关于自动歌唱评估与歌唱信息处理 30+ 余年的综述

预印本

Arthur N. dos Santos
电气与计算机工程学院 坎皮纳斯州立大学 坎皮纳斯，巴西 arnics@unicamp.br

Bruno S. Masiero
电气与计算机工程学院 坎皮纳斯州立大学 坎皮纳斯，巴西 masiero@unicamp.br

arXiv:2601.12153v1 [eess.AS] 17 Jan 2026

摘要

过去三十年来，自动歌唱评估与歌唱信息处理不断发展，为歌唱教学法、表演分析和声乐训练提供支持。前者通过计算指标对歌者的表现进行客观评估，涵盖实时视觉反馈、声学生物反馈，以及先进的音高跟踪和频谱分析；后者则将预测人声信号与目标参考进行比较，以捕捉歌唱声音中蕴含的细微信息。重要进展包括交互系统的开发，大幅改善了实时视觉反馈；以及机器学习和深度神经网络架构的应用，提升了语音信号处理的精度。本文通过批判性考察相关文献，梳理这些技术的历史演变，并指出和探讨其中的关键空白。分析表明，仍存在一些长期挑战，例如缺乏标准化评估框架、难以从各种噪声源中可靠地分离人声信号，以及未能充分利用先进的数字信号处理和人工智能方法来捕捉艺术表现力。本文详细阐述这些局限及相应的技术进展，说明解决这些问题如何弥合客观计算评估与近似人类判断的主观歌唱表现评估之间的差距，最终提升自动歌唱评估系统的技术准确性和教学适用性。

关键词 自动歌唱评估 · 实时视觉反馈 · 声学生物反馈 · 歌唱信息处理 · 卡拉 OK 评分

1 引言

歌唱评估根据技术标准和表现力标准，对歌者的表现进行评估。传统上，这项工作由人工评估者完成；如今，计算方法能够实现客观且便捷的歌唱表现评估。此外，这些方法还通过显示屏提供实时视觉反馈（VFB），引导歌者调整其表现。采用生物传感器的系统则通过采集生理信号并将其转化为视觉提示，进一步提升这一过程，使歌者能够感知其发声器官中原本无法察觉的某些方面 [1, 2]。

尽管具有这些优势，自动歌唱评估（ASA）仍主要依赖基于元数据的分析，例如提取并比较音高值，而非处理原始人声信号。这些系统既不会增强或分离录制音频中的歌声，也无法支持与参考人声轨道直接比较，例如与商业发行歌曲中的原始人声进行比较。因此，歌唱信息处理（SIP） [3] 的兴起标志着一种范式转变：它将歌唱视为数据，通过分析实际的人声信号，而非仅仅依赖提取参数来处理歌唱。这一转变带来了重大挑战，尤其是人声分离方面，算法必须准确分离出歌者的声音，同时尽量减少伪影。诸如将伴奏误分类为人声、多个

人声源重叠，以及原始音轨中的残余干扰等错误，都可能显著影响分析准确性 [4, 5]。

本综述采用结构化框架分析自动歌唱评估和歌唱信息处理的发展。历史分析聚焦于1988–2010年。在此期间，Welch等 [6, 7, 8, 9, 10, 11], Rossiter等 [12, 13, 14, 15], Howard等 [16, 17, 18, 19, 18, 20, 21, 22, 23], 以及Goto等 [24, 25, 3] 的开创性贡献，几乎占据了该主题全部文献的40%。他们的研究至关重要，因为在严峻的技术限制下，这些研究奠定了相关领域早期的方法论和技术基础。这些开创性研究不仅界定了这一时期的研究范围，也为后续发展铺平了道路，使最初的方法得以扩展到更广泛的应用领域，并融入当代技术进步。

随后，本文以更广泛的视角回顾截至2022年发表的其余文献。本阶段并非进行深入的历史分析，而是通过考察预期应用、特征提取方法、目标变量类型和评估方法等关键方面，着重识别发展趋势。这种研究方法的转变优先关注新兴模式的识别，而非其历史意义，从而有助于发现这些领域中可能存在的空白及未来研究方向。
本文其余部分结构如下：第2节介绍所采用的方法，包括时间范围、关键词、信息来源、排除标准及指导性研究问题。第3节进行历史分析，第4节考察自动歌唱评估和歌唱信息处理领域的当前研究状况。第5节讨论这些领域的新兴趋势及潜在研究空白，第6节则总结全文。

2 方法论

本综述考察自动歌唱评估和歌唱信息处理系统从创立到该领域最新进展的发展历程。综述时间跨度从20世纪80年代末（即有记录以来最早的自动歌唱评估系统出现之时）至21世纪20年代初，涵盖本分析中发现的最新研究。

本研究采用基于关键词的检索策略进行了全面的文献综述，使用的检索词包括：“自动歌唱评估”“实时视觉反馈”“声学生物反馈”“歌唱信息处理”和“卡拉OK评分”。这些关键词有助于从 Google Scholar[1], Scopus[2], ResearchGate[3], 和 Academia[4]等数据库中检索相关学术成果，包括期刊论文、会议论文集和高等学位论文。尽管 ResearchGate 和 Academia 收录了一些未经同行评审的材料，最终综述仅纳入经过严格同行评审的出版物。

所综述的文献涵盖多个学科领域，包括音乐教育、心理学、声学、语音通信和计算学习。具有代表性的期刊包括 *Canadian Journal of Music Education*、*Psychology of Music*[5], *Music Education Research*[6], *Applied Acoustics*[7], *The Journal of the Acoustical Society of America*[8], *Speech Communication*[9], 以及*IEEE Transactions on Audio, Speech, and Language Processing*[10]。

此外，还考察了跨学科及计算领域的来源，包括 *Journal of Computer Assisted Learning*[11], *Logopedics Phoniatrics Vocology*[12] *Journal of Voice*[13],以及 *Journal of Music, Technology & Education (JMTE)*[14]。

[1]https://scholar.google.com/[2]https://www.scopus.com/home.uri[3] https://www.researchgate.net/[4]https://www.academia.edu/[5] https://us.sagepub.com/en-us/nam/psychology-of-music/journal201640[6] https://www.tandfonline.com/journals/cmue20[7] https://www.sciencedirect.com/journal/applied-acoustics[8] https://pubs.aip.org/asa/jasa[9] https://www.sciencedirect.com/journal/speech-communication[10] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=10376[11] https://onlinelibrary.wiley.com/journal/13652729[12] https://www.tandfonline.com/journals/ilog20[13]https://voicefoundation.org/[14] https://www.atmimusic.com/journal-of-music-technology-and-education/

此外，还纳入了与声学、语音处理、音乐技术和人工智能相关的重要会议论文集，包括 声学学会会议论文集[15], 国际声学、语音与信号处理会议 *(ICASSP)*[16], *INTERSPEECH*[17], 音频工程学会 *(AES)* 大会[18], 以及国际音乐信息检索学会会议 *(ISMIR)*[19]。

此外，本研究还回顾了以下多学科会议：国际计算机音乐多学科研究研讨会（*CMMR*）[20], 声音与音乐计算会议（*SMC*）[21], *IEEE*视觉通信与图像处理会议（*VCIP*）[22], 以及国际音乐表达新界面会议（*NIME*）[23] 。

本研究还考察了来自以推动音乐技术和声学发展而闻名的机构的理学硕士和哲学博士学位论文，例如约克大学[24] 和皇家理工学院（*KTH*）[25]。
尽管卡拉OK评分系统与本综述相关，但相关资料主要记载于专利中，而这些专利往往缺乏理论基础、全面的评估方法或实验验证 [26]。因此，专利文献不纳入分析。

本研究基于以下关键研究问题对所选语料进行分析：

1. 预期的应用是什么？ 2. 对预测信号采用了哪些特征提取方法？
3. 目标变量以数据还是元数据的形式表示？ 4. 评估是如何进行的（客观还是主观）？

3 历史分析

最初，自动歌唱评估（ASA）是主要方法，因为在当时的技术条件下，捕捉歌手的声音、提取有意义的特征并将其与参考值进行比较，已经是一项颇具挑战性的任务。然而，随着技术进步，歌唱信息处理（SIP）应运而生，标志着范式转变。与依赖基于元数据的比较的自动歌唱评估不同，歌唱信息处理会将人声预测信号直接与人声目标信号进行比较，将歌唱视为数据，而非派生参数。本节将探讨这两个相互交织领域的历史起源。

3.1 自动歌唱评估（ASA）

自动歌唱评估领域通常被认为起源于四个奠基性系统，每个系统都代表了前代系统的发展。这些系统是在作者密切合作、共享相互重叠的方法的基础上开发而成，从而形成了一个相互关联的框架。以下小节将详细介绍这些开创性的自动歌唱评估系统。

3.1.1 歌唱评估与发展（SINGAD）系统

SINGAD（歌唱评估与发展）系统被广泛认为是最早的自动歌唱评估软件， [6]旨在通过提供人声音高的实时视觉反馈来支持小学音乐教育。该系统采用了改造自人工耳蜗的峰值拾取设备，使儿童能够在显示器上将自己的基频（f_0）与显示的目标音高进行直观对照 [7]。

实证研究 [8, 9] 表明，即使没有成人监督，使用 SINGAD 的学生在歌唱能力方面的提升也比采用传统师生教学方法的学生更为显著。该系统后来

[15]https://www.ioa.org.uk/conference-series/institute-acoustics-proceedings[16]
https://ieeexplore.ieee.org/xpl/conhome/1000002/all-proceedings[17]
https://www.isca-speech.org/[18]https://aes2.org/event/conventions/[19]
https://www.ismir.net/[20]https://link.springer.com/conference/cmmr[21]
https://smcnetwork.org/[22]
https://ieeexplore.ieee.org/xpl/conhome/1800602/all-proceedings[23]
https://nime.org/[24]https://www.york.ac.uk/[25]https://www.kth.se/en

升级为支持乐器数字接口（MIDI）的音频播放，并扩展了功能 [10, 11]。

3.1.2 实时声学与喉部生物反馈增强（ALBERT）

SINGAD 仅能提供 f_0上的实时视觉反馈，而实时声学与喉部生物反馈增强（ALBERT）软件则在其基础上进行了扩展，纳入了更广泛的发声参数。该软件基于这样一个假设而开发：有效的声音反馈需要实时监测更多特征，尤其是喉部活动。其增强功能包括显示喉部闭合商（CQ）、频谱比、声压级（SPL）、微扰和抖动 [12, 13, 14]。
作为博士研究项目一部分开展的实验 [15] 表明，使用 ALBERT 总体上会提高声压级（SPL）和闭合商（CQ）比值，而这两者分别与“歌者共振峰”和音域密切相关。此外，ALBERT 通过颜色映射优化的可视化效果，对注意力持续时间有限的人尤其有帮助。

3.1.3 Windows SINGAD（WinSINGAD）

在 SINGAD 和 ALBERT 的基础上，开发人员推出了 WinSINGAD。这是一款面向自动歌唱评估（ASA）的软件，兼容微软 Windows 操作系统 [16, 17, 18]。它提供了一套全面的显示面板，包括输入声压波形、 f_0、短时频谱、窄带声谱图、频谱比、声道面积变化，以及用于视觉评估的侧向网络摄像头画面。

尽管 WinSINGAD 提升了自动歌唱评估系统的可及性，但其主要贡献并非技术创新，而是促进了歌唱教育研究 [19], , 尤其是在工作室环境中 [18,20, 21, 22] ，以及嗓音治疗领域的研究 [23]。利用 WinSINGAD 开展的研究凸显了实时视觉反馈的重要性；侧向网络摄像头画面有助于评估歌唱时与姿势相关的方面，例如下颌位置和头颈对齐。
然而，这些研究也指出了若干实验方面的挑战。离轴放置麦克风会导致混响过强；在歌者嘴唇附近使用领夹式麦克风时会出现信号过载；音乐伴奏产生的声学干扰也会造成测量误差，并可能导致显示面板上的结果不准确。

3.2 歌唱信息处理（SIP）

在 WinSINGAD 开发时期，Sing&See [27] 等其他自动歌唱评估（ASA）应用也展示了实时定量 VFB 的有效性。然而，这些系统依赖符号表示，例如乐谱，或机器可读元数据，例如 MIDI 文件。

为克服这些局限，歌唱信息处理（SIP） [3]应运而生，旨在促进对歌唱人声的分析，并将其与商业发行的 CD录音中的人声成分联系起来。

3.2.1 MiruSinger

这一新领域的开创以 MiruSinger [25], 的开发为标志。该系统旨在评估用户歌唱声音中的f_0 和颤音特征。系统通过识别歌唱输入中最主要的谐波结构，实时估计 f_0 。对于 CD录音中的歌唱部分，则使用 PreFEst算法 [24], 进行 f_0 估计，从而能够对两条 f_0 轨迹进行比较分析。通过对 f_0的一阶有限差分应用短时傅里叶变换（STFT），量化颤音的可能性，该结果作为一项关键特征。随后，使用支持向量机（SVM）分类器将颤音质量分类为较差或良好。在可视化显示中，目标特征轨迹与用户输入特征叠加显示，便于评估二者的相似性或差异。

4 当代研究成果

自 SINGAD、ALBERT、WinSINGAD 和 MiruSinger 以来，众多研究拓展了自动歌唱评估和歌唱信息处理的方法，并催生了多种应用。本节根据这些研究的主要应用领域对其进行分类，并系统概述该领域的当代进展。

4.1 综述研究

其他综述研究　探讨了实时视觉反馈工具在歌唱训练中的发展及其教学影响　g.
Hoppe 等人 [28] 评述了四种视觉反馈系统——SINGAD、ALBERT 和 Sing&See，分析了它们的功能演变及教学应用。该研究通过评估训练成果的定量和定性指标，考察了这些系统在辅助歌唱学习者和青少年学生方面的有效性。研究结果证实，视觉反馈系统有助于提升歌唱能力，同时强调需要进一步研究，以优化视觉呈现的细节程度、用户技能水平与这些工具的教学整合之间的平衡。
在此研究基础上，Lã 和 Fiuza [29] 全面综述了实时视觉反馈的应用，重点关注发声中的呼吸和振荡方面。他们的分析涵盖了关键的生理和声学监测技术，包括呼吸模式跟踪（例如肺容量和声门下压力）、声带振动分析（通过电声门图和逆滤波）以及声学可视化（通过声谱分析）。研究结果突显了生物反馈在教学上的优势，并表明实时生理和声学信息有助于为学生和教师提供更加科学严谨的声乐训练方法。

4.2 教学研究

一系列教学研究评估了实时视觉反馈在歌唱教学中的有效性。这些研究主要侧重于评估现有技术，而非推动新的技术框架发展。
Callaghan 等人 [30] 探讨了计算机辅助视觉反馈在歌唱教学中的应用，并考察了其对声乐训练的影响。研究对象包括歌唱学生和教师，并通过五种不同的显示类型，分析音高跟踪、元音质量和共振可视化等关键特征。评估重点关注声学分析、视觉呈现的有效性及其在教学中的适用性。研究结果强调了视觉反馈准确性和任务相关性的重要性，并指出，针对语音优化的算法应用于歌唱时，在更宽的音高和强度范围内存在局限。

在此基础上，Callaghan [31] 进一步研究了视觉反馈在歌唱教学中的作用，并围绕三个关键方面展开分析：(a) 提取与感知相关的嗓音特征（如音高、元音身份和音色）；(b) 开发有助于歌者调整发声输出的有效视觉显示；(c) 将视觉反馈融入结构化歌唱课程的教学实践。研究使用了 Kay Elemetrics[26] 的商用工具，包括用于音高跟踪的 Visi-Pitch、用于元音身份识别的 Sona-Match，以及用于音色分析的计算机化语音实验室（CSL），这些工具均安装在一套专用系统上。为确保一致性，研究使用了高质量的头戴式麦克风，以避免因姿势调整而影响固定麦克风设置。研究结果有助于深入理解如何有效利用实时视觉反馈来支持嗓音发展和教学法。

Wilson 等人 [27] 通过比较两种反馈模态，考察了实时视觉反馈对提升歌唱音高准确度的影响：一种是提供基于目标的视觉引导的基于网格的系统，另一种是显示实时音高但不记录历史数据的键盘式显示屏。研究还将这些方法与未接受反馈的对照组进行了比较。以音分为单位测量的音高准确度显示，两个实验组均有显著提升，而对照组则未见可测量的进步。该研究展示了增强反馈在技能习得中的关键作用，进一步表明视觉支持对于提升歌唱中的音高准确度不可或缺。
Martínez 等人 [32] 对歌唱教学法中的发声进行了客观分析，重点关注 f_0 这一主要特征。研究分析了持续元音发声在三个时间区段——初始、中间和末尾——的表现，以评估颤音特征和非预期振荡。研究使用 Praat[27] 提取 f_0，并使用 MATLAB[28] 进行可视化。结果表明，结合使用听觉刺激和 VFB 显著提升了对发声的解读，这说明，与仅凭听觉感知相比，实时视觉强化能促进更有效的学习。

4.3 歌唱练习

除教学应用外　实时视觉反馈已广泛应用于歌唱练习，重点在于　n
提升歌唱表现　通过声学分析、音高跟踪和交互式训练工具提升表现。

[26]https://www.kayelemetrics.com/[27]
https://www.fon.hum.uva.nl/praat/[28]
https://www.mathworks.com/products/matlab.html

Lundy等人 [33]对歌唱学生的歌唱与说话嗓音进行了声学分析，采用歌唱功率比（SPR）作为主要指标来区分两者。研究对象包括不同训练阶段的歌手，研究人员使用快速傅里叶变换（FFT）和多维语音程序（MDVP）软件分析数字录音。提取的特征包括 SPR、 f_0、抖动、微扰和噪声谐波比（NHR）。与此前针对专业歌手的研究结果相反，本研究发现，歌唱音调与说话音调之间的 SPR 没有显著差异。不过，语音样本的微扰和 NHR 值较高，表明说话嗓音的频谱变异性更大。

Singing Coach 系统 [34] 旨在通过实时音高跟踪与可视化，帮助歌手改善音高和节奏。该工具面向初级、中级和高级用户，集成了音高跟踪、可视化记谱、节拍器指导和播放功能，用于自我评估。评估采用客观方式进行，系统提供自动评分和即时反馈。训练难度逐步提高，以促进声乐表演的更高准确度。

Lin 等人 [35] 推出了一款面向 iOS 的实时交互式音高训练工具，旨在提升歌手的音准和时序准确度。该系统采用基于 FFT 和倒谱的音高估计，并集成了音准分类器、评分机制和交互式训练模块。该评估面向希望提升音高准确度的用户，采用实时 VFB，并根据预定义阈值内的音高偏差使用通过/未通过评分系统进行评估。客观评估显示，用户的表现得到了显著改善，训练后平均音准准确度达到 94.81%。

Cantus [36], 是一款用于实时声乐训练的基于网络的应用，采用听唱法，并配有同步的 VFB。用户对着麦克风演唱，同时观察屏幕上亮起的音符，使自己的演唱与显示的参考音对齐。音高准确度评估依靠 YIN 算法 [37] 进行 f_0 检测，将用户输入与目标音符进行比较。评分范围为 0 到 10 分，音高偏差在半音的三分之一以内可获得满分。该评估方法客观地运用频率分析，对演唱准确度进行量化。

Rosenzweig 等人 [38] 开发了 TuneIn，这是一种基于网络的界面，旨在通过实时音准反馈支持合唱排练。该系统通过 f_0 估计和偏差分析评估音高准确度，尤其针对钢琴伴奏下的合唱演出。评估过程采用用于实时音高估计的卷积表示音高估计算法（CREPE） [39] ，并在钢琴卷帘界面中可视化音高偏差。客观评估以音分测量音准偏差，并提供颜色编码的反馈，帮助歌者改善音高准确度。

4.4 游戏与虚拟现实

交互式游戏化系统和虚拟现实（VR）应用也被探索为声乐训练的创新工具，提供沉浸式和交互式环境，以提升用户参与度和表演评估效果。

Brereton 等 [40] 开发了虚拟歌唱工作室（VSS），该系统旨在模拟不同的声学环境，并分析歌手如何调整其表演以适应不同的空间条件。该系统提供实时听觉反馈，使歌手能够感知自己的声音，仿佛正在进行现场表演。研究邀请专业歌手在真实和虚拟空间中进行表演，并通过客观的信号处理技术进行评估，测量节奏、颤音和音准，同时依据用户反馈开展主观评价。尽管歌手表示真实环境与虚拟环境之间存在明显差异，但客观结果显示表演变化有限，其中颤音幅度是受影响最大的参数。

Al Kork 等 [41] 推出了一种旨在教授、表演和评估人类节奏口技的交互式游戏化系统。该系统利用实时传感器数据分析舌头动作、面部动作和音高，从而提供引人入胜且结构化的学习环境。研究采用了头盔式设备，将超声波换能器、Kinect 摄像头和麦克风集成在一起，用于追踪发音动作并将其与专家模型进行比较。评估框架结合了运动行为和音高准确度等客观测量，以及来自虚拟导师的主观反馈；虚拟导师通过口头和书面指导帮助改进用户的表演。

张等人 [42] 提出了一种实时歌唱评分系统，利用虚拟现实、 f_0 分析和动态时间规整（DTW）进行音频比较。该系统旨在帮助歌手在虚拟现实环境中练习，以提升其舞台表现力并缓解表演焦虑。系统通过将用户的声乐表演与参考音轨进行比较来评估表现，实时评分会影响虚拟舞台上的角色动画。客观评估采用了高评分方法，而主观反馈则表明，该系统具有很强的沉浸感，并能有效减轻用户对虚拟表演的焦虑。

4.5 卡拉OK评分

尽管关于卡拉OK评分的科学文献相对较少，但仍有一些值得关注的研究探讨了自动评估方法，重点关注统计建模、音高准确度、节奏和动态分析。

Qiu [43] 开发了一种卡拉 OK 评分算法，采用由高斯混合模型（GMM）训练的隐马尔可夫模型（HMM），评估音高、重音、过零率和功率等音乐特征。该系统使用 YIN 算法 [37] 进行音高检测，并引入半音容差参数来调整难度级别。训练所用数据集由一位业余但训练有素的歌手录制，并使用 HTK 工具箱[29]开发音符内模型。在评估过程中，系统将歌唱输入分割为单个音符，并根据错误率对准确度进行评分；错误率定义为误分类帧数与总帧数之比。该方法依靠统计建模，提供客观的表演评估。

Tsai 等人 [26]提出了一种用于卡拉 OK 表演的自动歌唱评估系统，整合了从卡拉 OK VCD 录音中提取的音高、音量和节奏特征。该系统采用多层级评分方法：基于音高的评分使用 MIDI 音符和动态时间规整确定；基于音量的评分根据通过动态时间规整分析的对数能量序列得出；基于节奏的评分则使用隐马尔可夫模型计算。客观评估结果表明，系统自动评分与人工评分之间存在较强相关性（皮尔逊系数为 0.82），验证了该系统在表演评估中的有效性。

Narang 等人 [44] 重点研究了流行摇滚演唱中的力度标记自动转录，并将响度作为主要评估特征。该研究分析了商业演唱录音及其对应的卡拉 OK 版本，提取源分离人声轨，并与原始人声分轨进行比较。研究使用皮尔逊相关系数评估响度曲线的相似度，并辅以基于峰值的直方图分析来考察局部动态，同时计算全局动态范围。评估采用了客观标准（单尺度响度指标）和主观标准（对信息量与准确性的专家评估）。尽管这种方法成功捕捉了力度方面的结构相似性，但源分离过程产生的伪影限制了其准确性。

4.6 才艺表演与音乐学院入学评估

尽管才艺表演与音乐学院入学评估的目的各不相同，但二者采用了共同的方法：由考试委员会评估候选人的歌唱能力。在此背景下，一些研究探索了人工评估的自动化替代方案。

Lal [45] 开发了一种面向电视才艺表演的电话歌唱评估系统。该系统将用户对预录歌曲片段的演唱与原始版本进行比较，使用稳健音高跟踪算法（RAPT） [46] 进行音高估计，并通过维特比对齐测量音高差异。随后，使用欧几里得距离度量这些差异，并将结果转换为十分制评分。系统根据音高曲线对齐情况评估用户，并对替换错误作相应扣分。研究同时进行了客观评估（算法评分）和主观评估（人工评委排名），结果显示用户参与度很高，来电中有相当一部分来自重复参与者。

Bokshi 等人 [47] 提出了一种自动歌唱质量分类系统，用于评估电视播出的美国歌唱比赛中的参赛者。该系统同时利用音频和视觉特征：音频部分采用梅尔频率倒谱系数（MFCC），视觉部分则通过唇部／眼部运动分析构成。系统根据评判结果，将由受过训练和未受过训练的歌手组成的候选者分为合格与不合格。评估采用了多种机器学习分类器，包括逻辑回归、朴素贝叶斯和 k近邻（k-NN），并应用于单模态（音频、唇部或眼部）及多模态（视听）特征。结果表明，基于音频的分类准确率达到 95%；加入唇部和眼部特征后，性能略有提升，最高提高 2%，而单独使用这些视觉模态时，准确率为 82%。

Bozkurt 等人 [48] 提出了一种专为音乐学院入学考试设计的歌唱声音评估系统，以f_0 分析作为主要评估特征。该系统从包含 1018 条歌唱录音和 2599 条钢琴录音的数据集中提取 f_0 序列，并采用 Melodia 算法[30]进行音高跟踪。表演评估使用动态时间规整和以 Julia 实现的基于机器学习的分类系统进行。客观评估通过基于动态时间规整的距离直方图训练分类器，平均交叉验证准确率达到 0.74。由于隐私限制，仅共享了提取的 f_0 序列数据，而未共享原始音频录音。

[29]https://htk.eng.cam.ac.uk/[30] https://www.upf.edu/web/mtg/melodia

4.7 稀有歌唱

稀有歌唱包括在特定文化、历史或仪式背景中不常见、具有专门特色或得以保存的声乐传统、风格或曲目。一些研究探索了如何利用计算工具和交互系统分析、教授并保存这些独特的歌唱实践。

Elmer and Elmer [49] 提出了一种计算机辅助方法，以音高和时间为主要分析特征，对非常规歌唱风格进行分析和表征。该研究聚焦于歌唱表演，运用频谱分析和 f_0 提取进行评估，并采用一款与CoolEdit[31]集成的定制音高分析程序。客观评估结合了声学分析和听觉分析，既能可视化随时间变化的音高模式，也能捕捉音高稳定性与不稳定性等特征。该系统还借助一种新颖的符号记谱系统，对音高组织、滑音、呼吸模式、合唱及教学要素进行表征，从而为非标准发声技巧提供了一种结构化方法。

Stavropoulou 等人 [50] 研究了基于视觉反馈的歌唱软件在提升希腊小学生（6—9岁）的音高准确度和声音质量方面的效果。研究使用Singing Coach Pro作为训练工具，通过录制参与者演唱希腊传统歌曲和西方古典乐曲的表现，并分析训练前后的音高准确度，评估其歌唱能力。参与者被分为三组：初始歌唱表现良好、提升幅度有限者；有所提升者；以及音高偏差显著者。结果表明，视觉反馈，尤其是与教师指导相结合时，能够提高参与者的积极性和歌唱准确度；希腊传统歌曲以及年龄较大的儿童，特别是女孩，提升更为显著。

Chawah 等人 [51] 开发了一个教育平台，旨在通过专用“超头盔”采集多模态发音与声学特征，促进稀有歌唱技巧的学习与分析。该系统面向演唱科西嘉*CantuinPaghjella*和拜占庭圣咏的专业歌手，评估依据是对声带运动、呼吸、舌部和唇部运动以及音频信号进行同步录音。数据处理采用 RTMaps[32] 和基于 MATLAB 的工具完成。尽管该项目仍处于数据收集阶段，但其长期目标是通过提供自适应歌唱模块和三维声道可视化来促进学习，从而支持稀有声乐传统的保护。

Moschos 等人 [52] 推出了 FONASKEIN，这是一款用于实时歌唱声音练习并提供 VFB 的交互式软件应用，尤其适用于微分音阶与非平均律音阶。该系统使用 Max/MSP 开发，使用户能够研究微分音调律系统，例如拜占庭音阶和古希腊音阶，同时还支持导入以音分定义的自定义音阶。该平台包含两个核心组件：(1) 使用 fiddleMax 对象对˜麦克风输入进行实时分析与转换，以实现音高检测；(2) 将 MIDI 文件转换为带有微分音回放控制的乐谱。音高估计通过基于离散傅里叶变换（DFT）的方法完成，并采用最大似然方法提取 f_0。通过提供实时听觉与视觉反馈音准，FONASKEIN 帮助歌手掌握微分音调律系统并提升音高准确度。

4.8 声学生物反馈

在 ALBERT 和 WinSINGAD 奠定的基础上，一项研究探讨了声学生物反馈在歌唱教学法中的应用。这种方法通常使用可穿戴设备或侵入式传感器监测生理和声学参数，并提供实时反馈，以加强声乐训练。

Piao 和 Xia [53] 提出了一种多模态歌唱辅导系统，该系统整合了配备压力传感器的可穿戴呼吸检测器，可监测呼吸状态并进行音高分析。该系统面向初级歌唱学习者（18 至 26 岁），通过交互式乐谱界面直观显示音高轮廓和呼吸状态，从而在练习过程中提供实时反馈。研究采用 2×2 析因设计，招募了 14 名参与者（10 名具有音乐背景，4 名没有），让他们在两种条件下学习两首歌曲：仅提供音高反馈，以及提供音高 + 呼吸反馈。客观评估表明，对于受过音乐训练的参与者，加入呼吸监测后音高准确度提高了 21.25%；而没有接受过正规音乐训练的参与者则未见显著改善。

4.9 表现力与表达性

尽管这两个术语常被交替使用，但表现性指的是声乐表演中不同的方面。表现性涉及表演的时间动态，包括起音、衰减、延音和释音（ADSR）包络等参数。相比之下，表达力与表演的情感和诠释特质有关，

[31] https://www.adobe.com/special/products/audition/syntrillium.html[32] https://intempora.com/products/rtmaps/

根据卡拉OK世界锦标赛（KWC）的指导方针[33],，表演需要通过演唱方式、肢体语言和面部表情传达情感与热情。

Bonada 等人 [54] 介绍了歌唱导师，这是一种自动评估工具，可在音符和音符内部两个层面评估音准、节奏和表现力。该系统采用未经训练的基于隐马尔可夫模型的分割算法，并辅以启发式规则，将表演指标与MIDI参考对齐。系统还可识别颤音、起音特征和音符过渡等表现力元素。研究使用业余歌手演唱的三首商业流行歌曲进行评估，分析了 1000 多个音符。与人工标注相比，该分割算法的准确率达到 95%（容差窗口为 30 毫秒）。由专家对表现力转录进行人工审查后发现，对于正确分割的音符，其准确率超过 95%。

Yu 等人 [55] 提出了一种歌唱声音表现评分系统，该系统整合了使用 openSMILE 提取的声学特征和情感特征[34]。系统通过将业余演唱与专业演唱版本进行比较来评估其表现，使用 DTW 进行对齐，随后提取特征（如音高、响度和梅尔频率倒谱系数）。分类采用支持向量机和深度玻尔兹曼机（DBM）完成。DBM 模型的表现略优于支持向量机分类器，加权召回率最高达到 0.5854，而后者为 0.5444。不过，研究指出，受数据集规模有限的制约，这两种模型的整体表现均受到限制。

4.10 其他自动歌唱评估与歌唱信息处理

其他研究侧重于利用机器学习和信号处理技术开发自动歌唱评估与歌唱信息处理系统，以评估声乐表演。这些方法旨在通过分析音高准确度、音符过渡和整体音乐表现力，客观地量化歌唱质量。

张等 [56] 提出了一种基于深度度量学习的自动歌唱评估系统，该系统使用三元组网络，将高质量的歌唱表现映射到更接近参考音轨的位置，同时区分表现不佳的歌唱。该系统利用带有自注意力机制的卷积神经网络（CNN）处理频谱图、色度特征、 f_0和节奏图。模型使用独唱片段进行训练和测试，片段主要来自中国声乐表演。评估时，将相对于参考音轨的高质量和低质量歌唱的嵌入表示进行比较，结果显示其与人工评分高度相关。与传统的基于规则的方法相比，该系统在预测歌唱质量方面表现更佳；客观评估也证实了所学习的嵌入空间具有稳健性。

法吉赫等 [57] 提出了一种用于检测独唱中的音符起始、止音和过渡的新型算法，适用于实时和离线应用。该方法依据音高轮廓轨迹的变化来识别音符边界。研究使用带标注的声乐数据集（埃尔科迈什维利[35] 和SVNote1[36]）评估该系统，并通过F值和统计分析（方差分析）将其与八种现有起音检测算法进行基准比较。该算法表现优异，较其他方法提升了2%至36%，其中在高质量音频录音上的提升最为显著。此外，实时评估显示，根据数据集特征不同，处理延迟介于46至230毫秒之间。

5 讨论

本讨论有两个主要目标：（1）综合现有文献的优势与主要贡献；（2）找出值得进一步研究的潜在研究空白。

5.1 现有文献综述

第4.1节回顾的研究追溯了VFB工具的发展历程及其在当代歌唱教学法中的整合应用。这些研究强调，完善反馈机制对于提升教学效果至关重要，最终有助于改善声乐训练的技术表现和表达表现。

第4.2节证实了VFB在声乐教学中的教育价值。这些研究强调，视觉呈现必须具有感知意义，并凸显了技术进步在优化视觉反馈系统、采用更有效的训练方法方面所发挥的作用。

[33]https://www.kwc.fi/[34]https://www.audeering.com/research/opensmile/[35] https://www.audiolabs-erlangen.de/resources/MIR/2019-GeorgianMusic-Erkomaishvili[36] https://zenodo.org/records/7061507

第4.3节的研究表明，基于VFB的歌唱练习工具能够有效提升音高准确度、时序准确度和歌声精确度。研究结果进一步强调了实时反馈在构建声乐训练体系中的重要性，使个人表演者和合奏场景都能受益。

第4.4节探讨了虚拟现实和游戏化在声乐训练中的融合，展示了它们在提供沉浸式反馈、完善发声技巧和提升学习者参与度方面的能力。通过将实时分析与互动组件相结合，这些系统为数字声乐教学提供了颇具前景的途径。

第4.5节回顾的研究考察了卡拉OK评分的多种方法，从统计建模到基于特征的表演评估。这些研究成果为开发更先进、更可靠的自动歌唱评估系统铺平了道路，其应用涵盖音乐教育、游戏和娱乐领域。

第4.6节讨论了自动歌唱评估的进展，尤其是在才艺竞赛和音乐学院招生中的应用。通过结合音高跟踪、统计建模和基于机器学习（ML）的分类，这些方法为传统人工评估提供了客观且可扩展的替代方案，从而提高了声乐表演评估的一致性和效率。

第4.7节的研究聚焦于稀有歌唱传统的计算分析、训练与保护。通过运用音高跟踪、多模态分析和交互式学习工具，这些研究展现了技术在研究和传承非主流发声技巧方面的潜力，确保这些传统持续保持其文化影响力。

第4.8节强调了声学生物反馈系统在声乐训练中的作用，展示了它们在提升音高准确度、呼吸控制和整体教学成效方面的有效性。通过实时生理监测和交互式反馈机制，这些系统能为发声技巧和表演优化提供有价值的见解。

第4.9节的研究通过开发评估歌唱中的表演动态、情感内容和表现细微差异的方法，推动了对歌唱表现力的计算分析。这些研究成果有助于开发更完善的数据驱动评估工具，用于声乐训练和表演评估。

最后，第4.10节介绍了自动歌唱评估方面的进展，重点关注音符分割准确度的提升以及用于表演评估的机器学习模型的改进。这些进展推动了精确、高效的计算工具的发展，为声乐训练和音乐教育提供支持。

5.2 未来研究贡献的潜在空白

现有文献凸显了技术在声乐训练、表演评估以及多元歌唱传统传承中的作用日益增强。随着数字工具不断发展，未来研究应着力完善这些方法、提升可及性，并将其融入主流教学框架。尽管已取得显著进展，数字信号处理（DSP）和人工智能（AI）的现代进展在自动歌唱评估（ASA）和歌唱信息处理（SIP）框架中仍未得到充分应用。这些技术的整合有望推动该领域取得重大进展。

1. **预测信号与目标信号分离**：

   自动歌唱评估和歌唱信息处理中的一个基本挑战，是减轻各种干扰人声信号分析的噪声形式。这些噪声包括：

   a) 卷积噪声：在现实环境中，声波会从墙壁和家具等表面反射，从而引入回声和混响等声学现象 [58, 59]。由于直达声与反射成分之间的相互作用，这些影响会使录制信号发生失真 [60, 61]。b) 加性噪声：歌唱表现通常发生在伴有背景音乐的环境中，背景音乐一般通过公共广播系统（PA系统）或伴奏乐器播放。即使使用定向麦克风（如心形指向动圈麦克风） [62], ，录音中也不可避免地会捕获一定程度的背景噪声，从而引入不需要的成分。c) 竞争性噪声：观众的存在可能会以欢呼、鼓掌甚至跟唱的形式引入进一步的干扰。这是一种尤其具有挑战性的加性噪声，因为它与主要人声信号共享频谱和时间特征，使分离过程更加复杂。

   这些噪声源会在预测信号中引入不需要的伪影，而预测信号理想情况下应代表歌手经过分离的声乐表演。因此，需要采用先进的增强或分离算法来滤除噪声，并保持人声录音的完整性。

SIP 中的目标信号也面临类似挑战：其目标是从商业制作的歌曲中提取分离的人声，以便与预测信号进行比较。这需要开发强大的源分离算法，以便从给定的音乐曲目中有效分离人声。

在这一领域，人们采用了维纳滤波、波束形成和时频掩蔽等传统信号分离技术。然而，近年来数据驱动方法，尤其是人工智能驱动方法的进步，显著提升了相关效果。新兴的“机器听觉”领域 [63], 类似于计算机视觉，利用机器学习和深度学习（DL）实现高层次的听觉场景分析。

一种值得关注的方法是 Wave-U-Net [64], ，该架构受时序卷积网络（TCNs）启发 [65] ，可直接在时域中执行端到端源分离。将其配置为单输入多输出（SIMO）系统后，便可分离人声、鼓、贝斯及其他乐器等不同音乐组成部分。

较新的方法采用二维深度 U-Net 架构 [66] 进行人声与伴奏分离，并使用基于短时傅里叶变换的幅度谱图。这些模型采用带有瓶颈层的编码器-解码器结构，通过使用 ℓ_1 损失函数，预测人声和伴奏部分的谱图掩码。持续改进这些架构，尤其是提升其在不同音乐风格和录音条件下的泛化能力，仍是一项尚待解决的研究挑战。

2. **特征、计算指标与非侵入式主观评估**:

有效且标准化的歌唱评估框架应兼顾技术熟练度与艺术表现力。当前的计算分析主要聚焦于与技术执行相关的低层特征，包括音高准确度、节奏精确度、力度、呼吸控制和颤音，这些因素均有助于形成高层次的音乐诠释。典型的计算描述符包括 f_0 轨迹分析、动态时间规整、 ℓ_2 距离度量和基于隐马尔可夫模型的分割。

然而，更全面的评估策略应纳入音乐信息检索（MIR）中常用的更广泛的低层特征。这些特征包括能够捕捉高层次谐波关系的谐波表示（如色度图），以及基于听觉的频谱表示，例如 Bark、Mel 和 Gammatone 频谱图。此外，相应的倒谱变体——Bark 频率倒谱系数（BFCC）、梅尔频率倒谱系数（MFCC）和 Gammatone倒谱系数（GTCC）——可提供更精细的声学描述符，进一步提升计算歌唱评估的稳健性。

为建立严谨且标准化的评估方案，定性研究应在精心构建的测试集上系统评估这些计算特征的适用性。这有助于确保不同评估范式下表演评估的一致性。此外，还应将艺术表现力纳入计算评估模型，其中包括情感传达、音色变化和歌词吐字等因素。

先进的基于人工神经网络（ANN）的方法为非侵入式主观评估提供了有前景的途径。例如，可采用自动语音识别（ASR） [67] 评估歌词准确度，而歌曲情感识别（SER） [68] 模型则可量化情感表现力。鉴于当代人工神经网络在分类方面达到的先进性能，将这些方法纳入自动歌唱评估和歌唱信息处理框架，将显著提升声乐表演评估的深度和可靠性。

6 结论

实时视觉反馈是自动歌唱评估和歌唱信息处理框架的基本组成部分。卡拉OK系统采用的评分机制等方式虽然能够对歌唱者的表现进行总体评估，但要实现实时自我纠正并提升技能，对错误进行即时反馈至关重要。尽管这些方法已在教学法和声乐训练等领域展现出成效，但要充分利用数字信号处理和人工智能领域近期的技术进步，仍需取得重大进展。当前最迫切的挑战之一是

人声信号的增强与分离，因为现实录音环境会引入各种噪声，从而影响自动歌唱评估和歌唱信息处理的准确性与有效性。尽管音乐信息检索中的计算特征提取在人声分析中发挥了关键作用，现有研究对丰富多样的可用描述符才刚刚开始探索。还需开展更多研究，以确定最适合实现稳健可靠人声评估的特征。此外，人工智能驱动的方法具有巨大潜力，可通过对歌唱表现中的主观方面（如歌词准确度和表现力）进行建模，补充传统评估方法。整合基于人工智能的技术，能够弥合计算评估与类人评估之间的差距，从而提升自动歌唱分析系统的全面性与准确性。

7 致谢

谨向 Mateus Prauchner Enriconi 工程师、理学硕士致以诚挚谢意，感谢他提出的深刻见解、建设性讨论和宝贵建议。

8 数据可用性声明

本研究发现所依据的数据可向通讯作者 dos Santos, A. N. 提出合理请求后获取。

参考文献

[1] Bryan E. Nichols. 为音乐课堂设计歌唱评估。《General Music Today》，30:13–17，4 2017年。[2] Sebastian Silas、Daniel Müllensiefen 和 Reinhard Kopiez。歌唱能力评估：基于项目反应理论的歌唱测试及用于歌唱数据的通用开源软件环境的开发与验证。《Behavior Research Methods》，56:4358–4384，9 2023年。[3] Masataka Goto、Takeshi Saitou、Tomoyasu Nakano 和 Hiromasa Fujihara。基于歌声建模的歌唱信息处理。载于《2010年IEEE国际声学、语音与信号处理会议》，第5506–5509页，2010年。[4] Laure Prétet、Romain Hennequin、Jimena Royo-Letelier 和 Andrea Vaglio。歌声分离：训练数据研究。载于《ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)》，第506–510页，2019年。[5] Shoko Araki、Nobutaka Ito、Reinhold Haeb-Umbach、Gordon Wichern、Zhong-Qiu Wang 和 Yuki Mitsufuji。30+ 源分离研究三十余年：成果与未来挑战，2025年。[6] Graham Welch、David Martin Howard 和 C RUSH。SINGAD（歌唱评估与发展）系统：针对音准较差歌者开展的课堂试点研究。《Canadian Journal of Music Education》，1988年。[7] Graham F Welch、C Rush 和 DM Howard。SINGAD（歌唱评估与发展）系统：在课堂上的首次应用。《Proceedings of the Institute of Acoustics》，10(2):179–185，1988年。[8] Graham F. Welch、D. M. Howard 和 C. Rush。实时视觉反馈对歌唱中人声音高准确度发展的作用。《Psychology of Music》，17(2):146–157，1989年。[9] David M. Howard 和 Graham F. Welch。基于微型计算机的歌唱能力评估与发展。《Applied Acoustics》，27(2):89–102，1989年。[10] David M. Howard 和 James A. S. Angus。8至11岁儿童与受训成年歌者歌唱音高策略的比较。《Logopedics Phoniatrics Vocology》，22(4):169–176，1997年。[11] Howard Daw'd M 和 James AS Angus。实时视觉反馈与小学生歌唱音高技能的发展。《Proc.》l。《*OA Vol*》，20(第6部分)，1998年。[12] David M Howard。语音与歌唱中的实时视觉显示。《Defence Science Journal. Delhi》，43(3):211–221，1993年。[13] David Rossiter 和 David M Howard。Albert：用于交互式分析和显示声源与声学参数的系统。《Proc. lOA Vol》，16(第5部分)，1994年。[14] David Rossiter、David M. Howard 和 Mike DeCosta。接受训练时，有无实时视觉生物反馈对声音发展的影响。《The Journal of the Acoustical Society of America》，99(5):3253–3256，05 1996年。

[15] David Paul Rossiter。用于声音教学的实时视觉显示。博士论文，约克大学，1995年。[16] David M Howard。量化儿童歌唱声音的发展变化。载于《歌唱生理学与声学国际会议》第一届会议论文集，荷兰格罗宁根，第2–5页。Citeseer，2002年。[17] David M Howard、Graham F Welch、Jude Brereton 和 Evangelos Himonides。探索用于歌唱训练的新型实时视觉显示。载于第三届《生物医学应用中的声发射模型与分析国际研讨会》，2003年。[18] David M. Howard、Graham F. Welch、Jude Brereton、Evangelos Himonides、Michael DeCosta、Jenevora Williams 和 Andrew W. Howard。WinSINGAD：用于歌唱教室的实时显示。Logopedics Phoniatrics Vocology，29(3):135–144，2004年。PMID: 15370645。[19] David M. Howard。歌唱课程中的实时视觉反馈技术。音乐教育研究，24(1):40–57，2005年。[20] Graham F Welch、Evangelos Himonides、David M Howard 和 Jude Brereton。Voxed：作为歌唱教室中有效教学辅助工具的技术。载于《跨学科音乐学会议》（cim04）论文集，2004年。[21] Evangelos Himonides、Graham F. Welch、David M. Howard 和 Jude Brereton。歌唱教室中的实时反馈：一项运用新型声音技术的创新行动研究项目。Music Education Research，7(2):225–249，2005年。[22] David M. Howard、Jude Brereton、Graham F. Welch、Evangelos Himonides、Michael DeCosta、Jenevora Williams 和 Andrew W. Howard。实时显示对歌唱教室有益吗？一项探索性研究。Journal of Voice，21(1):20–34，2007年。[23] D.M. Howard、A. Jackson 和 A.W. Howard。嗓音治疗实时视觉显示的发展。载于《2005年第三届IEE信号处理医学应用国际研讨会》（参考编号：2005-1119），第141–146页，2005年。[24] Masataka Goto。实时音乐场景描述系统：通过主导基频估计，从真实世界音频信号中检测旋律和低音线。Speech Communication，43(4):311–329，2004年。专题：真实世界声音的识别与组织。[25] Tomoyasu Nakano、Masataka Goto 和 Yuzuru Hiraga。MiruSinger：一种利用实时反馈和音乐CD录音作为参照数据的歌唱技能可视化界面。载于第九届IEEE多媒体国际研讨会专题会议（ISMW 2007），第75–76页，2007年。[26] Wei-Ho Tsai 和 Hsin-Chieh Lee。基于音高、音量和节奏特征的卡拉OK歌唱自动评估。IEEE Transactions on Audio, Speech, and Language Processing，20(4):1233–1243，2011年。[27] Pat H Wilson、Kerrie Lee、Jean Callaghan 和 C William Thorpe。学会音准歌唱：实时视觉反馈有帮助吗？Journal of interdisciplinary music studies，2，2008年。[28] Dave Hoppe、Makiko Sadakata 和 Peter Desain。歌唱训练中实时视觉反馈辅助技术的发展：综述。Journal of Computer Assisted Learning，22(4):308–316，2006年。[29] Filipa MB Lã 和 Mauro B Fiuza。歌唱教学法中的实时视觉反馈：当前趋势与未来方向。《应用科学》，12(21):10781，2022年。[30] Jean Callaghan、William Thorpe 和 Jan van Doorn。视觉反馈技术在歌唱教室中的应用。载于澳大利亚音乐教育研究协会第二十三届年会论文集。纽卡斯尔，9月。墨尔本：AARME，2001年。[31] Jean Callaghan、William Thorpe 和 Jan van Doorn。歌唱与视觉的科学。载于跨学科音乐学会议论文集，奥地利格拉茨，2004年。[32] Luis Alberto Martinez、Gisela Gracida、Rafael Ángel Urrutia、Eladio Cardiel、Manuel Mauricio Lara 和 Pablo Rogelio Hernández。歌唱教学与学习背景下的音频发声分析。载于2019年第16届电气工程、计算科学与自动控制国际会议（CCE），第1–4页。IEEE，2019年。[33] Donna S Lundy、Soham Roy、Roy R Casiano、Jun W Xue 和 Joseph Evans。歌唱学生歌唱与说话嗓音的声学分析。Journal of Voice，14(4):490–493，2000年。[34] Martha Randall。Singing coach unlimited。《美国音乐教师》，55(4):99，2006年。[35] Kin Wah Edward Lin、Hans Anderson、MHM Hamzeen 和 Simon Lui。自学歌唱音高训练应用中实时交互式用户界面设计的实现与评估。载于ICMC，2014年。

[36] Manuel Pérez-Gil、Jesús Tejada、Remigi Morant 和 A. Pérez-González De Martos。Cantus：用于实时声乐训练和音乐音准评估的软件解决方案的构建与评估。《Journal of Music, Technology & Education》，9:125–144，2016年7月。[37] Alain de Cheveigné 和 Hideki Kawahara。Yin：用于语音和音乐的基频估计器。《The Journal of the Acoustical Society of America》，111(4):1917–1930，2002年4月。[38] Sebastian Rosenzweig、Lukas Dietz、Johannes Graulich 和 Meinard Müller。Tunein：用于练习合唱声部的基于网络的界面。载于《国际音乐信息检索学会会议（ISMIR）》演示与最新研究成果，2020年。[39] Jong Wook Kim、Justin Salamon、Peter Li 和 Juan Pablo Bello。Crepe：用于音高估计的卷积表示，2018年。[40] Judith Brereton 等。Singing in Space (s)：真实与虚拟声学环境中的歌唱表现——歌手评价、表演分析与听众感知。博士论文，约克大学，2014年。[41] Samer Al Kork、Deniz Ugurca、C Sahin、Patrick Chawah、Lise Crevier-Buchman、Martine Adda-Decker、Kele Xu、Bruce Denby、Pierre Roussel-Ragot、Benjamin Picart 等。一种用于学习、表演和评估现代当代歌唱的创新人机交互游戏化应用——“人声节奏盒”。载于 VISAPP（第2卷），第640–650页，2015年。[42] Pei Zhang、Yi Han、Weidan Liang 和 Jingyi Chen。一种基于虚拟现实技术的实时歌唱评分系统。载于《物理学杂志：会议系列》，第1982卷，第012170页。IOP出版社，2021年7月。[43] Dongping Qiu。卡拉OK电脑游戏评分算法的开发。硕士论文，KTH，信号处理，2012年。[44] Jyoti Narang、Marius Miron、Ajay Srinivasamurthy 和 Xavier Serra。使用响度度量分析声乐表演中的音乐力度变化。Evangelista G、Holighaus，编。《DAFx20维也纳会议论文集》，第3卷（DAFx20in22）；2022年9月6日至10日；奥地利维也纳，第33–39页，2022年。[45] Partha Lal。歌唱评估算法比较。载于 Interspeech。Citeseer，2006年。[46] David Talkin 和 W Bastiaan Kleijn。一种稳健的音高跟踪算法（RAPT）。《语音编码与合成》，495:518，1995年。[47] Mangona Bokshi、Fei Tao、Carlos Busso 和 John HL Hansen。基于视听特征的歌唱质量评估与分类。载于《 2017 IEEE 视觉通信与图像处理会议（VCIP）》，第1–4页。IEEE，2017年。[48] Baris Bozkurt、Ozan Baysal 和 Deniz Yüret。歌声评估数据集与基线系统。载于《国际计算机音乐多学科研究研讨会（CMMR）论文集》，葡萄牙马托西纽什，第25–28页，2017年。[49] Stefanie Stadler Elmer 和 Franz-Josef Elmer。一种分析和表征歌唱的新方法。《Psychology of Music》，28(1):23–42，2000年。[50] Sofia Stavropoulou、Anastasia Georgaki 和 Fotis Moschos。视觉反馈歌唱声乐技术在希腊小学中的有效性。载于 ICMC，2014年。[51] Patrick Chawah、T Fux、Martine Adda-Decker、A Amelot、N Audibert、B Denby、G Dreyfus、A Jaumard-Hakoun、C Pillot-Loiseau、P Roussel 等。用于采集、可视化和分析稀有歌唱的教育平台。载于 ISCA、INTERSPEECH 2014：展示与讲解贡献，第2128–2129页，2014年。[52] Fotios Moschos、Anastasia Georgaki 和 Georgios Kouroupetroglou。FONASKEIN：用于歌声练习的交互式软件应用。载于《声音与音乐计算（SMC）会议论文集》，9月，第326–331页，2016年。[53] Ziyue Piao 和 Gus Xia。感知呼吸：带有呼吸引导的多模态歌唱辅导界面。载于《国际音乐表达新界面会议》。PubPub，2022年。[54] Jordi Bonada、Alex Loscos 和 Oscar Mayor。歌唱辅导系统：歌声的表现力分类与分割。载于第121届音频工程学会年会。音频工程学会，2006年。[55] Yang Yu、Weisi Lin、Dong-Yan Huang、Minghui Dong 和 Haizhou Li。歌声表演评分。载于《2015年亚洲语言处理国际会议（IALP）》，第119–122页。IEEE，2015年。[56] Huan Zhang、Yiliang Jiang、Tao Jiang 和 Hu Peng。通过参照学习：面向歌唱评估的深度度量学习。载于 ISMIR，第825–832页，2021年。[57] Behnam Faghih、Sutirtha Chakraborty、Azeema Yaseen 和 Joseph Timoney。一种用于实时和离线环境中歌唱起音与止音检测的新方法。《应用科学》，12(15):7391，2022年。

[58] Frank Filipanits。采用基于频谱的时域处理优化的可听化系统设计与实现。迈阿密大学硕士研究项目，*5*月，1994年。[59] Ennio Cruz da Costa。技术声学。Blucher，2003年。[60] Jont B. Allen 和 David A. Berkley。用于高效模拟小房间声学的镜像法。*The Journalof the Acoustical Society of America*，65:943–950，1979年4月。[61] Alan V. Oppenheim、Alan S. Willsky 和 S. Hamid Nawab。信号与系统（第*2*版）。Prentice-Hall, Inc.，美国，1996年。[62] Franz Zotter 和 Matthias Frank。全景声，第19卷。施普林格国际出版，2019年。[63] Richard F. Lyon。人类与机器听觉。剑桥大学出版社，2017年5月。[64] Daniel Stoller、Sebastian Ewert 和 Simon Dixon。Wave-U-Net：用于端到端音频源分离的多尺度神经网络。*arXiv*预印本 *arXiv:1806.03185*，2018年。[65] Shaojie Bai、J. Zico Kolter 和 Vladlen Koltun。用于序列建模的通用卷积网络和循环网络的实证评估，2018年。[66] A. Jansson、E. Humphrey、N. Montecchio、R. Bittner、A. Kumar 和 T. Weyde。使用深度 U-Net 卷积网络进行歌声分离。未发表，2017年10月。[67] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：一种用于语音表征自监督学习的框架。*CoRR*，abs/2006.11477，2020年。[68]Pedro Benevenuto Valadares、Karen Gissell Rosero Jácome、Arthur Nicholas dos Santos 和 Bruno Sanches Masiero。歌曲情感识别：音频特征与人工神经网络之间的性能比较。计算机科学电子学术入门期刊，20(4)，2022年12月。