在智能哼唱查询系统的设计中引入全变差正则化

Shivangi Ranjan
shivangiranjan14@gmail.com 印度理工学院坎普尔分校 印度

Vishal Srivastava
vishalsrivastava177@gmail.com
印度理工学院坎普尔分校 印度

arXiv:2302.04577v1 [cs.SD] 9 Feb 2023

摘要

哼唱查询（QBH）系统是音乐信息检索中的一种特殊情况，其输入是用户哼唱的旋律，输出则是包含该旋律的原始歌曲。典型的哼唱查询系统由旋律提取和候选旋律检索两部分组成。

在旋律提取中，准确的音符转录是关键的使能技术。然而，当前的转录方法无法确切地捕捉旋律，也无法解决用户哼唱查询中的不准确问题。本文采用全变差正则化（TVR）对查询进行去噪。该方法能够处理用户哼唱中的错误，同时不丢失有意义的数据，并可靠地捕捉底层旋律。

在候选旋律检索中，我们采用基于全卷积神经网络的时间序列分类深度学习方法。训练后的网络会将输入查询分类为属于某一首目标歌曲。

在实验中，我们使用 Roger Jang 的 MIR-QBSH 数据集，该数据集是标准的 MIREX 数据集。我们证明，将 TVR 去噪查询纳入训练集，可将系统的整体准确率提升至 93%，高于其他最先进的 QBH 系统。

CCS 概念

• **音乐信息检索** → **哼唱查询**；音乐转录；全变差正则化； • **时间序列分类** → 全卷积神经网络。

关键词

数据集、神经网络、QBH、时间序列分类

1 引言

近年来，Spotify、iTunes 等基于网络的音乐获取系统相继出现。这类系统的兴起推动了各种音频数据库的快速增长。因此，我们需要开发能够适应音频数据类型特定特征的基于内容的检索方法[3]。哼唱查询就是这类基于内容的检索系统的一个实例，它可以将用户哼唱的曲调作为输入，对音频数据库进行查询。作为一种颇具吸引力的音乐检索范式，QBH 被视为音乐领域中标准文本查询系统更自然的替代方案，因为它不要求用户事先了解歌曲的相关元数据。QBH 研究主要致力于解决两个问题：a）旋律提取；b）候选旋律检索。

可以分析任何查询或候选旋律，提取出与各个时刻对应的音符序列或音高值。旋律轮廓是歌曲旋律的整体形态，以时间序列音符序列表示时，也称为音高向量。旋律轮廓是识别和编目旋律最重要的特征之一。典型的QBH系统通过有效处理用户哼唱查询中的不准确之处来运行。由于用户缺乏专业经验，且没有专业录音设备，用户哼唱的查询除了存在噪声和失真外，还可能因意外的麦克风点击声、录音不当等诸多因素而产生瑕疵，导致出现伪峰。因此，首先平滑输入的查询音高向量十分重要，这样系统便只会考虑有效的音符转换。值得注意的是，由于目标数据库中的候选旋律已有可用的标注表示（采用MIDI 格式），因此无需对这些旋律执行相同的去噪过程[16]。本文借助全变差正则化（TVR）处理含噪查询的平滑问题。我们的方法基于这样一个观察：通过全变差对查询进行去噪，比中值滤波或线性平滑等基本方法更有效，因为这些方法无法保留锐利边缘。相比之下，TVR 在平滑含噪平坦区域的同时，能够有效保留边缘信息。对输入音高向量进行TV正则化可消除伪峰。除了TVR之外，我们还会进一步对音高向量进行斜率滤波，最终得到经过降噪的分段常数音高向量。这种两步旋律提取过程能够有效应对用户哼唱查询中出现的不准确之处。

查询经过处理以去除噪声并变为分段常数后，会被纳入用于候选旋律检索的神经网络架构训练集。传统上，在目标数据库中检索候选旋律的方法是结合局部敏感哈希与动态时间规整 [9, 13]或通过序列比对，并以编辑距离衡量两首歌曲之间的相似度 [11]。然而，为了基于 MIR-QBSH 数据集设计我们的 QBH系统，我们尝试采用深度学习方法。借助全卷积神经网络（FCNs），我们对时间序列进行分类，以确定输入查询属于哪一首目标歌曲。我们发现，将 TVR去噪查询纳入训练集以训练神经网络进行时间序列分类后，能够以 93% 的准确率预测目标歌曲，这一结果优于最先进的系统。

2 相关工作

在哼唱查询领域的一项开创性研究中，Ghias 等人 [6] 于1995年采用三种音高跟踪方法——自相关、最大似然和倒谱分析，构建了一个 QBH系统，并发现改进版自相关方法的表现最佳。随后，他们将音高轮廓转换为由字母 U、D、S 组成的帕森斯编码字符串序列，分别表示当前音符高于、等于或低于前一个音符。此外，为了检索候选旋律，他们采用了字符串对齐方法。由于其模式匹配算法基于动态规划，歌曲检索耗时较长。

在其 QBH系统中，Ito 等人 [7] 旨在解决自动音乐转录导致基频估计不准确的问题。该系统采用多个基频候选值，以弥补单个频率估计中可能出现的误差；其检索算法则是一种独特的评分函数，会考虑哼唱查询与目标歌曲之间可能存在的差异，即节奏归一化和调性匹配。实验结果表明，在音符转录中考虑多个基频候选值，确实能提升检索系统的整体性能。

为了解决用户哼唱查询中常见的不准确问题，Valero-Mas 等人 [16] 分析了三种不同的旋律抽象技术，用于编码估计得到的旋律音高轮廓：符号聚合近似（SAX），这是一种用于时间序列的编码方法，既能实现降维，也能为距离计算提供下界；采用半音量化的 PAA 时间分段（PAA-ST），这是对 SAX 的改进，以固定网格半音量化取代基于统计分布的方法进行垂直量化；以及采用半音量化的音高变化分段（PC-ST），它会在音高发生变化时对旋律轮廓进行动态分段。主要结论是，尽管 SAX 已被证明在一系列时间序列问题中表现良好 [18],，但它不足以用于设计 QBH系统；改进版本 PC-ST 的性能提升便证明了这一点。

Ranjan 等人 [11] 提出了一种 QBSH 系统，采用全变分去噪的方法从查询中提取旋律。该方法基于这样一个事实：全变分去噪在降低噪声的同时，能够有效地捕捉底层信号，因为其设计目的就是保留信号的边缘信息。在候选旋律检索方面，他们采用了一种独特的方法，将生物信息学领域的半全局序列比对技术融入其中。这种方法能够将较短序列与较长序列中的子序列进行比对。他们在 MIR-QBSH 数据集上评估了该系统，系统的总体准确率达到 71%。

N. Mostafa 等人 [9]提出了一种结合卷积神经网络（CNN）和隐马尔可夫模型（HMM）的音符转录技术。他们将原始音频数据作为卷积神经网络的输入，使其直接从原始音频数据中学习相关特征向量，从而无需手动进行特征工程。通过采用隐马尔可夫模型（HMM），他们对音符转录的时间维度进行了建模。他们的 HMM 中有三个状态，分别表示音符转换：音符开始时的第一个状态、中间的稳定音符状态，以及音符衰减时的最后一个状态。每个 HMM 状态都有一个后验概率，该概率由卷积神经网络根据哼唱的原始音频样本进行建模。在候选旋律检索中，他们使用局部敏感哈希（LSH）返回最接近的近似匹配结果，随后使用动态时间规整（DTW）和推土机距离（EMD）计算这些候选结果之间的距离。总体而言，他们的 QBH系统达到了 92% 的准确率，优于提交至 MIREX 2017 的现有方法。这些实验在 MIR-QBSH 数据集上进行，该数据集是标准的 MIREX 数据集。

3 方法

3.1 数据集

为了训练我们的神经网络执行时间序列分类，我们使用了两个标准数据集：

(1) MIR-QBSH：该语料库 [8] 由用户哼唱的童谣录音组成，这些录音是在 7 年间从国立清华大学（台湾）计算机科学系“音频信号处理与识别”课程的学生中收集的。该语料库由 Roger Jang 提供，包含两部分：

- MIDI 文件：48 个单声道 MIDI 文件作为真实标注
- WAVE 文件：4431 个与这 48 首 MIDI 曲目对应的 WAVE 文件查询。每个查询时长为 8 秒，采样率为 8kHz。

(2) MTG-QBH：该语料库 [14]由 17 位不同受试者演唱的 118 段旋律组成，其中 9 位女性、8 位男性，音乐经历各异，从毫无经验到业余音乐家不等。录音时长各不相同，最短为 11 秒，最长为 98 秒，平均时长为 26.8 秒。

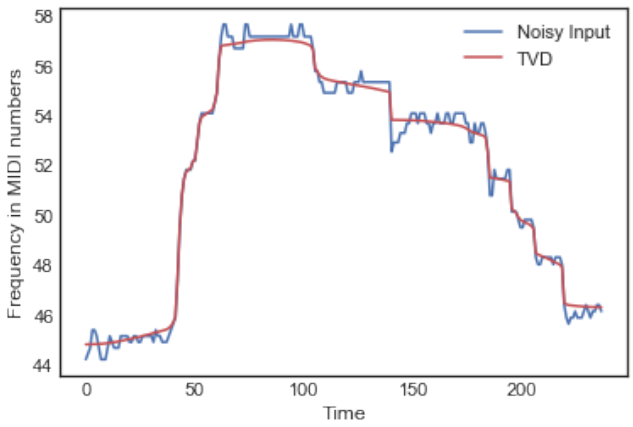

图 1：通过全变分去噪（TVD）去除噪声，同时保留信号边缘信息的示意图

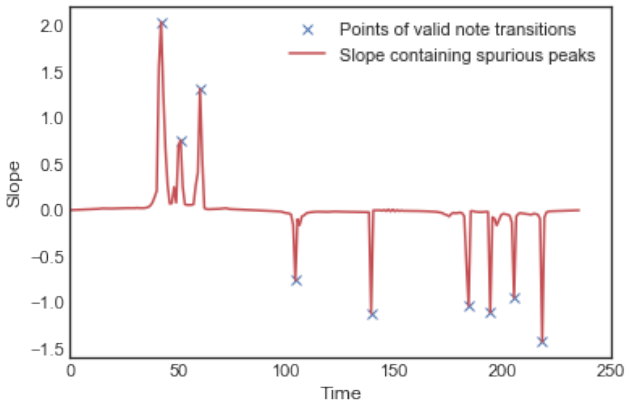

图 2：斜率滤波识别音符急剧转换点的示意图

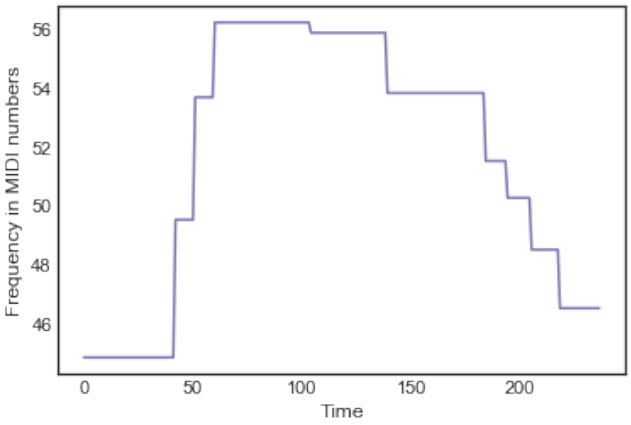

图 3：重建的旋律轮廓

3.2 旋律提取

稳健的旋律提取流程是面向音乐数据库的任何内容检索系统不可或缺的组成部分。首先，我们使用 Praat 工具中包含的实现来提取每个查询的旋律轮廓。Praat 是一款用于语音学领域语音分析的免费软件应用程序 [20]。从 Praat 提取的转录音符序列存在不准确之处，并含有用户哼唱时引入的旋律失真。因此，我们需要一种高效的技术来降低查询中的背景噪声并识别底层旋律。信号处理中常用的一种噪声降低方法是 Rudin 等人提出的全变差正则化 [12]。该方法旨在保留给定含噪信号中的边缘信息 [15]。我们采用了 [11]中所述的方法。与典型的低通滤波器不同，TV正则化被描述为一个优化问题，也就是说，它等于给定信号梯度的 L1 范数。这意味着，TV正则化的基本原理是：输入信号若包含过多且可能多余的细节，就会呈现较大的全变差全变差，即信号梯度的绝对值，其积分值会很高。根据这一原理，降低信号的全变差可以去除不需要的噪声和伪细节，同时保留信号中的边缘等必要特征。与线性平滑或中值滤波等基本方法相比，TV正则化的效果更好，因为TV正则化在去噪时能有效保留平坦区域中的边缘信息，而这些方法并非为保留锐利边缘而设计，因此会导致音符转换信息丢失。对于包含噪声和其他伪细节的图像 i，我们可以采用所提出的噪声降低模型，在二维空间中求解一个最小化问题，从而确定对应的图像 j [5, 11]:

$$min\|u\|_{TV(\Omega)} + \frac{\lambda}{2}\int_{\Omega}(f-u)^2 dx, u\epsilon BV(\Omega)$$

其中，BV(Ω) = 定义域上的全变差， λ = 惩罚项。

求解上述方程的最小化问题后，即可得到所需的去噪图像。

在所提出的QBH系统中，全变差正则化作为scikit-image多功能图像处理例程集合中的一项功能使用 [19], ，并将其扩展到了1-D时间序列数据；在本例中，该数据就是哼唱查询。因此，将输入音高向量通过全变差去噪处理，可以消除伪峰。

然而，输出结果仍未达到所需的完美分段常数形式，因此还需要经过额外的处理步骤。需要增加一个斜率滤波过程，才能将输入查询重构为分段常数信号。该滤波过程会在信号中识别出最有可能发生音符转换的时刻。通过评估和检查平滑音高向量中每个点的斜率，可以去除较小的峰值，仅保留幅度较大的斜率。这是因为只有这些斜率具有较大的幅度，才能表示急剧的音符转换。此外，如果两个峰值彼此接近或重叠，则意味着两个音符转换发生在极短的时间间隔内，而人类哼唱的查询中不可能出现这种情况。因此，将局部最大值视为有效的转换点。最后，我们得到与有效音符转换对应的斜率索引。这一新的峰值列表用于重新计算最终的旋律轮廓，使其按照预期呈分段常数形式。实现旋律提取所遵循的步骤如图1、图2和图3所示。

3.3 候选旋律检索

查询经过适当处理后，会被发送到目标数据库进行候选旋律检索。候选旋律检索是 QBH 系统的关键组成部分。本文提出了一种基于FCN的候选旋律检索系统。我们的系统以时间序列分类（TSC）[1], 为基础。时间序列分类是深度学习的一个通用领域，研究如何借助监督学习，为来自众多来源之一的输入时间序列分配标签。设计快速、准确且可扩展的时间序列分类系统至关重要，并可应用于许多不同领域，例如

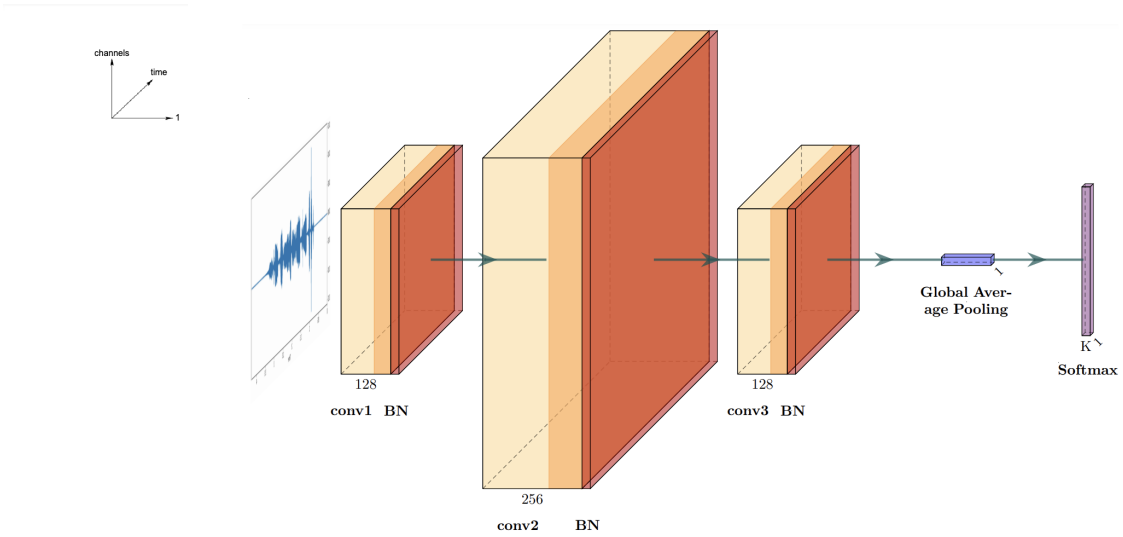

图4：全卷积神经网络架构 [17]

零售广告和云服务 [2]。在 QBH 的语境下，问题陈述的正式描述如下：

设 **(X, y)** 为数据集，其中**X** 表示查询的旋律轮廓或时间序列表示，**y** 是一个离散类别变量，由用于将每个查询归入正确目标歌曲的标签组成。**X**有 m 个观测值，**y**有 n 个可能取值。训练数据集**T**由 k 个此类训练实例构成：$\mathbf{T} = \{(X_1, y_1) ..., (X_k, y_k)\}$。

时间序列数据分类问题包括在**T** 上学习一个分类器，将输入向量**X** 映射为关于**y**的概率分布。

Wang等人 [17] 首次提出全卷积神经网络（FCN），用于时间序列数据分类。FCN不包含局部池化层，因此时间序列的长度在所有卷积操作中保持不变。FCN的另一个特点是使用全局平均池化（GAP）层取代传统的最终全连接层，从而显著减少网络中的参数数量。研究表明，FCN能够有效地进行时间序列分类，适用于文档分类和语音识别等领域 [4]。我们的系统采用FCN将输入查询分类为目标歌曲之一。系统架构如图4所示。

所提出的架构由三个卷积块、一个GAP层和一个传统的Softmax分类器层组成。每个卷积块执行三项操作，即先进行卷积，然后进行批归一化，并将结果输入ReLU激活函数。三个卷积层分别包含128、256和128个滤波器，滤波器长度分别为8、5和3。为保持时间序列的长度，每个卷积均采用零填充，步幅为1。最后一个卷积块的输出进入GAP层，在整个时间域上进行平均。最后，Softmax分类器生成所需的输出。

4 实验

4.1 数据预处理

为了训练用于候选旋律检索的全卷积网络，我们主要使用了 MIR-QBSH 数据集。此外，我们还将MTG-QBH数据集添加到 MIR-QBSH 数据集中，以构成训练集。这不仅增加了训练集中的数据点数量和多样性，也提高了网络训练的鲁棒性。MIR-QBSH 数据集包含4431个查询和48首目标歌曲，而 MTG-QBH数据集包含118个查询和81首目标歌曲。为了提高训练后网络的准确率并保持数据库内部的一致性，每个查询都被划分为等长帧，窗口大小为 = 5 秒，跳跃大小为 = 2 秒。

4.2 基线系统

我们设计了一个快速且可扩展的基线系统，以展示所提出的检索方法在并列比较中的效率。基线系统的架构是一个简单的多层感知机（MLP），由 5 层组成，每层均与前一层全连接 [10]。网络的最后一层是一个 softmax 分类器，与其前一层的输出全连接。每个隐藏层包含 300 个神经元，激活函数为 ReLU。

4.3 所提出的 QBH 系统的评估

我们使用了第 3.1 节中提到的数据集的两个版本进行了实验。首先，网络仅使用原始含噪数据集进行训练，未加入总变差正则化。随后，我们将去噪后的查询添加到训练集中，使检索准确率从 67% 大幅提升至 93%。

训练全卷积网络时，初始学习率为 0.005，在前 30 个训练轮次中保持不变，此后学习率

表 1：所提出的 QBH 系统与其他最先进系统的评估

算法	准确率
采用全变差正则化+ 全卷积网络（所提出的方法）	0.93
未采用全变差正则化+ 全卷积网络	0.67
采用全变差正则化+多层感知机（基线）	0.78
Mostafa 等人	0.92
BS1	0.86
WHLX1	0.47
TYCX4	0.93
ZH1	0.89

每个训练轮次降低 0.7，直至交叉验证准确率趋于稳定。网络使用数据集的 25% 作为验证集。我们证明，在数据集中通过TV正则化纳入去噪查询后，对于任何给定查询，所提出的全卷积网络系统都能以 93% 的准确率返回正确的目标歌曲，高于 Mostafa 等人提出的当前最先进系统 [9]，也高于所有其他提交至 MIREX 的系统。我们的系统准确率与性能强大的 TYCX4 相同。不过需要注意，TYCX4 是基于帧的哼唱查询系统，其计算开销远高于我们的基于音符的系统。评估结果列于表 1。

5 讨论与未来方向

可以看出，先使用 TV正则化对输入信号进行去噪，再使用全卷积神经网络对处理后的查询进行时间序列分类，这种组合方法取得了最高的准确率，达到 93%。因此，我们得出结论：所提出的系统优于其他最先进的系统。当数千种可能的输入查询映射到一小组目标歌曲时，这种方法尤为有效。在这些情况下，采用全卷积神经网络对时间序列数据进行分类，是一种能够迅速提升系统可扩展性的高效方案，同时还能提高准确率。尽管性能有所提升，所提出的系统仍有改进空间。基于音符的哼唱检索系统涉及的特征工程相对简单，且速度更快，因此比基于帧的哼唱检索系统更高效；而基于帧的系统通常计算开销更大。不过，逐帧分析输入查询和目标歌曲能够提供丰富得多的信息，进而提高检索准确率。还可以通过考虑更多特征来进一步提升准确率，例如每个音符哼唱时的力度，以及对特征工程进行其他改进。

尽管众多学者投身于哼唱查询领域，该领域仍在积极发展中。旋律提取和候选旋律检索这两个组成部分都还有许多方法有待测试。

（1）一项重要工作是设计一种系统，能够从原始音频数据中进行自动音乐转录，并提供多个基频候选值。这种方法有双重优势：通过提供多个候选值来吸收单一频率识别中的误差，同时也有助于设计端到端哼唱检索系统。

（2）通过在频域中可视化歌曲，可以获得关键洞见并进一步提高准确率。例如，众所周知，色度特征在音频翻唱歌曲识别任务中表现出色。然而，它们在 QBH系统中的应用仍有待探索。

(3) 对歌曲进行谱分析，不仅能够提供时间序列表示中原本会丢失的信息，还能让开发人员利用计算机视觉研究中涌现的强大模式识别方法，并将其应用于音乐领域。这种跨学科方法有望拉近这两个领域的距离。

6 结论

在本研究中，我们提出了一种高效、智能的哼唱查询系统，并在其设计中融入全变差正则化，从而提升了候选旋律检索的准确率和性能。针对后者，我们的系统采用了全卷积神经网络，该网络已被证明能够有效执行时间序列分类。因此，我们的系统实现了 93% 的准确率，优于其他最先进的系统。值得一提的是，全卷积网络的一个显著特点是，对于长度可变的时间序列，除最后一层外，各层所使用的参数数量不变。因此，该网络支持迁移学习：可以先使用某个数据集训练模型，再利用不同的目标数据集对其进一步优化。这样便无需大量新数据，并能提升整个神经网络的性能。

参考文献

[1] Witali Aswolinskiy、René Felix Reinhart 和 Jochen Steil。2018 年。储备池空间与模型空间中的时间序列分类。《Neural Processing Letters》。第 48 卷。789–809。[2] Anthony Bagnall、Jason Lines、Aaron Bostrom、James Large 和 Eamonn Keogh。2017 年。近期算法进展的大型时间序列分类评测：综述与实验评估。《Data Mining and Knowledge Discovery》。第 31 卷。606–660。[3] Stephen Downie。2003 年。音乐信息检索。《Annual Review of Information Science and Technology》。第 37 卷。295–340。[4] Hassan Ismail Fawaz、Germain Forestier、Jonathan Weber、Lhassane Idoumghar 和 Pierre-Alain Muller。2019 年。用于时间序列分类的深度学习：综述。《Data Mining and Knowledge Discovery》。第 33 卷。917–963。[5] Pascal Getreuer。2012 年。使用 *Split Bregman* 的 *Rudin-Osher-Fatemi* 全变分去噪。《Image Processing On Line》。第 2 卷。74–95。https://doi.org/10.5201/ipol.2012.g-tvd [6] Asif Ghias、Jonathan Logan、David Chamberlin 和 Brian C. Smith。2002 年。音频数据库中的哼唱查询：音乐信息检索。*2002* 年自然语言处理经验方法会议论文集。231–236。[7] Akinori Ito、Yu Kosugi、Shozo Makino 和 Masashi Ito。2010 年。一种基于多个基频候选值、从音频信号中进行哼唱查询的音乐信息检索方法。*2010* 年国际音频、语言与图像处理会议。1 5 10 1109 2010 5685029 – . . /ICALIP。. Jyh-Shing Jang、[8] Nien-Jung Lee 和 Chao-Ling Hsu。2006 年。用于 MIREX 2006 的简单但有效的 QBSH 方法。[9] Naziba Mostafa 和 Pascale Fung。2017 年。一种使用卷积神经网络的基于音符的哼唱查询系统。*Interspeech* 会议论文集。3102–3106 10 21437 2017 1590 . . /Interspeech。- Tuyen Pham、Gi Nam、[10] Yong Kwang、Kang Shin、Park 和 Ryoung Kang。2017 年。一种通过估计匹配位置实现的新型歌唱/哼唱检索方法

基于多层感知机。KSII《互联网与信息系统汇刊》。第 7 3102 3106 10 3837 2013 07 008 卷。- . . /tiis。 . . Shivangi Ranjan 和 Vipul Arora。. 一种用于哼唱查询的半全局比对生物信息学方法。*2020* 年 *IEEE* 第四届信息通信技术会议（*CICT*）。1 5 10 1109 51604 2020 9312085 – . . /CICT . . [11] 2020

Leonid Rudin、Stanley Osher 和 Emad Fatemi。. 基于非线性全变差的噪声去除算法。《物理学D：非线性现象》。第 60 卷，第 1–4 期。259–268。[13] Matti Ryynänen 和 Anssi Klapuri。2008 年。使用局部敏感哈希对 MIDI 和音频进行哼唱查询。《2008 年 IEEE 国际声学、语音与信号处理会议论文集》。2249–2252。拉斯维加斯，内华达州，美国。[14] Justin Salamon、Joan Serra 和 Emilia Gómez。2012 年。音乐检索的音调表示：从版本识别到哼唱查询。《国际多媒体信息检索杂志》，混合音乐信息检索特刊。[12] 1992

[15] Ivan Selesnick. 2012. 全变分去噪（*TVD*），纽约大学坦顿工程学院。[16] Jose Valero-Mas、Justin Salamon 和 EmiliaGómez. 2015. 在基于音频的哼唱查询背景下，分析音高量化和音符分割对歌声对齐的影响。第 *12* 届声音与音乐计算会议论文集。[17] Zhiguang Wang、Weizhong Yan、Tim Oates. 2017b. 使用深度神经网络从头开始进行时间序列分类：一个强有力的基准。国际神经网络联合会议。1578 1585 1611 06455 4 – 。arXiv: 。v Li Wei、Jessice Lin、Eamonn Keogh 和 Stefano Lonardi. 。ExperiencingSAX：一种新颖的时间序列符号表示。*Data Mining and Knowledge Discovery*。卷。15, 107 144 10 1007 10618 007 0064 2021 04 18 – 。[18] 2007

https://doi.org/ 。/s - - -z - - https://scikit-image.org/docs/dev/api/skimage.restoration [19]

https://www.fon.hum.uva.nl/praat/ [20]