时序上的上下文关系也很重要，因此，我们将显著性函数输入一个具有残差结构的 CRNN，希望能够得到更精确地显著性函数图。

对于旋律定位子任务，我们在谐波和卷积层之后，设计了一层以显著性函数图为输入的全连接层。原因是我们在显著性函数图中，基频及其泛音位置的能量会比较强，形成比较清晰的谐波结构，而非基频的谐波结构会比较弱，我们希望全连接层能够学习到这种差异，从而判别旋律是否存在。我们在 3 个公开测试集上对本章提出的模型进行评测。结果表明，本章提出的模型在三个测试集上均超过了目前性能最好的算法。