链条完整可追溯。第四，平台运营需实施“传播日程矩阵”管理，赛前推出预热专题与选手短片，赛中主轨直播同步辅助平台实时花絮，赛后分发战术解析与幕后专访，实现赛事全周期传播。最后，应确立“内容差异＋风格统一”准则，各平台可在语速、语态、视觉结构上做微调，但应统一核心解说风格与术语体系，由中心团队负责审核与标准发布，以维持跨媒介内容一致性和传播协同效能。

3.3 设立文化内涵与解说故事化叙述框架体系

在新媒体语境下，为了提升体育解说的深度与情感共鸣可构建以文化内涵为核心且可执行的赛事叙事与故事化解说框架，应首先建立标准化“赛事叙事模板库”，模板明确信息单元（运动员成长档案、历史脉络、地域文化符号、战术谱系与观众情绪节点），并以结构化元数据便于平台检索与 AI 辅助调用；其次在稿件流程中嵌入多层次校对机制，设置文化审校员与法务复核岗，利用关键词黑白名单与敏感度评分模型筛查地域刻板、意识形态偏差与可能触发的政治敏感点；第三强化现场解说的话语策划方法，采用叙事节律控制（引子—铺陈—高潮—回落）与情绪线索标注，使解说员在实时播报中插入短篇式人物叙述与战术演变微叙事以增强代入感；第四将运动员与受众的多源内容（运动员自媒体、粉丝社群素材、历史影像）纳入解说素材库，建立版权与合规快速确认通道，支持二次创作与跨平台重构，以维持叙事一致性与文化延展性；第五在培训体系中引入叙事写作与民族文化素养模块，通过案例实操、口语转化为书面语训练及实时模拟演练提升解说员在新媒体场景下的跨文化传达能力，并引入基于受众画像的叙事适配与 AI 辅助标签化工具，用传播效果指标追踪叙事节律与情绪响应，实现模板迭代与平台沉淀 [4]。

3.4 构建数据驱动的受众反馈分析与内容再生产机制

在新媒体情境下，体育解说的传播优化还应构建数据驱动的受众反馈分析与内容再生产机制，具体为：首先，数据采集层面，针对弹幕、评论、点赞、转发、停留时长等多项互动信息搭建统一的数据接口予以采集，随后经过标准化、清洗算法转化成可追踪行为数据库，接着便可利用时间序列的模型分析受众内容关注点、互动节奏及其变化趋势，从而给动态调整体育解说传播策略提供数据支持。其次，语义情绪与热词提取层面，可以采取 BERT 或 ERNIE 等基于深度语义理解的情感分析模型对用户弹幕与评论中的关键词词性标注、依存句法关系与情绪极性分类，以生成不同平台的情感热度曲线、主题词频率图表，从而检测受众体育赛事过程中情绪和议题变化走向。再次，在受众画像构建环节，应融合静态属性（性别、年龄、地域）与动态行为（观看时段、互动密度、内容偏好）构建多维画像模型，利用聚类算法与协同过滤方法区分专业观众、普通观众及新手群体，为解说团队提供语言风格与知识深度的个性化调整依据 [5]。最后，在内容 A/B 测试与再生产环节，应设计多版本解说脚本（如专业化与娱乐化版本）并通过算法推荐并行推送，监测点击率、完播率、互动率等指标差异，筛选表现最佳版本纳入标准模板，同时由系统自动标注高热度片段生成短视频、专题回顾及知识化延伸内容，实现数据反馈—内容优化—传播再生产的动态闭环。

4 结语

综上所述，受新媒体环境的影响，当前体育解说呈现出在时效性、互动性、碎片化以及品牌化等四方面显著的传播特点，在这种情况下为实现更优的传播效果可采取建构差异版体系、搭建跨媒介联动、设立文化内涵与解说框架体系以及运用数据驱动四项优化策略。同时我们还应看到新媒体环境下体育解说传播仍存在着各网络平台版面协同困难、人才供应匮乏以及机制重建艰巨等问题，但若能科学设计传播框架并强化平台协同将体育解说打造成为新时代体育传播的重要媒介触点，从而推动我国体育话语生态更具文化底蕴及活力。

参考文献

[1] 任捷.场域理论视角下新媒体体育解说场域结构与媒介环境研究[J].新闻研究导刊,2025(10).

[2] 陈悦盈.5G背景下我国主流媒体在体育新闻传播中的新角色定位分析[C]//第七届贵州省体育科学大会论文摘要集.2024.

[3] 袁崧浩.塑造人,影响人:新媒体环境下体育解说的立场把控[J].科技传播,2020,12(11):2.

[4] 何欣键.新媒体语境下电视体育解说员的机遇与挑战[J].冰雪体育创新研究,2025(11).

[5] 张研.新媒体语境下体育解说倾向性的问题研究——大学生篮球《北京化工VS中国矿大》解说作品创作阐述[D].上海体育学院,2022.

 DOI: https://doi.org/10.12345/whyscx.v8i6.34835

FireRedTTS-2:An Introduction to Multi-Speaker Dialogue Speech Synthesis Technology and its Application

Qing Liu

Kunming Media Convergence Center, Kunming, Yunnan, 650118, China

Abstract

FireRedTTS-2, as an advanced open-source streaming Text-to-Speech (TTS) system, exhibits particularly outstanding performance in multi-person, multi-turn dialogue scenarios, bringing revolutionary breakthroughs, especially to the generation of talk shows. This article will focus on the core functionalities of FireRedTTS-2 in multi-person, multi-turn dialogue generation, delve into the key technical principles supporting its powerful performance, and analyze its application value in the production of multi-role talk shows, podcasts, and traditional radio programs, by integrating with the actual needs of radio stations and media convergence centers. The aim is to provide valuable references for developers and creators in related fields.

Keywords

TTS; Speech Synthesis ;Multi-Speaker Dialogue; FireRedTTS-2,Podcast

FireRedTTS–2 多角色对话语音生成技术与应用研究概论

柳晴

昆明市融媒体中心，中国·云南 昆明 650118

摘　要

FireRedTTS-2 作为一款先进的开源流式文本到语音（TTS）系统，其在多人多轮对话场景下的表现尤为突出，尤其为谈话类节目的生成带来了革命性的突破。本文将聚焦 FireRedTTS-2在多人多轮对话生成方面的核心功能，深入剖析支撑其强大表现的关键技术原理，并结合广播电台、融媒体中心节目制作实际需求，详细分析其在多角色谈话类、播客类、传统电台类音频制作领域的应用价值，以期为相关领域的开发者和创作者提供有价值的参考。

关键词

TTS；语音生成；多角色对话；FireRedTTS-2；播客

1 引言

传统的文本到语音技术主要聚焦于将书面文本转换为清晰、自然的单人语音。尽管在音质和自然度上取得了长足进步，但其应用场景始终受限于“独白”形式。对于播客、广播剧、有声书对话、视频解说等需要多人互动的内容形式，传统 TTS 技术显得力不从心，通常需要预先录制不同角色的语音再进行后期剪辑合成，流程繁琐、成本高昂。

FireRedTTS-2 通过其创新的模型架构和训练策略，有效解决了传统语音合成在长对话、多说话人交互中的韵律不连贯、说话人切换生硬、情感表达不足等难题，为高效、高质量地创造沉浸式音频内容提供了强大的技术支撑。

FireRedTTS-2 的出现，标志着 TTS 技术从“独白”向“对话”的范式转变。其内置的多人多轮对话功能，能够根据标注的文本脚本，自动生成一场由多个虚拟角色参与的、韵律丰富、情感起伏、且交互自然的语音对话。这不仅是技术能力的提升，更是为音频内容的生产方式带来了革命性的变革，为 AIGC 领域开辟了全新的疆域。

2 FireRedTTS–2 的核心功能

FireRedTTS-2 在多人多轮对话生成方面的核心功能，旨在彻底革新播客等音频内容的创作和生产模式。主要体现在：

零样本或低样本声音定制：允许用户提供少量目标说话人的声音样本，即可合成具有该说话人特有音色和风格的语音，极大地降低了对专业播音员的依赖。

高效率的对话脚本转化：能够将书面对话脚本高效地转化为听起来逼真、生动的音频内容，大幅缩短制作周期。

自然流畅的多人对话合成：能够生成包含多个对话者（如主持人、嘉宾）之间自然、连贯的对话，目前支持 3 分钟的 4 说话人对话，并且通过扩展训练语料库可以轻松扩展到更长对话时间和更多的说话人。

精准的说话人身份识别与切换：能够准确识别并区分

【作者简介】柳晴（1979–），男，中国云南昆明人，本科，高级工程师，从事广播电视工程研究。

不同说话人，实现流畅、无缝的身份切换，让听众能清晰辨别对话参与者。

上下文感知的韵律与情感表达：生成的语音韵律能够准确捕捉并反映对话的整体情感走向和上下文信息，使得对话更具感染力和表现力。

流式生成：降低首包延迟，逐句或逐段生成语音，确保用户在收听过程中获得流畅无阻的体验，不会因等待合成而产生不适。

3 技术原理深度解析

FireRedTTS-2 之所以在多人多轮对话生成方面表现出色，得益于其精巧的技术设计，尤其体现在以下几个关键组件上：

3.1 12.5Hz 流式语音分词器 (Tokenizer)

多人多轮对话的核心在于理解和维持对话的上下文。传统的 TTS 系统在处理长对话时，由于序列过长，容易丢失长距离依赖信息，导致韵律断裂和语义模糊。FireRedTTS-2 的 12.5Hz 流式语音分词器正是解决这一问题的关键：

序列长度压缩：将语音信号以 12.5Hz 的帧率编码成离散的语音分词（Audio Tokens），加速了训练和推理，延长了最大对话长度，编码了更丰富的语义到分词建模，并支持实时应用的高保真流式生成。

增强长距离依赖建模：更短的序列使得 Transformer 模型能够更容易地捕捉到跨越多个话语甚至整个对话的长距离依赖关系。这对于理解对话的主题、人物关系、情感发展至关重要。

提高计算效率：Transformer 模型在处理序列时，计算复杂度与序列长度成正比。序列长度的缩短直接带来了训练和推理时间的显著降低，为处理复杂的对话场景提供了可能。

降低内存占用：更短的序列也意味着更少的内存消耗，这使得在有限的硬件资源下处理更长的对话成为可能。

3.2 双 Transformer 架构

为了有效处理长对话中的文本和语音信息，FireRedTTS-2 采用了独特且高效的双 Transformer 架构（见图一）。

3.2.1 文本 - 语音交错序列

格式设计：[Speaker Label] <Text> <Audio Tokens> [Speaker Label] <Text> <Audio Tokens> ...

信息融合：这种格式强迫模型学习文本内容如何与语音表现相结合，以及说话人标签如何指示身份。

长距离依赖：将所有话语信息编码在一个长序列中，使得模型能够直接访问和利用历史话语的信息。

3.2.2 主干 Transformer (Backbone Transformer)

定位：这是一个大型的、基于 Qwen2.5 的仅解码器 Transformer，负责处理整个交错序列。

任务：它对整个输入序列进行一次自回归推理，编码文本语义和对话上下文，并预测 第一层 的语音分词。同时，它产生丰富的隐藏状态，这些隐藏状态包含了全局的上下文信息。能够捕捉到对话的全局信息，为后续的精细化生成提供全局指导。

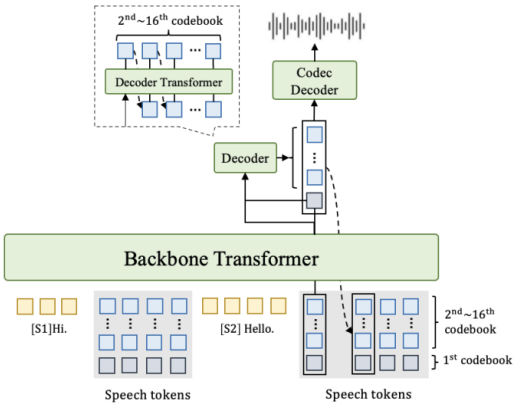

图 1 文本到语音模型

3.2.3 解码器 Transformer (Decoder Transformer)

定位：一个相对较小的 Transformer 模型。

任务：它接收主干 Transformer 产生的全局隐藏状态，以及 已生成的第一层分词。然后，它在此基础上进行 N-1 次自回归推理，逐层预测剩余的语音分词（例如，预测第二层、第三层 ...）。

3.2.4 双 Transformer 架构的技术优势

高效上下文利用：它能在每个时间步访问到主干 Transformer 提供的完整对话上下文，确保了生成语音的连贯性和一致性。

计算效率提升：相较于传统的深度自回归模型（每一步都需要完整的模型推理），这种“主干 + 解码器”的设计大幅减少了计算量，尤其是首包延迟。主干 Transformer 仅需推理一次，解码器 Transformer 的推理成本相对较低。

与传统方法的对比（以“延迟模式”为例）：传统的 TTS 模型会将整个模型堆叠 N 层，每一层的输出都需要前一层的输出，在生成第 k 步时，模型需要进行 k 次自回归推理。FireRedTTS-2 仅需一次主干 Transformer 的完整推理，和 N-1 次较小的解码器 Transformer 的推理。这显著降低了计算负担，特别是在处理长对话时，其优势更为明显。

3.3 训练策略

FireRedTTS-2 采用了精心设计的训练策略，以确保模型能够逐步掌握复杂的对话生成任务：

3.3.1 第一阶段：预训练 (Pre-training)

数据：使用大规模的独白语音数据集（例如，1.1M 小时的语音）。

目标：让模型学习基本的文本到语音映射能力，掌握

不同说话人的基础声学特征，建立对语音信号的初步理解。这是模型“发声”的基础。

3.3.2 第二阶段：后训练 (Post-training)

数据：使用大规模的多说话人对话数据集（例如，300k 小时的对话，包含 2-5 个说话人）。

目标：专注于提升模型在多人对话场景下的表现，学习说话人身份切换、多轮上下文的依赖关系、以及对话中的韵律和情感模式。此时模型开始具备“对话感知”和“互动能力”。

3.3.3 第三阶段：监督微调 (SFT, Supervised Fine-tuning)

数据：使用特定应用场景或目标说话人的少量数据。

目标：将模型“定制”到特定的应用需求。例如，模仿特定播客主持人的语速、语调、停顿习惯；调整情感的细微差别，使播客对话更具真实感和表现力。

4 应用分析

FireRedTTS-2 在多人多轮对话生成方面的强大能力，为谈话类音频内容创作带来了前所未有的机遇和变革。

4.1 颠覆传统制作流程

4.1.1 痛点

传统的谈话类音频制作高度依赖人工，需要主持人、嘉宾、录音工程师、后期剪辑师等多个角色。录制、降噪、多说话人声音分离、混音、语速调整、情感化处理等环节耗时耗力，成本高，技术门槛高。

4.1.2 FireRedTTS-2 的解决方案

脚本即成品：用户只需撰写详细的对话脚本，并提供少量目标说话人的声音样本，FireRedTTS-2 即可直接生成高质量的播客音频。

“零样本”创作：对于没有现成录音的场景，甚至可以通过简单的文本描述或模仿，让模型生成“全新的”人物声音，实现真正意义上的“零样本”播客创作。

极大提高效率：制作周期从数天甚至数周，缩短到数小时甚至数分钟。

降低成本：显著减少了对专业配音演员、录音棚、后期制作人员的依赖，降低了播客制作的门槛和成本。

4.2 提升生成语音的质量与表现力

4.2.1 自然度与连贯性：

精准说话人切换：论文评估结果显示，FireRedTTS-2 在说话人相似度、说话人切换的平滑度上表现优异，听众能够清晰地辨别不同说话人，且切换过程不显突兀。

上下文韵律：通过 12.5Hz 分词器对长对话上下文的有效编码，以及双 Transformer 的全局上下文建模，生成的语音韵律更加自然连贯，符合人类对话的节奏和情感起伏。这避免了传统分段合成时常出现的“硬生生”的感觉。

4.2.2 情感丰富性

情感微调：通过 SFT，模型可以被训练以模仿特定的主持人的情感表达风格（如热情、幽默、严肃、思考等），使得语音听起来更具“人格魅力”。

情境感知：模型能够根据对话内容，适当地调整语速、语调和重音，营造出更具代入感的听觉体验。例如，在讨论严肃话题时，语调会更沉稳；在轻松互动时，语速会加快，语气更活泼。

4.2.3 可定制性与多样性

个性化声音：播客创作者可以为自己的播客“量身定制”主持人的声音，使其与品牌形象或内容风格高度契合。

多样化角色：轻松合成多位主持人、嘉宾的语音，构建结构更复杂、信息量更大的多人对话语音节目。

4.3 拓展应用场景

自动化新闻播报：将文字新闻稿快速转化为多主持人播报的有声新闻，实现新闻内容的即时传播。

教育与知识科普：创作高度专业化、信息量大但需要多人互动的教育谈话节目，让复杂知识变得易于理解和接受。

有声书与长内容阅读：将书籍、文章转化为带有生动对话的有声内容，满足用户碎片化时间获取信息的需求。

个性化主持人：用户可以根据个人喜好，定制虚拟主持人的声音、风格甚至口头禅，创造属于专属的虚拟语音数字人。

4.4 开源优势

经济性：开源模型允许用户免费使用，大大降低了广播电台、融媒体中心的应用成本。

安全性：开源模型可以部署到自己的网络或者云服务器上，纳入自身的系统安全管理体系中。

可拓展性：可以基于开源模型进行二次开发、微调，创造出更具特定功能、更符合特定领域需求的应用。

5 结语

FireRedTTS-2 在多人多轮对话生成，特别是多角色谈话类节创作领域，展现出了前所未有的强大能力。其创新的 12.5Hz 流式语音分词器和高效的双 Transformer 架构，有效解决了长对话、多人多轮交互的核心技术挑战，使得生成的内容在自然度、连贯性、表现力和效率上都实现了质的飞跃。

在语音节目制作中，它带来了“脚本即成品”的革命，使得自动化新闻、谈话访谈、教育科普、有声读物等应用得以快速实现，并为未来个性化、互动式播客类型节目奠定了基础。

展望未来，FireRedTTS-2 为代表的语音生成技术将继续向着更深层次的上下文理解、更精细的情感模拟、更实时的互动编辑、更广泛的风格以及更完善的播出安全方向演进。FireRedTTS-2 不仅仅是一个语音生成工具，更是未来