文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

人工智能在视频通信中实时翻译和字幕生成的应用

李福霞　付学涛　李明慧*
博鼎实华（北京）技术有限公司
DOI:10.32629/acair.v3i4.17901

[摘　要] 全球化远程协作深度普及的当下,语言障碍已成为制约跨区域视频会议效能的关键桎梏。本文系统梳理AI驱动的实时翻译与字幕生成技术研究进展：首先阐述语音识别(ASR)、神经机器翻译(NMT)及低延迟边缘计算的技术机理,重点解析Transformer架构在声学建模与语义映射中的创新实践；其次深入剖析光学字符识别(OCR)字幕识别、多语言翻译引擎与动态渲染的全流程技术细节,结合数据对比主流平台在不同噪声场景、语种组合下的性能表现；最后指出复杂噪声干扰、小语种语料稀缺、实时性与准确率平衡及数据隐私合规等核心挑战,提出基于自监督学习的鲁棒性优化、联邦学习驱动的小语种扩展、轻量化模型部署及同态加密隐私保护等发展方向。研究可为视频会议系统测试工程、技术标准化建设及跨语言协作生态构建提供重要参考。
[关键词] 人工智能；视频通信；实时翻译；字幕生成；边缘计算
中图分类号：TP18　**文献标识码**：A

The Application of Artificial Intelligence in Real-time Translation and Subtitle Generation for Video Communication

Fuxia Li　Xuetao Fu　Minghui Li*
Potin(Beijing)Technology Co.,Ltd

[Abstract] With the in-depth popularization of global remote collaboration, language barriers have become a key constraint restricting the efficiency of cross-regional video conferences. This paper systematically sorts out the research progress of AI-driven real-time translation and subtitle generation technologies: firstly, it expounds the technical mechanisms of Automatic Speech Recognition (ASR), Neural Machine Translation (NMT) and low-latency edge computing, focusing on analyzing the innovative practice of Transformer architecture in acoustic modeling and semantic mapping; secondly, it deeply dissects the full-process technical details of OCR subtitle recognition, multilingual translation engine and dynamic rendering, and compares the performance of mainstream platforms in different noise scenarios and language combinations based on measured data; finally, it points out core challenges such as complex noise interference, scarcity of minority language corpora, balance between real-time performance and accuracy, and data privacy compliance, and proposes development directions including robustness optimization based on self-supervised learning, minority language expansion driven by federated learning, lightweight model deployment and homomorphic encryption privacy protection. The research can provide important references for video conference system testing engineering, technical standardization construction and cross-language collaboration ecosystem construction.
[Key words] artificial intelligence; video communication; real-time translation; subtitle generation; edge computing

引言

数字化转型进程加速推进的背景下,视频通信已从传统“可视化沟通工具”升级为全球企业协作、政务交流、学术研讨的核心基础设施。据Fortune Business Insights发布的《2024-2032年视频会议市场规模与趋势报告》,2024年全球视频会议市场收入预计达330.4亿美元,2025年将增至372.9亿美元,对应年增速约12.9%[1]。随着企业全球化步伐加快,跨国公司的语言障碍不仅造成会议时间被浪费,还导致关键信息在传递过程中出现失真,显著拖慢项目进度并增加“隐性成本”。人工智能技术的突破性进展为解决这一痛点提供了关键路径。近年来,深度学

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

习在语音识别、机器翻译领域的应用大幅提升了实时跨语言沟通的可行性：端到端ASR模型的词错误率(WER)已明显下降，边缘计算技术则压缩端到端延迟[2]。在此背景下，腾讯会议、Zoom、微软Teams等主流平台陆续集成AI实时翻译与字幕生成功能，用户渗透率持续快速攀升。本文从技术原理、应用现状、核心挑战及发展方向四个维度展开分析，旨在为相关技术研发与产业应用提供系统性参考。

1 核心技术原理与架构

1.1实时翻译技术链与关键模块

实时翻译系统采用“前端信号处理－核心语义转换－边缘云协同传输”的三级协同技术架构，各模块联动实现从语音输入到目标语言输出的低延迟转换。在前端语音信号处理阶段，系统先通过声电转换设备采集音频信号，经预加重、分帧加窗等预处理步骤后，提取梅尔频率倒谱系数、感知线性预测系数等声学表征特征。为应对复杂声场干扰，自适应噪声抑制模块采用最小均方误差算法动态消除稳态噪声，盲源分离技术则通过独立成分分析实现多说话人语音分离。

ASR是技术链的核心环节，当前主流方案已从传统隐马尔可夫模型+高斯混合模型演进为基于深度学习的端到端架构。其中，Transformer-ASR模型凭借并行计算效率高、长序列建模性能优异的特性，成为当前语音识别领域的研究焦点，其通过多头自注意力机制捕捉全局声学依赖关系，配合位置编码解决时序信息丢失问题。

NMT承担源语言文本到目标语言文本的转换功能，其核心在于构建深层语义映射模型。NMT通过Encoder-Decoder架构自动学习语言规律。Encoder模块由6-12层Transformer编码器构成，将源语言文本转换为固定维度的语义向量；Decoder模块则通过注意力机制动态聚焦源语言语义向量的关键部分，生成目标语言序列。为平衡翻译质量与实时性，当前系统普遍采用知识蒸馏技术[4]。

低延迟传输是实时翻译的关键保障，边缘云协同架构为此提供了有效解决方案。系统采用“终端预处理+边缘计算推理+云端模型更新”的协同模式：将ASR的特征提取、NMT的初步解码等轻量级任务部署于终端或边缘节点，而模型训练、大规模语料更新等重计算任务则在云端完成[3][4]。

图1 实时翻译技术链

1.2字幕生成技术流程与优化策略

字幕生成系统实现从视频内容到多语言字幕的自动生成，涵盖“视频帧文字识别－文本翻译－时间轴对齐－动态渲染”四个核心环节。在视频帧文字识别阶段，OCR技术先对输入视频帧进行预处理：通过自适应二值化增强文字与背景对比度，采用形态学滤波消除椒盐噪声，再经轮廓检测和文字区域定位提取文本区域。

对于低质量视频场景采用多模态融合识别策略：结合相邻3-5帧的文本信息进行交叉校验，同时引入语音识别结果辅助修正OCR错误。例如，当OCR识别结果为“人工智能x用”时，系统通过比对ASR输出的“人工智能应用”，自动修正错误字符，使低质量视频的字幕识别错误率降低。

时间轴同步是字幕生成的核心技术难点，可采用音视频帧对齐算法：通过提取视频帧的音频特征与ASR输出的语音特征进行动态时间规整，确定每个字幕片段的起始时间戳和结束时间戳。为应对网络抖动导致的音视频不同步问题，系统实时监测RTP(实时传输协议)的时间戳偏移量，当偏移超时则自动调整字幕显示时间。在渲染显示阶段，系统提供多种显示模式供用户选择：单语言模式仅显示目标语言字幕，采用半透明黑色背景确保文字清晰度。

2 应用现状与多维度性能对比

2.1国内外主流平台技术实践

国内视频通信平台在实时翻译与字幕生成领域呈现“本地化深耕+垂直场景创新”的发展态势。例如，腾讯会议构建了覆盖17种语言的实时翻译系统，其普通话识别准确率可达97%[4]，其“智能语义断句”技术通过分析语音停顿和语义边界，自动优化字幕分段，控制单句字幕长度，阅读舒适度提升。钉钉整合多语言翻译引擎，实现多语种互译，重点强化企业协作场景的功能适配：支持会议纪要自动翻译、白板内容实时字幕化，提升跨国团队的文档协作效率。

国际视频通信平台以“多语言覆盖+生态化扩展”为核心竞争力。Zoom的AI翻译功能支持多种语言，涵盖英语、西班牙语、阿拉伯语等主流语种及斯瓦希里语、豪萨语等小语种，其通过开放API接口，实现会议内容的实时本地化处理。微软Teams整合语音识别、翻译、合成全链路技术，支持多种语言的实时字幕与语音翻译，“实时语言检测”功能可自动识别发言者语言，无需手动切换翻译方向。

2.2多场景性能对比分析

为客观评估主流平台的技术性能，选取识别准确率、端到端延迟、语种覆盖度及噪声鲁棒性多个测试指标，测试场景涵盖安静办公室(信噪比≥30dB)、嘈杂会议室(信噪比15-20dB)、移动场景(信噪比5-10dB)三种典型环境。测试结果显示，在主流语种识别准确率方面，国内平台表现更为突出，在多语种覆盖上，国外平台表现亮眼，但小语种翻译质量仍有提升空间[5]。值得关注的是，所有平台在移动场景下的性能均出现不同程度下降，WER较安静环境均上升，这一结果表明复杂噪声环境仍是技术优化的重点方向。

3 核心挑战与未来发展方向

3.1当前技术瓶颈与应用挑战

复杂声场环境下的鲁棒性不足是当前ASR技术面临的首要挑战。实际应用中，视频会议场景存在多种噪声干扰：会议室混响导致语音信号拖尾，移动场景的风噪掩盖有效语音，多人交替

发言时的语音重叠进一步增加识别难度。此外,方言与口音差异也严重影响识别效果,如汉语的粤语、四川话等方言场景。

小语种语料稀缺制约多语言翻译系统的覆盖能力。全球7000多种语言中,拥有完善标注语料非常少,多数小语种存在“数据荒漠”问题[6]。语料稀缺导致小语种翻译模型的双语评估替换分数(BLEU)值很低,难以满足实用需求[7]。

实时性与准确率的平衡是系统设计的核心权衡难点。为提升翻译准确率,需要增加模型参数量和计算复杂度,但这会导致推理延迟上升；反之,过度追求轻量化会牺牲翻译质量。此外,多模态信息融合的深度不足也影响用户体验,现有系统多独立处理语音和文本信息,未能充分利用视频画面中的表情、手势等视觉信息辅助语义理解。

数据隐私合规要求日益严苛对技术架构提出新挑战。视频会议内容包含大量敏感信息,欧盟GDPR、中国《数据安全法》等法规对数据收集、传输、存储提出严格要求。现有系统多采用“终端－云端”数据传输模式,存在数据泄漏风险；而纯本地化部署虽能保障隐私,但受终端计算能力限制。

3.2前沿技术方向与解决方案

基于自监督学习的鲁棒性优化是提升复杂场景性能的关键技术路径。自监督学习通过利用海量无标注数据(如公开演讲视频、播客音频)进行预训练,大幅降低对标注数据的依赖。当前研究热点包括:采用对比学习构建噪声不变的声学表征,使模型在不同噪声场景下的泛化能力提升；引入扩散模型进行语音增强,通过逐步去噪生成清晰语音。此外,针对方言与口音问题,采用“通用模型+方言适配器”的架构：通用模型学习语言共性特征,方言适配器通过少量标注数据微调特定参数。

联邦学习驱动的小语种扩展为解决语料稀缺问题提供了创新性解决路径。联邦学习通过“数据不动模型动”的方式,在保护数据隐私的前提下实现多机构数据联合训练:各参与方(如高校、研究机构)在本地使用自有小语种语料训练模型,仅将模型参数更新上传至联邦服务器,服务器聚合参数后反馈给各参与方,反复迭代直至模型收敛。同时,跨语言迁移学习技术也在快速发展,基于大语言模型的零样本翻译能力,可实现未见过语种的初步翻译,为小语种系统构建提供“冷启动”方案。

轻量化模型与硬件加速技术助力实现实时性与准确率的平衡。模型压缩方面,采用结构化剪枝移除冗余的卷积核和注意力头。硬件加速方面,采用FPGA(现场可编程门阵列)或ASIC(专用集成电路)实现翻译引擎的硬件化部署。此外,边缘计算节点的分布式部署进一步缩短传输延迟。

多模态融合与隐私计算技术推动系统向更智能、更安全方向发展。多模态融合方面,通过Transformer架构融合语音、文本、视觉信息:语音模态提供时序特征,文本模态提供语义特征,视觉模态(人脸表情、手势)提供情感和意图特征,三者协同使歧义句消歧准确率提升。隐私保护方面,采用“同态加密+联邦学习”的融合方案：同态加密实现数据在加密状态下的计算,确保原始数据不泄露;联邦学习实现分布式模型训练,避免数据集中存储风险。此外,差分隐私技术通过向模型参数添加噪声,防止攻击者通过模型反推原始数据,满足隐私保护要求。

4 结束语

AI驱动的实时翻译与字幕生成技术正处于快速发展阶段,不仅重构了视频通信的跨语言协作模式,更成为推动全球数字化协作的重要基础设施。当前技术已在主流语种、安静场景下实现实用化,但在复杂噪声鲁棒性、小语种覆盖、实时性－准确率平衡及隐私保护等方面仍需突破。未来,随着自监督学习、联邦学习、多模态融合等技术的深入应用,视频会议系统将逐步实现“全场景鲁棒”“全语种覆盖”“零感知延迟”“强隐私保护”的目标。

建议行业层面加快技术标准化建设:制定语音识别准确率、翻译延迟、字幕同步精度等关键指标的测试规范；建立多语种语料共享机制,推动小语种数据资源池建设;加强隐私保护技术的合规性认证,构建安全可信的技术生态。通过产学研用协同创新,最终实现“无障碍全球沟通”的愿景,为全球化协作的深度发展提供技术支撑。

[参考文献]

[1]Fortune Business Insights.Video conferencing market size, share & trends analysis report 2024 - 2032[R].Pune:Fortune Business Insights,2024.

[2]A Survey on State-of-the-art Deep Learning Applications and Challenges[R].arXiv:2403.17561,2024.

[3]天翼云开发者社区.边缘计算与云服务器协同:低延迟场景下的分布式架构设计[EB/OL].2025-05-16.https://www.ctyun.cn/developer/article/676520525463621.

[4]Niehues J,et al.Low-latency neural speech translation[C]//Interspeech 2018:1298-1302.[Online].

[5]Wang,G.,Zhao,Q.,Zhou,Z.,& Liu,Y.(2025).Research on Real-time Multilingual Transcription and Minutes Generation for Video Conferences Based on Large Language Models[J].International Journal of Innovative Research in Engineering & Management,11(6):8-20.

[6]冯笑,杨雅婷,董瑞.基于回译和集成学习的维汉神经机器翻译方法[J].兰州理工大学学报,2022,48(5):99-105.

[7]陆雯洁,谭儒昕,刘功申.基于半监督学习的小语种机器翻译算法[J].厦门大学学报(自然科学版),2019,58(2):200-208.

作者简介：

李福霞(1983--),女,汉族,山东省烟台人,硕士,博鼎实华(北京)技术有限公司,工程师,研究方向：多媒体通信。

付学涛(1975--),女,汉族,北京人,本科,博鼎实华(北京)技术有限公司,工程师,研究方向：密码通信。

***通讯作者：**

李明慧(1975--),女,汉族,河北秦皇岛人,硕士,博鼎实华(北京)技术有限公司,高级工程师,研究方向：多媒体通信。