附件三：

推荐资源：歌曲识别与歌唱评价数据集

腾讯音乐天琴实验室联合清华大学人机语音交互实验室（THUHCSI）、音频语音与语言处理研究组（ASLP@NPU）、CCF 计算艺术分会及北京市智慧广电（网络视听）重点实验室四家机构共同发布片段翻唱、哼唱和歌唱评价三套开源数据集。面向行业、高校专业研究人员提供数据开放服务，夯实产业发展基础。

一、哼唱识别开源数据集介绍

腾讯音乐天琴实验室哼唱识别开源数据集（Lyra-Query by Humming (Lyra-QBH Dataset)）为解决哼唱识别行业开源数据集的短缺而构建。该数据集的样本收集由多位用户在真实场景下所录制，从歌曲选择、录制环境、设备上更满足线上真实的哼唱样本数据。同时，在哼唱识别场景，研究人员也可加入干扰数据，能够帮助研究人员更好地测评哼唱识别算法的性能。Lyra-QBH 包括了 1005 个哼唱样本数据，100 首歌曲的 MIDI 文件以及对应的歌曲信息。该数据集采用微信小程序进行录制，由 97 名用户所参与（男性 38 名，女性 59 名），有效的片段时长范围在 9～10s，人工处理筛选后共计 1005 个哼唱录音片段，同时哼唱的录音片段覆盖 100 首曲目。

申请者需要在 Lyra-QBH 数据集主页(https://lyracobar.y.qq.com/hum_dataset.html)上点击申请按钮后，填写申请信息，填完后确认同意“使用条款”。工作人员将会在 3 个工作日内通过邮件发出下载链接。

二、歌唱评价数据集介绍

腾讯音乐天琴实验室歌唱评价数据集（Lyra-Singing Assessment Dataset (Lyra-SA Dataset)）是国内首个整曲演唱的歌唱评价开源数据集，其样本数据主要来源于全民 K 歌，对音乐教育、线上卡拉 OK 及线下赛事具有非常高的研究与应用价值。该数据集致力于提供真实场景下的歌唱数据及标签，帮助研究人员测评或建立歌唱评价模型。歌唱评价数据集包含 10 首歌曲的 100 个演唱干声，和对应的 MIDI、歌词文件。样本来源于全民 K 歌的授权用户。数据集中不存在相同演唱者提供多个样本的情况。在基于“普通听众的歌唱评价有着相关性”的观念的基础上，我们邀请了一些普通听众对干声进行评分，并标注出音色性别与音

色年龄。请注意，这些粗略标签仅供参考，我们计划后续联合高校推出更加严格和精准的标签。

申请者需要在 Lyra-SA 数据集主页(https://lyracobar.y.qq.com/singvoicedataset.html)上点击申请按钮后，填写申请信息，填完后确认同意“使用条款”。 工作人员将会在 3 个工作日内通过邮件发出下载链接。

三、片段翻唱识别数据集介绍

腾讯音乐天琴实验室片段翻唱识别数据集（Lyra-CoverSegment Dataset(Lyra-CS Dataset)）来自于 QQ 音乐曲库满足开源授权条件的歌曲，其中包含不同语言、流派、歌手的歌曲原唱及对应的翻唱或 live 版本片段，可用于听歌识曲、片段翻唱识别实验。Lyra-CS 打破了目前只有全曲翻唱开源数据集的局面，进一步促进听歌识曲技术的发展。Lyra-CS 总时长 399.7 小时，包含 539203 个录音片段。其中每个片段为长度 15 秒以下不等长的 wav 文件（8 kHz，16bit）。此外，还提供片段歌曲所对应的歌曲名和演唱者，供开发者参考。

申请者需要在 Lyra-CS 数据集主页(https://lyracobar.y.qq.com/hum_dataset.html)上点击申请按钮后，填写申请信息，填完后确认同意“使用条款”。 工作人员将会在 3 个工作日内通过邮件发出下载链接。