分类号 G21/156　　密级
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 AI语音合成技术在网络音频平台中的应用与发展策略研究

研 究 生 姓 名：刘思捷
指导教师姓名、职称：李 艳 副教授
学 科、专 业 名 称：新闻传播学 新闻与传播
研 究 方 向：网络与新媒体

提 交 日 期：2023 年 6 月 10 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：刘思捷 签字日期：2023.6.12

导师签名：李艳 签字日期：2023.6.12

导师(校外)签名：王玉明 签字日期：2023.6.12

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：刘思捷 签字日期：2023.6.12

导师签名：李艳 签字日期：2023.6.12

导师(校外)签名：王玉明 签字日期：2023.6.12

Research on the Application and Development Strategy of AI Speech Synthesis Technology in Network Audio Platforms

Candidate : Liu Sijie

Supervisor: Li Yan

摘 要

随着媒介技术的不断更迭，“声音”作为辅助性媒介的地位正在发生改变，声音与音频的价值逐渐被挖掘出来。在移动信息传播的背景下，网络音频开始进入大众视野，各大音频平台的勃兴为音频行业带来了新的发展机遇。近年来，网络音频头部平台纷纷加入使用AI语音合成技术的行列，通过智能化、人性化、专业化服务，为听众带来不一样的听觉体验。而在听觉文化转向背景下，AI语音合成技术形塑了全新的人工声音景观，所以探究网络音频内容中AI声音要素的呈现很有必要。

本研究首先对AI语音合成技术与网络音频平台进行溯源，发现深度学习方法是当前语音合成的主流方法，而AI语音合成技术在网络音频平台中的应用已进入蓬勃发展期，其实AI作品在视频平台早已有之，如今借助网络音频发展的东风，拥有了不可小觑的发展潜力。本文针对目前市场上具有代表性的网络音频AI服务展开研究，选取了以新闻资讯为主的云听平台、以综合性音频服务为主的喜马拉雅和以儿童教育内容为主的恐龙贝克平台中的AI传播内容进行具体研究，总结其应用功能和应用实践结果。本文在对典型案例进行分析的基础上，结合SPSS软件进行问卷分析，总结出AI语音合成技术在网络音频平台中的应用困局，即市场准入门槛低、AI应用功能缺位、合成语音质量难保证、存在侵权风险等。针对以上问题，本文给出强化有声市场管理、优化AI音频服务功能、提升合成语音质量、坚持伦理与法律原则的解决对策，希望通过本研究研究，可以为AI语音合成技术在网络音频领域的正面应用与发展提供参考，构建出的全新声音环境。

关键词：AI语音合成技术 合成语音 网络音频平台

Abstract

With the continuous changes in media technology, the status of "sound" as an auxiliary medium is changing, and the value of sound and audio is gradually being explored. In the context of mobile information dissemination, online audio has begun to enter the public's perspective, and the flourishing of major audio platforms has brought new development opportunities to the audio industry. In recent years, online audio header platforms have joined the ranks of using AI speech synthesis technology, providing listeners with different auditory experiences through intelligent, user-friendly, and professional services. In the context of the auditory culture shift, AI speech synthesis technology has created a new artificial sound landscape, so it is necessary to explore the presentation of AI sound elements in online audio content.

This study first traces the origin of AI speech synthesis technology and online audio platforms, and finds that deep learning is the mainstream method of speech synthesis at present. However, the application of AI speech synthesis technology in online audio platforms has entered a vigorous period of development. In fact, AI works have already existed on video platforms, and now, with the help of the development of online audio, they have significant development potential. This article conducts a research on representative online audio AI services in the current

market, selecting AI voice works from cloud listening platforms that focus on news and information, Himalayas that focus on comprehensive audio services, and Dinosaur Baker platform that focuses on children's educational content, to conduct a specific study and summarize their application functions and practical results. Based on the analysis of typical cases and questionnaire analysis using SPSS software, this article summarizes the application difficulties of AI speech synthesis technology in network audio platforms, including low market entry barriers, lack of AI application functions, difficulty in ensuring the quality of synthesized speech, and risk of infringement. In response to the above issues, this article provides solutions to strengthen sound market management, optimize AI audio service functions, improve the quality of synthesized speech, and adhere to ethical and legal principles. It is hoped that through this research, it can provide reference for the positive application and development of AI voice synthesis technology in the field of network audio, creating a new sound environment.

Key words： AI speech synthesis technology; synthetic speech; network audio platform

目 录

1绪 论

1. 1研究背景及意义

1. 1. 1研究背景

（1）音频市场前景广阔

目前，中国网络音频产业规模达到亿元，可谓处于高速发展阶段。随着收听场景逐渐增多，围绕“听”的传播媒介工具制造和音频内容生产也进入发展快车道。多元化的适用场景和用户需求，创新性的音频内容，使得以声音为核心的音频经济市场不断扩大，呈现快速上升态势。艾媒咨询数据显示[①]，年中国在线音频市场规模达到220.0亿元，同比增长67.9%。

网络音频平台涵盖多种有声服务类型，播客、有声书、广播剧、音频直播、新闻播报等。其中，广播剧作为传统广播电台内容传播的重要组成部分，随着传统媒体的优势逐渐消失和移动互联网的蓬勃发展，在网络音频平台上播放的广播剧开始兴盛起来，可见网络音频内容已经占领相当一部分以“听”为主的经济市场。同时，有声书形式的出现广受欢迎，它不仅解放了用户的双眼和双手，还增加了便携性、易得性和吸引力，是网络音频有声服务的重要内容之一。喜马拉雅就将原本以文字内容为主的书籍，通过智能语音转换技术，将单一的文本信息转化成音频信息，并快速建立起庞大的用户社群，设置了听书专区。未来，音频业务将会演化为一个体量巨大的产业，智能技术与有声内容的结合也将为音频市场打造全新的内容生产模式。

（2）国家高度重视人工智能技术发展

人工智能技术作为当下发展的大趋势，已经孵化出一系列产业，这很大一部分原因在于我国将发展人工智能技术放在了重要的战略位置。国家主席习近平曾多次在大会上提到人工智能相关技术，从2017年国务院发布《新一代人工智能发展规划》，到2019年强调推动人工智能开发应用，再到2021年“十四五”规划中到的科技创新，国家在政策层面全力支持新一代信息技术等科技产业集群的崛

[①] 艾媒咨询. 2021年中国在线音频行业发展及用户行为研究报告[R]. 2021.

起，同时也取得了显著成就，在《数字中国发展报告（2021年）》中提到我国数字技术创新能力得到快速提升[①]，人工智能、云计算、区块链、量子信息等新兴技术有望跻身全球第一梯队，可见我国力求在智能化发展大潮中挺立潮头的决心。

人工智能技术作为学习、模拟和扩展人的智能的方法，已逐渐形成了自己独有的理论体系，重塑了人与人、人与物、人与环境之间的关系。AI语音合成技术作为语音交互技术中相对成熟的一部分，它通过深度学习、深度神经网络和智能算法训练等技术，将文字信号转换为语音信号，实现了多场景的语音播报和实时互动。目前，AI语音合成技术广泛应用于智能家居、智慧教育、智慧医疗、智能汽车服务等场景，人们对合成语音的接受度也在随着应用范围的扩大而逐渐提高，这是合成语音技术发展向好的重要标志。所以研究AI语音合成技术的应用与发展符合时代发展要求，也期望能在一定程度上为后续研究提供参考。

（3）技术赋能刺激听觉文化的回归

从最初的口语、文字、印刷传播到如今的新媒体传播时代，声音基本贯穿了人类传播的全过程。马歇尔·麦克卢汉将人类社会发展分为三个阶段，分别是部落社会、脱部社会和地球村[②]，在部落社会阶段，听觉生活占主导地位，听觉一定程度上压制着视觉的价值。到了口语传播的时代，声音也是个体之间交流的首选媒介，在原始传播交流中占有重要地位。来到文字和印刷为主流的传播时代，听觉跌落神坛成为辅助性信息传递渠道，人们开始向视觉转向。而在被文字、图片、视频充斥的新媒体传播时代，视觉文化受到前所未有的追捧，中国互联网络信息中心(CNNIC)发布的第50次《中国互联网络发展状况统计报告》显示[③]，截至2022年6月，中国的短视频用户数量达到了9.62亿，在互联网中所占比例达到91.5%，声音彻底沦为附属品。

但是AI语音合成技术的出现让声音信息找到了突破之路。美国计算机科学家尼古拉斯·尼葛洛庞帝(Nicholas Negroponte)曾预言：“20年后，你可能对着桌上一群身高八英寸的全息式助理说话，声音将成为你和你的界面代理人之间的主

[①] 国家互联网信息办公室. 数字中国发展报告（2021年）[R]. 2022.

[②] 马歇尔·麦克卢汉. 《理解媒介:论人的延伸》[M]. 何道宽译. 译林出版社, 2019:114.

[③] 中国互联网络信息中心(CNNIC). 第50次《中国互联网络发展状况统计报告》[R]. 北京:2022.

要沟通渠道[①]。”在信息爆炸的当下社会，用户获取有效信息需要消耗巨大精力，读屏模式更是加剧了这一现象，语音合成技术的出现正好解放了受众高度紧张的视觉感官，寻找听觉感官的回归。声音的伴随性使用户不需要投入太多注意力就能在无意识状态下获取信息，利于唤起听觉感官的觉醒，重新实现感官平衡。网络音频平台的勃兴，为声音重回公众视野提供了途径。近年来，喜马拉雅、蜻蜓FM等移动有声平台的崛起越来越让人们认识到声音的魅力，而合成语音的应用也将继续拓展人们对声音的想象力。因此研究AI语音合成技术现阶段在音频作品中的应用困局和发展路径，对听觉文化的回归与网络音频平台的发展都具有建设性意义。

1. 1. 2研究意义

（1）理论意义

关于语音合成、语音交互技术的研究基本集中于电信技术、工业信息经济、计算机软件应用、语言文字学和医疗领域，与新闻与传媒领域相结合的研究体现在新闻生产、智能播报和AI合成主播；而关于网络音频、有声阅读或有声读物的研究中，从学术层面的意义来看，学界目前还处于初期发展阶段，更多集中在内容生产、用户使用、媒介发展、盈利模式和平台宏观管理与发展的层面，语音合成技术与网络音频相结合的研究较少。本研究以网络音频平台中的合成语音作品为研究对象，探索网络音频平台中语音合成技术的应用现状、问题和发展路径，希望通过本次研究帮助大家认同和肯定听觉传播的价值。

（2）现实意义

在数字经济发展浪潮中，人工智能迅速崛起，在此背景下，语音交互技术在生活场景中得到普遍应用，家居设备、智能客服、车载语音、语音助手等都体现着AI智能语音技术应用的不断深入，已经与人产生紧密联系。语音合成技术在在线阅读平台、网络音频平台和网站语音播报中均有所涉及，其中在音频作品中的应用在很大程度上提升了平台的吸引力和竞争力，但技术引入带来的负面影响与问题也随着时间的推移逐渐浮现出来，给技术和平台的发展都带来了挑战。因此根据调查与分析AI语音合成技术在音频作品中的使用现状，既能为智能科技公司

[①] 尼古拉·尼葛洛庞帝.《数字化生存》[M].胡泳，范海燕译，电子工业出版社,2017:145.

的语音发展提供事实参考，也能补足技术应用的漏洞和合成语音作品的缺陷，为用户解决实际问题，对网络音频的内容生产和传播有促进作用。

1. 2研究问题及方法

1. 2. 1研究问题

本文的研究问题主要聚焦于 AI 语音合成技术在网络音频平台的应用现状，智能时代声音的表现方式和特点，具体分析了相关的 AI 语音技术使用案例，明晰了现阶段 AI 有声内容的缺点和问题，并针对性的从技术、伦理和法律等各个方面提出了人工智能时代使用 AI 语音合成技术创造有声作品的优化方法，以促进 AI 语音合成技术与网络音频平台的融合发展。

1. 2. 2研究方法

（1）案例研究法。这一方法主要体现在第二章、第三章、第二章在公开数据平台移动观星台中选取了4个典型网络音频平台作为研究样本，从宏观层面阐述了AI语音合成技术在网络音频领域的引入、支持与价格情况，对合成语音生成网络音频的基本情况有所了解；第三章则挑选了AI语音合成技术在有声领域应用市场中最突出、最新颖的三个平台作为案例，分别为喜马拉雅、云听和恐龙贝克，分析合成语音在不同平台的应用现状，为问卷调查提供一定支撑。

（2）问卷调查法。本研究将这一方法用在第四章，利用问卷星平台设计、发布问卷，通过互联网渠道进行了为期近两个月的问卷回收工作，问卷内容主要围绕用户对技术的接触与需求、用户对合成语音的评价与满意度、AI作品风险等。此外，本研究尝试使用社会科学统计软件SPSS26. 0讨论变量关系，针对性分析语音合成技术的应用难题，也让发展策略更具说服力。

1. 3理论依据

在本世纪初，以声音和听觉为主要对象的研究开始进入学者视野，国外学者将其称为“Sound Studies”，翻译为“听觉文化研究”。最早提出听觉文化概念的是加拿大学者雷蒙德·默里·谢弗(R.Murray Schafer)，他在其著作《为世界调

音》中提出了声音景观理论，他认为声音景观除了噪音，还包含具有艺术欣赏价值的音乐作品。部分学者认为，声音景观拥有三个思想来源，除谢弗对音乐的哲学思考外，提出声音景观也是为了保护生态环境，因为20世纪工业时代的成熟，机器噪音增多；同时维护听觉感官的感受。理论是从现实问题中孕育和发展而来的，为了解决噪音污染和对视觉感官过于倚重等问题，声音景观理论应运而生。在本研究中，主要强调声音景观是对听觉感官的重建，人类社会过于注重视觉感官而忽略了听觉感官，而AI语音合成技术的赋能正好是让听觉感官重回人们视野的最佳时机。

新闻传播领域的部分学者将声音景观的研究与传播学联系在一起，认为声音景观的研究实质上是在处理传播相关的问题，在人们身处的声音环境之中，声音景观成为了一种媒介，研究人们对声音环境的感受和体验，声音景观中介了人与人的关系，人与世界的关系。此外，一些学者将声音景观的研究与声品质联系在一起，提出追求高保真的声音景观。高保真的声音景观是指声音之间不会重叠，既有信号声又有基调声的具有透视性的声环境，其优势在于声音之间不会相互消散和遮蔽，各种各样的声音都可以被清晰地听到。

历史上从录音机、电视到互联网技术，声音的传播效果越来越生动活泼，传播速度不断加快。当前随着AI语音合成技术的迅猛发展，语音合成以假乱真，声音存储空间不断扩大，声音的应用途径也不断增多，声音的商业价值不断凸显，智能音箱、手机语音助手、汽车导航系统、游戏系统音等语音产品层出不穷，声音的应用范围越来越广，传播速度越来越快，声音侵权现象也时有发生。目前学术界对于声音景观的研究还停留在传统媒体时代和互联网时代，这实际上为本文的研究提供了一个可以深入的方向，本文与时俱进，研究人工智能时代声音景观的具体呈现，通过对现阶段网络音频平台中AI有声作品的梳理总结，力图展现AI语音合成技术在声音领域的创新和应用，讨论AI语音合成技术给声音带来了何种便利，同时又存在何种问题，提出相应的解决策略，达到强调听觉感受的重要性，塑造良性声音景观的目的。

因此，本研究尝试以听觉转向为背景，以声音景观为理论视角进行相关研究。声音景观作为文化景观的一部分，是不同时间、地点与文化的体现，而AI语音合成技术用建构的形式为听众呈现一种虚拟景观，是先进技术的媒介实践产

物。声音景观分为三个维度：声音、听众和环境，从声音维度而言，AI 语音合成技术的发展使声音生产方式发生了显著变化，从以往的通过身体器官产生的声音转变为现在的依托语料库和算法模型产生的拟人声音；从听众角度而言，技术赋能刺激了听觉感知的回归，正如麦克卢汉所说的电子技术时代声音变得越发重要，声音的使用和声音产品让听众开始重视听觉感官；从环境角度而言，AI 语音合成技术对人声的深度学习，逐渐模糊了现实和虚拟的边界，AI 语音开始融入生活场景。可见 AI 语音合成技术对声音景观产生了重要影响，甚至重塑了声音景观。所以，在 AI 语音合成技术形塑的全新声景下，探究网络音频内容中的具体应用，以及技术驱动下的声音所带来的问题，并对问题提出针对性意见，利于 AI 语音合成技术和网络音频行业的长效发展，也利于丰富听觉文化相关研究成果。

1. 4文献综述

1. 4. 1AI语音合成技术研究进展

语音合成技术与语音识别技术是实现人机语音通信的两项关键性技术，共同构建出语音交互技术。与语音识别相比，语音合成技术发展更为成熟，更具研究价值。

（1）国内研究进展

2022 年 12 月 25 日，通过对中国知网数据库中已经发表的研究论文进行检索，以“语音合成技术”“TTS 技术”“语音合成”“合成语音”为关键词，共找到约 3000 条结果，国内研究已经呈现规模。其中，专门研究语音合成技术的文献约 1700 篇，与 AI 语音识别技术相结合的文献约 500 篇，在 AI 语音交互技术中讨论合成技术的约 800 篇。

能检索到的关于语音合成技术最早的研究是学者邓国亮在 1981 年发表的《几种重要的语言信号数字处理技术》，文章将语言信号数字处理技术分为两大类，其中一类是以语言信号模型为基础的分析——合成法，并简要介绍了合成法的三种重要处理技术，这可以算是国内首次进行的合成技术研究[①]。从此，对语音合

[①] 邓国亮. 几种重要的语言信号数字处理技术[J]. 南京航空航天大学学报, 1981, (03) : 83-94.

成技术的研究开始步入初级阶段，主要体现在数据压缩、技术调制、语音编码方面。但具有实用意义的研究是在计算机技术和数字信号处理技术的发展基础上发展起来的，从1985年开始就陆续出现了语音合成技术相关的应用研究，直到1993年，关于语音合成技术的应用方案如同雨后春笋般涌现，研究重点转向语言文字领域，学界开始探索中文语音合成、简体汉字语音合成以及汉语语音拼接，学者朱耀庭、李霞（2002）指出基于波形拼接技术所合成的汉语普通话的可懂度、清晰度已经达到了很高水平，并借此使汉语语音合成技术走出实验室、走入市场[①]。但他们同时提出语音合成系统不仅应该发音清晰自然，还应该具有模拟特定人发音的能力。2007年到2013年，关于语音合成技术的应用研究达到巅峰，研究重点开始放在与各学科领域的融合发展，讨论了合成语音在家电科技、智能玩具、车载系统、盲人图书馆、智能教学、医学器械领域的应用实践。2015年，AI语音合成技术开始与新闻传播领域相结合，记者刘胜男采访科大讯飞平台事业部总经理于继栋后发布《当传媒业遇到智能语音技术》，文中提出语音合成技术不但可以使传统媒体“发声”，还能让自媒体创作内容拥有更多可能性，其中科大讯飞与喜马拉雅建立合作伙伴关系，共同推进语音技术与音频内容的结合[②]。AI语音合成技术悄然改变着人们的生活习惯和方式，也为传媒行业带来更多改变与机遇。

经过梳理发现，关于AI语音合成技术在新闻与传播领域的研究内容主要有以下三个方向：合成语音新闻生产与传播效果、虚拟主持人与AI主播、播音主持领域的应用发展。

第一，合成语音新闻生产与传播效果。在知网数据库中进行文献搜索后，经过精炼浏览，最终研读了最具代表性的核心期刊文献。其中，卢维林、宫承波（2020）指出智能媒介运用语音合成等关键技术而“知冷热”，提醒了新闻生产与新闻内容也应有“温度”，人工智能设备给基于语音的交互式新闻生产与呈现带来更广阔的发展空间[③]；喻国明教授（2021）采用脑电(EEG)技术对不同语速下

[①] 朱耀庭, 李霞. 中国计算机产业的下一个亮点——汉语语音合成的实用化[J]. 世界科技研究与发展, 2002, (05) : 49-54.

[②] 于继栋, 刘胜男. 当传媒业遇到智能语音技术[J]. 中国传媒科技, 2015, (07) : 14-17.

[③] 卢维林, 宫承波. 智能音箱中的新闻生产与呈现逻辑[J]. 青年记者, 2020, (13) : 50-51.

受众使用合成语音新闻产品的传播效果与影响进行研究[①]，考察用户体验，得出以下结论：当合成语音新闻为 1.5 倍速时，男性受众的信任度高，当合成语音新闻为 1 倍速时，女性受众的信任度高。此外，还有关于自动化新闻生产与农业新闻生产的研究，以及智能语音技术在新型主流媒体中的应用，依据智能语音技术的应用场景，提出了具体建议，助力实现新闻播报智能化。

第二，虚拟主持人与 AI 主播。关于这一维度的语音合成技术研究，大多集中在 AI 语音的发声特点和语音考察，其中具有代表性的是姜泽玮（2021）提出智能语音合成技术的研究应先重点把握 AI 播音中的具象层面问题，从语流速度、音节调值、句中停顿和重音形式入手，针对性完善 AI 播音质量[②]。此外，栾轶玫教授（2022）的研究强调了技术的潜在风险，其文中提出利用语音合成技术模仿新闻主播的声音生成新闻语音，可能引发受众认知风险和 AI 主播伦理风险，所以技术使用者和受众都需合理掌握技术应用边界[③]。

第三，播音主持领域的应用与发展。在 AI 语音合成技术的应用方面，广电行业主要将其运用到节目制作上，韩冰（2020）等人提到语音合成应用场景实现了编辑读报审稿，支持自定义配置，用户据文本语种选择发音人及音色，自定义配置发音语速、断句停顿、数字与数值的读法等，可见语音合成应用到了报业内容采集、生产等业务流程中[④]；而发展研究多是针对真人播音员主持人所做的技能培养，强调语音合成技术生成靶向语音会放大真人主播原本已经存在的问题，从而对他们提出更高的要求。同时也指出真人主播拥有不可取代的优势，即富有人性与人格，这让情感合成成为智能语音技术的一大明确发展方向。

（2）国外研究进展

AI 语音合成技术的起源时间可以追溯到 18 世纪，人们使用机械装置来模拟合成人类发声；到 20 世纪初，具有代表性的是 Dudley 发明的“VEDER”电子

[①] 喻国明, 王文轩, 冯菲和修利超. 合成语音新闻的传播效果评测——关于语速影响的EEG证据[J]. 国际新闻界, 2021, 43(02):6-26.

[②] 姜泽玮. 收听人工智能语音播报与阅读文本的短时记忆效果差异——以新华社客户端新闻为个案的实验法研究[J]. 中国记者, 2021, (03):84-87.

[③] 栾轶玫. AI主播的媒介应用及伦理风险[J]. 视听界, 2022, (02):126.

[④] 韩冰, 张慧, 谢陶欣. 智能语音技术在融媒体业务中的应用[J]. 中国报业, 2020, (19):20-21.

发声器；1980 年，KLATT 发布了共振峰合成器；随着人工智能技术的不断发展，逐渐来到各种神经网络模型训练语音合成的阶段。

国外关于语音合成技术的研究可以总结为两个方面，一方面针对模型或算法改进，具有代表性的有 Naoto（2019）针对语音合成的不足，提出了一种新的分类器训练概念，该概念结合了代表器参数语音合成能力的正则化项。另一方面针对感知进行分析，Noé Tits（2020）提出了一种分析控制 TTS 系统参数对生成句子的感知影响的方法；Takuya（2019）则提出了一种情感语音转换方法，使用周期一致的声音生成对抗网络，从中性语音中产生情感发声；而 Kristen（2019）针对语音合成技术中语音捐赠的道德问题进行了研究。此外，Nagata（2020）尝试利用语料库进行笑声的合成，对自然笑声进行音标标注，并定义了合成声音所需的语境，通过声母或辅音的声学标记合成笑声并进行评价。

综上，虽然国外关于语音合成技术的研究明显早于国内，但国外研究方向大体上还是集中于计算机信息系统等理科领域，强调如何从技术层面提升语音效果。而国内研究方向已经开始从专门的技术研究向语言文字和新闻传媒等人文社会学科转变。AI 语音合成技术在新闻传播领域中的研究多是作为人工智能技术的一部分辅助讨论 AI 主播的新闻内容生产与应用，而缺乏纯粹的针对“声音”的研究。

1. 4. 2网络音频研究现状述评

网络音频平台在经历播客时代和移动时代后，来到了音频收听设备和用户收听场景的持续拓展的全场景时代。在“PGC+UGC”的生产模式下，各大音频平台内容呈现差异化特点，用户数量保持稳定增长，网络音频的发展逐渐趋于成熟。2020 年 1 月，荔枝作为中国在线音频市场首家上市公司，正式登陆纳斯达克证券交易所。目前，中国网络音频市场已正式进入全场景发展时代。

（1）国内研究现状

因为网络音频行业在我国发展历史较短，国内在此方面的研究还有所欠缺。截至 2022 年 12 月 25 日，中国知网数据库中 “新闻与传媒”学科分类下，以“网络音频”“网络音频”“有声书”“有声读物”为关键词，共检索到 400 多条结

果。从研究成果来看，将其分为媒介形态的转变、付费音频现象和移动网络音频的融合发展三个维度。

第一，媒介形态的转变。关于这一维度的研究，其中较多讨论的是传统广播媒体的转型，汪艳（2016）指出网络音频的移动性、伴随性特征加上语音交互、可穿戴设备技术的发展，使大众化广播走向“窄播”音频，移动网络音频媒介展现出无法阻挡的发展趋势[①]。牛沛媛（2018）以中美两国各具代表性的广播流媒体客户端阿基米德 FM 和 iHeartRadio（心动广播）为例[②]，探讨从传统广播到网络音频客户端的媒体演化之路。到 2020 年，智媒融合更是加速推进了媒介形态的革新与转变。段宇、温蜀珺在其文中提到智媒时代的媒介生态发生巨大变化，而 5G 更加催化媒介变革，网络音频应借助技术力量致力于场景创新[③]；张路琼（2020）将音频平台分为三类：综合性音频平台、垂直有声阅读平台和音频直播平台，同时提出技术的发展创造新的媒介，重现声觉空间，为网络音频的场景化和智能化发展开辟新路[④]。

第二，付费音频现象。付费音频是网络音频领域一个十分值得讨论的话题，他的良性发展可以很大程度促进高质量声音产品的产生，而针对付费音频的研究主要集中于用户付费意愿与付费平台发展进路。李武、胡泊在《新闻大学》发表的《声音的传播魅力：基于音频知识付费情境的实证研究》结果表明[⑤]，声音吸引力对用户付费意愿有显著正向影响，以此证明声音效果的重要性。栾轶玫教授（2018）分析了付费音频产品与平台的现状和问题，提出建构良性移动付费“音频生态圈”，强调了平台意识与版权意识[⑥]。

第三，移动网络音频的融合与发展。在网络音频的媒介融合方面，赖黎捷（2020）从广播网络化发展趋势出发，结合广播音频与互联网音频各自的特点，梳理二者互融的现实路径，提出以人工智能技术为支撑，打造智慧型广播[⑦]；骆

[①] 汪艳. 移动互联网时代的媒介产品新形态[J]. 新闻研究导刊, 2016, 7(01) : 177+193.

[②] 牛沛媛. 传统广播向网络音频客户端的转化——以阿基米德FM和iHeartRadio为例[J]. 传媒, 2018, (19) : 48-50.

[③] 段宇, 温蜀珺. 智媒时代下音频节目的结构嬗变与内容创新[J]. 视听, 2020, (04) : 13-14.

[④] 张路琼, 崔青峰. 网络音频的传播特征及媒介演变[J]. 青年记者, 2020, (29) : 75-76.

[⑤] 李武, 胡泊. 声音的传播魅力：基于音频知识付费情境的实证研究[J]. 新闻大学, 2020(12) : 49-60+120.

[⑥] 栾轶玫, 周万安. 传统广播转型新方向:移动付费“音频生态圈”[J]. 新闻与写作, 2018, (10) : 44-47.

[⑦] 赖黎捷, 颜春龙. 广播音频与互联网网络音频的融合发展[J]. 中国广播, 2020, (08) : 32-36.

蓓娟（2019）以听听 FM 与百度智能音箱的合作为例，提出其应以 AI 智能加速助力媒介融合。在网络音频的发展进路方面，刘志国和路金玉（2022）从媒体深度融合背景出发，讨论了网络音频平台的技术应用缺失问题，并提出了改进与优化策略[①]；李霞飞（2017）提出网络音频 APP 应与车联网联动发展，同时开发可穿戴设备和智能家居，使信息获取更加人性化和智能化。

另外，关于专门研究“语音技术+网络音频”的成果为零，考虑到有声读物和有声书是移动网络音频中的重要服务内容，所以“语音技术+有声读物”和“语音技术+有声书”的研究也具有重要参考价值。刘一鸣、高玥（2019）通过讨论有声读物与智能语音的融合发展方法，尝试提高有声读物的制作效率与质量[②]；任子寒、姚瑶和余人（2021）提到语音交互技术为人们提供良好交互体验的同时也存在潜在风险，侵权现象时有发生甚至有加剧之势，会对有声领域的健康发展造成危害[③]。

（2）国外研究现状

国外对网络音频产业的研究主要集中于“播客”，“播客”(Podcast)一词源于“iPod”与“广播”(Broadcast)的结合，它是一种在互联网上发布文件并允许用户订阅从而自动接收新文件的方法，或用这种方式所产生的无线电台节目。播客作为以互联网为载体的个人电台，实现了自由度极高的广播，也让音频内容的价值开始凸显。国外相关研究分别讨论了播客现状、播客用户使用习惯以及播客商业发展模式。全球最大的流媒体平台 Spotify 在 2018 年收购了 Gimle Media 和 Anchor，添加了播客服务，为用户提供创作平台，创新生产模式，对传统播客产生了重要影响。Toni（2019）分析了西班牙播客的发展现状：西班牙媒体集团 PRISA 借助互联网和数字化技术推出了 Podium 播客，这可以说是传统播客形式发生根本性转变的关键节点。部分美国学者首次从播客的使用动机和方式上对用户群体进行了大规模调查，其研究探讨了影响消费驱动和行为的因素，最后得出娱乐和信息优势是激励美国音频消费的重要动因。此外，该研究还发现听众的行为动机会影响他们的收听选择，例如收听习惯、收听内容、收听的设备设置等。Andrew 认为学界对于音频的研究正在减少，他指出已经发表的有限研究大多是

[①] 刘志国, 路金玉. 媒体深度融合背景下我国网络音频的发展路径新探[J]. 出版广角, 2022, (09) : 83-86.

[②] 刘一鸣, 高玥. 人工智能语音在网络音频中的应用研究[J]. 出版发行研究, 2019, (11) : 35-39.

[③] 任子寒, 姚瑶, 余人. 语音交互技术在网络音频中的应用风险与防范策略[J]. 编辑学刊, 2021, (04) : 18-23.

在播客出现初期进行的，以当前的研究情况而言急需更新。虽然有大量研究正在分析教育教学技术和新闻广播生产的播客，但播客文化、播客受众、技术属性、播客美学等却很少被注意。

可见，关于网络音频的研究正处于发展阶段，虽然有不少文献都提到可以借助科学技术推动音频内容的生产和创新，但无论是国内还是国外，针对语音合成技术与网络音频相结合的研究仍比较稀缺。

1. 5创新点

一直以来，视觉都是大众普遍关注的方向，视觉感官也一直占据主流，听觉作为重要的感知器官却难有一席之地，本研究结合听觉相关理论研究技术驱动的网络音频平台，强调听觉回归的重要性，通过事实案例证明声音信息的应用价值。从研究内容来看，人工智能技术的突飞猛进确实拓展了技术与新闻传播领域的研究范围，但是将AI语音合成技术与网络音频相结合的研究寥寥无几。从研究对象来看，本研究以AI音频作品为研究对象，通过查找大量资料，如AI语音技术原理、技术引入时间、技术供应公司、技术使用价格，统计分析AI应用功能、AI语音质量、AI音频内容，丰富学术资料，对技术赋能音频的相关研究具有重要价值。

2AI语音合成技术与网络音频平台的相关概述

AI语音合成技术作为人工智能技术中的关键一环，经过长时间的发展，已经形成了较为成熟的技术模型与原理。而网络音频平台是新闻传播范围内除视频平台外，最值得研究的新媒体平台。两者所属领域看似相差甚远，但随着技术的不断更迭，AI语音合成技术与网络音频平台的交融逐渐加深。在此之前，本章对AI语音合成技术和网络音频平台的相关内容进行梳理，以了解技术与平台的发展演变过程，以及现阶段语音合成技术在网络音频市场中的基本情况。

2. 1AI语音合成技术相关概述

语音合成技术作为一项前沿的智能信息处理技术，融合了多个学科领域，其所要解决的主要问题是如何将可视文字信息转化为可听的声音信息。语音合成可以划分为三个级别：Text To Speech（文字到语音的合成）、Concept To Speech（概念到语音的合成）、Intention To Speech（意向到语音的合成）。目前的 AI 语音合成技术还处于 TTS 技术阶段，即从文本到语音的转换技术，利用电子和机械的方法产生人工语音。

2. 1. 1AI语音合成技术的发展历程

关于语音合成技术的研究已有近百年历史，早在现代电子信号处理技术被发明之前，人们就开始努力制造能够生产人类声音和语言的机器或机械装置。1779年，克拉森斯坦(Kratzenstein)研发出一种机械声学合成器，这种合成器是用簧片、风箱和皮革等材料制作而成，用于模拟人类的声带、肺部和声道，可以通过改变共振腔的形态来合成各种不同的声音要素。这可以说是人类历史上第一次人工声音合成技术。

到了 19 世纪，随着电子器件问世，语音合成技术有了长足的进步。其中，最具代表性的是贝尔实验室 H.Dudley 于 1939 年设计的“VODER”电子合成器，该系统利用类似于白噪音的激励产生非浊音信号，利用周期性的激励产生浊音信号，模拟声道的共振器是一个 10 阶的带通滤波器建模，模型增益通过人来调节。

1980 年，D. Klatt 设计出串/并联混合型共振峰合成器，它的结构模型如图 2. 1 所示，该模型可以模拟不同的嗓音，经过精心调整参数可以合成出非常自然的语音。美国 DEC 公司的 DECtalk 系统正是采用了 Klatt 的串/并联共振峰合成器，使其能够与标准的接口和计算机联网或独立的电话网络相连接，从而实现多种声音信息，其发音更加清楚，有 7 种声音可供使用者选用。

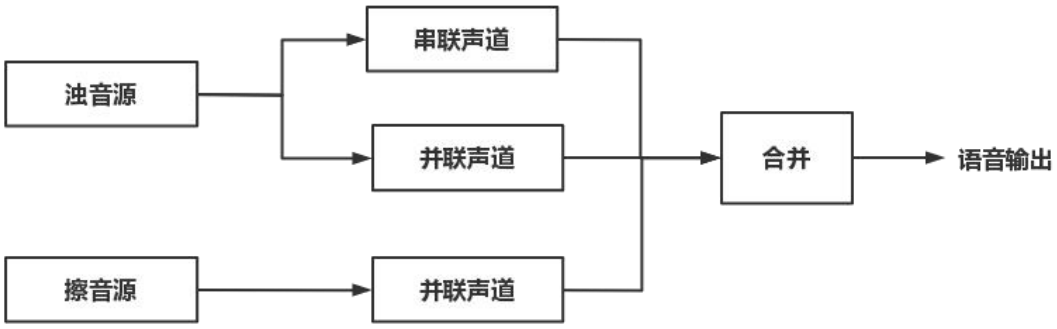

图2. 1 Klatt共振峰合成器结构模型

多年的研究与实践发现，尽管利用共振峰合成器可以获得多种具有真实感的人工合成语音，但共振峰合成机中的峰值参数很难被精确地提取出来，且合成声音的总体质量很难满足文语变换的实际需求。20 世纪 80 年代末，基音同步叠加方法(PSOLA)的提出使基于时域波形拼接方法合成的语音在音色和自然度方面得到了极大的改善，有效地解决了语音片段的拼接问题。

1990 年，由于计算机的运算量、存储量等方面的巨大进步，使得基于大语料库的单元挑选与波形拼接合成方法逐渐成熟。它的核心思想是从预先录制和标注好的语料库中选取合适的单元，通过细节调整（或不调整），拼接得到最终的合成语音，其优势在于保持了高质量的原始声音。

随着 AI 技术持续发展，基于深度学习的语音合成技术逐渐被人们所知，如 DNN/CNN/RNN 等多种神经网络结构都可以用于对语音合成系统进行训练，而深度学习的算法可以更好地模拟人声变化规律。自微软研究院在 2011 年推出的一种基于背景关联的深层神经网络(DNN)和隐马尔可夫(HMM)的声学方法在大规模语料库中取得了很好的效果后，许多学者将目光投向了深度学习的智能语料库。

2. 1. 2AI语音合成技术的相关原理

从语音技术的发展历程来看，语音合成技术的原理主要可以分为三大类，分别为波形拼接、参数合成和深度学习，这三类语音合成方法各有优劣势。

（1）波形拼接方法

波形拼接方法是将预先录制的语音音节存储在机器中，在合成语音的时候按规则挑选出已存储的语音音节，再通过拼接算法将它们进行组合，最终输出合成后的连续性语音，其技术原理如图 2. 2 所示。

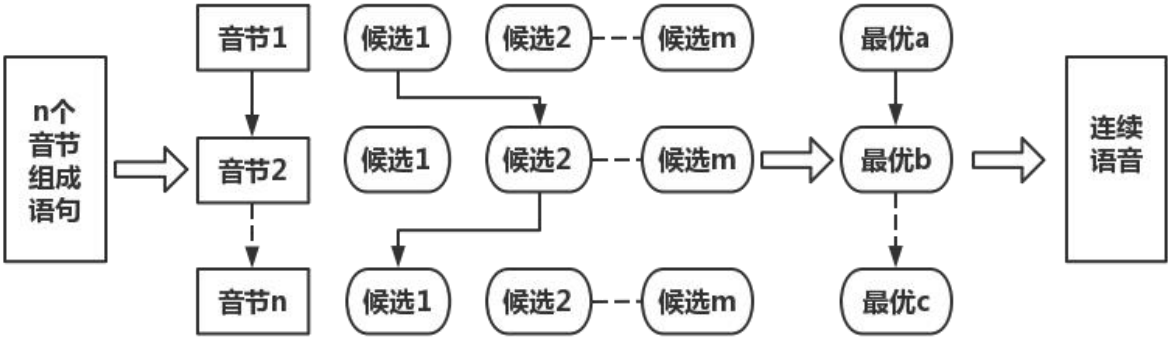

图2. 2 波形拼接语音合成技术原理

PSOLA 算法属于典型的波形拼接方法，它按照上下文拼接单元，调整韵律特征，合成波形保留了主音发音段的功能，可以获得具有很高可懂度和自然度的合成语音。波形拼接方法的优势在于可以对原始语音单元进行直接拼接，若音库的容量足够大，或当已有语音库与合成文本属于相同领域时，合成语音的自然度会更好。但该方法也存在一些缺陷：一是依赖大数据库的波形拼接方法需要预先录制一个语音库，该语音库会占用大量存储空间，因此在智能移动终端设备上的应用存在局限；二是此方法合成的语音相对比较单一，无法及时满足用户不断更新的使用需求；三是拼接时的韵律调节范围受到较大限制，当调节幅度过大，则会影响合成语音的质量。

（2）参数合成方法

参数合成方法充分利用了数字信号处理技术，该系统在语音分析训练阶段，根据语音合成的特点，将语音波形通过声码器转换成频谱、基频、时长等语音参数。基于隐马尔可夫模型的参数合成系统结构如图 2. 3。

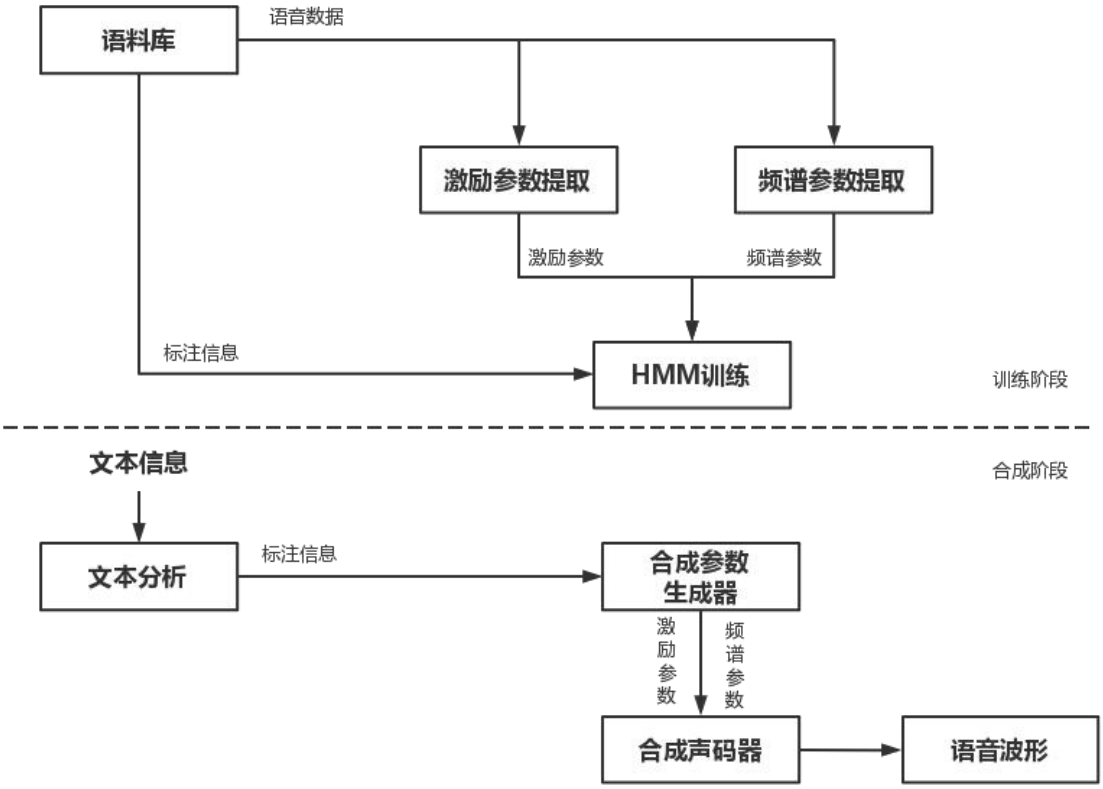

图2.3 基于隐马尔可夫模型的参数语音合成系统结构

这种方法首先对语音数据进行激励参数和频谱参数的提取，再与语料库中的标注信息进行结合训练，当激励源为周期脉冲序列时可以得到池音的声带振动，当激励源是随机的噪声序列时则得到清音。将训练得出的参数用于文本中的标注信息，通过声码器进行编辑最终得出语音波形，适当调整这些参数就可以发出不同的声音，而调整激励参数的周期或强度，就能改变合成语音的音调等。可见，只要激励参数与频谱参数合理，合成器就能自由地合成用户所需语音。参数合成方法最大的优势是灵活且易于转换，便于调整语音的时长、基音周期等参数，并通过最模型自适应方法进行参数切换。在转换过程中，无需重复录制音库，仅用训练小量数据就能完成语音合成。但参数合成方法的弊端在于经过反复的语音到参数再到语音的转换过程，合成语音的自然度不尽人意。

（3）深度学习方法

智能语音技术已经进入以深度学习方法为主导的发展阶段，该方法主要通过模仿人体的神经元机制来分析数据，利用多层的人工神经网络结构学习数据的内部规律，组合低层的具体特征，从而得到高层的抽象特征，该方法包含前馈神经

网络(FNN)、卷积神经网络(CNN)、循环神经网络(RNN)、长短时记忆网络(LSTM)等。其中，卷积神经网络是最常见的深度学习网络架构，受生物自然视觉认知机制启发而来，主要用于提取局部特征，其突出特点包括：局部感受野、权值共享和池化操作。CNN 网络能够从大量的样本中自主学习新的特征，并将其扩展到同类样本中。该算法将所有的信息处理成一个叫做“卷积核心”的“滑动窗口”，这个“滑动窗口”可以通过多个特征间的共享来达到对神经网络参数进行有效压缩的目的。为了提升网络的健壮性，通常在卷积操作之后会增加一步池化操作。基于上述特征，卷积神经网络已经成为当下最受关注的深度学习模型之一。

伴随着 Tacotron、WaveNet 上线，端到端的语音合成进入人们的视野。以 WaveNet 为例，它就是以深度神经网络方法为主导，不需要依赖任何发音理论模型，也不用对声音信号本身做各种理论模型及简化假设，因此受到的语言学限制更少，可以使神经网络的学习能力得到充分发挥。通过 WaveNet 还原的语音发声自然，细节丰富，其效果与人类产生的真实声音不相上下，缺点在于生成语音的效率低、速度慢。为此，研究人员基于深度学习提出了新的语音合成模型，主要分为以下三类：

频谱预测模型如图 2. 4 所示，从文本信息中预测声音的某种表示特征，通常使用线性频谱或梅尔频谱。

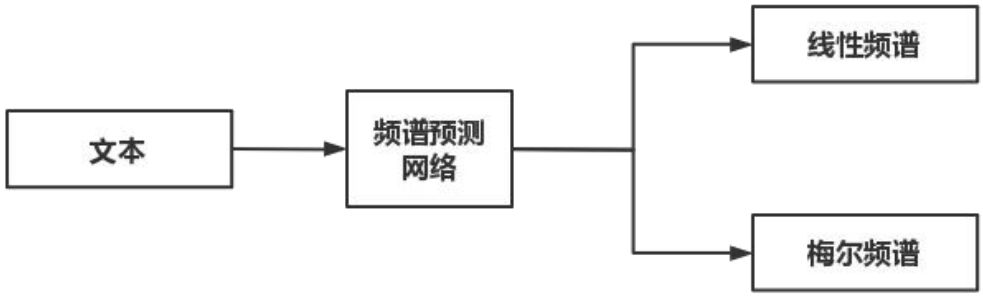

图2. 4 频谱预测模型

声码器的声学模型如图 2. 5 所示，根据语音信号的特征参数恢复声音的原始波形，在深度学习中通常使用频谱预测模型生成的线性频谱或者梅尔频谱作为输入。

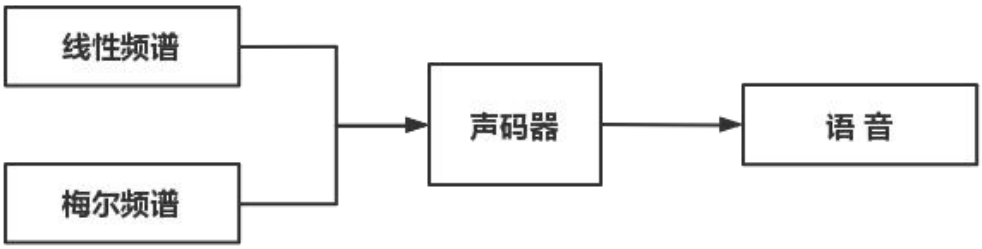

图2. 5 声码器声学模型

完全端到端的模型是将频谱预测模型和声码器相结合，直接从文本信息得到声音的原始语音波形，如图 2. 6 所示。

图2. 6 完全端到端模型

当前，语音合成模型以深度学习方法为主，以上提到的基于深度神经网络的语音合成模型也是各音频领域使用最为广泛的模型，所以本研究所提及的案例皆是以此类模型为基础，展开合成语音在网络音频平台的使用与发展研究。

2. 2网络音频平台的整体概况

网络音频是互联网时代的新兴产物，是传统广播在新媒体时代的转型之作。网络音频虽同时包含电脑端设备和移动端设备的音频内容，但由于移动端已成为当下主流，现在广泛将其定义为：以智能手机（包括平板设备）、车载、智能家居、可穿戴设备等终端为音频载体，通过在线和离线的形式为用户提供音频收听、个人录制与分享等服务，包括广播电台、音乐娱乐、新闻资讯、相声评书、财经教育等音频内容的平台。

2. 2. 1网络音频平台的发展历程

每一种后继的媒介都是对先前媒介不足的补救或补偿[①]，在满足人类不断变化的需求的过程中演进和进化，网络音频平台的出现印证了这一说法。

网络音频从传统广播衍生而来，很大程度上得益于汽车无线电技术的发展。21 世纪以来，随着中国经济体量一路腾飞，中国汽车保有量也在不断增长，巨大的车载音频广播市场为网络音频平台提供了潜在的发展空间。

最初，网络音频必须在电脑终端上在线收听，但由于台式机器的不便携带性，限制了网络音频内容的发展和传播。到 2010 年，互联网与手机等移动通讯设备的普及为网络音频平台的快速扩张提供了技术支持，网络音频平台进入起步阶段，覆盖率低，用户数量有限，彼此之间竞争性不强。各音频平台依然保持着传统广播媒介的模式，只不过将传统电台的内容迁移到网络平台上，缺乏革新，对音频平台的早期发展造成一定影响。

2011 年 9 月，蜻蜓 FM 正式上线，成为国内首家网络音频应用平台。随后，喜马拉雅和荔枝也纷纷上线，用户通过手机终端就能录制、上传内容，开启了网络音频平台 UGC 生产模式的新局面，自此积累了大量用户和内容创作者。三大音频平台也进入了网络音频市场的差异化竞争阶段。

各网络音频平台以“内容为王”，在带来日益丰富的内容的同时，也不可避免地导致了庞大的内容泛滥，这使得高质量的优质音频内容成为越来越重要的稀缺资源。而如何在新的网络音频市场环境下保留并吸引更多的用户成为各大音频平台迫切需要解决的问题。基于此，各大音频平台开始转变发展理念，拓宽经营范围。2016 年，各平台开辟了新的内容类型，实现对内容的精准投放，创新盈利模式，展开知识付费的尝试。网络音频平台相互促发的媒介环境变化为行业发展提供了契机，使资本对这种新型媒介形态产生兴趣，大量资本的注入使网络音频平台得到进一步发展与扩张。

从2017年开始，各音频平台开始谋划自身发展方向，头部平台更是竞争激烈，争先打造音频界的超级平台。2018年8月，蜻蜓FM上线智能语音功能，首次将语音合成技术引入音频平台。2019年，喜马拉雅也引入语音技术，随着近几年的不

[①] 保罗•莱文森. 数字麦克卢汉——信息化新纪元纪指南[M]. 何道宽译. 社会科学文献出版社, 2001:16.

断更新与发展，喜马拉雅平台的AI语音服务已经较为普及。可见，音频平台除了发展自身业务之外，也开始布局跨界合作。

2.2.2网络音频平台的种类

在网络音频出现的数十年时间里，其平台逐渐朝精细化、垂直化的方向发展。目前市面上的网络音频平台大致有三大类别[①]：综合性音频平台、垂直类音频平台、音频直播平台。

（1）综合性音频平台

在所有网络音频平台中，综合性音频平台是数量最多、用户使用量最大的一种平台类别，它打破了专业音频生产者与普通用户之间的界限，打破了主流电台与商业电台的壁垒，结合了互联网的移动性、共享性、即时性等特点，把音频产品的生产、推广、营销等聚合在一个平台中。综合性音频平台最大的特点在于音频内容种类繁多，涵盖了娱乐、儿童、广播剧、人文历史、财经知识、情感讲座在内的数十种音频内容类型。同时，它还在超出内容本身之外的音频社交、主播孵化、衍生产品等方面进行布局，满足用户不同的需求。目前市场占有率最高的综合音频平台主要以喜马拉雅 FM、蜻蜓 FM、荔枝 FM 等为代表，它们作为用户使用与接触最为频繁的音频应用，在用户规模和用户黏性方面都占据绝对的优势地位。

（2）垂直类音频平台

区别于综合性平台广泛而完整的音频内容，音频市场中出现了一系列专注于细分领域的音频产品，主要为用户提供精准化、个性化的音频服务。其中，具有代表性的为教育类音频平台和有声阅读平台。

教育类平台的音频内容主要以儿歌、童话寓言和课文教材为主，在市场竞争压力下开发出诸如儿歌伴唱、音乐助眠等多种功能，以满足听众的个性化需求。除此之外，部分平台引进知名少儿主持人录制音频内容，有效增强了音频产品对低龄听众的吸引力。

有声阅读平台的音频内容主要包括小说、动漫、图书、杂志等，以微信阅读、懒人听书等为代表，此类平台兼具了娱乐性与文学性，在我国一直拥有着庞大的

[①] 张路琼, 崔青峰. 网络音频的传播特征及媒介演变[J]. 青年记者, 2020, (29) : 75-76.

市场空间。在互联网时代，有声阅读平台的诞生满足了用户随时随地收听文学作品的需求，听众的阅读载体、知识获取方式、内容的分享和交流渠道也都发生了翻天覆地的变化。

（3）音频直播平台

与视频直播不同，音频直播平台的主播不需要出镜，而是采取“声音为主、图片为辅”的直播形式，将与用户的社交互动作为重点，在直播内容上更具生活化和随意性，致力于通过优质音频内容与听众建立紧密联系。荔枝 FM 自 2017 年尝试音频直播形式后，就逐渐转型成为直播类平台，其音频直播内容丰富，覆盖面广，是网络音频直播领域占有率较高的平台之一。随着听众趋于个性化的收听要求，克拉克拉成为该领域的后起之秀，它的传播内容以二次元文化为主，强调音频内容的趣味性，是一款内容定位更加迎合年轻听众的轻偶像直播互动平台。同视频直播的发展趋势类似，音频直播也正从综合类直播向更加精细化的音频直播平台类型发展。

2. 3AI语音合成技术在网络音频中的基本情况

AI 语音合成技术与网络音频内容的融合源起——AI 导读，这是有声平台最早的 AI 语音作品形态。AI 导读作为国内率先实现人工智能浓缩书的产品，通过提取一定比例（一般为 10%）的全书干货，利用 AI 语音速读方式，发挥导读作用。据统计，喜马拉雅平台的 AI 导读作品合计 283 部，上传时间集中于 2019 年 1 月至 6 月。其中最早的作品是 2019 年 1 月 29 日上传的《AI 导读：全球通史（中）》和《AI 导读：全球通史（下）》。

直到 2019 年 12 月，AI 导读系列作品陆续停更，但这并不代表 AI 语音合成技术退出网络音频领域，相反它开始迎来更多元化的作品形式，如 AI 读书、AI 电子书、AI 单播和多播等。第一个非导读 AI 作品为主播“华章有声读物”于 2019 年 6 月 5 日分享在喜马拉雅平台的《学会决断 AI 版》。在这之后，不断有 AI 主播入驻网络音频平台，其中不仅有专业团队，还有许多个人用户。AI 语音合成技术渐渐被应用于广播、小说、商业、生活等各类主题的作品中。

而在这背后支撑语音合成技术顺利进入网络音频领域的是各大语音合成技术公司。科大讯飞是我国规模最大、覆盖面最广的智能语音厂商，在智能语音技

术领域拥有较高话语权，它不仅成立了专门的讯飞开放平台，还开发了一款专业的手机语音识别类产品——讯飞有声，通过语音来操作设备以及通过语音输入来代替各项服务。语音合成服务在150万次—1亿次不等，价格在每年4060元—7万元之间，同时提供免费测试包、3个自选发音人和音库定制2次，深度定制个性化声线。

百度在人工智能浪潮下，进军智能语音领域，于2019年4月成立百度智能云，力求在人工智能、大数据和云计算服务等领域领先全球。百度语音音库分为基础音库和精品音库，主要以次数包和字符包规格为单位进行计费，价格在1200元-15万元不等，2022年9月上线长文本在线合成功能后，可以一次性合成10万字以内的文本。

腾讯在2016年4月创建AI Lab（人工智能实验室），以其社交数据大平台为依托，累积了数十万小时的语音数据，拥有海量语言信息，可以轻松处理各类语音问题。腾讯云平台的语音合成技术主要采用字符数计费的方式，分为语音合成标准音色和语音合成精品音色，不同服务的付费价格也不相同，而针对首次使用语音合成服务的用户，腾讯云平台提供了800万字符的免费调用额度。企鹅FM就是以腾讯为后盾，按作品收取费用，每部作品定价在10—30金豆不等（10金豆为1元），购买整本专辑就可以永久收听。

同时，喜马拉雅在语音合成技术领域投入大量精力，建立智能语音实验室，专注于对语音识别、语音合成、信号处理、编码解码和智能音效等方面的研究。2022年初，喜马拉雅自主研发的跨语言语音合成创新技术论文，及其与中国科学技术大学合作的说话人日志技术研究论文，均被世界顶级学术会议“2022年IEEE国际音频、语音与信号处理会议”（2022 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022）收录。喜马拉雅平台的AI语音作品开通限时免费服务，不同作品的免费时长均不相同，在7—18天之间，到期后该作品会改变“限免”标识为“VIP”标识，即会员免费听。

根据对技术引入情况、技术供应商以及技术使用价格进行简单梳理后，现对网络音频平台中AI语音合成技术的基本情况已有大致了解。在此基础上，讨论AI语音合成技术应用于网络音频平台的实践情况、问题及对策具有一定意义。

3AI语音合成技术在网络音频中的应用实践

AI语音合成技术塑造了一个人工声音景观，使AI语音打破传统声景思维，它不仅包含声音的物理属性，更多强调了是声音所表达的内容、收听声音的方式、企业或受众对声音元素的应用，本章将基于此着重探讨合成语音在网络音频中的应用功能和应用现状。如今，AI语音合成技术已在应用市场上大有作为，海量音频内容接入多种智能终端设备，如智慧课堂、智慧家居、智慧医疗等，这些智能终端大部分内置了语音助手，如百度旗下AI助手和苹果搭载的Siri，都能够以语音方式实现多场景的人机交互。而网络音频平台更多强调的是用户的听觉感受，所以AI语音合成技术应以服务用户的“耳朵”为基准，带来一场听觉盛宴。

3. 1AI语音合成技术在网络音频中应用的功能特点

计算机系统将复杂的文字信息转化为可听的声音信息，使其逐渐达到模仿人说话的效果。但在这一过程中，用户需求不断发生变化，他们不再满足于简单的文字到声音的转变，而是开始提出更高的要求，于是语音合成技术在网络音频中的应用开始具备基本的功能特点，即改变语速、改变音色和选中文段进行播报。

3. 1. 1调节语速的功能

碎片化阅读时代，用户专一地收听某一档广播节目或听取某一本书的时间越来越少，要在有限的时间内获取更多的声音信息，调节语速的功能就显得尤为重要。最初的合成语音调节语速档位较少，相邻档位之间差距过大，用户难以选择到最合心意的阅读速度。为此，网络音频平台细化了调节语速的功能，以此达到优化用户阅读体验的目的。具体情况如表3. 1所示。

表3.1 网络音频平台语速调节功能统计

网络音频平台	语音合成技术公司	调节语速功能	档位数量
荔枝	\	0.5-2倍速	6
喜马拉雅	科大讯飞、 喜马拉雅智能语音	0.5-3倍速	8
蜻蜓FM	百度语音	0.5-3倍速	9
企鹅FM	腾讯语音	0.5-2倍速	自由调节

根据表中所列情况，网络音频平台普遍采取了0.5－2倍速和0.5－3倍速的调节方式。其中科大讯飞不仅以传统的数字化方式进行呈现，同时以文字化表达为辅助，将0.5－2倍速表示为缓慢－很快。

虽然网络音频平台都以0.5倍速作为基础语速，但观察档位数量就能知道他们的间隔单位并不相同。荔枝和喜马拉雅都以0.25个单位依次递增，但前者缺少1.75倍速这一档位，后者在2倍速后直接增加1个单位为3倍速。蜻蜓FM的语速调节功能分为两个阶段，第一阶段从基础语速开始以0.25个单位增加到2倍速，第二阶段从2倍速开始以0.5个单位增加到3倍速，共分为9个档位，是目前以固定值划分语速档位的最大值。

企鹅FM以进度条方式呈现调节语速功能，以0.5倍速为起点，2倍速为终点，设置1倍速和1.5倍速两个刻度值，用户可以在这一区间内任意滑动滑块从而选择最合适的语速，最大程度满足用户的个性化需求。可见，网络音频平台的调节语速功能已经呈现出细分化特点。

3.1.2多项音色选择的功能

为了展现出语音合成技术的更优成果，网络音频平台会格外看重音色服务功能。当下文字作品题材愈加多元化，所以需要多种类的音色与之相匹配，女频中较受欢迎的甜宠言情类作品适合选用少女音，而男频中较火的题材多为都市玄幻类，更适合选用温暖男音。网络音频平台提供的音色服务如表3.2所示。

表3.2 网络音频平台的音色服务统计

网络音频平台	语音合成技术公司	音色种类	音色数量
喜马拉雅	喜马拉雅智能语音	AI主播：知性女声－喜小华、温柔御姐－喜小迪、温暖男声－喜小玖； 真人声音采集：喜小道－喜道公子AI、苏小刀－一刀苏苏AI	5
蜻蜓FM	百度语音	小度、小爱同学	2
企鹅FM	腾讯语音	标准男声、标准女声	2

企鹅FM的音色种类较少，只简单区分为标准男声与标准女声用以播读不同题材的作品，在这里不过多讨论。蜻蜓FM的主播小度和小爱同学均为女声，以讲述短文为主，小度以优雅女声讲述古今异事和怪诞杂谈，小爱同学以可爱女音讲述稚嫩童话和校园故事。

随着网络音频平台的持续发展，音色也丰富多变，喜马拉雅在这一方面表现得尤为突出。喜马拉雅平台发布了3种普通音色，分别为知性女声－喜小华、温柔御姐－喜小迪和温暖男声－喜小玖，虽然平台会根据作品内容自动匹配合适音色，但用户也可按自身意愿选用。与此同时，平台还采集了签约主播“喜道公子”和“一刀苏苏”的声音信息，推出了具备真人主播声音特点的男声喜小道和女声苏小刀，一定程度上突破了时间和空间的限制，更加接近真人播读场景。值得一提的是，喜马拉雅在音色服务功能上还添加了音效功能，包含三类：一是清澈人声，人声更突出明亮，感染力十足；二是剧院混响，提升空间感，让声音空灵悠扬；三是超重低音，声音浑厚深邃，富有弹性。这一功能作为一种新尝试目前还处于试听阶段，需要进一步调整，但可以肯定的是它能更好满足用户的娱乐和精神需求。

3.1.3选中文段进行语音播读的功能

除了语速和音色的多种选择，部分网络音频平台还可以选中文段进行语音播报，这一功能主要是针对提供文稿的音频作品存在的，沿用了在线阅读平台的类似功能，即用户点击任意文字内容，合成语音可以立即跳转实现智能语音播报，这使语音合成的过程更加机动灵活。

目前只有喜马拉雅平台具备选中文段播报的功能，且以句为单位，将句号作为分隔标志，用户通过选择不同句子实现音频播放。在部分在线阅读平台中，其播放界面会出现一个“文稿”选项，点击查看AI文稿，以句为单位选中文字后会弹出“播放”选项，点击播放就可以实现音频与文字的对应。喜马拉雅则不需要这些繁琐步骤，它的电子书AI朗读自带文稿，点击文字就可以直接播放指定句子的音频。

选中文段进行语音播报的功能效用虽不大，但它拓宽了用户收听书籍的选择范围，不仅限于书籍题材、语速和音色，更加细节到先播放哪一句话、后播放哪一句话和何时播放，最大程度给予听众选择的自由。

3. 2AI语音合成技术应用于网络音频平台的典型案例

在声音景观中，技术发挥了巨大的决定性力量，声音内容和场景凭借语音技术实现了改变和重构，为了感受技术手段所构建的虚拟声音场景，本研究采用案例分析法，直观地展现AI语音合成技术在网络音频中应用的方式、内容和影响，选取了综合性音频平台“云听客户端”“喜马拉雅”和垂直类音频平台“恐龙贝克”中的典型案例进行多角度分析，分别涉及新闻播报、有声听书和儿童伴读等音频内容，能够更全面地呈现AI语音合成技术的应用行为。同时，这些案例多发生于2020—2022年，具备相应的研究价值。

3. 2. 1云听客户端的相关应用

2020年3月5日，“云听”正式上线亮相，它是继“央视频”上线后中央广播电视总台推出的基于移动端发力的声音新媒体平台，是贯彻习近平总书记关于“守正创新，把新媒体新平台建设好运用好”的批示所做的一项重大战略举措。

云听客户端在原有“中国广播”基础上改造而成，激活用户规模早已超过千万人。同时，云听联接总台5G智能新媒体中心，对央视视频内容进行音频化再生产，对播音员主持人、资深编辑、记者、制作人以及总台独家版权资源进行了深度挖掘。云听基于总台“5G+4K/8K+AI”等新型技术，将人工智能和5G网络等技术运用到平台的研发建设中，为广播总台的频率升级及传统广播向网络音频转型提供技术和平台支撑。AI语音合成技术的应用就是云听客户端迈出的第一

步。

（1）AI主播实现全国两会报道

在2022年两会期间，云听AI主播与中国之声记者合作实现了全国两会报道的首次AI播报，其中超过85%以上的播报内容皆由AI主播提供。在两会期间，云听AI主播与中国之声的记者团队组成“两会报道融媒联盟”，该联盟可以用200字/分钟的速度将代表委员们的意见回传至云听，大幅提升制播效率。

在云听客户端中搜索两会相关内容，第一条精选搜索结果就是“两会快报”，平台对他的简介是“AI播报，快听两会”，截至到2022年3月11日，该档栏目已经更新246集，更新频率集中在3月4日至9日，即会议高峰阶段，平均每集时长在20秒—5分钟之间不等，皆为两会短讯，相关咨询由中国新闻网、央视新闻、人民日报等提供，具有时效性和权威性。总体播放量达到97.03万，可见听众对AI合成语音制成的新闻接受度远比想象的高。

此次两会新闻播报的声音信息皆来源于中央广播电视总台的播音员主持人，通过采集主持人声音特点和语言习惯，利用AI语音合成技术打造而成的主播，声音流畅、调值准确、音质清晰，呼吸的气声也完美再现。

在此基础上，如何保留原本声音的“人性温暖”成为最值得关注的话题。因此，模拟主持人原生声音的AI数字化开发成为重中之重，此举一方面可以盘活原有的声音资源，凸显主持人原生IP的价值；另一方面可以通过引入AI技术，形成“人声”与“人工声”共存的独特IP，使优质的声音得到最大化的开发和价值体现。云听则通过对主持人声音的模型训练和深度学习，实现了词汇语句、情感逻辑和语言方式的个性化表达，从而打造出适用于各类资讯板块的AI主播。在此次报道中，云听推出AI主播团体IP“云小天团”，打造定位于政策要闻的“云小琦”、定位于社会热点的“云小宇”，以及定位于行业报道的“云小江”，多角度覆盖两会议题，全面布局两会报道。

云听借力AI技术不断优化技术手段和报道方式，将与生活关系密切的“四个创新”落到实处，打造出富有科技感的两会音频频道，用技术强化媒体融合，用AI拓宽传受渠道，以声音为时代切片，使其成为主流媒体新闻播报的一大亮点。

（2）AI音频赋能H5

驾驶场景下的封闭环境为音频提供了绝佳空间，赋予用户更强烈的音频内容

消费需求，与广播的先天基因高度契合。在此前提下，云听与主流汽车厂商合作开发车联网产品，使其在虎年春节期间携手比亚迪汽车推出春节拜年H，将拜年和贺卡等春节经典形式相结合，加入音频和AI元素，极大提高了祝福的趣味性。

这支H5的AI语音合成元素体现在新年贺卡生成的过程中，用户可以选择女主播、男主播、童声和方言这四种AI语音。云听通过这样的方式帮用户“说出”祝福，尤其是童声和方言这两种选项，让用户的祝福语变得生动有趣。这些AI合成语音同样是基于总台海量节目内容资源和播音主持人才资源，通过对真人声音的采集与学习制作而成的，使用户在祝福编辑环节挑选相同或不同的句子，最后也呈现出各不相同的语音效果。其中，可选的祝福对象覆盖到多个年龄阶层，例如送给银发一族的“福星高照”、为上班族准备的“工作顺心”和契合Z世代沟通语境的“不会be”等。这些文案配上H5提供的句式模板，重新组合生成的语音为用户提供了春节祝福的新思路。云听这支H5还呈现出定制化特点，从祝福对象、祝福语、AI语音的选择和背景音的匹配，再到全家福的合成，用户在每一环节都拥有绝对的选择权，不仅让接收祝福的用户产生被重视的情感认同，用户收到祝福后的主动分发和二次传播概率也会随之大幅提高。

（3）AI主播主持《数说中国故事》

2022年9月19日，由经济之声和云听共同打造的声音纪录片《数说中国故事》正式开播，7:30在经济之声《天下财经》、12:00和18:30在《环球新财讯》持续滚动播出。目前共有1151人订阅。节目具体内容如表3.3所示。

表3.3 声音纪录片《数说中国故事》具体节目内容梳理

	主题	节目数量/集	播放量/万	播放内容(节选第一集)
具体内容	创新	6	93.75	小镇1天为何涌入近50万游客？
	协调	6	80.94	这里的“新社区工厂”， 如何吸纳就业3万多人？
	绿色	6	66.03	1吨好空气为什么能“卖”50元？
	开放	6	47.22	进口游戏机 1天到货的“妙招”是什么？
	共享	6	51.16	半夜直播4小时 怎样卖货上千美元？
总计	\	31	353.09	\

该声音纪录片以“创新、协调、绿色、开放、共享”五大新发展理念为主题，分为对应的五个篇章，每集选取一组相关数据，从数字视角讲述国人新生活。节目共计31集，除去1集总预告，每一篇章均为六集，在总预告片中有“AI主播，讲给你听：我是云听AI主播‘云小听’”的表述，合成语音以女主播的声音形象出现，为听众讲述数据故事。其中，每期节目通过5—6分钟的短音频，围绕数据，以点带面，以近十年内发生在身边的小故事折射我国经济社会发展的大成就，运用大量真实声音素材，呈现出鲜活震撼的故事场景。此外，节目通过收听打卡与听众形成每日互动，每集均有10条评论，从“继续打卡AI主播”“好节目还想听”和“怎么就播完了，感觉没听够”等评论中可以看出，听众对AI主播的满意度和期待值并不低，是云听利用AI技术的一次成功尝试。

此外，云听和倒映有声科技公司在人工智能语音领域也展开了长期深入合作。倒影有声作为一家TTS(Text To Speech)科技创新企业，早在2020年就上线了网络音频制作平台，以每天单机生产500万字的速度，配合真人主播，不仅能节约90%以上的录制成本，同时能满足多种内容音频化需求。2021年3月，双方签署战略合作协议推动更多基于总台IP的人工智能语音产品，现在云听平台中已经可以搜索到由倒影有声制作而成的语音合成新闻。

云听作为主流媒体音频平台，对AI语音合成技术的应用起到了领头作用，有效促进了行业优质音频服务发展。

3. 2. 2喜马拉雅的相关应用

喜马拉雅自成立以来，一直秉承“用声音分享人类智慧、用声音服务美好生活”的初心，以丰富的音频内容连接了数亿人，搭建了一个内容创作者和听众共同成长的网络音频平台。一方面，创作者用声音分享故事、收获粉丝、增加收益；另一方面，听众从中获得陪伴、获得心灵和精神需求的满足。

现如今，喜马拉雅的成就已经进入“让技术加持声音、让声音打开想象”的阶段。前文提到，喜马拉雅在AI语音技术领域研究多年，在说话人日志和跨语言语音合成方面获得大量认可。在实际应用中，喜马拉雅的AI语音合成技术已经体现在评书、新闻、故事、小说等多种内容的创作中，帮助喜马拉雅在现有的“UGC + PGC + PUGC”内容生态基础上，进一步拓展AIGC的可能性。

具体而言，喜马拉雅语音团队利用生成式对抗网络研发出声码器“PhaseGAN”，这种基于生成对抗网络的声码器拥有比WaveNet更高的生产效率，WaveNet作为谷歌DeepMind用于生成原始音频波形的深层神经网络模型，本身在短时间内能将原始模型效率提高近千倍，PhaseGAN的效率远在它之上。而喜马拉雅的TTS模型系统更是占有其他网络音频平台所不具备的独特优势，在语音合成、语音识别、语音编解码以及语音信号处理等技术上，喜马拉雅进行了深度研发，通过基于BERT模型的多任务建模，在文本正则化、多音字识别和韵律预测等任务上，取得一定成就。例如TTS前端文本处理分析模块，高精度、全自动地对文本进行韵律预测，同时采用并行解码器，生成语音合成序列，改进语音合成后端模型的结构。下文将对喜马拉雅的AI语音合成技术应用实践成果进行详细阐释。

（1）AI实现单田芳声音重现

单田芳先生作为我国著名评书艺术表演大师，他的从艺生涯表演录制完成了111部共1.5万余集广播电视评书作品，书迷遍布全国各地。单田芳作为中国评书事业承上启下的关键性人物，他不仅促进了评书艺术的发展，同时开创了评书市场先河。

2021年，在北京单田芳艺术传播有限责任公司授权下，喜马拉雅团队拜访采集了单田芳生前演出声音，通过AI语音合成技术完美复现了单田芳先生苍劲沙哑的独特嗓音，在逝世三周年之际推出了“单田芳声音重现”系列专辑，让已故评书大家重回大众视野。

与普通的合成音频相比，评书中含有更多情绪表达，尤其单田芳先生擅长用声音刻画人物角色，声音韵律节奏变化较大，如果仅靠现有的语音合成框架模型做提取，最终合成评书的整体情绪会趋于平淡，缺乏原作的高低起伏。针对这一难题，喜马拉雅开发出一套可以支持多种情感类型和语音风格的技术模型，不仅可以解读不同的情绪文字，还可以分辨叙述和对白，极大丰富了合成语音所能表达的情感和节奏。所以，无论单田芳评书中的声音多么丰富多变，AI合成语音都能提取复刻。除此之外，单田芳评书中还存在很多区别于普通话发音的白话发音。如“这个”中的“这”字，普通话读作“zhè”，但在评书中通常读为“zhèi”。为了解决发音问题，团队开创性地设计了口音模块并专门标注了这些特殊发音，

从而使单式AI腔调更好演绎听众耳熟能详的经典之作。

目前，喜马拉雅已上线80余张单田芳先生的评书专辑，超两万条声音，多张评书专辑长期位列相声评书热播榜前列，例如《十二金钱镖》播放量高达721万、《毛氏三兄弟》播放量达525.6万、《左宗棠全传》播放量达425.2万等，不少作品的声音完播率远超普通人声作品。

令听众期待的是，喜马拉雅已经申请的语音合成相关专利中，自研跨语言语音合成技术可以使没有任何英语原始数据的合成声音说英语，这意味着，未来将可能听到单田芳先生播讲英文内容。

（2）AI超拟真有声书《智能交通》上线

2022年4月21日，AI超拟真有声书《智能交通》在喜马拉雅平台上线，以百度创始人、董事长兼CEO李彦宏所作《智能交通：影响人类未来10—40年的重大变革》一书为基础，使用李彦宏公开语音数据，通过AIGC创作生成。

这是喜马拉雅与百度成功合作的结果，从20万字专业文本到超拟真音频作品，仅使用李彦宏300句语音信息就生成了堪比真人声音的音频内容，达到普通用户基本无法区分真人声音或合成声音的效果。该有声书采用自然语言处理技术（Natural Language Processing, NLP）对文本进行预处理并添加韵律信息，再生成声学模型，对声韵母韵律表征进行建模，支持发音内容、风格和音色的迁移，与单纯的NLP不同的是这需要文本和声学联合建模以实现不同语境下的声学变化，最后通过高质量声码器还原语音，见图3. 4。

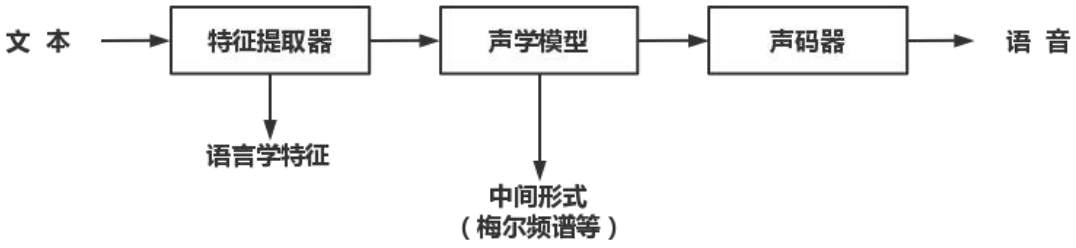

图3. 4 基于NLP技术的语音生成流程

目前，《智能交通》已更新102集，共分为智能信控、智慧停车和车路协同等12章，播放量达到254. 4万。从第一期节目标题“瑞萌：喜马拉雅声音AI虚拟人首次亮相”可以看出，节目设置了一位虚拟主持人，她的自我介绍是18岁的萌

妹子，作为喜马拉雅001号AI员工“瑞”的分身与听众分享咨询和知识，为其AI算法增加新的训练数据。在本节目中，瑞萌以AI互动官的身份与李彦宏进行虚拟对话，通过AI技术进行有关智能科技话题的十问十答，实现虚拟人与虚拟声音的首次问答模式，获取了李彦宏对人工智能和城市交通的战略思考和前沿观察。

除上述案例，新京报、潇湘晨报、时代周报等众多主流媒体进驻喜马拉雅，借助喜马拉雅TTS技术加速制作新闻类音频节目，为听众提供更多收听权威新闻的渠道。喜马拉雅与新京报合作推出的《鲸快报》专辑，连续几周位居喜马拉雅新闻专辑类榜首。

AIGC内容生产方式的兴起助力了AI语音合成技术在音频领域的应用，通过生成声音素材，以完成音频内容生产的自动编排和组合。喜马拉雅掌握这一方式将不仅实现简单的效率提升，还将激发出不曾有过的创作思路和创意认知。

3.2.3恐龙贝克的相关应用

AI智能技术在教育领域的应用是推动新一轮科技与产业变革的关键驱动力。据全球最大的企业增长咨询公司Frost&Sullivan预测，AI教育具有巨大的发展潜力，其国内市场规模将突破7000亿元。

目前，针对儿童早教启蒙的故事类有声平台众多，但都只是单纯的阅读形式和应用形式的转变，并不能最大限度满足儿童用户在成长过程中的需求。“恐龙贝克”作为智能亲子陪伴软件应运而生，它围绕“AI科技，智能早教”的理念，以小恐龙“贝克”为主要伙伴形象，通过讲述海量儿童故事，包括经典动漫、国学启蒙、童话寓言等众多不同题材音频有声故事内容，创造丰富的儿童成长体验。恐龙贝克还按照0—3岁、3—6岁和6岁以上三个不同的早教阶段，筛选了相对应的启蒙内容，以满足各年龄段儿童个性化的娱乐和学习需求。

（1）“留声机”复刻父母声音

复制真人声音是AI语音合成技术的一大特点。早在2017年，加拿大Lyrebird公司就开发了一款语音变声系统，将用户声音进行演算分析，以较高还原度模拟发声。国内首次推出这一技术的则是科大讯飞公司，通过10句话的简单录制，完整保留用户音色，将真人声音转换为AI主播声音，推动复刻技术进入人们视野。

在恐龙贝克平台中，声音复刻功能以“留声机”形式出现，作为平台核心功

能，原则上可以复制任意声音，但对启蒙阶段的儿童来说，他们对父母声音的需求是最大的，所以在这里着重讨论父母声音的复制与保存。

“留声机”声音复刻功能采用个性化语音合成技术方案，基于深度学习神经网络迁移等技术，Mos值（语音质量指标）接近4分，使用100句语料即可训练一个语音合成模型，声音达到99%的相似度。在恐龙贝克平台中进入“留声机”界面，根据操作提示选择喜欢的故事文本，点击“开始”即可进入声音录制环节，具体流程如图3. 5所示，经由录音采集、音频检测和模型训练、语音复制合成三个阶段。录音采集阶段对录音环境和录音质量提出一定要求，环境杂音过多或录音清晰度和流畅度过低将会重复采集声音语料。经过音频质量检测后，将进入模型训练和语音合成环节，平均等待两小时就能定制出专有声音。

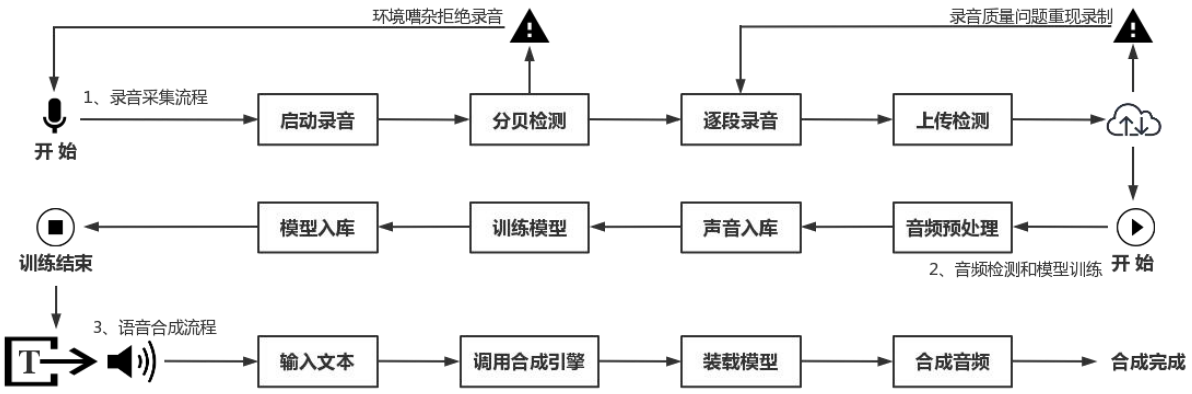

图3. 5 恐龙贝克平台声音复刻的具体流程

用户可以选择线上录音或线下上传语音包两种方式合成自己的专属音库，通过本地和云端调用定制AI语音呈现“故事屋”板块中的所有故事，既能解放父母压力，也增强了亲子熟悉度。在此基础上，定制语音还可以自定义文本内容，录入相关信息再一键输出音频，就能实现阅读陪伴，激发孩子在不同语境中主动创作和表达的能力。从“只能听”到“需要听”，到“想要听”，再到“我要谁讲”，内容生产的主动权已经逐渐交还给用户。

随着泛教育企业的不断发展，声音复刻功能的应用场景逐渐延伸到教育教学模式的改革创新中。试想用儿童熟知的声音来引导和督促学习进程，势必能有效激发学习者的自主性、自律性和自学能力，一定程度上弥补儿童缺乏情感交流的普遍状态，无疑是对启蒙教育的有利帮助。同时，通过复制声音，将人物角色及

关系体验与文本信息相结合，获得前所未有的独特体验，从而为音频平台注入个性化魅力。

恐龙贝克开放声音复刻功能，力求共同探寻这一功能的应用价值，将AI语音合成技术植入到更多场景中，打造出科技感和温度感并存的声音产品。

（2）卡通明星化身AI主播

初次进入恐龙贝克界面会听到亲切地打招呼“点点我，和我说话吧”，这是贝克形象在发挥语音交流和互动对话的作用。对于启蒙阶段的儿童来说，用语言来表达观点和获取知识的诉求十分强烈，语音交互功能正好切中这一诉求。

除语音交互功能外，恐龙贝克率先将语音合成技术与国内顶级卡通IP结合，独创个性化声音和IP语音为主的“声音超市”，提供多样化声音选择，打造了众多AI卡通明星主播。在恐龙贝克界面中，进入故事汇，点击AI专区，就可以挑选心仪的卡通明星形象，如超级飞侠乐迪、小猪佩奇、孙悟空、汪汪队等。听之前可以先进行角色试听，再选定喜欢的人物声音讲故事、唱儿歌以及语音对话。其中，“秋木叔叔”作为原广播电视台播音员主持人，入驻恐龙贝克成为故事主播，带来了近三千个经典故事，深受听众喜爱。他受邀利用个性化声音复刻，与卡通明星一起变身AI主播，实现24小时无限次反复播放，真正满足儿童“想听谁讲就听谁讲”的需求。

从心理学角度上来讲，选择用卡通明星AI主播讲故事，儿童用户的记忆点和接受度会更高，带来的视听刺激更明显。对他们来说听故事的环境和设备是次要的，谁讲故事才是重点，而卡通明星AI主播所塑造的声音形象正好能满足用户的这一需求，因此这也将成为儿童在启蒙阶段学习语言的一种优质方法和重要手段。

基于人性化需求，AI主播的音频内容支持单曲循环、顺序播放和后台自动续播。除此之外，更有轻松哄睡、益智教育、安全教育等多个主题和节日专题供选择，为家长和儿童提供便捷、全面、有价值地听书服务。

除儿童早教领域，标贝科技公司加大研发力度推出了“标贝阅读”平台，该平台有明星模仿音东北大叔、粤语女声阿紫、粤语男声小冬和四川话贝莹等方言主播，声音俏皮搞怪，感情色彩鲜明，富有感染力。在标贝阅读3.7版本上新TTS3.0声音后，再次升级灵柒、雾风华、猪小妹、小勤、冰儿、瑶瑶、阿志、小美、小

菲和龙妈妈10个声音主播，支持童声、青年男声和青年女声三种AI合成语音，涵盖知识科普、游戏配音和新闻播报等场景，多音字、停连停顿更加准确，使AI主播发音能力得到显著提升，用户在选择主播声音过程中通过操作“音效增强”按钮，可以明显体验比原发音更具表现力的声音效果。

AI语音合成技术的加持，使相关平台在提供优质内容的同时，最大可能地用AI赋能声音，为用户提供智能化亲子教育服务，也让孩子真正享受到求知和探索的乐趣。

4AI语音合成技术在网络音频中的应用困局

从声音维度而言，AI语音合成技术让以往通过身体器官产生的声音转变为依托语料库和算法模型产生的拟人声音，将声音从物理领域引入社会文化领域，而“人”在社会发展中起主导作用，离开人的发展如同无源之水。在网络音频领域中，听众扮演着相同的角色，听众对声音的听觉体验与感受尤为重要，了解听众的听觉感受是呼唤听觉回归必不可少的环节，所以要深入了解技术应用的问题，必须聚焦听众，探讨听众对AI语音合成技术的认识情况和满意程度以及对技术应用的态度，为AI语音合成技术的正面使用奠定基础。

本研究针对现阶段具备合成语音特征的网络音频客户端设置了问卷进行调查，利用问卷星平台进行相关题目的设置，生成链接和二维码海报，主要在微信群、QQ群、微博等多个网络渠道进行发放，问卷发放调研时间为2022年10月至2022年12月，为期约两个月。本次问卷作答者分布在全国多个省份，一共线上收集了601份问卷，对问卷进行筛选，出现以下情况视为无效问卷：答题时间过短（20秒以内）或过长（40分钟以上）；问卷出现明显无意义答案；针对“您是否使用过网络音频软件？”和“您是否在网络音频平台中收听过AI合成语音内容？”两个问题，答案为“否”。最终回收有效问卷数为348份。

4. 1问卷设计与测量指标

本研究问卷分为五个部分，共设计30题。第一部分是对受众的性别、年龄范围、学历以及职业四个方面的人口统计学特征进行调查。第二部分是关于受众对网络音频和AI语音合成技术的接触与使用情况，包括网络音频软件的使用频次、使用时间段、收听内容以及AI语音合成技术的了解与接触渠道。第三部分是关于受众对网络音频中AI语音合成技术应用的接受程度与评价测评，包括受众对AI音频作品的接受程度、对现阶段AI语音应用功能的评价以及对AI语音的流畅度、清晰度和出错率的评价。第四部分是关于受众对网络音频应用AI语音技术的风险认知与印象，包括受众对AI语音合成技术风险的了解情况、AI语音合成技术的应用存在何种风险、AI语音合成技术适用的社会情景进行评价。第五部分是针对现有风险与问题的应对建议和受众对AI语音合成技术大范围应用于网络音频领域

的态度。

测量指标一：被调查者的基本信息。问卷的第一部分是被调查者的基本信息情况，包含的变量指标有性别、年龄、学历、职业类型四部分。其中年龄的设计从18岁及以下到46岁及以上，以了解不同年龄段对AI语音合成技术应用的看法，学历与职业类型也采用相同的分类方法。这些变量指标构成了对受众基本情况的初步了解，以便构成后续研究的可行性。

测量指标二：对网络音频和AI语音合成技术的接触与使用情况。其中包含两个维度，第一维度是网络音频软件的接触使用：是否使用过网络音频软件、通常使用哪项网络音频软件、网络音频软件的使用频次、网络音频软件的使用时间段、网络音频内容收听情况；第二维度是AI语音合成技术的接触使用：了解AI语音合成技术的渠道、是否收听过AI语音合成的网络音频作品，其中 AI语音合成技术的了解程度包括的变量指标有从未听说、听说过但不了解、稍微了解、比较了解、非常了解，以此筛选出有效的研究样本。

测量指标三：网络音频中AI合成语音的接受程度与评价测评。对网络音频平台应用AI合成语音的接受程度包含的变量指标有完全不接受、勉强接受、无所谓、可以接受、完全接受。对调节语速、音色选择和选中文段播读功能的评价变量指标为非常不满意、不太满意、一般、比较满意、非常满意。对AI合成语音流畅度和清晰度的评价变量指标为非常生硬（非常模糊）、生硬（模糊）、一般、流畅（清晰）、非常流畅（非常清晰）。对错读误读现象的变量设置为从未遇到、偶尔遇到、经常遇到、不清楚，以便了解技术体验者的主观感受。

测量指标四：网络音频应用AI语音技术的风险认知与印象测评。风险认识与印象测评主要包含两方面，一方面是AI语音合成技术存在何种风险，变量指标为虚假信息泛滥；社会信任度下降；泄露个人隐私信息；冲击新闻真实性，挑战新闻专业机构权威；侵犯他人声音权、名誉权和财产权；侵犯网络音频著作版权；诱发技术犯罪。另一方面是对应用AI语音合成技术相关社会情境的评价，变量指标为复制明星声音并上传网络；影视行业利用合成语音进行人物模拟训练；模仿公众人物声音发表公开讲话；模仿亲人朋友的声音进行电话诈骗；未被鉴别的声音作为案件证据；使用声音信息进行身份验证；利用合成语音发表虚假新闻；复刻已逝人物声音。这些变量构成了受众对技术应用的担忧与认知情况考察。

测量指标五：AI语音合成技术应用于网络音频的风险应对测评。针对已有问题和风险的应对建议，包含规范有声市场管理；创新网络音频AI功能；提升合成语音质量；强化网络音频平台责任；立法监督侵权现象；培养公众信息素养等。同时包含受众对网络音频平台大范围应用AI合成语音的态度：支持，对AI语音合成技术的正面应用很有信心；支持，但对AI语音合成技术可能带来的风险表示担忧；反对，认为AI语音合成技术必定威胁个人和市场安全；保持中立四个指标。

4. 2AI语音合成技术应用的问题探析

4. 2. 1AI语音合成领域准入门槛低

如表4. 1，在关于如何应对技术应用问题的描述性分析中，选择规范有声市场管理的为329人，个案百分比占73. 4%，选择频率最大；选择创新有声读物AI功能的有164人，个案百分比为36. 6%；选择提升语音质量的人数为145人，个案百分比为32. 4%；选择强化有声读物平台责任的有192人，个案百分比为42. 9%；选择立法监督侵权现象的有209人，个案百分比为46. 7%；选择培养公众信息素养的为159人，个案百分比达到35. 5%；选择其他的有13人。

表4. 1 应对技术应用问题的描述性分析

	个案数	百分比	个案百分比
规范网络音频市场管理	329	27. 2%	73. 4%
创新网络音频产品AI功能	164	13. 5%	36. 6%
提升合成语音质量	145	12. 0%	32. 4%
强化网络音频平台责任	192	15. 9%	42. 9%
立法监督侵权现象	209	17. 3%	46. 7%
培养公众信息素养	159	13. 1%	35. 5%
其他	13	1. 1%	2. 9%
总计	1211	100. 0%	270. 3%

根据数据显示，“规范有声市场管理”占据绝对多数，可见样本对有声市场缺乏信心，这也侧面反映出语音合成领域必定存在问题。鉴于此，本研究通过亲

身使用AI语音合成软件、收听合成语音作品等方式，从宏观层面总结出现阶段AI语音合成领域的突出问题为准入门槛较低。

（1）语音合成软件便于操作

如今市面上能搜索到的AI语音合成软件大部分都是免费的，在各科技公司开放平台和下载网站上随处可见。在部分网络音频平台中，也提供了生成合成语音的功能，虽然有收取费用的现象，但价格都比较便宜，用户不需要花费过高价格也可以轻松合成高质量语音。站在用户角度上来说，低价使用语音合成软件可以使他们不顾后果地发挥主动性，即使收获了低于预期的结果也不会过于计较，但这并不是语音合成市场发展的长久之计。此外，用户使用AI语音合成软件不需要学习专业的计算机知识，也不需要具备专业计算机操作能力，只需根据软件界面的相关提示，进行“傻瓜”操作即可自由生成语音，这意味着AI语音合成技术向大众化推广成为必然。

随着科研公司在技术领域的不断深耕， AI语音合成技术的制作成本开始呈现下降的趋势，所需的声音素材减少，制作周期缩短，技术水平再上一个台阶。过去，合成一段虚拟语音的制作周期达半年之久，制作成本超过百万，但现有技术不仅突破了过去的瓶颈，在此基础上还带来了与原始语音相似度的提升。可随之而来的除了AI语音合成技术使用的简便快捷，还有用户准入门槛的不断降低，这给不法分子进入有声领域提供了极大便利。任何技术都是中立的，但若被有心之人利用，那无论是对技术本身还是技术相关领域都是不利的。

（2）合成语音样本易获取

AI语音合成技术包含文本分析、韵律生成和语音合成三个阶段，每一阶段都离不开深度学习算法，而深度学习的训练过程离不开大语料库[①]。大语料库的形成靠的是海量用户的声音数据，但许多科技公司收集数据的过程并不符合规制，反而是通过微信、QQ等社交媒体中仅有的语音功能窃取用户日常对话中的声音信息，达到丰富语料库资源的目的。用户沉浸于体验智能语音设备，警惕性较弱，在不知不觉中被他人盗取声音，泄露隐私信息。

而公共语料库的免费开放，让合成语音的样本获取变得更加容易。内容生产者可以在大语料库中寻找语音资源、下载语音数据，以此为深度学习的对象，经

[①] 魏伟华. 语音合成技术综述及研究现状[J]. 软件, 2020, 41(12):214-217.

过技术的反复训练，就可以生成自己想要的语音效果，像明星、政治家、企业家等公众人物的声音也能轻易被模仿。

近年来，生成合成语音所需的样本音频逐渐减少，很大程度上节约了语言样本的收集时间和成本。曾经需要上百小时语音数据样本量的局面已经发生改变，现在仅需几十秒或几分钟的原始语音样本量就可以训练出高质量的语音，甚至达到音色的高度近似，让人难辨真伪。

4.2.2合成语音应用功能缺位

如表4.2所示，在网络音频中AI语音应该增加哪项功能的调查中，选择增加语速调节档位的为156人，占有效个案总数的47.0%；选择音色种类的为169人，占有效个案总数的50.9%，选择的人数最多；选择特殊音效的为112人，占有效个案总数的33.7%；选择背景音乐的为119人，占有效个案总数的35.8%；选择角色间对话的为143人，占有效个案总数的43.1%；选择其他的为5人。从这一数据中可以看出，样本对语速档位和音色种类需求的频率最高，除此以外，未开发功能的频率也都在30%以上，这说明受众对已有功能提出更高要求，同时期望特殊音效、背景音乐和角色间对话功能的出现。

表4.2 网络音频中AI合成语音功能的描述性分析

功能类型	个案数	百分比	个案百分比
增加语速调节档位	156	22.2%	47.0%
音色种类	169	24.0%	50.9%
特殊音效	112	15.9%	33.7%
背景音乐	119	16.9%	35.8%
角色间对话	143	20.3%	43.1%
其他	5	0.7%	1.5%
总计	704	100.0%	212.0%

现阶段网络音频平台的合成语音主要有调节语速、多种音色选择和选中文段播读功能，但是要打造如真人主播一般的声音场景，仅有这三种功能是远远不够的。通过调查和收集AI语音网络音频作品的评论区，发现有“听得人想睡觉”“根

本听不出来谁是谁啊”“听起来都差不多”等留言，可以看出网络音频内容出现了枯燥乏味、风格单调等问题。根据问卷分析结果可以得出，这类问题的原因之一在于受众无法满足语速和音色功能，源于现有功能难以有效化解合成语音的同质化问题，反而呈现出同质化趋向；原因二则在于其他应用功能的缺位。

（1）合成语音风格同质化现象凸显

网络音频领域的风格同质化现象主要表现为合成语音的语速同质化和音色同质化。首先，为了印证语速同质化这一观点，本研究对AI语音作品的播音速度进行了统计，于2023年1月28日至2月28日，每日随机选取网络音频平台首页排行榜第一位的合成语音作品为样本，并以间隔一天的规律使用语速调节功能，统计稿件字数（标点符号计1字符，空格不计字符）与播读时间，测算平均速度，以“字符/秒”的形式进行分析，得出图4. 3。可以发现，AI语音合成作品的平均语速在4. 35字符/秒，标准差为0. 099，离散度低，合成语音风格存在同质化问题，语速调节功能没有缓解受众的听觉疲劳。

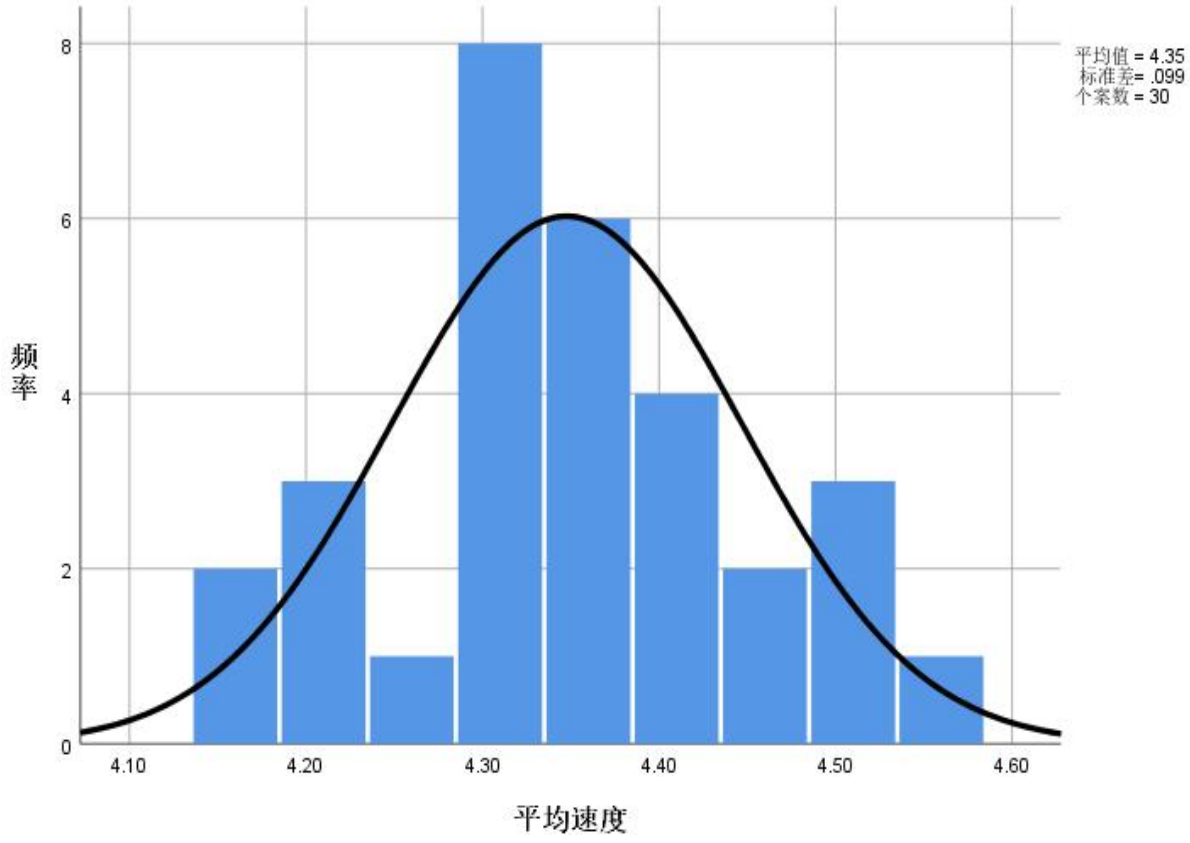

图4. 3 网络音频中合成语音语速的描述性分析

相比之下，音色同质化问题更为突出。AI语音的音色同质化现象最初是在短视频领域引发关注，以影视解说类视频为代表。以“三分钟看完电影”为噱头的影视解说短视频，通常会选择AI语音进行配音，这不仅可以提高视频制作效率，同时可以避免视频博主本身的口条问题和信息保护问题。但长期观看短视频的受众可以明显体会到该类视频配音的相似度极高，经过检索发现这一声音主要来源于牛片配音中单主播合成的“抖音电影解说一哥”音色，类似的问题也开始出现在网络音频领域。现阶段AI合成语音的音色功能主要以“量”为追求，无论是AI语音合成软件还是网络音频平台都提供了一定数量的音色种类，但排除以真人声音为样本的音色，利用率较高的只有其中2—3种，当某一种音色获得认可后，该音色将会在短时间内被应用于各类作品。同时，AI语音合成软件便于操作满足了主流大众“轻创作”的需求，而伴随“轻创作”而来的将是“低质创作”，即易于使用的语音合成工具必定会带来声音特性的同质化，而音色的同质化促成了受众在听觉感官上的空洞乏味。声音作为内容定位的一项参考因素，雷同的音色对于内容创作者而言百害而无一利，音色同质化带来的将会是更严重的内容同质化，受众在接受同质化音色的同时会无意识地将视频内容划分到同质化范围内，他们无法记住和区分创作者，最终只会模糊内容辨别能力，长此以往，网络音频内容将会失去技术加盟的竞争力，反遭侵害。

（2）合成语音功能缺失

首先，缺乏特殊音效功能。在现实生活中，即便是车载广播和新闻也会使用过渡音效来区别不同内容，但是合成语音作品却是平铺直叙地朗读，很容易产生催眠效应。由于网络音频类型的多样性，很多有声小说和广播剧中的环境音无法直接通过人声实现，需要依靠其他外力，例如在武侠小说中运用打斗音效、在搞笑故事中运用笑声音效，或者表现冬日的火焰声、表现天气的雨声和雷声、表现大自然环境的溪水流淌声等，声音特效可以很大程度提高网络音频内容的表现力，给予听众身临其境之感。

其次，缺乏背景音乐功能。背景音乐主要起到烘托氛围的作用，在真人配音的音频作品中，时常可以听到背景音乐或垫乐，以此来营造出符合文本内容的情景，即便没有人声讲述故事情节，也有音乐进行氛围的铺垫。在语音合成的网络音频中，氛围感的缺失加剧了听众的听觉疲劳，大篇幅的合成语音不可避免地会

引起单一的听觉感受，导致听众无法长时间收听，甚至会出现听完即忘的现象，最终听众花费大量时间却没有获取有效信息。但若选择与文风相匹配的音乐，则可以有效减少AI语音的单调感，提高网络音频的竞争力。

最后，缺乏角色对话功能。这一功能要求合成语音与音频作品内的每个人物进行匹配，而他的缺失主要体现在两方面。其一，这与制作步骤繁琐以及对技术要求过高有很大关系，合成同一声音运用于一部网络音频作品中已经是一项巨大的工程，若要加入多种不同声音则会大幅降低合成语音作品的生产效率，不利于行业发展。其二，这一功能的缺失形成了单一音色贯穿始终的语音合成现状，听众无法根据角色适配多元音色，导致作品中角色的辨识度低，听众的代入感弱，难以清晰分辨角色定位；同时，单一音色设置在同一作品中，AI语音与专业主播之间的桎梏将始终存在。

AI语音合成技术高速发展，但却忽略了在网络音频领域中最需要的沉浸式体验。单纯以声音传达故事本就不如“听觉+视觉”的双重效果，若还只停留于满足将文本转变为声音的层面，那么只会带来机械的文字转变机器。特殊音效、背景音乐和多角色对话功能是网络音频中必不可少的一部分，他们的缺失带来的不仅是故事的单调和枯燥，还会失去部分听众，阻碍AI语音合成技术在网络音频领域施展拳脚。

4. 2. 3合成语音质量参差不齐

如表4. 4， 在对网络音频中AI合成语音清晰度的描述性分析中，选择非常模糊的有8人，占有效个案总数的2. 3%；选择模糊的有11人，占有效个案总数的3. 2%；选择一般的有131人，占有效个案总数的37. 6%；选择清晰的有156人，占有效个案总数的44. 8%，选择的人数最多；选择非常清晰的有42人，占有效个案总数的12. 1%。在网络音频中AI合成语音流畅度的描述性分析中，选择生硬的为22人，占有效个案总数的6. 3%；选择生硬的有34人，占有效个案总数的10. 0%；选择一般的有132人，占有效个案总数的37. 8%；选择流畅的为138人，占有效个案总数的39. 5%，选择人数最多；选择非常流畅的为22人，占有效个案总数的6. 3%。由此可知，样本对一般和清晰的选择频率最高，说明网络音频中AI合成语音的流畅度和清晰度已基本满足受众需求。

对网络音频中AI合成语音的错读误读现象的描述性分析中，选择总是遇到的为105人，占有效个案总数的30.2%；选择偶尔遇到的为163人，占有效个案总数的46.6%，选择人数最多；选择从未遇到的为16人，占有效个案总数的4.6%；选择不清楚的为65人，占有效个案总数的18.7%。样本对总是遇到和偶尔遇到的选择频率最高，这说明AI合成语音的错读误读现象是现阶段语音质量的突出问题。

表4.4 网络音频中AI合成语音质量的描述性分析

语音质量指标	变量	频率	百分比	有效百分比	累积百分比
清晰度	非常模糊	8	1.0	2.3	2.3
	模糊	11	1.4	3.2	5.5
	一般	131	16.5	37.6	43.1
	清晰	156	19.6	44.8	87.9
	非常清晰	42	5.3	12.1	100.0
	总计	348	43.8	100.0	
流畅度	非常生硬	22	2.8	6.3	6.3
	生硬	34	4.3	10.0	16.3
	一般	132	16.5	37.8	54.2
	流畅	138	17.3	39.5	93.7
	非常流畅	22	2.8	6.3	100.0
	总计	348	43.5	100.0	
错读误读	总是遇到	105	13.2	30.2	30.2
	偶尔遇到	162	20.4	46.6	76.7
	从未遇到	16	2.0	4.6	81.3
	不清楚	65	8.2	18.7	100.0
	总计	348	43.8	100.0	

根据问卷分析结果，受众收听AI合成语音作品时，对声音的流畅度和清晰度已经呈现出比较满意的态势，但对声音的错读误读却表现出明显不满。所以，本研究针对错读误读现象的合成语音质量问题进行了更进一步的调查，主要选择了具有代表性的六个网络音频平台，选取2022年12月14日各平台首页的合成语音作品，每部作品审听约3000字，从多音字、停连、语气词、音调和吞音漏音五个方面进行质量测评，结果如表4.5所示。其中，出错率=出错量÷收听字数×100%。

表4.5 网络音频中合成语音质量结果统计

平台名称	收听字数	多音字出错量/出错率	停连出错量/出错率	语气词出错量/出错率	音调出错量/出错率	漏读出错量/出错率	总出错量/出错率
荔枝	2945	2/0.1%	22/0.7%	9/0.3%	4/0.2%	15/0.5%	52/2%
喜马拉雅	2788	2/0.1%	5/0.2%	7/0.3%	5/0.2%	4/0.14%	20/1%
蜻蜓FM	2623	3/0.1%	38/1.4%	11/0.4%	4/0.15%	3/0.11%	59/2.2%
企鹅FM	2612	1/0.03%	12/0.4%	0	7/0.3%	7/0.3%	27/1%
云听	2701	0	3/0.1%	0	1/0.03%	0	4/0.14%
恐龙贝克	2884	1/0.03%	21/0.7%	8/0.3%	2/0.06%	6/0.2%	38/1.3%

荔枝平台的样本是男频玄幻小说《完美世界：AI版》，多音字出错2处，分别是“一”和“血”；语气助词包含儿化音和轻声等问题，有9处出错，集中在“呼、咿呀、哦、喽”等字词中；音调有4处出错，为“模、五、老、翼”；还有停连出错22次和漏读15次，共计52处错误，出错率为2%。懒人听书平台的错读误读基本出现在停连和漏读上，这很可能会导致听众理解文字含义出现偏差，出现听不懂的问题。

喜马拉雅平台样本是搞笑穿越小说《怪盗基德：怪盗系统》，多音字有2处出错，分别为“地”和“应”；语气助词有7处出错，为“唔、着、哟、哦、啧、呼、啦”；音调出错5处，“七、只、唉、打、懵”；停连出错5次，漏读4次，共计20处错误，出错率为1%。喜马拉雅平台在以上五方面的语音质量测评中，错误分布比较平均，但偶尔会出现漏读整句话的现象，这会严重影响听书体验。

蜻蜓FM的样本为系列故事《智慧之旅》，其中多音字“中、还、地”出错3处；语气助词有11处出错，集中在“儿、呢、们、啊、了、么”；音调出错4处，为“奔、得、与、创”；漏读有3处，均出错在叠词“轰轰”上；停连出错38处，共计59处错误，出错率2.2%。出错率是六个平台中最高的，合成效果最不理想。

企鹅FM的样本为散文随笔《夏娃的花环》的AI导读版，多音字“应”出现错误；音调出错7处，体现在“学、以、咿呀、学、步、作”；停连出错12次，共计27处错误，出错率为1%。

云听客户端的样本为《两会快报》，仅在停连出错3次，音调“有”出错1

次，共计4处错误，错误率0.14%。因样本内容为新闻短讯，所以出错率极低，合成语音质量最佳。

恐龙贝克的样本为经典儿童故事《木偶奇遇记》，多音字“长”出现错误；语音助词出错8处，集中在“儿、了、么、哎哟、嗒、呀”；音调出错2处，分别是“扑”和“咱”；漏读6处，体现在叠词“嘻嘻、唧唧”上，共计出错38处，出错率为1.3%。

综上，虽然样本出错率在0.14%-2.2%（1.4‰-22‰）左右，但已远远超过图书质量合格标准（差错率≤0.2‰）[①]，整体质量并不达标。首先，针对多音字和音调问题，由于语料库的人工把关环节的缺位，当多音调字词出现在不同词组和语境中时，会按照错误的声音合成，这不仅给受众带来低级的使用体验，甚至会误导儿童、老人等特殊群体，使他们被迫接受错误知识。其次，语气词和漏读问题都受到同类因素制约。语气词问题多出现在拟声词中，合成语音的违和表现容易让人从语境中跳脱出来，心生尴尬。漏读也多出现在拟声词或叠词中，语速过快导致吞并同音字或尾字，减弱了合成语音的表现力。最后，停连作为合成语音过程中最常见的问题，一定程度影响了语音的自然度。停连主要是指一句话中的停顿和连接，适当的停连可以帮助听众轻松理解内容含义，而错误地停连只会增加受众理解难度。所以，合成语音的质量问题已不容忽视，虽然这只关乎微小的字词，但也需要网络音频平台和语音合成技术的供应商多从受众角度出发，切实解决问题。

4.2.4 AI语音合成技术可能带来的风险

如表4.6，在AI语音合成技术风险的描述性分析中，选择虚假信息泛滥的有294人，占有效个案数的65.6%，占比最大；选择泄露个人隐私信息的有232人，占有效个案数的51.8%；选择社会信任度下降和冲击新闻真实性、挑战新闻专业机构权威的为186人，占有效个案数的41.5%；选择的侵犯他人声音权、名誉权和财产权的有192人，占有效个案数的42.9%；选择侵犯网络音频著作版权的有122人，占有效个案数的27.2%；选择诱发技术犯罪的有159人，占有效个案数的35.5%；选择其他的有7人。从这一基本数据看出，样本对虚假信息泛滥、泄露个人隐私

[①] 孙艳华. 阅读听书平台智能合成语音的应用进展、质量现状和用户接受[J]. 编辑之友, 2021, (12): 81- 88.

信息和侵犯合法权利的频率最高，这也说明AI语音合成技术在这方面引起受众的担忧。

表4.6 网络音频中AI合成语音技术风险的描述性分析

风险类型	个案数	百分比	个案百分比
虚假信息泛滥	294	21.3%	65.6%
泄漏个人隐私信息	232	16.8%	51.8%
社会信任度下降	186	13.5%	41.5%
冲击新闻真实性，挑战新闻专业机构权威	186	13.5%	41.5%
侵犯他人声音权、名誉权和财产权	192	13.9%	42.9%
侵犯网络音频著作版权	122	8.9%	27.2%
诱发技术犯罪	159	11.5%	35.5%
其他	7	0.5%	1.6%
总计	1378	100.0%	307.6%

因此，为了更深入了解受众使用和收听AI语音作品的痛点，本研究对虚假信息和侵权现象做出了更进一步探究。

（1）合成语音导致虚假信息泛滥

AI语音合成技术的使用具有大众化特征，因此深度学习模型不需要过多人为协助和参与，只要输入样本声音信息就可以自动生成合成语音，它所需要的时间逐渐变短、成本随之降低，而收集到的样本和素材越来越多，使每个人都有能力合成非真实语音，并借助互联网和社交平台进行一定范围的传播，这毫无疑问会导致虚假信息泛滥成灾。

最初，AI语音合成算法停留于粗糙的拼接，语音不连贯、语流生硬等特点让人可以轻易区分出真人声音与合成语音。但随着波形拼接、单元选择和深度神经网络模型技术的形成，合成的语音开始呈现出流畅度和清晰度更高的趋势。但技术突破带来的不仅是语音效果的优化，更带来对真实性的冲击。在AI语音合成技术的日益普及下，人们难辨音频真假，真实世界和虚拟场景的边界变得模糊。随着虚假语音信息的不断增加，人们对公众人物、公开机构甚至专业媒体的信任感都会遭受巨大打击，新闻媒体机构发布的新闻将会逐渐失去它专有的权威性，这严重冲击了新闻客观性原则，摧毁了新闻真实客观的特性，最后陷入塔西佗陷阱。

虚假声音信息的存在，使人们离真相越来越远，最终演变成深度造假技术带来深度后真相的局面。一般而言，深度造假(Deepfake)是指通过AI换脸、面部表情合成、面部身份合成以及语音技术实现人脸的面部操作，俗称“换脸”，以达到混淆视听、娱乐受众和虚假宣传的目的[①]。深度造假催生了“眼见并不为实”的环境，进一步挑战着传统的真相定义，造成传播失序的恐慌。

借助深度造假技术，始自2016年的后真相内涵得以重获关注，真相瓦解的论调得到延续。中国传媒大学研究员姬德强将这一阶段称为“深度后真相”(Deep Post-truth)，他提到深度后真相的特征之一是视觉客观性的瓦解，借鉴这一概念，AI语音合成技术将会带来听觉客观的负面效应，它作为深度造假技术的一部分，创造高度真实性的音频，带来以假乱真的视听体验，最终带来深度后真相[②]。

在人工智能时代，AI语音合成技术形塑深度后真相主要体现在两个方面。一方面体现在与个性化推荐的结合。算法的个性化推荐根据用户的不同兴趣爱好推送符合其价值观的内容，长此以往形成的信息茧房使相同兴趣的网络社群处于信息孤岛中，高度的分裂和意见的片面化为深度后真相提供了天然沃土。另一方面，合成语音通过调动人们的情感进一步加速深度后真相的到来。合成语音不受时空限制，不受人物主体性和客观性约束，使用AI语音合成技术可以根据预设目的生产出代表主观意志的音频，配合舆论引导，从而影响和煽动人们的情绪，最终导致受众轻真相而重情感。

（2）合成语音造成的侵权现象

与指纹信息一样，世界上不存在完全相等的两种声音，每个人的声音都是独一无二的，音色、音高、音强、音长和频率、波长、振幅等物理性质都不相同，并且这些物理性质具有唯一性和稳定性特征，轻易无法改变，即使是经过刻意训练也很难摆脱原有的声音特点。

从这一层面上来说，声音属于个人生物信息识别范畴，是区分个人身份的重要生物标志，应赋予相应的法律属性。杨立新教授认为[③]，声音作为一种法益，已经得到部分国家的立法和司法保护，在现有技术背景下声音逐渐成为一种独立

[①] 刘建明. 深度伪造对媒体与人类的致命威胁[J]. 新闻爱好者, 2021, (04) : 8-13.

[②] 姬德强. 深度造假: 人工智能时代的视觉政治[J]. 新闻大学, 2020, (07) : 1-16+121.

[③] 袁雪石. 中国人格权法的创新与发展——杨立新教授人格权法思想研究[J]. 河南省政法管理干部学院学报, 2010, 25 (05) : 18-23.

民事权利，即声音权。声音权是指自然人自主支配自己声音的权利，决定对自己的声音进行使用和处分的具体人格权，具体包含：声音使用专有权、使用许可权、录制专有权和利益保护请求权。中央民族大学法学院学者梁震指出[①]，歪曲、偷录、剪接、失真处理不当等行为都属于侵犯声音权范畴，AI语音合成技术的一切侵权行为都离不开这一范畴，在此基础之上还产生了其他侵权现象。

我国网络安全保护法对个人信息保护做出了“谁收集，谁负责”的明确规定，要求隐私信息的收集方承担起数据安全保护义务。但个人声音信息只有在流通和共享的过程中才能发挥价值，这就导致网络音频平台等社交媒体很难对收集到的声音信息实施有效保护，侵权现象愈演愈烈。伪造者擅于利用AI语音合成技术，合成非真实音频，以他人身份发表各种不正当言论，如冒充政治家发表种族歧视、地域歧视、政治立场错误、破坏政治外交和民族团结等言论，丑化政治家形象，影响政治选举与裁决的公正性；或冒充企业家发表错误决策、性别歧视、色情暴力等言论，影响社会评价和公司声誉，失去股东信任，危害公司上市进程，最终侵犯他人名誉权。

除名誉权，AI语音合成技术对财产权的侵犯也显露端倪。现阶段人们通过签订合约的方式，将声音的使用权让渡他人并从中获得经济报酬。这就给了伪造者可乘之机，市面上的不法商家通过盗用和兜售他人声音，贩卖语音信息插件，按照固定价格进行私人声音定制，从中牟取暴利，形成完整的产业链。他们未经当事人同意就将合成语音在网络音频平台等多个领域进行商业化应用，妨碍他人通过自己的声音获取合法经济利益，忽视了声音的财产属性，变相侵害了他人的财产权。这一现象对明星来说尤为凸显，明星具有普通民众所缺乏的粉丝效应，有的内容生产者看重这一点，私自合成明星声音进行网络音频的录制，以此作为营销噱头，增强网络音频的吸引力，获得更多潜在用户的关注和购买。对于一些已经拥有大量粉丝并打造出品牌形象的头部声音主播来说也存在被恶意复制的风险。而粉丝的盲目追捧使他们不会去深究声音的来源与合理性，只会更大范围地使用合成语音，这不仅影响了明星的经济收益，也让明星自身的商业价值大打折扣。

此外，AI语音合成技术侵犯有声读物著作版权也是毋庸置疑的。喜马拉雅平

[①] 梁震. 探寻对声音利益的民法保护途径[J]. 法制与社会, 2017(01):256-258.

台曾发生过翻录网络音频作品，并设立网站进行盗版音频传播和售卖的事件，这一事件给当事人带来了巨额经济效益，也造成了喜马拉雅平台的严重损失。在当下，盗版音频侵权的现象只会更甚。

音频作品的形成牵扯到多方利益，包含原作者、录音主播、网络音频平台、网络音频的制作者等，需要经过他们的同意和授权，以邻接权的形态出现，受到著作权法的保护[①]。但在网络音频平台中，一些付费产品不可避免地会被制作成盗版音频，流传到网站或其他自媒体平台上，以吸引更多关注。这种盗用声音资源的行为在侵害原网络音频版权的同时，也侵害了版权所有人的权利。因为声音具有很强的可识别性，所以个性化声音是网络音频平台的一大卖点，而AI语音合成技术正好提供了便利，让用户通过相关软件就可以生成特殊声音，在网络音频平台中大规模传播。针对“凯叔讲故事”中的“宝拉”角色的配音，阿里云的语音合成技术就提供了这一角色的语音定制服务，用户上传文本后就可以生成“宝拉”的声音，若将其应用到其他平台，可能引发版权纠纷，扰乱网络音频市场，破坏有声内容出版生态。

[①] 郑聪. 使用AI转录网络音频字幕的法律界限——基于亚马逊被诉侵犯版权案视角[J]. 淮南师范学院学报, 2020, 22(06):14-19.

5AI语音合成技术在网络音频平台中的发展对策

AI语音合成技术虽塑造出以人工声音为主的全新声音景观，但在此过程中呈现出的困局和负面影响也不容忽视，以此为网络音频平台发展的着力点，如何通过规范市场、创新功能、改善质量、规避风险等措施，解决现实问题，实现AI语音技术在网络音频领域的高效发展，运用声音要素，对声音环境进行设计与构建，本章将展开具体研究。

5. 1强化合成语音市场管理，提高准入门槛

5. 1. 1健全用户注册使用制度

AI语音合成技术提供平台和网络音频平台作为合成语音作品的生产载体，应承担起相应的社会责任，健全和规范语音市场，避免用户的不当行为，所以需要建立注册和使用制度，给予一定限制，以此来提高用户进入AI合成语音领域的门槛。

第一，满足真人声音被复制的知情同意。虽然现在的公开语料库和平台所提供的真人合成语音是经过签订协议所获取的，但是在真实应用的过程中，平台应根据实际情况给予声音主人适当的知情权以及同意使用的权利。例如需要复制明星、政治家或企业家的声音用于新闻等音频作品中时，应通过短信、邮件等便捷通讯渠道让声音主人了解使用情景，若不允许使用则回复信息告知平台和用户，若同意使用则不需要回复，流程复杂但也相对安全，使用户在平台中选择声音样本时更谨慎细致。

第二，网络音频平台要构建新用户培训机制，增强用户的规范性。在用户注册平台时，进行指导性学习和规范宣传，让用户先熟悉AI合成语音作品的制作机制和使用规范。例如用户使用支付宝时，界面下的理财选项会弹出一份调查性问卷，用户只有完整填写问卷、通过调查才可以继续使用，以规避投资风险；B站通过答题的方式筛选用户，达到相应分数才能成为平台会员，对用户的规范性也有较高要求，在这方面技术提供平台和网络音频平台可以借鉴学习类似方法，对新用户进行AI语音合成技术的相关测试，测试要通俗易懂、覆盖面广，涵盖AI

语音合成技术的基础概念和应用行为。

第三，网络音频平台需要完善监督举报机制。针对用户举报的违规语音及时进行警告、删除、封号等手段进行处理，同时对算法的协同过滤设置感兴趣与不感兴趣的选项，一部合成语音作品在市场上的成功流通既依赖于算法的深度学习，也取决于算法推荐对内容价值的识别能力。此外需要用户进行实名制登记，一旦发现利用他人声音上传诽谤他人、破坏他人声誉、发表错误言论和色情暴力等内容的合成语音信息，或涉嫌违法乱纪的合成语音作品，实名登记使得相关平台可以迅速锁定身份，便于后续追责和惩罚，起到警示和预防的作用。

5. 1. 2技术设计者承担道德责任

AI语音合成技术设计者作为助力合成语音市场化的重要推手，对设计者提出更高要求则是让合成语音市场的规范化管理更上一步台阶，而这里所指的更高要求则是让技术设计者承担道德的责任。技术的设计性本身就是一种内在的道德属性，即便设计者没有改变人们认知和道德行为的明确目的，但设计者所设计出的技术却难以避免地对人们的道德判断和决策产生影响。虽然技术自身并不具有偏向性，但在与人发生关联时就具有明确的意向，与人一起成为道德共同体，以多种方式干预人的行为，或说服，或强制，或诱导，人的决定和实践越来越多地受到技术影响。所以，要提高AI语音市场的准入门槛，就需要重视AI合成语音技术设计者的道德意识，督促他们承担相应的道德责任。

第一，实现价值敏感性设计。彼得·保罗·维贝克指出价值敏感性设计方法，是在人工智能设计过程中主要关注人的价值[①]。这要求设计者在开发以AI语音合成技术为代表的智能语音技术时，要考虑人们对合成语音的喜好；合成语音会给人们带来何种影响；合成语音作品能否在网络音频领域良性发展；AI语音合成技术是否会影响人的自主意识；AI语音合成技术是否能满足人们的美好生活愿景等问题，利用AI语音合成技术造福人类生活。

第二，进行建构性技术评估。这需要技术设计者与所有技术利益相关者建立联系，如使用者、科技公司、网络音频平台等，询问他们的需求、意见和评价并将其反馈给设计者，便于设计者在设计过程中调整和修改，尝试在所有利益相关

[①] 彼得·保罗·维贝克. 将技术道德化：理解与设计物的道德[M]. 上海交通大学出版社. 2016：89-90.

者之间形成一种平衡。在技术研发阶段，AI语音合成技术的设计者应该询问用户使用合成语音时的体验、功能需求、需改进的问题，了解科技公司的产品定位、受众定位，以及合成语音作品的包装、宣传和盈利等方面的信息，进行针对性的修改，从而形成设计者与利益相关者共同决策技术未来的局面。

技术会影响人们的道德判断，所以作为技术的设计者，应该重视和提升自身道德意识，承担道德责任，重塑技术设计者形象，提高AI语音合成技术设计者的准入门槛。价值敏感性设计和建构性技术评估正好从技术设计角度规范了设计者的道德意识，这既是对设计者提出的更高要求，也是有效管理网络合成语音市场的一剂良方。

5. 2创新多元化功能，提升用户体验

5. 2. 1优化升级现有的语速和音色功能

因为选中文段播读的功能在网络音频平台还未大面积普及且受众需求不大，所以暂时只讨论语速调节功能和音色选择功能的优化趋向和方法。

语速作为一种声音处理方式，关系着有声作品的情节推动与人物刻画等多方面内容，不同的叙事情节需对应不同的语速，而在叙事之外，语速的变化还承担着表意的功能，所以对语速功能的优化升级很有必要。现阶段，调节语速的方式主要是通过固定倍速值调节，即按照0. 25倍或0. 5倍为间隔，设置多个倍速档位。目前，仅企鹅FM以滑动速度条的方式调节所需语速，实现了真正的自由选择，所以语速调节功能应以档位细分化或无档数限制的形态持续发展。此外，还需注重基础语速也就是1倍速的设置。速度可调方便了会调速和懂调速的听众，但针对视障人士或老年儿童等特殊群体，1倍速的设置就尤为重要。经过反复审听测算，部分音频作品确实出现语速过快的情况，其AI语音的1倍语速区间集中在290字/分-320字/分（标点符号按1字符计算）。专业播音员主持人的平均语速不超过300字/分，真人主播声音起伏、气息连贯，有利于信息的接收和理解。而AI语音以290字/分-320字/分的速度进行信息传达，频率基本相同，附带音调错误，难免会出现语速偏快现象。在参考多部AI音频作品的语速后，发现1倍语速区间为270字/分-290字/分时听众满意度最高。

而针对音色服务功能，目前网络音频基本具备基础男声和女声音色，喜马拉雅平台提供5种不同音色，从种类和数量上看这一功能必然是有所缺失的，所以它的优化路径就是提供更大范围选择音色的可能。可以结合语音合成技术公司提供的多元音色进行增加，如稚嫩童音、可爱萝莉音、甜美少女声、开朗少年音、成熟青年音、稳重大叔音等，各网络音频平台根据平台内容和需求，合成特色声音，提高平台的辨识度和竞争力。另外，方言音色的应用也能够有效提升用户体验、扩大受众群体。方言作为一种具有地域特色的语言体系，是与受众朝夕相处的语言，在社会风俗文化中占据重要位置。虽然语音合成技术公司提供了多种方言选择，如粤语、四川话、民族语言甚至英文，但这些音色在网络音频中均未体现。所以，要适当引入方言音色，增进听众的亲切感，这在一定程度上照顾了老年群体和思乡人群的感受，体现了丰富的音色对有声阅读的重要性。

需要注意的是，在AI语音作品构建出的声音景观中，具有记忆点的是不同音色所塑造的声音形象。音色作为固定场景中的声音标志(Soundmark)，音色等同于角色，是识别角色的唯一手段。因此，为避免大规模的音色同质化，不应盲目追求音色的数量而忽略质量要求，避免通过简单音色变频实现表面的音色多样，减少低质量音色给听众带来的繁杂的听觉体验，适当限制同一音色的使用频次。在此基础之上，AI语音作品应向个性化方向发展，寻找自身声景特色与重点，明确自身定位，寻找目标听众，研究声音与内容的关系，将AI音色塑造为品牌特色，形成受众记忆点，摆脱“同质化”标签。

5. 2. 2增加音效、背景音乐与角色间对话功能

音效、音乐和对话功能的缺失在很大程度上影响了受众的收听体验，所以添加以上功能是毋庸置疑的。但这并不是做简单的加法，而需要深入思考如何恰当添加、听众的真实需求、添加的负面效应等问题，尽量避免后患。

其一，添加特殊音效功能。音效的总体功能是塑造场景并完成叙事和转场，这里的音效主要指音频作品中所存在的环境音效和动作音效，以此探讨他们的场景和情节建构能力。雷蒙德在《声音景观》中将风、水、雨、土地等环境声音归为可以打破时间的第一声景，认为自然景观的差异可以体现声音景观的特殊性，起到奠定基调的作用。在网络音频内容中，恰逢其时的环境音效可以渲染气氛，

推动情节变动，这一点因为AI语音技术的加入显得更为重要。而动作音效的场景建构作用需与环境音效结合在一起，如出现频率最高的脚步声，与不同的环境音效配合会塑造出不同的声音场景，辅助语言承担部分叙事功能，为有声作品营造真实感。针对音效这一功能最重要的是要给予听众取消特殊音效的权利，在环境声音较少的音频作品中，特殊音效功能的确升华用户的听觉体验。但对于环境音较多的作品而言，过量音效会带来听觉上的冲击感，忽略内容呈现，甚至让人心生厌恶，无论是添加音效还是取消音效都是为了提升用户的使用体验，同时也推动AI合成语音功能向智能化、专业化和人性化方向发展。

其二，添加背景音乐功能。在声音景观视域内，背景音又称基调声(Keynote Sounds)，其作为其他声音的背景而存在，描绘当下场景中的基本声音特色。这一功能需要AI语音合成技术将文字信息转化成语音信号的同时，在线生成与作品类型相匹配的背景配乐。一方面，需以谨慎态度置入背景音乐。在真人主播发布的音频作品中，有音乐加以辅助的占绝大多数，但是部分置入了背景音乐的作品并未获得听众肯定，反而有不少抱怨背景音乐影响了收听体验的评价。因此，AI合成语音与背景音乐的结合更需谨慎，要均衡考量背景音乐与内容之间的促进作用和适配程度，以及背景音乐对内容理解和接收可能产生的负面影响；另一方面，需考虑音乐匹配的灵活性。固定唯一的音乐音量和曲调难以满足用户的个性化需求，所以应给予用户自主调节音量、曲调的能力，例如选择添加音乐或不添加音乐、选择音量减小或升高、选择抒情或激昂的音乐等，据个人喜好完成音乐的匹配。

其三，添加多角色对话功能。多角色对话主要是针对有人物对话的音频作品而言的，用于区分角色，提升用户体验。在搞笑穿越有声书《怪盗基德：怪盗系统》中，有主角白七茶、白七茶父母和班主任老师等人物，文章以大面积的对话形式呈现，但是同一音色的播读很容易让人脱离情景和故事情节，转而纠结于“谁是谁”的问题中，这对音频作品的呈现和表演毫无益处。多角色对话功能是解决这一问题的最佳方法，一方面可以通过AI算法自行匹配和调动合适的音色用于人物对话中，但这要求网络音频平台的音色存储容量足够大、音色种类足够丰富、音源足够支撑广大的受众群体；另一方面则是将选择权让渡用户，用户作为对某一有声内容最感兴趣的人，可以根据切身体会匹配最合适的音色，从而提升自身

的听书体验，实现“自给自足”。多角色对话功能作为一种优化功能，虽对AI音频作品提出了更高要求，但同时也会给听众带来耳目一新的听觉感受。

5.3平台与技术领域共同发力，提升语音质量

5.3.1网络音频平台夯实把关人职责，加大审核力度

当前网络音频平台以UGC、PGC或AIGC的内容生产模式运行，平台面对庞大的音频资料与作品，难以逐一进行审核，导致语音质量参差不齐，把关职责缺位。网络音频平台作为新媒体，早已参与到把关环节中，拥有一定的审核经验和能力，同时合成语音的错误都是有迹可循的，可见平台在语音质量方面的把关的确存在疏漏，因此要夯实把关人职责，充分行使好把关人角色，更加谨慎、智慧、严格地进行筛选和审视，开发多种把关方式，在这里主要提出三种方式。

一是建立字、词、句数据资源包。目前，百度智能语音公司已经提供多音字标注发音，可以解决部分语音错误，但智能语音公司中语音数据的更新迭代仍无法完美契合网络音频平台所需的语音资源，存在时间上的滞后性，所以为了确保发音、停连的正确率，网络音频平台有必要形成独立语音数据资源包，尽可能全面的设置发音规则和停连规则，建立实时更新系统，涵盖偏音怪音、古文用词、网络新词、时代热词等，及时填充词库，依据固定搭配调用正确发音。另外，由于语气词存在轻声、儿化音等特殊处理，还需为其制作单独的语音包资源，从而提高合成语音的质量和表现力。

二是委托第三方机构协助审核合成语音。几乎所有互联网平台都具备自己的审核机构，用于检查上传作品在各方面质量是否出现差错、是否违背主流价值观等，网络音频平台也不例外，上传至平台的语音合成作品可以在通过平台自身的初次审查之后，委托第三方机构对音频作品做深入筛查，降低合成语音的出错率，起到双重保险作用。同时，网络音频平台出于成本考虑，也可以直接聘请第三方平台完成合成语音的审核工作。在这里更推荐将第三方机构审查作为第二次审查，以提高合成语音质量，尽最大可能避免语音问题。

三是以显著标识区分合成语音作品与真人制作作品。网络音频平台应在内容生产者上传合成语音作品时，事先做好分类并提前告知上传合成作品的操作流

程，以便以显著的方式予以标识；同时针对已经上传的音频作品进行定时审查，筛选出非真实语音作品进行标记，避免漏标错标。平台进行标识的方法有两种，一种是在每一部音频作品的标题或作品头像上注明“AI版”或“智能合成”等文字，另一种是在音频作品内容的开头添加“xxx由AI人工智能合成”的音频，让听众清晰分辨合成语音作品和真人制作作品。

5.3.2情感语音合成技术规范发展，优化情感表达

声音景观设计离不开声音景观的创意手段，对AI网络音频内容而言，其实现声景创意的有效方法之一就是强化声音的情感表达能力。情感表现力弱作为AI语音的质量短板。究其原因，AI语音合成技术催生了见字出声的朗读模式，而情感是一项复杂的心理活动，真人主播之所以能够情绪饱满、富有变化地将其准确表达出来，是基于主播对不同语境中的字、词、句的深刻理解与真切感受。而AI语音唯有通过技术规范，添加情感化特点，才能提升语音质量，与听众产生共振。

其实早在20世纪初，就出现了以“情感+语音”为中心的交流会议，主要讨论了如何解释情绪状态本身，即如何建立情绪模型；如何在语音中体现情感，即怎么将情感置入语音[①]。经过多年研究，人们通过利用情感计算的概念，分析带有已知情绪状态的语音信号中情感和语音之间的关系，将分析得出的情感特征用于生成能模拟人类情感、自然清晰的声音。

情感合成语音技术虽早已问世，但目前业界急需一个公认的对合成的情感语音的客观评判标准，所以制定相应标准是当下技术优化发展的方式之一。通过统一标准可以对情感语音进行适当的调整和规范，例如情感计算数据达到某一数值是合成语音表达情绪的最佳状态，这能为情感语音合成技术的健康发展保驾护航。要注意评判标准只是用于优化情感表达，而不是画地为牢将其限制于固定框架内。另外，建议AI语音合成技术供应商强化标点符号的情感表现作用，如问号为上扬语气、叹号为表现强烈的语气、区别设置逗号和句号的停顿时间等，这对加强合成语音在网络音频中的情感化表达大有裨益。

2021年4月上旬，AI合成语音情感控制取得突破。微软发布了支持11个情感

[①] 高莹莹, 朱维彬. 面向情感语音合成的言语情感描述与预测[J]. 清华大学学报, 2017, 57(02):202-207.

类别的AI语音合成技术，将情感分为平静、温柔、撒娇、开心、不满、沮丧、尴尬、悲伤、严厉、愤怒、恐惧，以“平静”为零点，以1%为情感程度量化单元，轻松调节情感程度[①]。若将微软情感语音合成技术用在组合不同的上下文中，则可以使有声阅读具有丰富表现力，让听众更能体会到网络音频主人公的细微情绪变化，情感语音大有可为。

5. 4坚持伦理与法律的双重原则，适当规避风险

5. 4. 1强化伦理主体性责任

关于声音景观的伦理思考，一个基本立足点就是围绕发声者和听众所形成的声音特征及其关系的探究。两者构成了声音景观存在的基础，他们基于各种原因和目的进行发声和收听，与之相应的伦理规约则是作为传受双方各自的责任担当。虚假语音信息和侵权现象的出现，很大一部分原因是各主体在各自领域内并未遵循其伦理规范原则，形成了AI语音合成技术伦理风险，其伦理责任的主体内容涉及公众、算法工程师、媒体平台等。伦理主体协同合作则是规避技术风险的重要思维，薛宝琴在《人是媒介的尺度：智能时代的新闻伦理主体性研究》中将新闻伦理的主体拓展为个人、智能技术公司、媒体平台、算法工程师、编辑记者等[②]。借鉴这一思考维度，尝试从个人伦理、数字伦理、职业伦理三个方面进行阐述。

首先要规范个人伦理。个人伦理主体是数量庞大、风格鲜明、布局分散的受众群体，AI语音合成技术在网络音频领域所体现的娱乐性和实用性吸引受众遵循内心需求表达欲望，但并不意味其行为毫无限制，受众行为需要符合公序良俗和礼仪规范，建立自我伦理，避免过度主观性的表达。同时，由于思维的局限性，受众往往对自我认知有所偏差，因此需要进行媒介素养教育，通过提升用户媒介素养，在他们心中建立伦理防线，对于不确定的技术使用行为或违背道德的行为进行自我管理，减少对内心欲望的依赖，合理使用AI语音合成技术。

[①] 孙艳华. 阅读听书平台智能合成语音的应用进展、质量现状和用户接受[J]. 编辑之友, 2021, (12):81-88.

[②] 薛宝琴. 人是媒介的尺度:智能时代的新闻伦理主体性研究[J]. 现代传播(中国传媒大学学报), 2020, 42(03):66-70.

其次是要建立数字伦理规则。现在的数字网络无处不在，所以建立人们普遍认可的数字伦理规则是极其必要的。技术设计者和算法工程师作为AI语音合成技术依赖来源，应承担起相应的数字伦理责任。一方面应坚持技术透明原则，算法执行过程中的“技术黑箱”是不透明的，其系统认知与决策过程十分复杂，连技术人员都无法完全理解，所以设计数字伦理规则应先考量如何打开“技术黑箱”，公开虚拟语音制作流程，打消受众疑虑，让受众信任声音信息；另一方面数字伦理规则需具有普适性，技术设计者和算法工程师应增加与专业媒体人的交流，提高传播伦理主体意识和AI语音合成技术工具理性的社会价值，此外还应听取跨学科专家意见，如技术、哲学、法学等领域的专家学者，汲取经验，顺应规律，与时俱进，以期做出数字伦理的合理规制。

最后需要巩固职业伦理地位。职业伦理主要是针对新闻媒体机构及专业新闻人而言的，在人工智能技术风靡的当下，他们也成为虚假信息传播的平台，在这样的情况下，新闻媒体机构更应该遵守新闻专业主义，追求客观性和真实性原则。全面、客观、公正地进行报道，避免夸大或轻视AI语音合成技术相关事实，将技术的风险扩大化，造成公众对技术感到恐慌。另外，技术给新闻媒体机构带来挑战的同时也带来了机遇，在虚假信息泛滥成灾的环境中，坚持发布真实信息，将会大幅增强媒体权威，使其逐渐成为公众接收信息的首要途径。因此，新闻媒体需不断巩固职业伦理地位，明确伦理原则，给予AI语音合成技术风险客观真实的描述。

5. 4. 2立法规范技术适用限度

由于有声市场的特定属性，以及AI语音合成等人工智能技术的入场，导致现有的法律法规无法全方面涵盖，声音权、名誉权、财产权、著作版权等均受到侵害，相关法律有必要通过参考现有的侵权案例进行细化完善，明确权利界限，以法律条文的形式对使用AI语音合成技术的相关行为进行引导监督，保护主体合法权益。

在网络音频领域中，若简单对名誉权和财产权侵权行为进行治理，只能起到暂时的治标作用，因为名誉和财产受侵害本质上其实是在侵犯声音权，所以要根治侵权行为，就必须对症下药将声音权以一种具体的人格权进行立法。前文提到，

声音具有标识个人身份的作用，是一种新型的人格要素，属于个人生物信息识别的范畴，所以可以通过借鉴国外对个人生物信息的立法保护模式为声音权立法提供帮助。例如美国虽未进行联邦制的统一立法，但各州陆续对个人生物信息设立了单独的条例进行保护，得克萨斯州的《生物特征信息隐私法》规定私人存储利用个人生物信息时要以书面方式通知并获取信息主体的同意，并将其当作敏感信息，采取更严格的保护措施。可见，为声音权专门立法是最具针对性的保护方式。所以在我国《民法典》中，更应以具体人格权形式为声音权立法，将其与人格权中的隐私权、姓名权、肖像权等并列，尊重声音权个性特征，维护民法体系的稳定。

而保护网络音频作品的著作版权则需要考虑两方面内容：一是适当提高违法成本。《著作权法》第五十三条明确规定[①]，未经录音录像制作者许可复制发行录音录像制品的，违法经营额5万元以上，并处违法经营额1倍以上5倍以下罚款；违法经营额难以计算或不足5万元的，并处25万元以下的罚款。可见，其惩戒力度十分有限，难以真正起到打击此类违法行为的作用。所以应适当提高罚款金额或监禁年限，让违法者即使有违法之心也不会付诸行动。二是利用数字水印技术保护网络音频著作版权。在反制深度伪造技术的方法中有一项DeepTag数字水印技术，DeepTag可以将消息嵌入到图像中从而获取Deepfake视频的来源出处，以此来积极防御深度伪造[②]。AI语音合成技术也可以参考这一方法，在合成语音时加上数字水印。这些水印并不会影响受众的听觉感受，但是在对合成语音进行鉴别时，其水印痕迹能够比较容易地被发现，据此快速识别盗版翻录的合成语音作品，可以有效保护网络音频内容的著作版权。对于受众来说，运用特定的合成语音鉴别软件，也可以通过数字水印轻松鉴别出合成语音的真伪。

[①] 吴瑶, 张亚莉. 有声书版权的主体认定与权责链关系——基于 2021 版《中华人民共和国著作权法》[J]. 青年记者, 2021, (23) :80-83.

[②] 王旖旎, 邸娜. 移动终端的数字音频水印检测软件实现[J]. 现代电影技术, 2021, (02) :44-48.

结 语

在数字化、智能化的网络时代，AI技术的应用一直广受社会关注，AI语音合成技术作为现阶段人工智能技术的优秀成果，其如何在网络音频领域绽放光彩也是业内持续讨论的话题。AI语音合成技术与网络音频的融合应用能够更好地实现文字到语音的跨越，塑造了全新的声音景观，为受众打造虚拟与现实随意转换的听觉文化新纪元。但是任何事物的智能化发展路径都不是一片坦途，人们始终会对新技术的介入抱有差异化态度，所以对AI语音合成技术在网络音频中的应用现状和发展路径进行研究讨论，具有一定现实指导意义。

在以波形拼接、参数合成和深度学习为基础的AI语音合成技术支持下，多个网络音频平台陆续开始出现AI语音合成作品，随着深度学习的神经网络模型占据主流，AI语音从最初的声音生涩僵硬、机械感十足到现在的声音清晰流畅、让人难以分辨，网络音频中的合成语音形势大好。AI语音合成技术的实际应用方面，调节语音、选择音色的功能已经得到普遍应用，选中文段播读功能作为补充也满足了受众不同程度需求，而在网络音频平台中，AI新闻播报、卡通明星播报、复刻逝者声音、复制父母声音均已得到初步实现；便捷、普惠、多元正在成为音频作品的新特点。

而这些新特点并未经受住时间的考验，在问卷分析结果中，受众认为网络音频中的合成语音功能、合成语音质量都还存在发展空间，甚至对其可能带来的虚假信息、隐私泄露、版权侵犯等切身利益问题格外关注。因为AI语音合成技术与网络音频的结合还处于初级阶段，相当一部分人对“技术+声音”的认知还局限于智能家居产品或手机语音助手，所以当前技术的大范围应用还面临多种挑战。过于普惠带来过低门槛，这是新事物初期发展过程中最常见的问题，为了快速普及与成长，技术与网络音频领域都存在这一现象；在合成语音功能方面，当受众的最低需求得到普遍满足时，他们将开始追求更能让人耳目一新的感官体验；语音错误高频出现，合成效果自然度低，这都导致合成语音质量无法达到受众要求；同时伦理风险与法律风险带来的伤害重创了技术发展信心。

为此，本研究针对已经出现的负面现象进行了深入思考和针对性的探析，试

图通过提高准入门槛、增加应用功能、提升语音质量和化解风险的发展对策优化虚拟声音环境。但不可否认存在不足，如本研究尝试赋予技术道德性质，赋予技术伦理规范，这虽顺应当下潮流，但却忽略了技术温度泛滥也会导致应用领域的发展停滞，回归单纯的“人”的发展。因此，网络音频应用AI语音合成技术应把握好人与技术之间的平衡，共同致力于营造健康、有活力的网络音频生产生态，促进人工智能技术社会的正向前进，带领国内乃至国外AI音频市场顺利搭乘高科技列车，走上智惠、智慧、智能之路，同时勾勒出技术、人与音频内容共同构造的听觉文化新时空。

参考文献

国外文献

[1] Kristen M Scott, Simone R Ashby, David A Braude & Matthew P Aylett. Who owns your voice? ethically sourced voices for non-commercial tts applications[C]. CUI'19: Proceedings of the 1st International Conference on Conversational User Interfaces August. 2019:1-3.

[2] Naoto Umezaki, Takumi Okubo, Hideyuki Watanabe, Shigeru Katagiri & Miho Ohsaki. Minimum Classification Error Training with Speech Synthesis-Based Regularization for Speech Recognition[C]. SPML'19: Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning November, 2019:62-72.

[3] Noé Tits, Kevin El Haddad & Thierry Dutoit. Neural Speech Synthesis with StyleIntensity Interpolation: A Perceptual Analysis[C]. HRI'20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot InteractionMarch, 2020:485-487.

[4] Takuya Asakura, Shunsuke Akama, Eri Shimokawara, Toru Yamaguchi & Shoji Yamamoto. Emotional Speech Generator by using Generative Adversarial Networks[C]. SoICT 2019: Proceedings of the Tenth International Symposium on Information and Communication Technology December, 2019: 9-14.

[5] Hasan, Haya R.Salah, Khaled. Combating Deepfake Videos Using Blockchain and Smart Contracts. IEEE Access, 2019:2169-3536.

[6] Capital One Services LLC; Patent Issued for Arrangements For Detecting Bi-Directional Artificial Intelligence (AI) Voice Communications And Negotiating Direct Digital Communications[J]. Journal of Robotics & Machine Learning, 2020:165.

[7] Xin Yang, Letian Yu. Automatic Comic Generation with Stylistic Multi-pag

e Layouts and Emotion-driven Text Balloon Generation[J]. ACM Transactions on Multimedia Computing, Communications, 2021:1-19.

[8] Kwok, Andrei O.J.Koh, Sharon G.M.A social construction of technology perspective[J]. Current Issues in Tourism, 2020:1-5.

[9] Elliott David, Soifer Eldon. AI Technologies, Privacy, and Security[J]. Frontiers in Artificial Intelligence, 2022:5.

专著

[1] 彼得•保罗•维贝克. 将技术道德化：理解与设计物的道德[M].上海交通大学出版社, 2016：89-90.

[2] 保罗•莱文森. 数字麦克卢汉——信息化新纪元纪指南[M].何道宽译, 社会科学文献出版社, 2001：16.

[3] 马歇尔·麦克卢汉. 《理解媒介：论人的延伸》[M].何道宽译.译林出版社, 2019：114.

[4] 尼古拉·尼葛洛庞帝.《数字化生存》[M].胡泳，范海燕译.电子工业出版社, 2017：145.

期刊文章

[1] 巴丹.新媒体环境下移动音频传播研究[J].长春师范大学学报, 2016, 35(01)：198-200.

[2] 白海莉.情感语音合成技术或对声纹鉴定准确性产生影响[J].科技创新与应用, 2018, (36)：24+26.

[3] 卜洪漩.智能时代新闻传播的伦理思考、价值遵循与制度保障[J].皖西学院学报, 2022, 38(01)：103-107.

[4] 陈晶, 隗静秋.移动音频平台版权侵权问题与应对——基于移动音频侵权案例的实证分析[J].科技与出版, 2022, (04)：87-93.

[5] 段宇, 温蜀珺.智媒时代下音频节目的结构嬗变与内容创新[J].视听, 2020, (04)：13-14.

[6] 冯菲, 王文轩, 修利超, 喻国明.冷热媒介：合成语音与真人语音的不同传播效

应——基于 EEG 的实验证据[J].新闻与传播研究, 2020, 27(12):5-20+126.

[7] 高莹莹, 朱维彬.面向情感语音合成的言语情感描述与预测[J].清华大学学报(自然科学版), 2017, 57(02):202-207.

[8] 宫承波, 陈曦.优质音频内容样态新玩法——智媒时代音频传播观察之二[J].新闻论坛, 2018, (04):22-24.

[9] 管佻路, 顾理平.价值冲突与治理出路：虚假信息治理中的人工智能技术研究[J].新闻大学, 2022, (03):61-75+119.

[10] 管佻路, 顾理平.智能语音交互技术下的用户隐私风险——以智能音箱的使用为例[J].传媒观察, 2021, (06):17-24.

[11] 管海建.智能语音技术在广播领域的应用探索[J].电视技术, 2022, 46(06):139-141.

[12] 姬德强.深度造假:人工智能时代的视觉政治[J].新闻大学, 2020, (07):1-16+121.

[13] 季凌霄.从“声景”思考传播:声音、空间与听觉感官文化[J].国际新闻界, 2019, 41(03):24-41.

[14] 赖黎捷, 颜春龙.广播音频与互联网移动音频的融合发展[J].中国广播, 2020, (08):32-36.

[15] 李武, 胡泊.声音的传播魅力:基于音频知识付费情境的实证研究[J].新闻大学, 2020, (12):49-60+120.

[16] 李葵, 徐海青, 吴立刚, 梁翀.基于多情感说话人自适应的情感语音合成研究[J].湘潭大学自然科学学报, 2018, 40(04):39-44.

[17] 李涛.论声音权在人格权编中的确立[J].三峡大学学报(人文社会科学版), 2019, 41(03):92-96.

[18] 李彤.现代听觉转向中的有声阅读[J].视听, 2021, (12):172-173.

[19] 梁旭艳.耳朵经济兴起的表现及原因探析——兼论互联网经济从眼球经济到耳朵经济[J].编辑之友, 2021, (08):18-23.

[20] 林爱珺, 马瑞萍.人工智能时代声音权立法的前瞻性思考[J].青年记者, 2019, (34):72-73.

[21] 刘建明.深度伪造对媒体与人类的致命威胁[J].新闻爱好者, 2021, (04):8-1

3.

[22] 刘亮, 陈德楠.新媒体时代移动音频提质研究[J].传媒, 2021, (19) : 45-48.

[23] 刘茜芸.数字移动音频产业中的版权保护风险与应对研究[J].科技与出版, 2021, (01) : 123-129.

[24] 刘婷.网络时代移动音频的法律性质及版权问题[J].经济研究导刊, 2022, (32) : 154-156.

[25] 刘艳辉.大数据和 AI 技术在新媒体传播渠道中的应用分析[J].中国传媒科技, 2022, (05) : 70-72.

[26] 刘一鸣, 高玥.人工智能语音在移动音频中的应用研究[J].出版发行研究, 2019, (11) : 35-39.

[27] 栾轶玫, 周万安.传统广播转型新方向:移动付费“音频生态圈”[J].新闻与写作, 2018, (10) : 44-47.

[28] 梅凯, 刘鸣筝.应用、变革与反思:智能语音加持下的新闻生产[J].青年记者, 2021, (03) : 57-58.

[29] 牛存有.5G 为传统广播的智慧化赋能[J].视听界, 2019, (01) : 44-47.

[30] 牛沛媛.传统广播向移动音频客户端的转化——以阿基米德 FM 和 iHeartRadio 为例[J].传媒, 2018, (19) : 48-50.

[31] 任子寒, 姚瑶, 余人.语音交互技术在移动音频中的应用风险与防范策略[J].编辑学刊, 2021, (04) : 18-23.

[32] 邵羽, 陆晓燕.人工智能时代声音权的民法保护——必要性、法理基础与路径[J].商业文化, 2021, (11) : 28-29.

[33] 孙建敏.移动音频的传播现状与评价体系——移动音频相关期刊论文研究综述[J].中国传媒科技, 2021, (04) : 31-33.

[34] 孙艳华.阅读听书平台智能合成语音的应用进展、质量现状和用户接受[J].编辑之友, 2021, (12) : 81-88.

[35] 童云, 周荣庭.移动音频声音质量评价维度[J].中国广播, 2020, (01) : 40-43.

[36] 唐嘉楠, 宋歌.塑造“声音景观”移动音频平台的生态圈构建[J].现代视听, 2019, (07) : 15-20.

[37] 王晨.技术赋能:智能移动终端付费声音产品的困局与展望[J].新闻世界, 202

2, (06) : 39-42.

[38] 王旖旎, 邸娜.移动终端的数字音频水印检测软件实现[J].现代电影技术, 2021, (02) : 44-48.

[39] 王长潇, 刘瑞一.从播客到移动音频媒体:赋权的回归与场景的凸显[J].新闻大学, 2019, (06) : 71-80+123-124.

[40] 王斌, 王育军, 崔建伟, 孟二利.智能语音交互技术进展[J].人工智能, 2020, (05) : 14-28.

[41] 王国忠, 赵淑华.论声音的人格标识特性和声音人格利益保护[J].黑龙江省政法管理干部学院学报, 2015, (01) : 77-79.

[42] 王涵.数字移动音频版权保护现状实证研究——以“喜马拉雅 FM”为例[J].出版发行研究, 2019, (02) : 69-72.

[43] 王文博.新媒介移动音频的生存策略研究[J]. 出版广角, 2021, (03) : 43-45.

[44] 王雪玉洁, 杨宇鹤.移动音频的发展成因、传播样态及品牌塑造[J].传媒, 2022, (20) : 77-79.

[45] 王垚, 邓逸钰.人工智能时代的移动阅读:需求、内容及交互设计[J].现代出版, 2021, (06) : 76-79.

[46] 王峥.语音合成技术在声音修复上的尝试[J].现代电影技术, 2019, (07) : 46-50.

[47] 吴建宇.全媒体时代移动音频发展新思路研究[J].科技传播, 2020, 12 (18) : 49-50.

[48] 吴瑶, 张亚莉.有声书版权的主体认定与权责链关系——基于 2021 版《中华人民共和国著作权法》[J].青年记者, 2021, (23) : 80-83.

[49] 武婷婷.人工智能时代的声音侵权现象研究[J].新闻传播, 2020, (06) : 16-17.

[50] 薛宝琴.人是媒介的尺度:智能时代的新闻伦理主体性研究[J].现代传播(中国传媒大学学报), 2020, 42 (03) : 66-70.

[51] 薛俏.数字化视域下我国移动音频的发展探讨[J].记者摇篮, 2022, (01) : 119-120.

[52] 许加彪, 张宇然.耳朵的苏醒:场景时代下的声音景观与听觉文化[J].编辑之友, 2021, (08) : 12-17+23.

[53] 杨帅, 乔凯, 陈健, 王林元, 闫镔.语音合成及伪造、鉴伪技术综述[J].计算机系统应用, 2022, 31(07):12-22.

[54] 杨鑫悦.新华社手机客户端 AI 合成主播的语音考察[J].新闻传播, 2022, (12):4-8.

[55] 殷爽.移动互联网时代我国移动音频的发展探析[J].视听, 2017, (02):58-59.

[56] 余苗, 赵文聪.人工智能发展与有声阅读平台创新探析[J].传媒, 2022, (06):54-56.

[57] 喻国明, 王文轩, 冯菲, 修利超.合成语音新闻的传播效果评测——关于语速影响的 EEG 证据[J].国际新闻界, 2021, 43(02):6-26.

[58] 喻国明, 王文轩, 冯菲.智能传播时代合成语音传播的效应测试——以语速为变量的效果测定[J].当代传播, 2020, (01):25-29.

[59] 张丹烽, 李冠宇, 赵英娣.语音合成技术发展综述与研究现状[J].科技风, 2017, (22):72.

[60] 张欣瑞.声音与智能传播:关于智能语音新闻的探索思考[J].新闻研究导刊, 2022, 13(15):73-75.

[61] 张学海, 杨璐铭.合成语音的声纹鉴定分析——以两名 AI 虚拟主播语音为基础[J].中国司法鉴定, 2022, (02):69-72.

[62] 张艳, 陈瑶.场域理论视角下有声阅读中的伦理失范成因及规制研究[J].出版发行研究, 2021, (11):30-34+79.

[63] 张洁意.移动音频平台的类型化发展策略[J].青年记者, 2018, (26):24-25.

[64] 张路琼, 崔青峰.移动音频的传播特征及媒介演变[J].青年记者, 2020, (29):75-76.

[65] 郑聪.使用 AI 转录移动音频字幕的法律界限——基于亚马逊被诉侵犯版权案视角[J].淮南师范学院学报, 2020, 22(06):14-19.

[66] 周鹃鹏, 赵洁.区块链技术背景下数字音频商业模式变革的逻辑——基于云听、喜马拉雅 FM 和 CastBox 的对比分析[J].传媒, 2022, (13):56-58.

[67] 周伟红, 胡伟.数据视角下 AI 语音传播的发展探析[J].青年记者, 2021, (18):79-80.

[68] 周月玲.移动音频自出版打造数字阅读新模式[J].出版广角, 2021, (14):78-8

0.

[69] 朱飞虎, 张晓锋.全民阅读视域下的有声阅读:痛点问题、核心价值与未来进路[J].中国编辑, 2022, (08):92-96.

学位论文

[1] 王娟.Deepfake：智能传播的伦理风险及其治理研究[D]. 东北财经大学, 2021.

[2] 周净.融入情感表现力的语音合成方法研究与应用[D]. 电子科技大学, 2021.

报告

[1] 艾媒咨询.2021 年中国在线音频行业发展及用户行为研究报告[R].2021.

[2] 国家互联网信息办公室.数字中国发展报告（2021 年）[R].2022.

[3] 中国互联网络信息中心(CNNIC).第 50 次《中国互联网络发展状况统计报告》[R].北京:2022.

附 录

《听众对网络音频平台中AI语音合成技术的认知与满意度调查》问卷

尊敬的先生/女士：

您好！我是兰州财经大学商务传媒学院新闻与传播专业的学生，非常感谢您的帮助。本问卷采取匿名填写的方式，主要为了了解广大听众对网络音频平台中使用AI语音合成技术的认识情况、满意程度及相关建议。本问卷仅用于撰写毕业论文，请您按自身情况合理填写。再次感谢您的支持，祝您生活愉快！

1、您的性别：（）男

（）女

2、您的年龄：（）18岁及以下

（）19——25岁

（）26——35岁

（）36——45岁

（）46岁及以上

3、您的学历：（）高中及以下

（）专科

（）大学本科

（）研究生及以上

4、您的职业：（）在校学生

（）公职人员

（）企业工作人员

（）工人、农民

（）个体商户

（）自由职业者

（）其他_____

5、您是否使用过网络音频客户端（如喜马拉雅）：（）是

（）否

6、（多选题）您通常会使用哪项网络音频客户端？（）喜马拉雅

（）猫耳FM

（）蜻蜓FM

（）企鹅FM

（）荔枝

（）其他_____

7、您对网络音频客户端的使用频次：（）每周10次及以上

（）每周5——10次

（）每周1——5次

8、您使用网络音频客户端的时间段：（）早上通勤时间

（）午间休息时间

（）晚上入睡以前

（）其他碎片时间

9、（多选题）您使用网络音频客户端的内容收听情况：（）广播剧

（）有声小说

（）相声评书

（）电台广播

（）经典名著

（）儿童故事

（）其他_____

10、您对AI语音合成技术相关概念了解多少？（）从未听说

（）听说过，但不了解

（）稍微了解

（）比较了解

（）非常了解

11、您是否是通过网络音频内容接触到AI语音合成技术？（）是

（）否

12、（多选题）您了解AI语音合成技术/产品的渠道：（）新媒体平台

（）主流媒体报道

（）亲人朋友

（）科普讲座

（）智能语音设备

（）其他_____

13、您是否在网络音频平台中收听过AI语音内容？（）是

（）否

14、您对网络音频平台应用AI合成语音的接受程度：（）完全不接受

（）勉强接受

（）无所谓

（）可以接受

（）完全接受

15、请您对现阶段网络音频平台中AI语音合成技术的应用功能进行评价：

	非常不满意	不太满意	一般	比较满意	非常满意
调节语速功能					
音色选择功能					
选中文字播读功能					

16、您认为网络音频内容是否应该添加AI技术应用功能？（）是

（）否

17、（多选题）您认为网络音频内容应添加何种AI技术应用功能？

（）语速调节档位

（）音色种类

（）特殊音效

（）背景音乐

（）角色间对话

（）其他_____

18、您认为网络音频内容中的AI合成语音流畅度如何？（）非常生硬

（）生硬

（）一般

（）流畅

（）非常流畅

19、您认为网络音频内容中的AI合成语音清晰度如何？（）非常模糊

（）模糊

（）一般

（）清晰

（）非常清晰

20、您在收听AI语音合成的音频内容时，是否遇到过错读误读？

（）从未遇到

（）偶尔遇到

（）总是遇到

（）不清楚

21、您是否给AI合成语音作品留言评论：（）从未评论

（）偶尔评论

（）总是评论

22、您给AI合成语音作品评论的内容倾向是：（）不如真人声音

（）和真人声音一样

（）比真人声音效果更好

23、您认为AI语音作品是否需要合成情感语音？（）是

（）否

24、您认为情感语音是什么？（）模仿真人情感的语音

（）通过算法自主匹配的情感机制

25、您认为AI语音合成技术应用于网络音频领域是否存在风险？（）是

（）否

26、您对AI语音合成技术风险了解多少：（）完全不了解

（）不了解

（）一般了解

（）了解

（）完全了解

27、请您评价以下情况的风险危害程度：

	不清楚	没有危害	少许危害	较大危害	极大危害
复制明星声音并上传音频平台					
影视行业利用合成语音进行人物模拟训练					
模仿公众人物发表公开讲话					
模仿亲人朋友的声音进行电话诈骗					
未被鉴别的声音信息作为案件证据					
使用声音信息进行身份验证					
利用合成语音发表虚假新闻					
复刻已逝人物的声音					

28、（多选题）您认为AI语音合成技术存在何种风险？

（）虚假信息泛滥

（）泄露个人隐私信息

（）社会信任度下降

（）冲击新闻真实性，挑战新闻专业机构权威

（）侵犯他人声音权、名誉权和财产权

（）侵犯有声读物著作版权

（）诱发技术犯罪

（）其他_____

29、（多选题）针对网络音频平台和AI语音合成技术已经出现的问题与风险，您认为应如何应对？（）规范网络音频市场管理

（）创新网络音频产品AI功能

（）提升合成语音质量

（）强化网络音频平台责任

（）立法监督侵权现象

（）培养公众信息素养

（）其他_____

30、您对网络音频平台大范围使用AI语音合成技术的态度是：

（）支持，对AI语音合成技术的正面应用很有信心

（）支持，但对AI语音合成技术可能带来的风险表示担忧

（）反对，认为AI语音合成技术必定威胁个人和市场安全

（）保持中立

致 谢

随着致谢二字落下，匆匆三年的硕士生涯也逐渐浮现出句点的踪影。我从一个穿着花裙子的小女孩，成为了一名略有学识的年轻人。三年前，我怀揣激动的心情走入一个陌生的城市，三年后我坐在家里的书桌前，一字一句敲下我的感受与收获。

论文写作无疑是一种痛苦的磨炼，但我也并不能声称自己克服了多么了不起的困难，洋洋洒洒的几万字并不能说明什么，我甚至要感谢这一过程给予我的馈赠，让我学会量力而行，学会不纠结。当然，也学会感恩，感谢给予帮助的李艳导师，全程以科学、诚实、严谨的态度耐心为我指导；感谢永远站在我身后的父母，毫无保留地支持我、爱护我，让我有勇气去闯荡世界；感谢带给我无限快乐的室友，让我拥有可笑又美好的记忆；感谢无数深夜里陪伴我的小猫，也感谢走得很慢依旧向前的自己。

最后，好好的道个别吧，愿大家下一旅程明媚灿烂。