台湾歌仔戏电视剧中的定时文本提取

Tzu-HungHuang[1,4] **Yun-En Tsai**[1] **Yun-Ning Hung**[3]**Chih-WeiWu**[2]
I-ChiehWei[5] **LiSu**[11] 台湾中央研究院 [2] 独立研究者 [3] MusicAI，美国[4] 台湾国立台湾大学 [5] 奥克兰大学，新西兰b11902023@ntu.edu.tw, lisu@iis.sinica.edu.tw

arXiv:2601.00299v1 [cs.SD] 1 Jan 2026

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

摘要

台湾歌仔戏（Kua-á-hì）是当地戏剧传统的重要形式，曾被广泛改编为电视剧，其中以 Iûnn L-e-hua 等先驱的作品尤为著名。这些视频对于深入研究台湾歌仔戏具有潜在价值，但其画质往往较差，且数据准备需要投入大量人工。为简化这一流程，我们开发了一个用于实时光学字符识别校正的交互式系统，并提出一种两步方法，将 OCR驱动的分段与语音和音乐活动检测（SMAD）相结合，从存档剧集中高效、精准地识别人声片段。由人声片段及其对应歌词组成的数据集，有望支持歌词识别、曲调检索等多种音乐信息检索任务。代码可在 https://github.com/z-huang/ocr-subtitle-editor 获取。

1. 引言

歌仔戏 [1] 于20世纪初兴起，是一种以台湾话（Tâi-gí）演出的台湾本土戏曲。这种戏剧形式将表演与歌唱结合，以连贯的戏剧叙事呈现，传统上常在宗教庆典和庙会期间演出，以敬奉神明。20世纪80年代，随着电视广播的普及，歌仔戏开始被大量改编为电视剧。在这些改编作品中，加入台湾电视公司的名师杨丽花-深耕十余年，创作了大量作品，并被广泛视为电视歌仔戏的先驱。这些作品如今已可在线获取，为透过音乐信息检索（MIR）方法理解和探索歌仔戏提供了宝贵资源。

这些视频保留了研究台湾歌仔戏所需的所有核心要素，包括音频、视频，以及通过内嵌字幕呈现的对话/歌词。然而，由于视频年代久远且画质较差，传统光学字符识别技术无法达到令人满意的准确度，因此需要耗时进行人工校正。为了高效地为各种音乐信息检索任务和探索准备数据，我们开发了一款用于字幕提取与校正的交互式OCR工具。此外，我们提出了一种两步工作流程，将OCR驱动的分段与SMAD算法相结合，从而构建一个包含人声片段及其对应歌词的台湾歌仔戏数据集。详细过程将在后续章节中阐述。

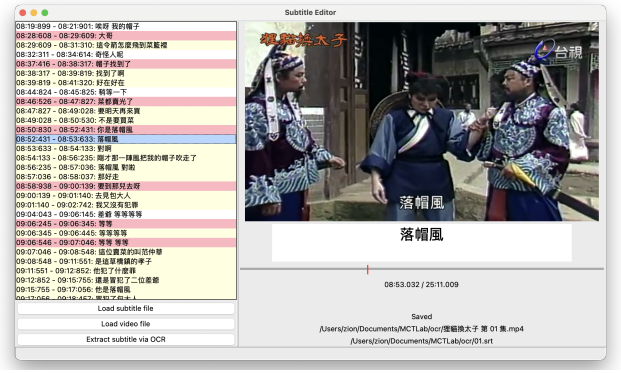

图 1。字幕编辑器的图形用户界面。

2. 字幕提取

2.1 通过光学字符识别进行文本提取

我们的字幕提取流程包含三个阶段：通过光学字符识别提取文本、时间合并，以及通过图形用户界面进行人工校正。为了以较高的准确率和较低的成本支持繁体中文识别，我们选择 EasyOCR [2] 作为系统的核心引擎。EasyOCR 不仅提供检测到文本的边界框坐标，还提供置信度分数，这些信息对于下游处理至关重要。为了兼顾计算效率与时间准确性，我们每隔 0.1 秒执行一次光学字符识别。此外，光学字符识别仅针对字幕通常出现的区域进行，从而减少其他屏幕元素（例如电视台标志）的干扰。我们保留 EasyOCR 检测到且置信度分数高于 0.01 的字幕字符串，然后按从左到右的顺序排列，并用空格连接。

光学字符识别阶段的输出包含大量

© T.-H. Huang、Y.-E. Tsai、Y.-N. Hung、C.-W. Wu、I-C. Wei 和 L. Su。依据知识共享署名4.0国际许可协议（CC BY 4.0）授权。**署名：** T.-H. Huang、Y.-E. Tsai、Y.-N. Hung、C.-W. Wu、I-C. Wei 和 L. Su，“台湾歌仔戏电视剧的定时文本提取”，载于第*26*届国际音乐信息检索学会会议迟交演示环节扩展摘要，韩国大田，2025年。

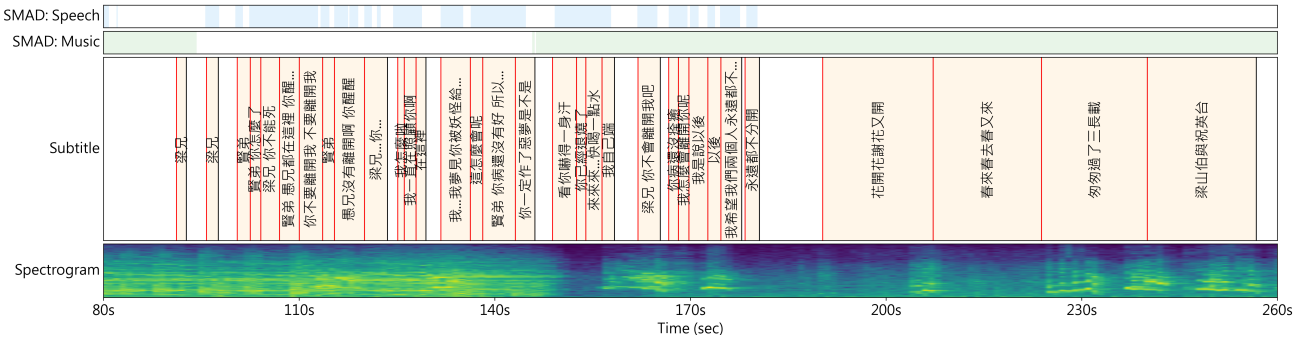

图 2。数据集中某一剧集的频谱图，以及估计得到的字幕、音乐和语音。

噪声，包括拼写错误、假阳性，以及各帧间断性的字符识别结果。为清理并整合文本序列，我们采用了以下时间合并机制。首先，仅保留时长超过预设阈值 t 的文本序列，以减少假阳性。此外，我们维护一份不太可能出现在字幕中的字符列表，并将这些字符从文本中删除。为对相邻的两个文本序列去重，我们根据编辑距离 [3, 4]和阈值进行相似度检查。具体而言，对于短字符串（$\leq$ 2个字符），要求完全匹配；对于中等长度的字符串（$\leq$ 6 个字符），允许编辑距离为 1；对于较长的字符串（$>$ 6 个字符），最大允许编辑距离为 2。如果两个相似的文本序列连续出现，则保留置信度分数较高的一个。此外，当基础文本（不包括省略号等细微差异）相同时，我们会将置信度分数更新为最高值，并保留较长的文本版本。如果上一条文本持续时间超过 t后出现新的、不相似的文本，则将先前的字幕片段定稿并保存。这种逐帧处理方式能够生成时间上稳定、一致性更高的字幕序列。目前这组参数的选择侧重于保留真正的阳性片段，即使因此产生一些假阳性也在所不惜；这些假阳性会在后续阶段予以过滤。

2.2 通过图形用户界面进行人工校正

为支持数据精炼最后阶段的高效人工校正，我们使用 Python tkinter [5]开发了专用的字幕编辑界面。如图 1 所示，该界面配备了键盘快捷键，轻量简洁，同时针对快速编辑进行了优化。界面采用双窗格布局：左窗格列出字幕片段，右窗格显示对应的视频帧、文本编辑器和时间轴控件。用户按下 e 键和 Esc 键，即可在查看内容时快速进入和退出编辑模式。为进一步简化合并流程，时间上相邻的字幕会以黄色背景标记，每个新序列的开头则以红色背景标示。这些视觉提示有助于用户快速识别并合并字幕片段。请注意，前两个阶段的所有功能均可通过“Extract Subtitle via OCR”按钮调用，从而在同一工具中实现三个阶段之间的无缝衔接。

3. 数据集准备

歌仔戏的声乐内容包含语音和歌唱两部分，我们的分析专门聚焦于歌唱。为分离歌唱片段，我们采用结合光学字符识别技术与 SMAD算法的两阶段方法，以加快处理速度。作为分离台湾歌仔戏表演核心元素的预处理方法，我们使用音乐源分离算法 [6]从原始混合音轨中提取人声轨。初步分析表明，由于声乐演唱中的旋律延长，歌唱片段的字幕显示时长（每个字符 0.9 秒）比语音片段（每个字符 0.4 秒）更长 [7], 因此，我们使用光学字符识别生成的字幕文件提取潜在的音乐片段：连续至少四个字符且显示时长超过 0.4 秒的字幕被判定为有效的歌唱内容。随后，我们结合 SMAD 结果，计算光学字符识别提取的片段与被判定为音乐内容的区段之间的时间重叠，以尽量减少慢速语音模式和器乐段落等造成的假阳性。图 2 展示了估算的 SMAD 结果和字幕信息。

结合 OCR + SMAD 的处理流程大幅加快了片段识别。从 220 集剧集（94.28 小时）开始，基于光学字符识别的分段生成了 1,798 个候选片段，将需要人工检查的素材缩减至 45.73 小时，减少了 51.5%。SMAD 筛选进一步将其缩减为 1,400 个片段（42.67 小时）。最终的听觉检查确认其中 841 个片段（35.68 小时）为真正的歌唱片段，时长介于 22.47–1,352.10 秒之间。现在，标注人员平均每集只需检查 6.4 个片段，而整个处理流程的精确率达到 60.1%。

4. 结论

本研究介绍了一种从低质量视频中提取字幕的交互式系统，并将其直接应用于制作歌仔戏数据集。通过所提出的方法，我们利用档案电视资料构建了一个数据集，为今后探索歌词识别和曲调检索等音乐信息检索任务创造了条件。除了歌仔戏之外，该框架也展现出处理中低质量历史电视内容的潜力，有助于推动数字人文和文化保护研究领域更广泛的应用。

5. 参考文献

[1] P. Hsu，“台湾歌仔戏：即兴、性别表演与传统选择”，博士论文，加州大学伯克利分校，2010年。[2] J. AI，“EasyOCR：即用型光学字符识别，支持多种语言和常见文字体系，包括拉丁字母、中文、阿拉伯文、西里尔字母、天城文等。”https://github.com/JaidedAI/EasyOCR，2025，访问日期：2025年8月3日。[3] V. I. Levenshtein，“能够纠正删除、插入和反转的二进制编码”，《苏联物理学报》，第10卷第8期，707–710页，1966年。[4] R. A. Wagner 和 M. J. Fischer，“字符串到字符串校正问题”，《美国计算机协会学报》（*JACM*），第21卷第1期，168–173页，1974年。[5]F. Lundh，“tkinter简介”，网址：*https:// www.pythonware.com/ library/ tkinter/ introduction/index.htm*，1999年。[6] “Moises ai”，https://moises.ai，2025，访问日期：2025年8月3日。[7]Y. Ozaki、A. Tierney、P. Q. Pfordresher、J. M. McBride、E. Benetos、P. Proutskova、……和 P. E. Savage，“全球范围内，歌曲和器乐旋律的速度比语音慢、音高更高，且音高更稳定：一项注册报告”，《科学进展》，第10卷第20期，文章编号 eadm9797，2024年。根据 APA第七版的建议，列出前6位作者，随后使用省略号，再列出最后一位作者。 [在线]。可访问：
h
t
tps://
www.science.org/doi/10.1126/sciadv.adm9797