人工智能技术在电影声音制作中的应用与展望

石宝峰　丁思立

北京电影学院声音学院，北京 100088

【摘要】为了能够对人工智能技术在电影声音制作中应用的现实情况和未来发展做出比较明确的判断，本文通过对目前国内外已有的、相对可行的技术方案和文献资料进行广泛搜集整理，列举了技术相对成熟的国内外公司产品技术特点，同时对人工智能技术在电影声音对白和效果类声音应用的前景进行了研究，提出了该技术在电影声音制作中应用的发展方向，目的在于对人工智能技术在电影声音中的应用作出客观评价。研究表明，人工智能技术虽然在短期内不会彻底改变电影声音制作流程，但随着技术水平的提升，该项技术的引入或将在一定程度上改变电影声音制作的现有形式。

【关键词】人工智能；电影声音；降噪；对白；效果

【中图分类号】J90-05；J91

【DOI】10.3969/j.issn.1673-3215.2024.10.008

1 引言

随着人工智能（AI）技术的发展，尤其是多模态大模型的快速进步，人工智能技术已经不限于在文字对话层面的简单应用，基于音视频的应用开始成为人工智能技术的一个重要发展方向，人工智能技术在电影声音制作过程中的应用成为可能。

北京电影学院声音学院于两年前开始关注人工智能技术在电影声音制作领域的应用前景，跟踪研究的目的在于探索人工智能技术对电影声音的制作观念、制作方法和工艺流程会产生哪些影响，影响程度有多深以及介入速度有多快，是否会颠覆目前电影声音制作的技术格局。

通过一年时间系统跟踪国际、国内人工智能技术发展状况可以发现，人工智能生成内容（AIGC）技术确实对传统声音制作观念和工艺流程产生了一定程度的影响，而且发展迅速。但同时，人工智能技术也并非十全十美。通过对其全面科学的研究评估可以发现人工智能技术长项与短处并存。

关于此类问题已有大量文献进行了相关研究，不同作者从各自角度做出了具有价值的研究。但是，能够从宏观层面总体分析研究人工智能技术在电影声音制作这个特定领域的文献数量相对较少。本文的研究目的是希望通过我们的阶段性研究成果，为人工智能技术未来大规模进入专业声音制作领域的发展应用提前做出研判。

2 人工智能技术在电影声音制作领域的应用现状

人工智能技术在电影声音领域的应用表现出明显的两极化特征。虽然人们希望借助人工智能技术解决制作过程中遇到的棘手问题，突破传统技术手

【作者信息】石宝峰（1970—），男，硕士，北京电影学院声音学院副教授，主要研究方向：电影声音；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。

段的羁绊，但暂不成熟的人工智能生成技术限制了其在电影声音制作各细分领域的广泛应用，而以“智能降噪”为代表的人工智能技术则已经在电影声音制作领域实现了规模化商业应用。近年推出的国产电影中，相当一部分影片的对白制作都会应用iZotope RX进行降噪处理，使用该软件进行对白降噪处理的方法已使用多年，成为声音制作工艺流程中不可或缺的重要环节。

成立于2001年的iZotope公司在其产品中很早就融入了以机器学习（Machine Learning，ML）为主的“智能”（Intelligent）技术，其理论基础是基于神经网络驱动滤波器的大数据算法。作为相对传统的效果器公司，iZotope融合智能技术的目的在于对输入声音进行信号分析并提供具有一定针对性的处理方案，核心目的是减少声音处理的复杂流程并提高声音处理的成功率。在RX10和RX11两个版本中，iZotope RX在保持核心程序传统降噪方法的基础上加强了人工智能技术的应用，但人工智能降噪技术依然是以提升质量和效率的辅助功能形式出现。iZotope的技术基础决定了其核心降噪功能并不完全依赖于人工智能技术的存在①，但在其发展过程中新增的，针对某些特定类型噪声处理的组件设计则呈现出选择性使用人工智能技术以强化降噪能力的特征。对于RX而言，人工智能技术的融入是“锦上添花”，是功能增强，人工智能技术并未撼动iZotope RX“以人为主”的应用模式。

Accentize系列产品诞生于2019年，采用了基于神经网络再合成的建模算法，是针对后期制作专业应用开发的产品。虽然iZotope与Accentize都采用了人工智能技术，但两者的技术路线并不相同。Accentize系列产品几乎完全依靠算法模型进行降噪处理，降噪能力更多取决于算法模型本身而非人为的参数调整，其中内置的多种算法模型还可以对应不同情况的带噪语音信号进行更优质、更适配的处理。同时，Accentize产品在降噪后依靠算法模型实现的信号合成和音色调整功能表现出较强的声音修复能力，对因降噪带来的音质损失进行了适度补偿，成为Accentize系列产品与iZotope现有产品异质化的重要特征。

相对而言，iZotope RX的设计思路和应用方式更偏向于传统降噪器的观念，功能强大、参数较多、调整灵活、适用广泛，对噪声的人为控制力较强，定位为“专业人员的专业工具”。反观Accentize，因人工智能技术的应用而降低了使用难度，噪声处理能力基于深度学习和智能信号合成技术。因此，Accentize的处理重点在于语音分离（Speech Separation）和分离后的信号合成，与iZotope RX侧重于降噪不同，Accentize产品的降噪与修复功能比较平衡，所以这类产品也常被称为语音增强（Speech Enhancement，SE）处理器而非传统意义上的降噪器。

近年推出的Supertone、Hush Pro以及iZotope VEA等软件采用不同算法的人工智能降噪器/语音增强处理器，虽然并非针对专业应用推出，但其部分性能完全能够达到专业应用需求。需要指出的是，人工智能降噪技术并非万能降噪技术，其降噪能力取决于算法模型和训练数据，两者的优劣共同决定了降噪结果的优劣。算法或训练数据的缺陷会造成人工智能降噪器对某些特定类型噪声的处理能力较弱或几乎无法处理。在这一点上，偏向传统降噪器的iZotope RX反而表现出对多种噪声类型的“普遍适用性”。

人工智能降噪技术之所以能够首先在专业领域规模化应用，一方面是因为在声学、电学、信号分析等方向的基础研究比较成熟，对各种噪声可以进行科学性较强的分类且能够从中提取出具有代表性的物理特征，这些相对明确的技术特征有利于对噪声成分的识别、分析和处理；另一方面，语音识别、语音分离和数字信号合成技术也能够在一定程度上重构语音信号。尽管语音和噪声信号的种类繁多、构成复杂，但通过有效的定性与定量分析以及针对性的处理，人工智能技术已经首先在电影声音制作的对白降噪领域取得了显著进展，成功实现了规模化应用。

3 人工智能技术在电影声音制作领域的发展方向

依靠算法生成电影对白和效果类声音是人工智能生成内容(AIGC)技术在电影声音制作领域最重要的发展方向,也是最值得期待的发展方向。与智能降噪技术的成熟应用不同,对白和效果声音的智能生成目前总体上处于试探性、实验性和创新性阶段。

3.1 人工智能技术在电影对白制作中的应用

人工智能技术对声音的分类方式与传统电影声音分类方式不同且在人工智能领域内部并未取得一致,这源于不同专业对同一事物分析研究的角度不同。目前,人工智能技术在声音信号应用方面依然部分沿用了传统的分类方式——以广播、电影、电视和游戏为主的专业领域和民用领域。可以预计,人工智能技术在电影声音制作领域第二个获得大规模应用的领域是“语音”(Speech),也就是电影中的“对白”(Dialogue)。

人工智能语音技术是人工智能技术的一个分支。目前人工智能语音技术在短视频配音、自媒体新闻播报、各行业语音客服等领域已经实现了一定程度的商业化应用,这是由语言的重要性、人工智能语音技术发展现状和研发应用价值等几方面因素共同决定的。

人工智能语音技术在电影对白制作领域的应用以“文本转语音”[②](Text-to-Speech,TTS)和“语音转换”[③](Voice Conversion,VC)两项技术为主,虽然两者都需要使用语音合成技术,但其底层原理和模型算法存在明显差异。

文本转语音的核心是自然语言处理(Natural Language Processing,NLP)和语音合成。转换过程中最关键的是需要对形成的语音进行语速、语调、断句、气息以及情绪的精准调节和灵活控制,这也是文本转语音过程中较难处理而又最具价值的环节。对于富含情绪因素的电影对白而言,如果不能在文本转语音过程中对语音的自然度和情感表达进行细化控制,则这项技术就不具备充分的实用价值。2024年推出或升级的几款大语言模型,如Llama 3.1、Claude 3.5 Sonnet、ChatGPT o1等较之前的语言模型在以上各个方面有所改进,对生成语音进行一定程度的控制已经成为可能。尽管技术上仍存在一定局限,尚不能实现对语音节奏、情绪、流畅度和自然度等多个核心要素的灵活调控,但整体表现已显著提升。与早期语言模型多采用封闭研究方式不同,2024年推出的模型中有相当一部分采用了开源方式,这为具有技术能力的用户定制化应用模型、降低应用成本及合规性应用提供了条件。因此,目前的文本转语音技术比较适合应用于动画片、科幻片等类型片角色的对白创造。

语音转换的结果主要依靠模型的数据处理能力,训练数据的质量、数量以及模型算法的优劣在其中具有决定性作用。这项技术的优势在于能够较好地继承原始语音中包含的情绪性特征,并在音色转换过程中保持原始语音表达的自然度。语音转换技术在剧情片等类型片的对白制作中具有较高的使用价值。

目前来看,语音转换技术在成熟度和实用性方面相较于文本转语音技术表现出一定优势,但两项技术总体发展进度并未呈现显著差异。在电影对白制作实践中,需要同时应用以上两种技术,针对不同类型影片进行针对性使用或组合使用,充分发挥两种技术各自优势,现阶段单一技术的应用往往无法全面满足电影对白制作的复杂要求。

3.2 人工智能技术在效果类声音制作中的应用

人工智能技术在语音应用领域获得成功后将迎来效果类声音(Sound Effects,EFX)的制作突破。虽然从目前情况看效果类声音的大规模生成应用可能会略晚于语音的应用,但两者成熟应用的时间差不会太大。

目前人工智能技术在效果类声音生成时主要采用“文本转声音”[④](Text-to-Audio,TTA)方式。效果类声音的多样性、变化性、复杂性和主观性、模糊性造成生成提示词(Prompt)与机器理解之间容易形成偏差,这是目前为止造成这类声音生成结果误差较大的一方面原因;另一方面,效果类声音的商业应用范

围远远低于音乐和语言的市场应用需求，再加上效果类声音的生成质量要求较高，因此现阶段专注于这一领域的公司数量相对较少，这也在一定程度上制约了人工智能在这一领域的快速进步。

文本转声音技术相对成熟的公司是Elevenlabs和Stable Audio。Elevenlabs公司的总体实力较强，产品类型涵盖了语音与效果类声音处理两方面；Stable Audio公司虽然整体偏向于利用人工智能技术生成音乐，但也同时提供了生成效果类声音的功能，并且生成时长上限达到了3分钟。另外，其提供的带有文本提示的声音转声音（Audio-to-Audio）功能也比较有特色。

以Nemisindo公司为代表的程序化音频技术（Percedural Audio）采用了依靠算法合成效果类声音的方式，提供了超过70种细分模型和700种预置程序，不同的程序和预置对应不同类型的效果类声音，每一个细分模型对应一种声音的合成算法并提供了详细而直观的声音调整参数。人工智能技术在程序化音频中应用的价值在于对算法的完善，通过对细分模型算法的优化使其能够合成出更加自然的声音。

使用人工智能技术生成效果类声音的公司中比较出色的是Krotos。Krotos应用人工智能技术的方式无法明确归类于“文本转声音”或“程序化音频”，这与Krotos技术发展沿革有一定关系。在引入人工智能技术之前，Krotos以声音合成技术为基础在特殊音响效果制作方向发展，在电影声音和游戏声音制作的细分领域推出了一系列优秀产品。Krotos产品的定位非常清晰，声音质量较高，引入人工智能技术后推出的Krotos Studio不再局限于特殊音响效果的研究与应用，将产品应用领域从相对小众的特殊音响效果扩展到整个效果类声音，虽然目前Krotos Studio并不能直接生成多声道环绕声格式的声音素材，但多个立体声分轨文件可以组合构成环绕声形式的素材用于混录。

与电影对白使用多模态大模型为基础不同，效果类声音的特点决定了这类声音的智能生成技术需要在通用语音模型的基础上结合使用专用模型才有可能获得更好的声音质量。目前来看，仅依靠通用大模型暂时还不能完全适应效果类声音制作的技术需求，这在Krotos的产品中已经有所体现，专用声音模型与通用大模型的结合、云模型与端模型的结合或定制化的开源模型更适应未来人工智能技术在效果类声音应用的发展方向，这也是我们认为在专业领域使用人工智能技术制作效果类声音的进度会略慢于语音的原因之一。

4 具有专业应用前景的部分人工智能公司

4.1 国外公司

表1中，Respeecher、Replica Studio和Altered的人工智能语音技术相对成熟。2020年12月，Respeecher公司使用人工智能技术为迪士尼（Disney+）的剧集《曼达洛人》（*The Mandalorian*）终季中年轻的卢克·天行者（Luke Skywalker）完成配音工作；2021年9月，在剧集《如果登月发生灾难》（*In Event of Moon Disaster*）中为理查德-尼克松制作配音；2022年9月，在《美国达人》（*America's got Talent*）中合成并发布了埃尔维斯·普雷斯利（Elvis Presley）[5]的声音等。从该公司制作的影视项目来看，可以说Respeecher是

表1 具有专业应用可能性的人工智能公司名单

序号	公司名称	国别	创立时间	主要处理方向
1	Respeecher	乌克兰	2018	TTS/VC
2	Replica Studio	澳大利亚	2018	TTS/VC
3	Altered	英国	2018	TTS/VC
4	ElevenLabs	英国	2022	TTS/VC/EFX
5	Speechify	美国	2017	TTS/VC
6	Stable Audio	英国	2019	Music/EFX
7	Krotos	英国	2013	EFX
8	Nemisindo	英国	2021	EFX
9	Cedar	英国	1988	SE
10	iZotope	美国	2001	SE
11	Accentize	德国	2019	SE
12	Supertone	韩国	2023	VC/SE

（注：公司数据截至2024年5月。表中所列的只是部分代表性公司，其中建立在英国或具有英国技术背景的公司居多，反映出英国在人工智能声音处理方面的优势。）

最早进入好莱坞声音制作领域的公司，也是目前在影视声音领域研究人工智能语音制作技术最为成熟的公司之一。

Replica Studios公司的产品最初主要应用于游戏领域，与主流的渲染引擎虚幻引擎（Unreal Engine）和Unity等能够完美衔接。Replica Studios在技术成熟后将应用范围扩大至包括电影、动画和有声书等在内的多个方向。2024年1月9日，Replica Studios公司与美国演员工会（SAG-AFTRA）签署了开创性的协议，该协议一方面确定了配音演员有权决定是否同意在游戏项目中使用自己声音的数字复制版[6]（Digital Replica）；另一方面也授权了Replica Studios公司在获得演员同意的前提下，可以使用配音演员的声音进行语言模型的训练以创造新的角色声音。合同中最关键的是加入了使用跟踪（Usage Tracking）、数字复制版未来应用的透明公开（Transparency）和二次补偿制度（Secondary Compensation）。根据这些制度规定，配音演员有权依据自己声音的数字复制版在最终完成项目中的使用数量获取报酬。该协议主要适用于人工智能语音技术在游戏角色方面的应用，对于广播、电影、电视和流媒体等领域的应用则以“开发者外部使用（External Use by Developer）”条款进行了严格约束。虽然该协议并未将传统的电影声音制作领域包含其中，但该协议或将作为美国演员工会的合同范本对后续该类协议的签署以及相关立法产生一定影响。

Replica Studios与美国演员工会签订的协议是我们目前在公开渠道看到的第一份有关人类语音在人工智能时代应用的、具有法律约束力的文件。协议条款中虽然没有明确规定人类语音具有知识产权属性，但通过对各个条款的详细研究，能够感受到在人工智能时代人类语音初步具备了一定的知识产权特征。虽然这份协议是行业协会与单独商业公司之间的法律约定，不具备广泛适用性，并未上升到国家立法层面，但不排除随着人工智能技术的发展，未来从法律层面会对人类语音做出进一步、具有知识产权性质的立法。我们认为该协议最大的价值在于“为人类语音与人工智能语音技术的潜在冲突提出了尝试性的、可实施的、具有法律约束力的解决方案”。

4.2 国内公司

2023年春节档上映的影片《流浪地球2》成功使用人工智能语音技术进行了部分角色的对白制作，标志着人工智能语音技术第一次在国产院线电影对白制作中成功应用。由爱奇艺研发的基于深度神经网络和大模型技术的“奇声影视剧智能配音系统”也已经开始商业化应用，“为超过300部海外电影制作了普通话配音版本，为50多部华语电影、800多集国剧制作了多国配音版本”[1]。

以上应用案例说明国内在人工智能语音技术的应用层面与国际领先水平差距不大，腾讯、字节跳动、科大讯飞、阿里云、百度等公司也都推出了相应的产品。目前，国内公司还是在语音的通用领域发展，主要以满足民用市场为主，而以Respeecher、Replica Studios等为代表的部分国外公司则已经进入电影、电视和游戏等专业应用的细分领域。迄今，暂时还没有看到国内有实力的公司明确宣布进入专业音频应用领域。

2024年6月，字节跳动推出的Seed-TTS语言生成基座模型具有较高的语言自然度和稳定性，大量训练数据中甚至包括了各地方言，具有较强的适用性，语音生成的质量较高，情绪相对饱满，技术上初步具备了电影对白需要的流畅性、差异性和自然度等要求。

5 人工智能技术在电影声音制作应用的特点及其局限性

5.1 人工智能技术在电影声音制作应用的特点

目前，人工智能技术在电影声音制作中还限制在局部应用层面，无论是生成内容还是声音修复，人工智能技术主要以克服传统制作方法无法实现的制作困难为主。即便在相对成熟的降噪应用方面，智能技术也无法完全替代人工操作。语音转换的使用前提也是首先要录制对白，之后才能进行音色替换。在以上应用场景中，传统制作方法依然占据主体地位且暂时不可替代。尤其在面对高标准应用时，人工智能技术还存在一定局限性，所以现阶段呈现出两种方法混合应用的情况，人工智能技术主要作为

传统制作方法的技术补充进入制作环节。

从目前来看，人工智能生成技术产生的效果类声音信号个性并不鲜明，相似性、趋同化的问题难以避免。

5.2 人工智能技术在电影声音制作应用中的局限性

5.2.1 技术指标

首先，绝大多数文本生成语音的声音技术指标偏低，其采样率通常局限在22.050kHz，难以达到电影声音制作所要求的24Bit/48kHz标准，虽然从专业应用角度可以适当降低对人工智能生成语音的技术指标要求，但依然希望其能满足16Bit/44.1kHz的最低标准。

2023年下半年至2024年上半年，部分国外公司声明可以达到48kHz的技术指标，但经过技术分析发现其原生音频的标称指标与音质主观评价并不相符，不排除所谓更高的技术指标是通过对原生音频进行频谱合成方式进行的适度补偿而非通过模型算法实现的、真正的高质量音频数据，这种声音在电影标准监听条件下使用还存在些许缺陷。

目前科大讯飞大模型语音合成的最高采样率能够达到24kHz，量化深度达到16bit[2]，字节火山引擎最新推出的双向流式接口则标明已升级至最高支持48kHz采样[3]；爱奇艺奇声使用语音转换方式为影视剧进行的配音目前能够达到32kHz的采样频率，有效声音频带达到了16kHz[4]，实现了较好的声音效果。总体而言，人工智能生成语音的有效频带在逐步向上扩展，从早期的4kHz逐步提升至6kHz、8kHz、12kHz至24kHz左右，这一方面受到训练数据的影响，另一方面也与模型算法的进步有关。

其次，目前文字生成语音内容的情绪表达存在不足，难以达到演员对台词细腻控制的程度。文字转语音过程中需要在生成过程中对音色、音调、速度以及情绪的多样性进行控制，虽然有些模型算法提供了上述调整功能，但总体效果还无法达到灵活控制的程度，而且各参数变化幅度不大，相对极端的参数设定会造成声音质量明显下降，语音表现力和情感表达略显不足。尤其是在激烈情绪状态下，人类语言表现出的复杂情感和复杂变化，人工智能语音技术暂时还难以有效模仿。对生成语音不能进行相对灵活地调整，在一定程度上影响了文本转语音在电影声音制作中的应用。在这一点上，语音转换技术的优势更加明显，也是现阶段人工智能技术最适合电影声音制作的方式。

最后，音响效果的生成质量有待提高。效果类声音在电影声音构成中的重要性仅次于对白，不但应用数量大，而且对声音质量有很高的要求，效果类声音的制作水平也是判定整部影片声音制作能力的重要标准之一。目前，人工智能生成的效果类声音一方面在技术指标上偏低，另一方面在音质主观评价层面也表现出动态不足、力度欠缺、信号劣化等问题。依据现有模型状况分析，如果希望效果类声音能够达到专业应用的水平，可能需要在多模态大语音模型的基础上，开发出面向效果类声音的专有模型，并进行高度针对性的数据训练，这将有可能生成满足电影声音制作要求的效果类声音，而这一过程也会引发研究团队对投入产出比的考量。

经过实测，Krotos生成的效果类声音技术指标能够达到24Bit/48kHz；部分人工智能公司生成的音响效果能够达到16Bit/44.1kHz的标准，已经普遍高于人工智能生成的语音。单纯从显性的技术指标分析，这样的结果是比较理想的，但音质主观评价结果认为在保持现有客观技术水平基础上的主观听感还有提升空间。

5.2.2 训练数据

限制人工智能生成内容技术指标和主观评价指标的因素主要源于两方面：一方面是模型算法的优劣，另一方面是训练数据的质量。随着模型技术的迭代、算力水平的提升和运算成本的降低，模型算法能够在可预见的时间内解决对高质量声音技术指标的支持问题，因此未来人工智能技术在电影声音制作领域应用的障碍不会是算力，也不会是模型算法等物理层面问题。生成高质量音频数据的前提是有足够多的高质量音频数据用于训练模型，而高质量音频数据是相对稀缺的资源，这种情况可能导致一个潜在问题：即便算法模型理论上能够支持较高技

术指标的内容生成，却可能因缺乏足够的高质量训练数据而造成发展受限。

我们通过与部分公司交流可知，大量高质量训练数据的获取是一个越来越突出的问题，而且是短时间内很难解决的问题。相对而言，高质量语音生成可以在一定程度上寄希望于模型算法迭代加以解决，但高质量效果类声音的稀缺则可能会影响人工智能技术在此类声音的应用进度。

5.2.3 知识产权

训练数据的质量与数量是人工智能生成技术的基础。早期大量训练数据无法律意识的应用虽然对人工智能技术的发展做出了极大贡献，但时至今日，人们对于具有知识产权的公开数据可否被无代价、无保留地应用于人工智能技术发展产生了质疑。

训练数据质量的价值大于数量。在满足数量的前提下，数据质量决定了模型的效果。2024年6月，Adobe公司在其Creative Cloud产品服务条款中硬性规定了对用户数据的使用权限，希冀将用户数据，尤其是用户的创意数据用于训练基于人工智能的创作软件Firefly Gen AI的研发。该条款的出台立刻引起轩然大波，造成Adobe公司在两天后被迫发表声明以图挽回声誉。该事件的出现不是孤立的，也不是Adobe一家公司面临的窘境。低质数据已经不能满足人工智能数据训练所需，各公司都需要优质数据用于提升人工智能的技术水平，而强烈的反对意见反映出人们对于数据的知识产权意识迅速提升。

知识产权意识的提升会进一步制约各公司获取高质量的训练数据，影响人工智能技术的发展，但这一问题的本质是商业利益问题，并非不可克服。

5.2.4 道德伦理和法律监管

2023年8月15日，由国家网信办会同发改委、教育部、科技部、工信部和公安部联合发布的《生成式人工智能服务管理暂行办法》在国内实施；2024年3月13日，欧盟议会通过了《人工智能法案》；2024年4月9日，美国田纳西州确立了《确保肖像、声音和图像安全法案》；2024年6月3日，欧洲数据保护监督机构（European Data Protection Supervisor，EDPS）编撰了《生成式人工智能与EUDPR：EDPS就生成式人工智能数据保护的首个指南》等，这些国家层面的立法行为标志着人工智能技术的监管逐步进入较完善的法制轨道。从现有情况看，有些现行法律条款的规定在人工智能时代背景下确实面临着挑战。

综上所述，人工智能技术在电影声音制作领域的应用面临着一定的局限性，但所谓的局限性，尤其是技术层面和商用层面的局限性也具有一定时效性。随着技术的迅速发展和法律制度的完善，当下的局限性会在人工智能技术的不同发展阶段得到不同程度的解决。

6 人工智能技术在电影声音制作中的定位

人工智能技术已经或即将进入电影声音制作的各个细分领域。虽然国内外已有一些成功案例，但距离普遍大规模应用，尤其是具有价格和时间费效比的商业性应用还有一定距离。目前人工智能技术的长项体现在各种声音素材生成、简化或完善传统制作工艺、降低制作难度和提升工作效率的层面，这些应用依然偏向于技术和工艺流程方面，处于辅助地位，虽然它能够在一定程度上参与并影响着艺术创作，但“人”在其中的主导性地位依然无法撼动，主观性、个性化的判断与实施依然需要以人为主体做出。截至目前，人工智能作为一种新兴的技术手段还摆脱不了制作工具的本质属性。

音乐虽然是电影声音重要的组成部分，但音乐制作和内容生成以作曲家、演奏家和音乐录音师为主，在此不做过多研究。从目前获得的信息和实例来看，人工智能技术在音乐领域的发展领先于其在电影声音各细分领域的发展速度，这与其强大的市场支撑有重要关系。

7 未来发展

经过历时一年的跟踪研究可以发现，人工智能技术已经或即将在某些局部改变现有电影声音制作体系。文中所述各种技术问题都能逐步解决，并不构成该技术在电影声音制作中应用的障碍，声音的录制与人工智能生成作为两种声音获取方式将并行存在。随着技术的快速迭代，人工智能技术在电影

对白、动效、音响效果、环境和音乐等各类素材的获取方式以及声音编辑、声音处理和预混、终混等制作层面都有可能产生一定影响，虽然这种影响不至于从根本上颠覆传统的电影声音制作工艺，但人工智能技术的介入必将对电影声音的创作意识、创作理念和技术路径产生深远影响，而其中最为关键的是对从事电影声音工作人员的知识结构提出新的要求。❖

注释

① iZotope RX降噪器的核心功能为De-click、De-crackle、De-clip和Spectral De-noise。

②文本转语音也称为文语转换。

③语音转换也称为语音克隆（Voice Cloning）、音色融合（Voice Morphing）、音色替换或语音替换（Speech to Speech，STS）等。

④本文提到的文本转声音中的“声音”不包括语音。

⑤埃尔维斯·普雷斯利又称“猫王”，美国男歌手、演员，出生于美国密西西比州图珀洛。

⑥声音的数字复制版指通过人工智能技术合成后的配音演员声音。

参考文献

[1] 爱奇艺.“奇声影视剧智能配音系统”获评工信部“2024新型数字服务优秀案例”[EB/OL].(2024-06-19)[2024-07-01]. https://mp.weixin.qq.com/s/ATiw_j6kTevHkp7CxMonBw.

[2] 讯飞开放平台文档中心.超拟人合成简介[EB/OL].[2024-07-01].https://www.xfyun.cn/doc/spark/smart-tts-iOS.html#_1-%E8%B6%85%E6%8B%9F%E4%BA%BA%E5%90%88%E6%88%90%E7%AE%80%E4%BB%8B.

[3] 火山引擎文档中心.产品简介.功能特性[EB/OL].(2024-04-28)[2024-07-01].https://www.volcengine.com/docs/6561/1257543.

[4] 李海.奇声(IQDubbing)—面向影视剧的AI配音技术[EB/OL].(2023-04-04)[2024-07-01].https://cloud.tencent.com/developer/article/22557838.

[5] How AI is Changing Audio Post-Production[EB/OL].(2024-02-14)[2024-07-01]. https://www.production-expert.com/production-expert-1/how-ai-is-changing-audio-post-production.

[6] Dialogue Cleanup-AI Versus Audio Professional-The Results[EB/OL].(2023-03-20) [2024-07-01]. https://www.production-expert.com/production-expert-1/dialogue-cleanup-ai-versus-audio-professional-the-results.

[7] Krotos Ltd.Edinburgh.METHOD OF GENERATING AN AUDIO SIGNAL [P]. United States Patent Application Publication.Patent No.:US 10,606,548.

[8] SAG·AFTRA.Replica Digital Voice Replica Development Agreement[EB/OL].(2024-01-09)[2024-07-01]. https://www.sagaftra.org/files/sa_documents/Replica%20Studios%20Agreement%20for%20Digital%20Voice%20Replicas_0.pdf.

[9] Mascha D.AI Toos for Audio-an Overview of the Latest Applications for Sound Postprodution[EB/OL].(2023-09-11)[2024-07-01]. https://www.cined.com/ai-tools-for-audio-an-overview-of-the-latest-applications-for-sound-postproduction/.

[10] Wang D L, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10):1702-1726.

[11] Anastassiou P, Chen J, Chen J, et al. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. [EB/OL]. [2024-06-05]. https://arxiv.org/abs/2406.02430.

[12] Wang Z, Chen Y, Wang X, et al. StreamVoice: Streamable Context-Aware Language Modeling for Real-time Zero-Shot Voice Conversion. [EB/OL]. [2024-05-15]. https://arxiv.org/html/2401.11053v2

[13] Adán L B, Vassallo T , Reiss J D, et al. FXive: A Web Platform for Procedural Sound Synthesis[EB/OL]. [2024-06-05]. https://aes2.org/publications/elibrary-page/?id=19529

[14] Liu H, Chen Z, Yuan Y, et al. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. [EB/OL]. [2024-07-02]. https://doi.org/10.48550/arXiv.2308.05734.

[15] Liu H, Yuan Y, Liu X, et al. AudioLDM 2: Learning Holistic Audio Generation with Self-Supervised Pretraining[EB/OL]. [2024-06-06]. https://doi.org/10.48550/arXiv.2308.05734.

[16] Su J , Wang Y , Finkelstein A ,et al.Bandwidth Extension is All You Need [C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).IEEE, 2021.DOI:10.1109/ICASSP39728.2021.9413575.

[17] Serrà J, Pascual S, Pons J, et al. Universal Speech Enhancement with Score-based Diffusion[EB/OL]. [2024-06-12]. https://doi.org/10.48550/arXiv.2206.03065.

[18]Micaela M. ARTificial: Why Copyright Is Not the Right Policy Tool to Deal with Generative AI[J]. The Yale Law Journal,2024:133.

[19] Lv S, Fu Y, Xing M, et al. S-DCCRN: Super Wide Band DCCRN with Learnable Complex Feature for Speech Enhancement[J]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022:7767-7771.

[20] Andreu S, Aylagas M V. Neural synthesis of sound effects using flow-based deep generative models[J]. In Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'22),2022(18):2-9.

[21] Evans Z, Carr C J, Taylor J, et al.Fast Timing-Conditioned Latent Audio Diffusion[EB/OL].[2024-06-25]. https://arxiv.org/abs/2402.04825.

[22] Yao J, Lei Y, Wang Q, et al.Preserving background sound in noise-robust voice conversion via multi-task learning[EB/OL].[2024-06-01].https://www.semanticscholar.org/reader/6294114d38667033dcb10720e97ca194f3be6d22.

作者贡献声明：

石宝峰：设计论文整体框架、论文撰写与修订，全文文字贡献70%；

丁思立：撰写部分论文，全文文字贡献30%。