人工智能语音合成技术在电影对白制作中的应用效果与优化策略

丁思立 吴 双 张嘉玮

北京电影学院声音学院，北京 100088

【摘要】为探究人工智能（AI）语音合成技术在电影标准监听下的呈现和表现出的具体问题，本文以实验性动画短片《小司令》对白制作为核心，对当前主流的开源和闭源语音模型进行了广泛的调研、比对与实验，并在此基础上总结了现阶段技术的共性特征，提出了遵循技术原理、经验证可实施的优化方案。研究表明，当前AI语音合成技术在电影声音制作中的应用仍存在某些特定问题，但综合能力已表现出可投入使用的潜力。制作者需理解基本技术原理及其与所合成声音的对应关系，依据需求灵活组合使用不同技术，并针对合成语音表现出的具体问题进行恰当的技术优化以达到较好的应用效果。

【关键词】人工智能；语音合成；语音转换；电影对白

【中图分类号】J90-05; J91

【DOI】10.3969/j.issn.1673-3215.2025.04.006

1 引言

随着深度学习（DL）技术的发展和硬件运算能力的提升，计算机从大量可用的训练数据中受益而开辟了诸多新的可能，这使基于神经网络的AI语音合成成为一种新的声音获取方式，实现了优于传统合成方法的语音质量和合成效率。

尽管取得了显著进展，但AI语音技术仍然远非完美。当前诸多模型表现出的有限细节呈现和可控程度成为其应用于电影对白制作的重要桎梏，这也是专业音频应用场景下各类AI生成技术所面临的现实问题。

从应用角度来看，目前主流的语音合成方法主要分为文本转语音（Text-to-Speech, TTS）[1]和语音转换（Voice Conversion, VC）[2]两种①，二者均存在一定的技术局限性。

对AI语音合成的评判标准往往优先遵循传统的客观评价指标，如关于频谱特征的梅尔倒谱失真（Mel Cepstral Distortion, MCD）、关于韵律特征的皮尔逊相关系数（Pearson Correlation Coefficient, PCC）等。这些指标量化的是声学特征与目标间的差异，并不总能与人类感知相关。部分辅以主观评价的研究虽有所改善，但由于对评价目标的设定较为粗略[3]，同样缺乏足够说服力。

不难发现，上述验证方式通常难以满足电影标准监听条件下还放的精度要求，其研究结果自然也无法准确对应标准监听下的实际听感，更无法暴露其实际应用效果与潜在问题。因此，在AI语音合成技术被引入电影声音领域的初期，实验性的制作验

【作者信息】丁思立（2000—），女，北京电影学院声音学院硕士研究生在读，主要研究方向：电影声音；吴双（2001—），女，北京电影学院声音学院硕士研究生在读，主要研究方向：电影声音；张嘉玮（2001—），男，北京电影学院声音学院硕士研究生在读，主要研究方向：电影声音。

证是尤为必要的。这不但有助于制作者们了解技术前沿的效果边界，明确不同技术的使用特征，更重要的是能在实际应用中发现其特定的优势和缺陷，进而探索AI技术在对白生成、声音处理和混录等各阶段的应用方案与优化策略。

本文基于对AI语音合成技术现有商业产品和学术研究成果的对比实验，以及在此基础上制作的动画短片《小司令》的验证性应用展开。作为一部实验性作品，片中使用的对白、动效和音乐等声音素材均由AI生成，编辑和混录工作由人工完成。本文集中讨论其中对白部分的制作情况。

2 AI语音合成技术发展现状

语音合成领域的研究正逐步走向精细化，衍化出在高质量、多情绪、多语种、低延迟、零样本学习（Zero-Shot Learning, ZSL）等多个细分方向的进一步探索。

微软与中国科学技术大学、香港中文大学等高校在提升语音质量和韵律方面研究的NaturalSpeech 3[4]、字节跳动与浙江大学面对当前Zero-shot TTS技术局限进行优化的MegaTTS 3[5]、阿里巴巴为流式模型最新改进的CosyVoice 2[6]以及谷歌针对目前多语种高质量训练数据缺乏问题开发的无监督数据扩展框架[7]等都是此发展阶段的典型代表，模型整体表现较之前有显著提升。与此同时，部分公司将技术成果转化为可付费调用的API接口，或将其作为新增功能嵌入到已有产品中。目前，国内AI语音合成技术处于服务大众化应用的阶段，国外则已显现出关注专业市场需求的趋势，技术发展速度带来的市场活跃度差异进一步引发了国内外声音制作行业对这一新技术需求强度和开放程度的不同表现。

《人工智能技术在电影声音制作中的应用与展望》[8]一文中总结了截至2024年5月国内外AI语音合成技术的商业产品及行业应用情况，在后续半年的发展中，国外出现了更多的应用实例，如游戏《赛博朋克2077》（*Cyberpunk 2077*）的音色转换[9]、电影《女巫游戏》（*The Witch Game*）的语种转换[10]等，可见该技术在专业音频领域应用的巨大潜力。但从此类局部角色、特定情况的使用模式中仍能发现其在规模化应用上存在不小的困难，除基于费效比（Cost-Effectiveness Ratio）的考量外，我们认为目前其与AI语音合成在技术层面的局限更为相关。

虽然国内外已有部分成功应用案例，但专门针对AI语音技术应用于电影声音制作的研究几乎处于空白状态，迄今为止未见具有权威性的、具有实践指导价值的文献发表。

3 AI语音合成效果对比与应用实践

为探究当前AI语音合成技术在电影标准监听条件下的还放效果及优劣势，我们从电影声音制作角度对目前提供该技术的开源和闭源模型进行了广泛筛选、实验和比对，结合TTS和VC各自技术优势进行混合应用并在此基础上总结规律，完成了《小司令》的对白制作。

《小司令》是一部时长5分钟的动画短片，讲述了一位中年男性在回到拆迁的老房子中取回玩具时睹物思情，回忆起自己儿时扮演小司令带领两名玩具士兵抵抗邪恶的大公鸡、保护家园的故事。选择该片作为实验对象的核心原因在于其设定角色的对白在数量、情绪和风格多样性等方面均存在挑战，有助于更有效地检验当前AI技术在语音合成应用时的综合表现。

3.1 应用效果比对

判定AI合成语音在电影声音创作中能否有效应用的评价维度较多，不同的评价方法侧重点并不相同。现阶段我们在最基本也最显性的6个方面进行了对比实验和分析。

3.1.1 技术指标

技术质量是电影声音制作最基本的要求，也是我们初期筛选可行模型的首要依据。客观上可参考采样频率（Sample Rate）及量化位深（Bit Depth）两项指标做出初步判断。

在前期实验阶段，绝大部分模型的技术指标仍无法达到电影声音制作理想的不低于24 bit/48 kHz的要求。综合评估后，我们最终保留了标示及实测值均达到或超过16 bit/44.1 kHz的模型以待进一步验证，其中包括ElevenLabs、Replica Studios、Speechify、Respeecher、火山引擎、阿里云Sambert、谷歌的Google Cloud等。另外，我们额外保留了仅支持16 bit/24 kHz的ChatTTS。一方面是基于其在自然度上的良好表现；另一方面也考虑到开源模型所具备的可调节性和迭代潜力。

需要注意的是，部分模型合成语音的实际频率覆盖范围与其注明的采样频率上限并不相符，实测中无论客观频谱还是主观听感都不及标称指标。

3.1.2 情绪表现

TTS和VC由于不同的模型架构而拥有不同的技术特征和适用场景，其中最关键的差异在于语音情绪的来源不同，这也直接影响了我们对两种技术的选择和应用。对于TTS而言，情绪与自然韵律是最难处理的问题。当前大部分成熟的模型面向的是导航、有声书等通用市场，在满足音质需求的前提下会优先考虑运算效率与成本，对于强变化、多细节这类电影声音制作层面的需求并不侧重，因此Speechify、Murf.ai、Google Cloud、Sambert等模型没有被最终选用。虽然他们的技术指标基本达标，在情绪控制方面也提供了选择，但合成语音大多较为平稳甚至单调，缺乏愤怒、狂喜、感动、痛苦、喊叫、耳语等张力较大的复杂情绪，难以匹配电影表演所需。部分公司已针对此问题进行优化，如Replica Studios和Altered Studio，其情绪模块会将音色和情绪彼此独立以提供更多选择余度，这不但能在改变情绪的同时保持角色音色的相对统一，还可通过相关参数对情绪特征和强度进行进一步调整。Replica Studios甚至能在Voice Lab模式下支持用户使用文本描述目标音色特性，后将四个空槽与主导槽融合为一种音色，再为这个形成的全新音色赋予选定的情绪。这些设计在一定程度上弥补了TTS技术在情绪表现方面的局限。

相较之下，VC技术在自然度、变化性等方面具有天然优势，能较好地承袭输入语音的细节特征[11]，尤其是重音、停顿、语调等对于情绪呈现至关重要。目前Replica Studios、Replay在这方面做得较好，但ElevenLabs、Altered Studio和火山引擎等诸多模型在输入与输出信号间存在着明显的情绪误差，难以实现理想的转换效果。

3.1.3 语种选择

语种选择余度是我们选定制作模型的最终参考。《小司令》自2024年4月进入制作阶段，基于特定剧情与画面风格，我们将中文作为首选语种并进行了大量调研和效果比对，但实验结果并不理想。实验初期，如表1所示，各模型对中文的支持度普遍偏低，因技术指标达标而被纳入考虑的多个模型中仅有部分提供了对中文的支持能力，其中Murf.ai仅有几个中文音色备选并且语调稍机械；Speechify支持豫、辽、川、鲁等多地方言但偏向自然朗读风格；Replica Studios、Altered Studio、PlayHT均不支持中文。Respeecher较为典型，虽提供了中文输入选项，但仅有5种中文音色备选，而英语发音则多达100余种且音色变化多样，甚至专门提供印、韩、英等国别差异的发音，丰富性明显优于中文。

考虑到彼时AI技术在中、英文语音发展的成熟度及其在比对试验阶段表现出的明显质量差异，我们最终决定所有对白生成均以支持度更高的英语进行，以便尽可能利用更广泛的技术资源进行实验。

表1 典型语音合成工具的中英文支持情况

（截至2024年10月）

典型厂商/模型	中文支持	英文支持
ElevenLabs	√	√
Replica Studios		√
PlayHT		√
ChatTTS	√	√
Respeecher	√	√
Altered Studio		√
Speechify	√	√
Murf.ai	√	√

3.1.4 便捷性与可控性

人工智能生成语音所涉及的真正困难不是制造一个产生语音的装置，而是操纵这个装置[12]。从应用角度而言，“操纵”涉及便捷性与可控性两方面，分别对应实际制作中的效率和精度。

3.1.4.1 便捷性

就TTS技术而言，目前生成便捷性较为优秀的产品源自英国的Sonantic公司②。Sonantic AI的工作界面近似于数字音频工作站（DAW），横轴为时间，纵轴轨道对应不同音色，每条输入文本生成的对应语音被直接放置于固定轨道的指定时间点上，每次系统还会额外生成三条语调、节奏均有变化的备用版本存放于播放列表（Playlist）中[13]。对于无法实现的特殊发音或语调，用户还可手动绘制曲线指定读音，对于电影声音制作者非常友好[14]。虽未按照原始路线继续前进，但Sonantic早期的技术成果和行业关联性仍令人印象深刻。

ElevenLabs在便捷性上有待改进，实验周期内的

持续跟踪表明，其从早期简单的“一键生成模式”逐步向DAW工作方式演进，以更好地贴近用户需求。

相比之下，Respeecher等目前仍没有针对此问题的优化操作，对于大多声音制作者而言并不是理想的选择。与行业工作习惯的脱节一定程度上降低了相关工具在专业音频制作领域的接受度。

3.1.4.2 可控性

可控性是声音制作者关注的另一个重点，这关乎制作者是否能在音频文件生成之前将模型调节到近似理想的目标，从而实现相对准确的参数化控制。目前大部分公司都提供了不同强度和精度的控制参数，设置上各有侧重。ElevenLabs面向通用市场，参数设置也偏向主观通俗，例如对情绪夸张（Style Exaggeration）程度的调节、对合成语音与源语音相似（Target Speaker Similarity）程度的调节等。Altered Studio以其内嵌的音频编辑套件为特色，参数强调声音处理的客观细节，能在生成后继续对单条语音实现均衡、去齿音等处理；Replica Studios更多与游戏产业合作，其参数相应地更注重有效性，音色和情绪完全由预置确定并提供了音调（Pitch）、速度（Pace）、音量（Volume）三个参数供进一步调节。

开源模型的操控则呈现另一种风格。由于并不直接面向应用端而是科研人员、开发人员或具有技术能力的终端用户，因此其参数调节所指代的并不是声音制作语境下的调节，而是针对算法模型的调节。例如ChatTTS中用于控制生成随机性的Temperature、控制采样累计概率的Top P、控制截断采样的Top K③等[15]。这些参数是对模型算法的有效控制，对生成过程有重要的影响，但在合成声音的过程中并没有能与之准确对应的实际含义，这不仅会给声音制作者带来理解上的困难，还会进一步导致模型生成规律无法明晰，即这些参数在实际制作中无法起效，无疑在应用层面构成明显劣势。

从算法原理角度看，VC技术理应能实现跨语种的自如转换，可控性应该TTS方式更高。但在实际制作中我们发现，由于中文语音训练数据不足，很多模型转换后的音频会带有明显的英语发音习惯，甚至会将中文直接错误地转换为相似发音的英文，如将“闺蜜”转为“Green me”，将“快来吧”转为“Queen Lab”等。由此可见，VC的低可控性为实际制作带来了无法回避的问题。

3.1.5 频谱与瞬态表现

3.1.5.1 频谱

TTS和VC均依赖声码器将特征参数转换为信号波形并输出声音，因此在频谱与波形上存在诸多共性问题。缺陷首先体现在高频细节方面，这是当前AI技术的薄弱环节，也是最终导致合成语音技术指标和自然度不足的原因之一。大部分主流语音合成模型在12 kHz以上的高频能量比较弱或完全缺失，影响了声音的细节展现。这一问题的影响因素除声码器自身限制外，由于高频信息在自然语音中的能量本身较弱，且训练数据中高频信息较稀疏，因此，大多数声码器的频率建模侧重于基频和共振峰等对语音清晰度和可懂度影响更大的频段（500 Hz~3 kHz）。

另外，特征表示（如梅尔谱、线性谱等）的选择也会对高频生成质量产生影响。在使用梅尔谱作为声码器的输入时，频率对数尺度的压缩也在一定程度上限制了高频细节的恢复。

除高频缺失的技术缺陷外，高频冗余同样会导致音色不自然的问题。例如Replica Studios虽覆盖了20 Hz~20 kHz的完整频率范围，但从图1的声谱图对比可见，右侧TTS合成语音在高频段存在多余能量成分，在6 kHz、12 kHz两个频点以上仍不合规律地合成了能量更强的信号，并非如左图实录语音一样形成高频能量的自然滚降。虽然从模型角度已经认识到这种缺陷，也试图通过算法进行补偿，但截至2024年10月实验期间，高频的不足或冗余依然导致AI合成语音在听感呈现上自然度欠佳。

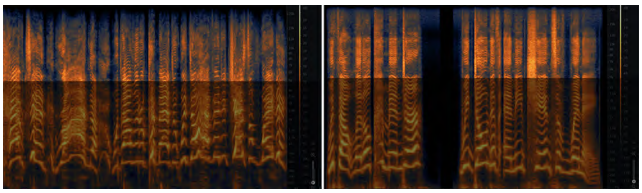

图1 高频冗余问题的声谱图对比

此外，AI合成语音信号还普遍表现出谐波结构不清晰的问题。以GPT-SoVITS[16]为例，在使用实录语音训练模型完成音色学习后，输入与录音内容相同的文本，让模型采用TTS方式以习得音色再次合成语音时，合成语音与实录语音虽在音色特征上能达到较高相似度，但声音的细腻程度仍存在差别，听

感较同等带宽的实录语音显得更加沉闷、模糊，尤其在清亮的元音（如/i/、/e/）或细节丰富的辅音（如/s/、/sh/）中表现尤为明显。从图2所示声谱图的谐波序列比对中也可明显看出，右侧经实录语音训练后所得的同音色合成语音存在低次谐波能量欠缺、共振峰强度不足，中频及以上谐波不清晰、频谱过于平滑，较左侧实录语音缺乏很多细节。

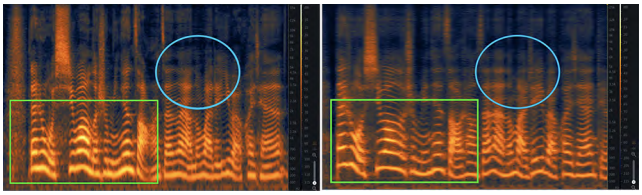

图2　声谱图谐波细节对比

3.1.5.2 瞬态

在瞬态展现上，AI合成的语音信号同样表现出劣势，包括尖音、齿音在内的音头瞬态部分力度明显不足。图3上下分别展现了实录语音和同音色TTS合成语音的波形情况。不难发现，合成语音字词间的能量分布较均匀，整体缺少明显的强弱对比。为提升瞬态特征的解析能力，在图4中通过调整短时傅里叶变换（STFT）参数，采用侧重时间分辨率的高斯窗函数，缩短窗长并增加窗口重叠率，能更明显地看出右图合成语音在音头瞬态的高能量区域并未表现出如左图实录语音一般应有的突出性，导致最终听感缺乏层次感，语音力度、清晰度、自然度都难以达到理想目标。除主观听感和视图判断外，此特征在表2的数值比对中也得到了证实，从两信号瞬态区域内200 Hz、500 Hz、1 kHz、5 kHz四个典型频点所在频率单元的能量差异中，同样能发现合成语音在瞬态方面的欠缺。这一问题目前在各公司的模型中普遍存在，但Replica Studios和ElevenLabs的表现稍好。

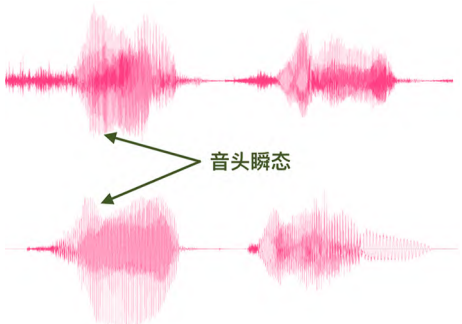

图3　瞬态问题的波形对比

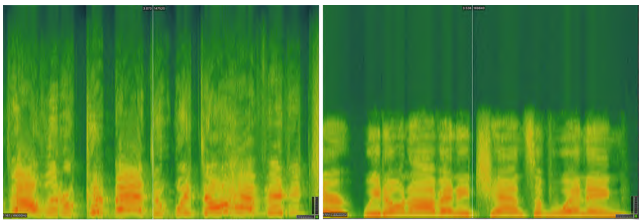

图4　瞬态问题的声谱图对比

表2　四个典型频点所在频率单元的能量对比

频率单元	实录语音	合成语音
200 Hz 所在单元	−64 dB ~ −62 dB	−72 dB ~ −70 dB
500 Hz 所在单元	−63 dB ~ −58 dB	−71 dB ~ −58 dB
1 kHz 所在单元	−73 dB ~ −63 dB	−86 dB ~ −73 dB
5 kHz 所在单元	−77 dB ~ −65 dB	−108 dB ~ −98 dB

3.1.6 音色统一性、稳定性

在当前的技术阶段，无论是使用TTS还是VC技术，合成语音的音色相似度与自然度均达到了一定水平，对统一性和稳定性的要求成为AI语音合成技术面临的下一个难题，这也决定了该技术能否真正大规模投入电影声音的工业化生产。

3.1.6.1 统一性

统一性问题体现在应用层面。目前单一技术或单一模型很难满足电影对白多角色、多变化的生成需求，需同时调用TTS和VC两类技术分属的多个模型混合应用，发挥各自优势以达到较好结果，而多模型混用客观上必然出现音色、音质等不统一的问题。

局部来看，即便在同一个模型中，不同输入信号也可能导致输出信号听感的不统一。在实验早期阶段，TTS偶尔会出现个别音节的音色跳变，而VC在统一性方面的问题则更加明显。面对不同输入音色，VC往往难以实现完全一致的输出，尤其以男、女声分别将相同录音内容输入同一模型时，输出语音在音色细节方面仍有差别。可见，跨性别VC的差异性仍是当前该领域的一个较大问题，大多数情况下需同性别语音输入才能实现更好的效果。

3.1.6.2 稳定性

稳定性体现在技术层面，主要指变化性强的输入语音是否会影响模型正常工作，这一问题通常出现在VC技术中。尽管VC理论上能准确传递原说话者的语气、语调和语言内容，但实践中发现，当以极端情绪的语音或喊叫、喘息等非语音成分作为输入时，几乎所有的VC模型均表现出了极强的不稳定性，输出语音存在明显的失真、间断和基频识别错误

问题。

与稳定情绪状态下的语音信号不同，极端情绪下的语音及非语音成分在频谱上有其独特性，谐波结构相较于稳定、连续、规律的常规语音存在一定差距，难以获得较好的算法实现。例如，喊叫声的基频波动范围大，能量集中在中频段且瞬态信息丰富；喘息声则几乎没有稳定基频，能量分布宽广但稀疏，而高频部分随机性强。二者在振幅、音调、共振峰等各方面都具有强烈的变化性，这种异常的动态特性导致模型难以适应并正确映射到目标特征中，最终表现为音色偏离、失真、信息丢失或随机噪声。如ElevenLabs频繁错误地转换基频且动态压缩严重；Altered Studio生成内容的情绪信息被明显削弱；Respeecher忽略了大量低电平信号且无法识别输出；Replica Studios则出现了音色跳变和失真等。上述问题与训练数据的多样性覆盖范围和模型处理能力有很大关系。

因此，目前的VC尚难以稳定输出高质量的目标语音，仅在情绪比较稳定的同性别声音转换中效果较好。相对而言，TTS输出音频质量的统一性和稳定性更可信，因此从实际应用角度而言更需调用两种技术协同工作，在统一音色和音质的前提下实现优势互补。

3.2 问题与优化

尽管迭代后的AI语音合成技术已取得巨大进展，但其对应的模型算法还面临着诸多问题。我们关注的是合成语音在声音制作阶段呈现出的共性缺陷并尝试总结有效的优化策略。

3.2.1 高频缺失优化

针对高频缺失的问题，传统解决思路是使用均衡器提升高频，但这种方式对AI合成语音的有效性较低。如果合成语音中存在高频信息，高频提升稍有作用；如果合成信号本身就高频缺失，提升高频则适得其反，往往导致噪声电平的提升。因此，可考虑使用激励器，依靠低次谐波的能量产生高次谐波成份，采用多段激励能针对指定的高频段进行谐波激励，弥补高频不足的问题。

使用饱和失真也是有效手段之一，微量的饱和失真能够在高频段产生适量的高次谐波。虽然与使用激励器弥补高频的原理不同，但最终都能获得一定的高频细节。上述两种方法的共同缺陷是所补充的高频成分自然度较弱，高频略显僵硬、呆板，与中低频的融合度不佳，变化不够细腻。

实践中可考虑的非传统优化策略是利用频带展宽技术（Bandwidth Extension, BWE）[17]进行高频信号的数字再合成。iZotope RX的Spectral Recovery模块及其他有效的信号合成工具均可作为选择，再辅以均衡等手段往往可获得相对自然的效果。

非传统优化策略的另一种思路则是利用AI技术进一步完善AI语音。Accentize公司的dxRevive及开源模型Resemble AI Enhance模块[18]、VoiceFixer[19]、AudioSR[20]等对高频信息都有一定的重构能力。图5为《小司令》某句台词高频重建前后的效果对比，通过频谱线性分析，左侧为原始TTS语音声谱图，右侧为TTS语音输入Resemble AI Enhance模块完成高频增强后的声谱图。图中圈定部分的声谱信息表明，频谱密度和频率响应均出现了明显改善，尤其是在圈定区域以上的频率成分还出现了增加的泛音能量，该部分能量表现出与自然语音相近的自然滚降趋势。虽然从整体听感呈现来看，这种方法对于AI语音高频信息的补充效果并不一定能达到完美的程度，但已是目前较为有效的优化手段。

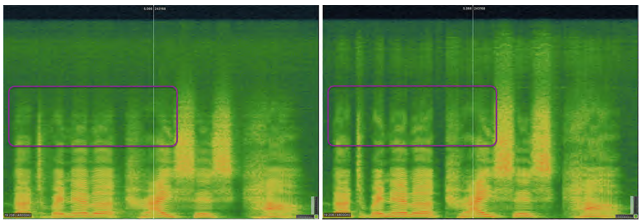

图5 高频重建效果的声谱图对比

无论采用何种高频补偿手段，都没有从本质上改变谐波间及音素间的相对比例关系，因此谐波序列自然度不足及瞬态缺失问题仍然存在。针对谐波序列的调整可考虑反常规使用iZotope RX进行处理，也可借助其他信号合成方式进行谐波序列微调。

3.2.2 瞬态优化

针对瞬态不足的问题，依靠传统手段的多段动态处理（Multiband Dynamics）是可选方案之一，目的在于针对性调整合成语音的子频带动态。对瞬态缺失的子频带进行动态扩展能在一定程度上强化AI合成信号的低电平细节，改善瞬态响应。

优化瞬态的另一种方法是瞬态整形（Transient

Shaping)，对音头瞬态进行单独调整，在强化音头细节的同时改善信号的动态响应方式。这种方法不但能弥补瞬态不足，还可对瞬态和延音进行适度的包络调整，可控性较多段动态处理更强，调整后的声音自然度也相对较高，图6为使用Transient Shaper工具进行瞬态整形前后的效果对比，从右图增强部分可见，利用该技术能够在音头部分对全频带都做出相对有效的补充。

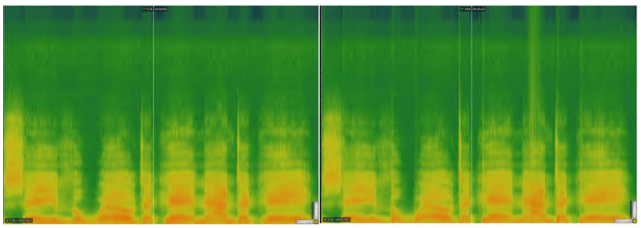

图6 瞬态整形前后的声谱图对比

3.2.3 情绪表现优化

3.2.3.1 TTS

自然语言中也包含了大量非语言内容，对于仅能以文本作为信息来源的TTS具有很大困难。关于语气词、呼吸等口语化内容，常规的文本输入方式通常难以达到目的，通过发音口型对应的拼音或英文字母（如“woah”“aeeeee”）进行补充是有效的优化方法之一。该方法对于让模型发出非常规词汇或特殊发音也有一定帮助。但从目前来看，如果语音情绪相对稳定，使用VC技术仍是相对理想的解决方案。

除了非语言内容的生成外，即便在同一文本、音色和情绪选项下，输出音频在语气、语调上依然会表现出差别。这些隐含在文字内容中的情绪细节是更难解决的问题，因为语音的节奏、停连、重音、速度、韵律等许多方面往往互相作用，共同形成了最终的情绪。

情绪细节的缺失可通过修饰对白文本和调节参数两种手段进行优化。标点符号的选择及使用频次可明显影响情绪，在句中加入合适的标点符号或空格还对改善语音节奏有所帮助。对于一些特定设置的模型，尝试以“小说体”文本输入也能提供额外的情绪信息提示[21]，如在ElevenLabs的使用中，“‘Don't give up your faith!’The captain shouted loudly.”就会比单纯的“Don't give up your faith!”带有更多呼喊的语调。

预处理参数的调整也可明显提升情绪表现的准确性。这是因为很多参数虽看似与传统声音处理参数无异，但在AI系统中，显性参数的调整往往会引发多个相关隐性参数联动，共同影响情绪表现。例如Replica Studios中的速度参数在调为正值时会为输出语音增加急促的情绪，其影响的是节奏、重音、语调等多方面而非仅仅体现在加速播放。在生成《小司令》队长角色的某段呼喊时，音色和情绪均可基本达标，但整体感觉仍不够急迫，此时提高速度和音调参数可产生较好效果。也就是说，这些过程性参数最终实现的是对情绪的附加调节能力。在使用提供此类参数的AI模型时，制作者可尝试大胆修改参数，这些调整不仅不会降低音频质量，在参数量值匹配良好的情况下反而能获得更加符合需求的情绪。

从根本上来讲，以上方法均限制在TTS系统内，尝试通过机器能够理解的语言为模型提供更多附加信息，进一步控制生成结果。

3.2.3.2 VC

针对VC的优化策略首先可考虑选用多个VC模型而非固守在某一模型上。不同模型内部算法并不完全一致，通过某一模型难以转换的声音在其他模型上并不一定失败；其次，增加合成数量并挑选适宜情绪而非执着于固定情绪同样有助于提升成功率。随着输入信号的变化和合成数量的增加，挑选出适宜情绪的可能性也在增加，形成以数量换取质量的局面；最后，在技术可接受的条件下，不妨适度降低技术指标以优先保证生成语音的情绪价值。只要技术指标能达到基本要求，即便后续对“富有情绪”的素材继续进行再合成或技术优化，也比生成一条技术指标优良却缺乏情绪的语音更具意义。

因此，对AI合成语言的情绪把控不能困于某一固定模式中，也不能以一个特定情绪的获取为最终目标，更不能因AI无法实现自己固守的情绪而轻易否定它。理性的方式应在已生成的且数量充足的声音素材中，挑选出最可行、最契合需求的声音，在技术指标与声音的真实、自然和情绪表达间作出权衡和妥协。

3.2.4 音色衔接优化

为从源头上避免听感不统一的问题，制作中应尽量选用同一模型完成同一角色语音的合成任务，而在不同角色的塑造上则可尝试利用不同模型，这

种模型差异带来的音色差异也是不同角色塑造所需要的。

多种技术手段或公司不同的组合使用带来的音色衔接问题相比前述问题稍好解决，在音色差异不明显的情况下，优化策略首选传统的声音处理手段，而差异明显时可考虑以TTS训练VC的方式完成补充，即继续对TTS输入额外的文本信息，并以该音色生成一批尽量符合目标情绪的语音素材，之后对其进行筛选、频带扩展、音色处理、段落拆分后作为数据输入VC完成音色训练。此时，仅需对个别语句进行实录，之后利用VC转换音色，便可获得与TTS音色相近且带有实录人声情绪的声音素材。这种方法是对两种不同AI语音合成方式的串联应用，是典型的以“AI训练AI”。

3.3 制作效果

经过对AI语音合成技术的研究、实验以及对以上优化方法的综合性运用，《小司令》制作过程中在一定程度上弥补了音色瑕疵、表现力不足等在目前阶段将该技术应用于电影声音制作时不可避免的缺陷，在多模型协同使用的条件下最终达到了相对一致的对白呈现。尽管在细节层面仍存在一定不足，但整体已达到较为理想的效果，基本实现了验证性制作的研究目标。

4 总结与讨论

事实表明，AI语音合成技术初步具备了投入应用的技术条件，模型也表现出与电影声音制作相匹配的行业关联性，规模化应用的大门正缓缓打开。在庆祝技术进步的同时，本文也揭示了AI技术以及脱离实际听感与标准监听环境的评估方式在电影声音制作中的局限。本文通过完整影片制作流程的应用实践，在真实的创作场景下，以符合电影录音工艺的标准检验了当下AI语音合成技术的发展情况与实际效果，更直接地暴露出AI语音在电影语境中的适应性问题。尤其是在面对情绪表达、有效控制、音色细节呈现等专业需求时，AI语音所暴露的一系列具体问题仍构成了其从“可用”迈向“应用”的核心障碍。

我们在短片《小司令》实践中总结的部分经验是基于传统声音处理观念和数字信号处理技术的尝试，其在AI合成信号中实现的良好效果也为我们提供了启示，即从听感呈现和问题本身出发，对AI合成语音的处理不能完全照搬传统方法和思维惯性，应以新的视角看待新技术表现出的各类问题，依据具体问题寻找可能适宜的技术手段，并在技术指标和声音艺术性之间做出适当妥协。

需要说明的是，在AI技术飞速发展的今天，文中提到的一切技术瓶颈均存在时效性，TTS与VC各自的特征与优劣也并不绝对。当《小司令》制作完成时，国内外多个AI模型的中文支持能力已大幅提高，可见技术层面的演进仍在加速，其在电影声音制作中的潜力远未被完全挖掘。未来，随着模型算法的优化和应用方面的探索，AI语音合成技术有望突破各类局限向更加精细化和个性化的方向发展。在这一进程中，AI语音技术能否真正融入专业声音体系，取决于其在艺术表达与技术控制之间的平衡能力，这一平衡的探索将不仅关乎算法优化，更关乎声音创作思维的重构。❖

注释

① TTS：一种将输入文本信息转换为自然语音的技术，是语音合成中最常见的形式。其通过分析输入的文本，将其转换为音素序列，再通过声学模型和声码器生成语音。VC：一种将输入语音信号中关于原说话者的相关特征（如基频、共振峰、语调、韵律）修改为目标说话者的相关特征，同时保持说话者无关信息（语言内容）不变的技术，主要应用于语音风格转换、情感转换等场景。

② Sonantic：成立于2018年的英国初创公司，专注研究具有表现力的语音合成技术，于2022年6月被Spotify公司以9300万英镑收购。

③ Temperature、Top P、Top K：概念源于自然语言处理（NLP）领域，是控制模型生成过程的三个关键参数，彼此间相互协调作用。Temperature（温度值）控制模型生成内容的随机性；Top P也被称为Nucleus Sampling，通过限制达到累积概率阈值的动态子集来控制输出的多样性；Top K则通过限制模型候选词数量来控制多样性。

参考文献

[1] TAN X. Neural Text-to-Speech Synthesis[M]. Singapore: Springer Singapore, 2023.

[2] BERRAK S, JUNCHI Y, SIMON K, et al. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021(29):132-157.

[3] SEYED H M, ALEXANDER K. An Overview of Voice Conversion Systems [J]. Speech Communication,2017(88): 65-82. DOI:10.1016/j.specom.2017.01.008.

[4] JU Z, WANG Y, SHEN K et al. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models[EB/OL]. (2024-04-23)[2025-01-22]. https://doi.org/10.48550/arXiv.2403.03100.

[5] JIANG Z, REN Y, LI R, et al. MegaTTS 3: Sparse Alignment Enhanced Latent Diffusion Transformer for Zero-Shot Speech Synthesis[EB/OL]. (2025-03-28)[2025-04-14]. https://arxiv.org/html/2502.18924v4.

[6] DU Z, WANG Y, CHEN Q, et al. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models[EB/OL]. (2024-12-25)[2025-04-08]. https://arxiv.org/abs/2412.10117.

[7] TAKAAKI S, GARY W, NOBUYUKI M, et al. Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data[EB/OL]. (2024-07-16)[2025-04-08]. https://doi.org/10.48550/arxiv.2402.18932.

[8] 石宝峰, 丁思立. 人工智能技术在电影声音制作中的应用与展望[J]. 现代电影技术, 2024(10):50-57.

[9] How Respeecher and CD PROJEKT RED Preserved the Voice of Cyberpunk 2077's Viktor Vekto[EB/OL]. (2024-08-07)[2024-12-07]. https://www.respeecher.com/case-studies/tag/game-development.

[10] JJ Dorfman. Harry Potter-Inspired Horror Movie Dubbed With AI Gets Picked Up for Release[EB/OL]. (2024-12-03)[2025-03-24]. https://www.cbr.com/harry-potter-inspired-horror-movie-dubbed-with-ai/.

[11] CHILDERS D G, WU K, HICKS D, et al. Voice conversion[J]. Speech Communication,1989.8(2):147-158.

[12] STEWART J Q. An electrical analogue of the vocal organs[J].Nature, 1922 (110):311-312.

[13] Sonantic. How Sonantic AI Voices Work[EB/OL].(2021-03-02)[2024-10-28]. https://www.youtube.com/watch?v=fNtwg-lXie8.

[14] ZEENA QURESHI. Q&A: The future of synthetic audio[EB/OL]. (2020-11-07) [2024 - 11 - 03]. https://www. samsungnext. com/blog/q-a-with-zeena-qureshi-sonantic.

[15] LIBUKAI. Awesome-ChatTTS[EB/OL].[2024-06-21]. https://github.com/libukai/Awesome-ChatTTS/tree/en.

[16] NAOZUMI520. RVC-Boss/GPT-SoVITS[EB/OL].[2024-12-25]. https://github.com/RVC-Boss/GPT-SoVITS.

[17] ARCHITT G, BRENDAN S, YANNIS A, et al. Speech Bandwidth Extension with Wavenet[J]. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2019:205-208.

[18] Resemble AI. Resemble Enhance[EB/OL]. [2024-06-22]. https://github.com/resemble-ai/resemble-enhance?tab=readme-ov-file.

[19] LIU H, KONG Q, TIAN Q, et al.VoiceFixer: Toward General Speech Restoration with Neural Vocoder[EB/OL].(2021-09-28)[2025-04-10].https://arxiv.org/abs/2109.13731v3.

[20] HAOHE L, KE C, QIAO T, et al. AudioSR: Versatile Audio Super-resolution at Scale[C].2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024:1076-1080.

[21] ElevenLabs. Speech Synthesis Prompting[EB/OL].[2024-10-22]. https://elevenlabs.io/docs/speech-synthesis/prompting.

作者贡献声明：

丁思立：设计论文整体框架与实验，论文撰写与修订，全文文字贡献50%；

吴双：参与讨论与实验，撰写部分论文，全文文字贡献30%；

张嘉玮：参与讨论与实验，全文文字贡献20%。

Application effects and optimization strategies of AI speech synthesis technology in film dialogue production

◎ DING Sili, WU Shuang, ZHANG Jiawei (Sound School, Beijing Film Academy)

Abstract: To investigate the presentation and specific issues of artificial intelligence (AI) speech synthesis technology under standard film monitoring conditions, this research focuses on the dialogue production of the experimental animated short film *Little Commander*. A comprehensive survey, comparison, and experimentation were conducted on both mainstream open-source models and proprietary models. Based on these findings, the research summarizes the common characteristics of current technology and proposes validated, technically grounded optimization strategies. The results indicate that while AI speech synthesis technology still presents certain challenges in film sound production, the overall capability demonstrates its potential for practical application. Filmmakers need to understand the fundamental technical principles and their correspondence to the characteristics of the synthesized speech, flexibly integrate different technologies according to production needs, and apply targeted optimizations to synthesized speech to achieve optimal results.

Keywords: Artificial Intelligence; Speech Synthesis; Voice Conversion; Film Dialogue