通过去噪预处理器和混合估计模型进行音高估计

洪裕誠國立臺灣大學
通信工程學系臺灣臺北
r
09942162@ntu.edu.tw

Ping-Hung Chen通信工程学系，国立台湾大学台湾台北
r07942055@ntu.edu.tw

丁建君
通信工程
国立台湾大学
中国台湾台北
jjding@ntu.edu.tw

摘要— **音高估计旨在估算基频和 MIDI 音符编号，在音乐信号分析和人声信号处理中发挥着关键作用。在这项工作中，我们提出了一种新架构，该架构基于学习式增强预处理器，并结合多种传统音高估计方法和深度学习音高估计方法，以便在噪声和干净场景下都能取得更好的音高估计性能。我们测试了 17 种不同类型的噪声和 4 SNR_{db} 种噪声级别。结果表明，所提出的音高估计方法在噪声和干净场景下均表现更佳，且响应时间较短。**

关键词— 音乐信号处理、人声信号增强、基频、音高估计

I. 引言

一般而言，音高估计的流程如下：对输入音频进行短时分析，或分析音符时长。1）通过起音检测获取音符时长。2）使用不同的方法计算每一帧的音高。3）获取若干个较高的峰值，这些峰值包含基频（f_0）及其谐波，以及位于f_0 之前的一个小峰值，即次谐波。

谐波和次谐波表示为

$$n \times f \text{ and } \frac{1^{th}}{n} \times f \qquad (1)$$

其中，f 是f_0 ，n 是整数。在干净信号场景下，大多数时域方法，如自相关函数（ACF）和归一化动态频谱特征（NDSFs）[1]，以及频域方法，如倒谱 [6] 和谐波乘积谱（HPS）[2]，都能很好地发挥作用。

NSDF 和 ACF [1] 的主要概念是原始帧 $x(t), t = 0, \ldots n-1$ 与延迟 $x(t-\tau)$的帧 $x(t)$ 移位后重叠部分的内积。τ 是原始信号的时延。

HPS [2] 用于计算音频频谱与其压缩频谱的乘积。HPS 的特性是，每个压缩频谱都会在其f_0附近出现一个峰值，因此所有频谱的乘积会使f_0 的值更高。短时傅里叶变换（STFT）[3] 在音频和图像处理领域很常见，因为 STFT 会输出如图 1 所示的频谱图。如果沿着转置频谱图的频率轴对能量求和，就可以获得单音在频率上的能量分布。最大似然（ML）方法是在f_0及其 4 个八度音上使用冲激函数。然后，将冲激函数与一个音符的频谱进行卷积，并对所得值求和，以获得该音符的基频f_0 。

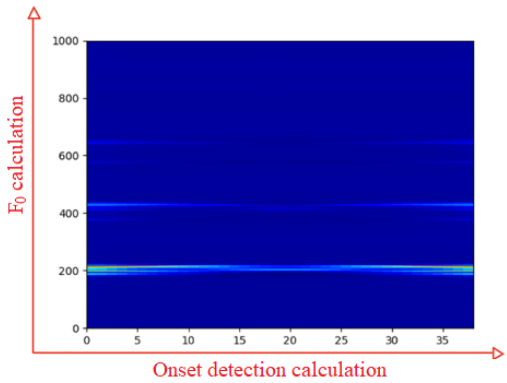

图 1. 用于音高估计和起音检测的频谱图

表 I. 不同方法的比较

方法s	公式
HPS	$HPS(f) = \prod_{i=1}^{n} \lvert X(nf) \rvert$, $f_{f0} = \max_f(HPS(f))$
STFT	$E(f) = \sum_t X(t,f)$ ' $f_{f0} = \max_f(E(f))$
Max 似然	$f_{f0} = \max_f \sum_k X(f)\delta(f-nk)$, $n \in N$
倒谱	$C(n) = real(\int_{-\frac{1}{2}}^{\frac{1}{2}} \text{对数} \lvert X(f) \rvert \, e^{j2\pi fn} df)$, $f_{f0} = \frac{fs}{\max_n Cep(n)}$
SRH	$SRH[f] = E[f] + \sum_{k=2}^{N_{harm}} [E[kf] - E[(k-\frac{1}{2})f]]$ $f_{f0} = \max_f(SRH[f])$

残余谐波求和（SRH）[4] 基于语音信号。在语音分析中，语音可以解卷积为激励信号和共振峰结构。该方法使用线性预测分析来估计激励信号。深度学习（DL）方法 Crepe 方法[5] 是数据驱动的，训练的计算成本较高，因此我们仅使用预训练模型来改进基频分析。基频方法研究的概览见表 I。

II. 所提出的方法音高估计

在这项工作中，我们在低信噪比场景下提出了一种稳健且响应时间合理的方法，该方法结合了多种传统方法和一种深度学习方法。这里，我们假设起始位置已知，并测试不同方法在干净音频、不同含噪音频下的表现以及响应时间。

每种方法都有其优点和缺点。机器学习和短时傅里叶变换方法容易受到低频噪声的影响，例如办公室声、火车站和

表 II. 不同音高估计方法的不同设置

方法	Min 频率	Max 频率	谐波数目
HPS	0	样本长度 / 2	3
STFT	20	1000	4
最大似然	20	800	5
SRH	80	500	5
CREPE	33	3951	-

通风声。我们运用机器学习集成的理念，结合多种在干净音频和含噪音频中音高误差较小、且响应时间在 2 秒以内的方法，以获得稳健的f_0 估计。

我们的想法是找出以下特性，并将额外的算法应用于这些方法。1. 在干净音频和含噪音频中，音高误差都很小。2 5、0、10、20 17 30。含噪音频包括信噪比为 - 的情况，以及不同类型的噪声。信噪比几乎接近真实值，而 -10 则无法听见。3. 每首歌曲的平均响应时间较短，在 2 秒以内，歌曲时长为 10 ~ 20s。我们取 HPS、短时傅里叶变换、机器学习、SRH 和 CREPE 的中位数，并采用微型预训练模型。我们为不同方法设置不同的频率范围，以避免 1000 Hz 以上的高频噪声，如表 II 所示。未来的工作可以进一步研究不同的频率范围。在本研究中，我们仅凭经验设定最低频率、最高频率和谐波数，然后将所有方法集成起来。

VI. 仿真结果

对于噪声数据，我们使用 UBC 噪声数据集，其中包含 17 种不同的噪声，涵盖了人们能够想到的 QBH 使用场景，噪声级别为 [-5, 0, 10, 20]。我们实验室的歌曲数据包含约 209 首歌曲。每首歌曲都需要考虑 68 种场景。表 III 概述了表 IV 和表 V 的内容，数值为噪声 1 至 17 的平均值。

我们提出的方法在 14 种场景中优于其他方法，在 3 种场景中表现相当。音高估计测量通过计算音高差来进行，具体如下：

$$error = \frac{1}{N}\sum_{n=1}^{N}\sqrt{|f_{method(n)} - f_{groundtruth(n)}|}, \quad (2)$$

其中，n 是歌曲中的音符总数。

在实验中，我们对表 IV 和表 V 中描述的所有方法进行了概览比较，并采用式 (2) 中的误差函数；实验涵盖 17 种不同的噪声，包括 1. 白噪声、2. 人群嘈杂声、3. 昆虫声、4. 海浪声、5. 地铁声、6. 校园声、7. 通风声、8. 汽车声、9. 火车声、10. 温室声、11. 展览馆声、12. 高斯噪声、13. 荒野声、14. 餐厅声、15. 机场声、16. 街道声、17. 办公室声。与表 IV 和表 V 中的结果一致，我们提出的方法在 13 种场景中优于其他方法，在 4 种场景中与其他方法表现相当。测量值为噪声级别 [-5 ~ 20] 下的平均误差。

五、结论

本文结合传统方法与深度学习方法，提供了新的见解。通过将不同方法的结果取中位数，我们可以避免大部分噪声对音高估计造成的影响，并获得稳健的音高估计。未来，我们可以进一步研究不同频率设置与不同方法之间的关系。

表 III。不同算法的不同测量结果。

	SRH	ML	HPS	STFT	Crepe	Ours
干净音频错误	1.6	2.34	2.05	2.38	2.57	**1.48**
含噪音频错误	2.59	3.6	3.12	4.07	3.13	**2.46**

表 IV. 噪声 1~9 SNR_{db} = [5, 0, 10, 20] 下的音高估计误差，采用 - .

方法 \ 噪声	1	2	3	4	5	6	7	8	9
SRH	2.06	2.93	1.79	3.03	2.61	2.49	3.26	2.83	2.74
YIN	8.06	7.79	5.32	8.05	7.81	8.06	8.03	8.15	8.41
NSDF	2.67	3.42	4.31	4.12	3.23	6.76	5.57	6.58	7.95
HPS	2.09	2.69	2.06	2.84	2.08	4.35	4.16	3.82	5.09
ML	2.51	3.29	2.86	3.76	2.83	4.87	4.49	4.32	4.99
STFT	2.63	3.45	2.39	3.45	2.90	6.15	5.11	4.90	6.43
CREPE	2.85	3.34	2.89	3.60	3.05	2.85	3.16	3.07	2.99
倒谱	12.95	13.56	13.50	13.43	13.64	13.70	13.55	13.33	13.51
所提出的方法	**1.59**	**2.19**	**1.50**	**2.34**	**1.69**	**3.68**	**3.15**	**3.15**	**4.07**

表 V. 噪声 10~17 SNR_{db} = [5, 0, 10, 20] 下的音高估计误差 - 。

方法 \ 噪声	10	11	12	13	14	15	16	17
SRH	2.48	2.67	2.04	2.97	2.96	3.45	3.00	2.28
YIN	8.06	7.95	7.49	8.02	7.78	8.10	7.69	7.89
NSDF	3.83	3.40	2.82	4.25	3.88	5.80	4.48	8.88
HPS	2.53	2.08	2.09	2.87	2.38	3.82	2.83	5.34
ML	3.08	2.52	2.50	3.75	3.07	3.93	3.82	4.55
STFT	3.48	3.09	2.68	3.46	3.22	4.79	4.39	6.69
CREPE	3.13	2.88	2.84	3.41	3.32	3.41	3.59	2.93
倒谱	13.18	13.37	13.15	13.61	13.57	13.57	13.52	13.65
所提出的方法	**1.94**	**1.62**	**1.57**	**2.35**	**2.**			

参考文献

[1] O. Babacan、T. Drugman、N. d'Alessandro、N. Henrich 和 T. Dutoit，《多种歌唱声音上的音高提取算法比较研究》，发表于*IEEE*国际声学、语音与信号处理会议，第7815–7819页，2013年。[2] P. De La Cuadra、A. S. Master 和 C. Sapp，《面向交互式音乐的高效音高检测技术》，发表于*ICMC*，2001年。[3] V. Pannala、G. Aneeja、S. R. Kadiri 和 B. Yegnanarayana，《使用单频滤波方法稳健估计基频》，发表于 Interspeech，第2155–2159页，2016年。[4] T. Drugman 和 A. Alwan.《基于残余谐波的联合稳健浊音检测与音高估计》。arXiv 预印本 arXiv:2001 00459，2019年。.[5] J. W. Kim、J. Salamon、P. Li 和 J. P. Bello，《Crepe：用于音高估计的卷积表示》，发表于*IEEE*国际声学、语音与信号处理会议，第161–165页，2018年。[6] B. Qian 和 Y. P. Li，《基于 LPC 残差倒谱的音高检测研究》，计算机工程与应用，第210–213页，2007年。