语音大模型:架构、训练与挑战分析

张亚洲[1,2],刘祈蒙[1],戎 璐[3],赵 彬[4],李爱军[4*]

(1. 郑州轻工业大学软件学院,河南郑州 450000;2. 天津大学智能与计算学部,天津 300350;3. 天津大学教育学院,天津 300350;
4. 中国社会科学院语言研究所,北京 102488)

摘 要: 大型语言模型(Large Language Models,LLMs)凭借其卓越的指令跟随能力与上下文学习能力在众多下游自然语言处理(Natural Language Processing,NLP)任务上取得巨大成功. 鉴于人类智能的多模态属性,这种研究热态自然地蔓延到其他模态,特别是视觉模态和语音模态. 在视觉领域,以 GPT-4V、LLaVa 为代表的视觉大模型使用基础语言模型作为"大脑"执行视觉理解和视觉推理任务,展现出跨越"任务壁垒"的能力. 对比而言,语音大模型(Speech Large Language Models,SLLMs)研究同样受到学术界与工业界的高度关注. 涌现出以 Whisper、Qwen-Audio 为代表的一系列模型,在语音识别、语音理解和语音合成等任务上不断突破性能边界,展现出令人瞩目的发展潜力. 本文旨在系统梳理和总结语音大模型的最新研究进展. 文章深入阐述语音大模型的基本框架,并详尽探讨相关核心概念,包括模型组件、训练策略、数据构建以及评估方法. 在此基础上,本文进一步分析了当前研究中的主要挑战,并展望了未来可能的发展方向.

关键词: 语音大模型;大型语言模型;指令跟随;语音理解;预训练

基金项目: 国家自然科学基金(No.62006212)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3454-19

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250367

Speech Large Language Models: Architecture, Training and Challenges Analysis

ZHANG Ya-zhou[1,2], LIU Qi-meng[1], RONG Lu[3], ZHAO Bin[4], LI Ai-jun[4*]

1. *Software Engineering College*, *Zhengzhou University of Light Industry*, *Zhengzhou*, *Henan* 450000, *China*;
2. *College of Intelligence and Computing*, *Tianjin University*, *Tianjin* 300350, *China*;
3. *School of Education*, *Tianjin University*, *Tianjin* 300350, *China*;
4. *Institute of Linguistics*, *Chinese Academy of Social Sciences*, *Beijing* 102488, *China*)

Abstract: Large language models (LLMs) have achieved outstanding success across a wide range of downstream tasks in natural language processing (NLP), thanks to their remarkable ability to follow instructions and learn from context. As human intelligence is inherently multimodal, the momentum of this research has naturally expanded into other modalities, particularly vision and speech. In the realm of vision, large-scale models like GPT-4V and LLaVa employ foundational language models as the "brain" enabling them to perform complex tasks in visual understanding and reasoning. These models have shown impressive abilities to break down task barriers, transcending traditional boundaries in vision-related tasks. In a similar vein, speech large language models (SLLMs) have attracted significant interest from both academia and industry. Notable models such as Whisper and Qwen-Audio have emerged as frontrunners, setting new performance records in speech-related tasks, including speech recognition, understanding, and synthesis. Their development demonstrates significant potential for further breakthroughs. This paper aims to provide a comprehensive review of the latest advancements in SLLMs research. It delves into the foundational architecture of these models, thoroughly exploring key concepts such as model components, training strategies, data construction, and evaluation methods. Furthermore, it addresses the primary challenges that researchers face in this rapidly evolving field and discusses possible future directions for research and development in speech-based large models.

收稿日期:2025-05-08;录用日期:2025-08-25;责任编辑:覃怀银

*通讯作者:李爱军

Key words: speech large language models; Large Language Models; instruction following; speech understanding; pre-training

Foundation Item(s): National Natural Science Foundation of China (No.62006212)

1 引言

ChatGPT 和 GPT-4 的问世标志着自然语言处理(Natural Language Processing,NLP)正式迈入大型语言模型时代. ChatGPT、Qwen、LLaMA、Mistral、Gemini 等一系列代表性大型语言模型展现出卓越的指令跟随能力与上下文学习能力,在文本生成、翻译、开放域问答、阅读理解等众多下游任务中取得突破性进展. 然而,人类智能的多模态属性表明,语言仅是人类认知系统的一个组成部分. 随着人工智能向通用智能迈进,语言、视觉与语音等模态间的融合成为关键趋势. 近年来,多模态大模型(Multimodal Large Language Models,MLLMs)研究热潮迅速蔓延,特别是在视觉领域已涌现出一批代表性成果,如 GPT-4V 和 LLaVa[1]展现出强大的视觉理解与跨模态推理能力. 与此同时,语音大模型(Speech Large Language Models,SLLMs)作为多模态智能的重要支柱,也在快速发展中[2]. 包括 Whisper[3]、Qwen-Audio 等在内的一系列模型,在语音识别、语音理解、语音合成等任务中不断突破性能边界,展现出显著的发展潜力(图 1).

图 1　语音大模型发展时间线

本文所讨论的语音大模型主要指代统一建模语音与语言的“端到端语音大模型”范式[4],不包括传统的声学-语言模型分离式自动语音识别(Automatic Speech Recognition,ASR)系统[5]. 相比模块化系统,端到端架构具备更强的语义建模、跨模态协同与泛化能力. 围绕该范式,本文将系统性综述语音大模型的研究现状,从架构设计、核心组件、训练策略、数据构建方法、评估方式五个方面展开分析,进一步梳理当前的主要挑战,并重点指出未来发展的三个核心方向:(1)语音理解(如情感识别、语音问答);(2)语音生成(如个性化语音合成、多语言说话人保留);(3)多模态语音推理(如语音-图文联合任务、语音条件生成).

2 语音大模型的定义与发展范式

2.1 语音大模型的发展范式

从发展范式来看,语音大模型大致经历了两个阶段,反映了从传统串联式设计向端到端深度融合的演进过程,其架构演化如图 2 所示:

第一阶段:专家模型协同范式[图 2(a)].

该阶段的模型设计延续了传统语音处理的流水线

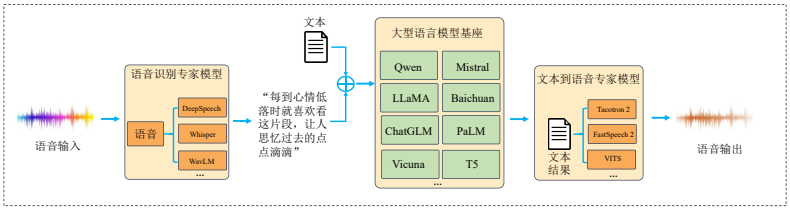
模式，即："语音输入→自动语音识别(ASR)→文本→LLM推理→文本→TTS合成→语音输出"的模式.

在这种范式中，语音识别模块(如Whisper、DeepSpeech[6])首先将语音信号转录为文本；随后文本输入大型语言模型(如Qwen、LLaMA、ChatGLM等)完成语义理解；最后由语音合成模块(如Tacotron2[7])将输出文本转化为语音. 该范式便于模块独立优化，但存在误差累积与副语言信息缺失等问题.

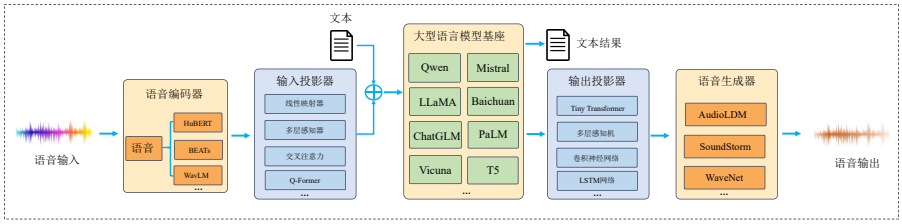
第二阶段：端到端语音大模型范式[图2(*b*)].

为解决第一阶段的局限，研究者提出了统一建模语音输入到文本/语音输出的端到端范式. 该范式通过引入语音编码器、输入投影器与LLM基座模型，使语音信号能够直接进入语言模型进行联合建模. 输出由投影器和语音生成器进一步解码成目标文本或语音[8].

代表性模型包括：(1)Whisper(OpenAI)：多语言语音识别与转录系统，采用端到端训练；(2)Qwen-Audio(阿里巴巴达摩院)：基于Qwen系列的多模态语音大模型；(3)VALL-E(微软)[9]：支持个性化说话人保留的语音生成大模型. 图2展示了这两个阶段的典型架构，其中图2(*a*)为串联式的专家模型协同范式，图2(*b*)为统一的端到端语音大模型范式. 可以看出，从"语音—文本—语音"的阶段式流程逐步演化为"语音→token→LLM→语音"的闭环建模，反映了语音大模型从功能拼接到能力融合的发展趋势[10].



(*a*) 第一阶段：专家模型协同范式



(*b*) 第二阶段：端到端语音大模型范式

图2 语音大模型的两阶段架构演化

2.2 语音大模型定义

在本文中，语音大模型主要指以自回归结构为基础、具备语音输入与输出能力的大型模型(传统判别式结构将不再纳入讨论范围). 它能够处理并生成语音序列，利用上下文进行动态理解，支持多种模态转换任务：如语音输入-文本输出(Speech2Text)、文本输入-语音输出(Text2Speech)、语音输入-语音输出(Speech2Speech).

语音大模型旨在构建一个统一的框架，实现对语音、文本及其两者交织数据的端到端处理与生成. 该框架不仅支持单一模态的处理，更重要的是能够实现跨模态的无缝转换与融合建模. 具体而言，一个语音波形 $A=\left(a_1,a_2,\cdots,a_Q\right)$ 是由长度为 Q 的语音特征组成的序列. 类似地，一段文本 $T=\left(t_1,t_2,\cdots,t_K\right)$ 是由长度为 K 的文本词元(如单词，子词或字符等)组成的序列. 我们将输入和输出多模态序列表示为 $M_{\text{in}},M_{\text{out}}=\left(M_1,M_2,\cdots,M_N\right)$，其中每个元素 $M_i\in\left\{a_i,t_j\right\}$，$\theta$ 为模型参数. 因此，语音大模型可以被表示为以下形式：

$$M_{\text{out}}=\text{SLM}\left(M_{\text{in}};\theta\right) \tag{1}$$

这一定义概括了语音大模型的核心任务，如图3所示[3,6,9,11~30].

3 主流架构

本节系统梳理了当前主流语音大模型的架构设计，并在综合多个典型模型公开文献与开源系统基础上，抽象归纳出一种广泛采用的"五大模块"架构范式包括：语音编码器(Speech Encoder)、输入投影器(Input Projector)、LLM基座模型(LLM Backbone)、输出投影器(Output Projector)和语音生成器(Speech Generator)[31,32]. 尽管当前尚无统一结构标准，但该模块化思路在多个代表性SLLMs(如VALL-E[9]、SpeechGPT[28]、SeamlessM4T)中均有体现，体现出"听—理解—说"流程的有效分工. 表1系统对比了典型模型的核心部件与本文抽象架构的对应关系[9,28,33,34].

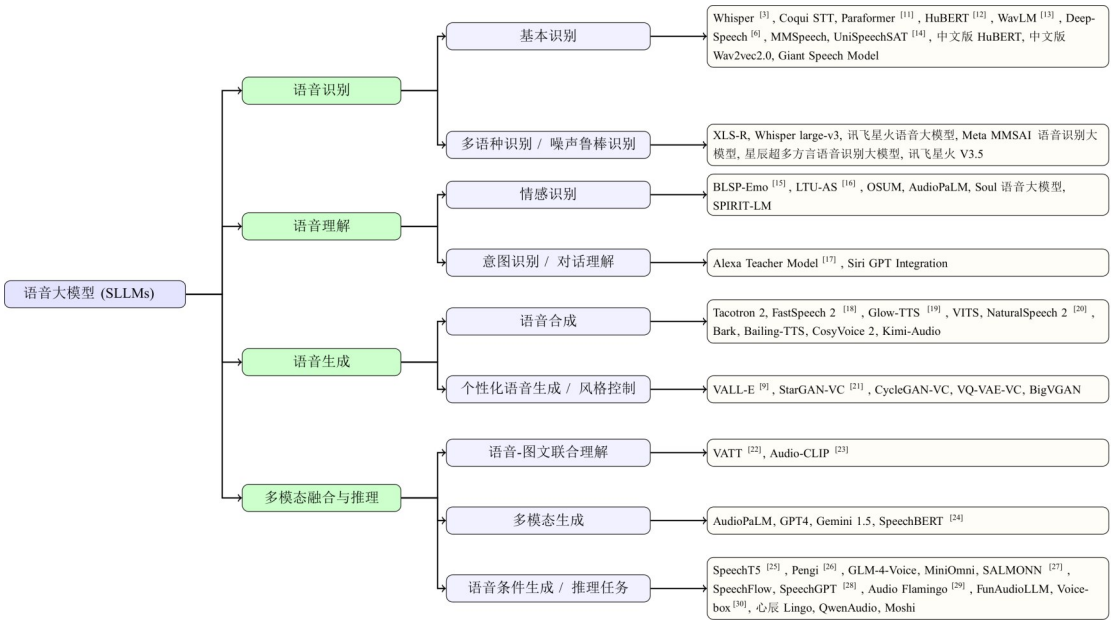


图3　语音大模型 (SLLMs) 分类体系

表1　典型语音大模型与核心模块对应关系

模型	语音编码器	输入投影器	LLM 基座模型	输出投影器	语音生成器
VALL-E[9]	EnCodec	Discrete token 嵌入	GPT-like Transformer	Token解码器	自回归vocoder
SpeechGPT[28]	Whisper编码器	Q-Former	LLaMA	线性或token映射	HiFi-GAN[33]
SeamlessM4T	wav2vec 2.0	Linear/adaptor layer	多语种Transformer LLM	线性投影	HiFi-GAN
VoxtLM[34]	EnCodec	Discrete token embedding	Decoder-only Transformer	Token projector	自回归vocoder

3.1　架构合理性与缺陷分析

3.1.1　合理性

近年来,主流语音大模型普遍采用“语音编码器—输入投影器—LLM基座—输出投影器—语音生成器”的分层模块化设计,这一范式已成为业界共识,被多项代表性工作反复验证[35]. 此架构有效模拟了人类“听觉—认知—表达”的感知流程:首先,语音编码器完成信号的声学及副语言特征提取;其次,输入投影器实现模态映射,使语音表征能够适配预训练语言模型输入空间;随后,LLM主干作为语义建模与推理的核心,承担复杂上下文理解与决策生成任务;最后,通过输出投影器与语音生成器将抽象语义转化为自然语音信号,实现完整的“听—理解—说”闭环[36].

以VALL-E为例,该模型首次将离散语音编码EnCodec与GPT结构结合,引入Discrete token作为语音表示,大幅简化了以往端到端语音合成系统的建模流程. 在架构设计上,VALL-E沿用“五大模块”模式,其中语音编码器输出离散词元(token),使得输入/输出投影器仅需完成Token ID与嵌入空间的映射,从而有效复用原生文本LLM模型结构.

3.1.2　局限性

尽管该架构在工程上结构清晰,但仍存在以下几个核心问题:

(1)线性流程割裂模态交互:模型结构通常采用“编码→投影→语言建模→解码”的线性处理流程,如LLaVA、Bark等模型. 这种方式虽然利于模块化实现,但容易割裂语音与文本之间的深层交互,难以模拟人类感知中并行整合的机制.

(2)对齐过于依赖浅层映射:输入投影器与输出投影器多采用线性映射或轻量多层感知器(MultiLayer Perceptron,MLP)[37]实现,缺乏深层对齐机制. 这会导致模型更关注模态间的表层特征匹配,而非真正理解模态间语义与概念的关联.

(3)缺乏认知机制建模:最新研究显示,当前语音大模型主要依赖大规模数据拟合,尚未在知识注入、情境推理、因果建模等结构化认知机制方面取得实质性突破[38,39].

3.2　语音编码器

语音编码器是语音大模型的核心前端模块,主要功能是将原始语音信号(波形)映射为紧凑而信息丰富的表征形式,为后续模态对齐与语言理解提供输入. 根据输出形式不同,当前编码器主要分为两类:基于连续嵌入的编码器与基于离散token的编码器.

(1)嵌入与词元的区分:嵌入(Embedding),将音频

的连续特征投射到低维向量空间中，使相似的语音片段在向量空间中邻近. 模型如HuBERT[12]、WavLM等输出的就是嵌入表示，适合情感、语义识别任务. 词元(Token)，通过量化等方式将连续语音离散化为token序列. 这些token并不对应语音中的音素，而是模型学习得到的抽象结构单元，适合生成任务，如Whisper、VALL-E等模型.

从认知角度看，语音编码器模拟了人类听觉系统的早期声波处理，将复杂声学信号压缩为便于神经系统理解的中间表示. 其公式表达如下：

$$F_x = \text{Encoder}(A_x) \tag{2}$$

其中，A_x表示输入音频波形；F_x表示输出的嵌入或token序列.

(2)主流编码器方法对比：表2[40~45]列出了当前常用的神经语音编码器及其特性. 这些编码器广泛应用于识别、生成和压缩等语音任务，并成为语音大模型的关键前端模块.

表2 常用的语音编码器

编码器	输出类型	特点	应用/代表模型
SoundStream[40]	嵌入	多层RVQ + GAN,低比特率高保真	Google语音压缩
EnCodec	嵌入	多级量化,兼容多语种	Meta Voicebox、SeamlessM4T
HiFi-Codec	嵌入	主观音质优,低延迟	腾讯合成引擎
Lyra[41]	嵌入	超低码率,高推理效率	移动端语音通信
VQ-VAE[42]	词元	首个提出离散音频token化	模拟人类认知离散表示
VQ-VAE-2[43]	词元	多层级嵌套token学习	更好建模音频结构
BEATs[44]	嵌入/词元	自监督预训练加多任务建模,支持理解任务	多语种语音理解任务
EAT[45]	嵌入	融合文本与音频的端到端Transformer建模	跨任务音频理解与生成

这些编码器(如EnCodec、SoundStream)预训练于LibriSpeech、VCTK等数据集，可依据下游任务需求选择性替换或与其他模块组合使用. 下面将按其技术发展脉络与编码方式的演进具体介绍这些语音编码器(Codec).

(*a*)SoundStream[40]：由Google提出，属于较早期的嵌入式编码器方案. 其核心结构为多层残差量化结构(Residual Vector Quantization，RVQ)[46]，并结合生成对抗网络(Generative Adversarial Networks，GAN)[47]进行训练，以在低比特率下实现高保真的语音重建. 该方法广泛应用于语音通信与音频压缩等场景.

(*b*)EnCodec：Meta开发的多级量化编码器，作为Voicebox[30]与SeamlessM4T等多模态语音模型的标准前端模块. EnCodec在继承SoundStream残差量化结构的基础上引入更细粒度的层次化压缩机制，显著提升了语音还原质量和多语种兼容性.

(*c*)HiFi-Codec：由腾讯推出，针对实时语音合成任务设计，采用多级RVQ结构结合HiFi-GAN[33]增强模块，兼顾低延迟与主观音质，在对话系统与语音合成场景中具备良好表现.

(*d*)Lyra[41]：Google面向移动端语音通信推出的超低码率编解码器. 该方法在压缩比与推理效率间实现了平衡，适用于资源受限设备的实时语音传输.

(*e*)VQ-VAE(Vector Quantized Variational Autoencoder)：由Google提出，首次将矢量量化机制引入语音编码，具备将连续潜在空间映射为离散代码本的能力，从而实现结构化语音表征. 该模型为语音生成类任务(如TTS与风格迁移)提供了基础性编码方式.

(*f*)VQ-VAE-2：DeepMind在VQ-VAE基础上引入多层级嵌套结构，提升了token建模能力与上下文捕捉范围，适用于复杂语音结构的学习与建模.

(3)BEATs与EAT结构解析：融合结构创新的自监督编码新范式.

BEATs(Bidirectional Encoder representation from Audio Transformers)由Meta提出，采用CNN前端+Transformer主干，通过预测掩码音频单元进行自监督预训练[44]. 其设计借鉴BEiT的离散化token建模，在SUPERB基准上取得领先性能，广泛适用于ASR、Speaker ID、Emotion等任务，已成为语音表征学习的重要模型.

EAT(End-to-end Audio Transformer)基于纯Transformer架构，取消传统卷积前端，直接对音频波形进行建模，结合掩码建模与多任务学习实现语音-文本联合训练[45]. 其在长语音依赖捕捉和跨模态对齐上具备优势，既可作为通用编码器独立应用，也能嵌入端到端语音理解系统.

模块总结与选型建议：从任务适配角度来看，嵌入式编码器更适合语音识别与理解等任务，其输出为连续向量，便于下游语言模型进行上下文建模. 如HuBERT与WavLM广泛用于情感识别与语义理解场景，EnCodec也被集成于多模态系统Voicebox和SeamlessM4T中作为特征前端，体现其良好的兼容性与可迁移能力.

3.3 输入投影器

大型语言模型(LLM)以文本模态为核心，通过离散符号序列建模语言结构与语义. 然而，语音是一种连

续、高维的信号,其编码方式与文本存在显著差异,主要体现在:(1)表示形式上,文本由离散词元构成,而语音需经编码器转化为连续特征向量;(2)时间结构上,文本为线性序列,而语音包含韵律、节奏、停顿等多层次特征. 由于LLM缺乏对原始语音特征的建模能力,直接输入语音特征往往效果有限,因此必须引入桥接机制,实现语音与语言模态的对齐与融合.

一种常见且有效的解决方案是在预训练语音编码器与LLM之间引入一个输入投影器(Input Projector),该模块亦被称为"模态对齐器""中间连接器"或"语音-语言接口". 该设计思路已在SpeechPrompt[48]和SALM[49]等典型工作中得到应用. 其核心目标是学习一个映射函数,将语音编码器输出的特征表示F_x投影至文本特征空间T,其中$P_x \in T$,从而生成语言模型能够理解和处理的提示向量序列:

$$P_x = \text{Project}(F_x) \tag{3}$$

其中,F_x表示语音编码器提取出的连续语音特征;Projector(·)表示输入投影器的映射函数,可以是线性层、非线性神经网络或更复杂的深度模块;P_x表示对齐后的文本模态特征,通常作为prompt(提示)传递给LLM.

输入投影器的实现方式多种多样,按模型能力的演进顺序,常见的输入投影器方式包括:(1)线性投影器(Linear Projector)[50],通过单层线性变换将语音特征映射到文本特征空间;(2)多层感知器(MLP)[37],使用多层非线性变换实现更复杂的特征对齐;(3)交叉注意力(Cross-attention)[51],利用一组可学习的查询向量与语音特征进行交互,从而提取关键特征表示;(4)Q-Former[52],基于查询向量机制,选择性地提取任务相关的语音特征作为提示;(5)MQ-Former[53]:通过多尺度特征提取与对齐机制,实现对不同时序粒度语音信号的建模. 根据近年来语音大模型的设计趋势,常见的投影器方式如表3所示[37,50~53],相比于直接拼接语音-文本输入,使用投影器更易训练、更灵活,特别适用于少样本任务或低资源语言场景.

表3　常用的输入投影器方式

方式	描述	优点	缺点
线性投影器[50]	单层线性映射	计算效率高,训练简单	仅适用于线性特征对齐,表达能力弱
多层感知器[37]	多层非线性映射	可建模复杂关系	训练成本高,易过拟合
交叉注意力[51]	查询向量与语音特征交互对齐	适应性强,选择性强	参数多,训练资源大
Q-Former[52]	选择性提取任务相关特征	去除冗余,提升下游性能	泛化性偏弱
MQ-Former[53]	多尺度交叉对齐机制	支持不同时间粒度建模	计算复杂,部署难度大

3.4　LLM基座模型

在语音大模型架构中,预训练大型语言模型(LLM)作为"语义理解核心",负责统一建模语音投影特征与文本上下文,承担语义理解、任务决策与语言生成等关键功能. 其作用类似人类在语音理解过程中调动语言知识与推理结构,是支撑语音生成质量和语义合理性的核心模块.

相较于从零训练语言模型,当前主流做法是直接复用已有LLMs(如LLaMA、Qwen、ChatGLM等)作为基座,并通过输入投影器实现语音模态与语言模态之间的桥接. 这种方式能够显著降低训练成本,同时保持高质量的语言建模能力与多任务泛化性能.

通常,语音特征由编码器抽取为token表示$\boldsymbol{F}_x$,再通过输入投影器映射为语言模型可接收的提示向量序列$\boldsymbol{P}_x$,最终由LLM生成输出文本:

$$t, S_x = \text{LLM}(\boldsymbol{P}_x, \boldsymbol{F}_T) \tag{4}$$

其中,t为模型生成的语言输出,决定了下游语音生成的语义与风格S_x表示语音信号词元(speech tokens),用于指导后续语音生成器的语音生成. 当前主流的LLM模型构成语音大模型的"认知中枢",表4对其结构、能力与适用范围进行总结[54,55]. 以LLaMA、Qwen等主流LLM为例,式(4)实际对应的计算流程如下:

(1)对于语音理解与生成任务,输入投影器将语音编码器输出的特征F_x映射为LLM可识别的提示嵌入$\boldsymbol{P}_x$(例如,VALL-E[9]中F_x为EnCodec token,$\boldsymbol{P}_x$为线性映射后的d维向量,与LLaMA词向量维度一致);

(2)LLM主干(如LLaMA、Qwen)将$\boldsymbol{P}_x$与可选的文本输入(如$\boldsymbol{F}_T$,即文本prompt)拼接组成完整输入序列,并通过自回归机制生成目标文本t或语音token序列S_x;

(3)具体而言,LLaMA结构支持输入任意长度的嵌入序列,因此LLM($\boldsymbol{P}_x$,$\boldsymbol{F}_T$)实为"[语音提示向量|文本提示向量]"的拼接输入,输出为token级的文本或语音表征,适配下游的语音生成或TTS模块.

现有主流大型语言模型如LLaMA,Qwen,DeepSeek等,均具备强大的文本生成与上下文建模能力,成为构建语音大模型的基础组件.

以下简要介绍几种代表性模型:

(1)LLaMA是Meta开发的自回归语言模型系列,包含多个规模变体(7 B至400 B参数)以及多次更新(例如Llama/Llama 2/Llama 3/Llama 3.1/Llama 4). 该模型采用标准Transformer架构,通过高效的预训练策略在模型性能和计算开销间实现平衡. 其显著优势在于

表4 常用开源语言大模型主要参数与架构对比

模型	发布日期	预训练数据规模	参数量/B	支持语言	架构
Qwen	Sep-2023	3 T tokens	1.8/7/14/72	en,zh	Causal Decoder
Mistral	Oct-2023	0.25 T tokens	7/16	en	Causal Decoder
LLaMA	Feb-2023	3 T tokens	7/13/70	en,zh,fr	Causal Decoder
Baichuan	Apr-2023	0.1 T tokens	7/13/53	zh,en	Causal Decoder
ChatGLM	Jun-2023	1 T tokens	6/130	en,zh	Encoder-Decoder
PaLM[54]	Apr-2022	0.78 T tokens	8/62/540	en	Causal Decoder
T5[55]	Oct-2022	16 T tokens	3/11	en,fr,de	Encoder-Decoder

较小参数量下仍保持强大的性能,具备优秀的迁移学习能力. 然而,在复杂推理任务上的表现相较GPT系列仍有提升空间.

(2)Qwen最新版本迭代到Qwen 2.5,基于改进的Transformer架构,特别优化了多语言处理能力. 该模型在中文及多语言场景下表现突出,支持灵活的任务扩展. 但在非中文任务特别是低资源语言场景下的效果仍需改进.

(3)ChatGLM是清华大学与智谱AI联合开发的对话式语言模型,采用双向注意力机制增强的Transformer架构. 模型在中文对话生成与理解方面表现优异,且具有较好的部署效率. 其局限性主要体现在非中文处理能力较弱.

3.5 输出投影器

输出投影器的主要作用是将语言模型生成的抽象语义表示(通常为词元序列)转换为语音生成器可接收的连续特征表示,从而完成从“语言理解”到“语音合成”的模态过渡. 在语音大模型中,LLM输出的token通常缺乏直接的模态信息(如韵律、音色、语速等),因此需要通过投影器完成语义到语音特征的对齐映射:

$$H_x = \mathrm{Project}(S_x) \tag{5}$$

其中,S_x为LLM输出的语义token序列;H_x为输出特征,作为语音生成器的输入. 这一模块的设计确保了语言模型生成的内容能够被转化为目标模态(如语音)的具体表现形式.

当前输出投影器的主流设计方式包括以下4类.

(1)小型自注意力网络(Tiny Transformer)[56]:基于轻量化Transformer架构,能够建模token之间的全局依赖,适用于多轮对话、情感合成等上下文敏感任务.

(2)卷积投影器(Convolutional Projection)[57]:利用一维卷积结构建模token序列中的局部时间结构,具备高效、低延迟的优点.

(3)多层感知器:最简单的结构,仅由若干全连接层组成,适合低资源、移动端推理等对复杂性要求不高的场景.

(4)条件生成网络[58]:将输出投影器与语音生成器联合训练,以重构目标语音模态特征为优化目标,提升上下文一致性与情感传递能力,以最小化目标模态特征之间的差异为目标进行训练:

$$\min \| H_x - H_x^* \|_2^2 \tag{6}$$

其中,H_x^*为语音生成器期望的特征表示. 该方式生成质量高,但训练资源需求大,适合高质量语音合成系统(如VALL-E). 具体如表5所示[37,56~58].

表5 输出投影器设计对比与适用性分析

方式	优点	局限	典型应用
小型自注意力网络[56]	能建模上下文,参数少	相比CNN计算复杂	多轮对话语音生成
卷积投影器[57]	局部结构强,效率高	全局依赖弱	实时语音生成与流媒体系统
多层感知器[37]	实现简单,速度快	表达能力弱,缺乏序列建模能力	简易语音任务、移动端推理
条件生成网络[58]	高生成质量,上下文敏感	资源消耗大,训练复杂	高保真语音合成系统(如VALL-E)

由表5可知在多个代表性语音生成系统中,Tiny Transformer被广泛用于输出投影模块,其结构通过压缩版Transformer Block实现上下文建模,同时避免了全尺寸模型带来的资源开销. 例如在SpeechGPT框架中,Tiny Transformer结构作为输出桥梁,有效提升了语义到语音的生成一致性与节奏感. 该方法兼具表达能力与效率,适合部署于具备对话、多语义合成能力的应用场景. 然而,该结构的全局建模能力相较于完整Transformer仍有限,未来研究可考虑引入局部感知增强机制、与模态对齐模块(如Cross-Attention)联合设计,以进一步提升其适应性.

3.6 语音生成器

语音生成器是语音大模型的最后一个核心组件,负责将输出投影器生成的连续特征表示$\boldsymbol{H}_x$(如token序列)转换为可感知的语音波形或频谱. 完成从语义理解到信号合成的模态转换. 与人类语言表达中“由思维到发声”的过程相似,该模块对语音自然度、节奏连贯性、情绪表达与风格控制起到决定性作用. 语音生成本质上是一个高维连续、时序强约束的生成问题.

当前主流语音生成方法大致沿着以下三类路径演进:自回归方法(Autoregressive,AR)、非自回归方法(Non-Autoregressive,NAR)与扩散建模方法(Diffusion Models). 不同方法在建模机制、合成效率、语音质量与控制能力等方面各具优势,表6提供了典型代表模型的对比概览[18,33,59,60].

近年来,语音生成器的研究持续演进,其核心目标

是在保障语音自然度与表达力的基础上,进一步提升生成效率与控制能力. 早期代表模型WaveNet基于自回归架构,通过逐帧建模实现高精度语音生成,能有效捕捉语音序列中的细粒度时序依赖,表现出极高的自然度. 然而,其串行推理机制导致生成速度缓慢、计算开销较大,难以满足实时语音合成的需求. 为缓解这一瓶颈,Tacotron 2提出两阶段端到端语音生成方案,先将文本编码为Mel频谱,再通过声码器(如MelGAN[59]或HiFi-GAN)还原为语音波形,在自然度和通用性之间实现较好平衡,成为现代语音合成系统的重要基础结构.

为提升推理效率,非自回归(NAR)方法逐渐成为研究主流. FastSpeech 2[18]在Tacotron系列基础上引入持续时长、音高与能量控制器,采用全并行解码方式显著加速生成过程,同时增强了对语速与风格的建模能力. 在声码器设计方面,HiFi-GAN[33]凭借多尺度判别器与残差生成模块,在保证实时性的同时显著提升了音频质量,已成为构建高质量语音合成系统的重要组件(例如FastSpeech 2+HiFi-GAN的组合架构).

表6　常用语音生成器方法及适用性对比

方法	类别	描述	特点	应用场景
WaveNet	自回归(AR)	自回归卷积生成器	自然度高但速度慢	高质量语音、对话克隆
Tacotron 2	自回归(AR)	频谱-声码器两阶段TTS	输出自然,可配多声码器	通用TTS、语音助手
MelGAN[59]	非自回归(NAR)	基于生成对抗网络(GAN)的声码器模型	生成速度快,适合流媒体	流媒体合成、语音增强
HiFi-GAN[33]	非自回归(NAR)	高质量声码器,端到端支持	质量高,速度快	高质量TTS、实时语音合成
FastSpeech 2[18]	非自回归(NAR)	非自回归的TTS模型	控制力强,生成效率高	实时语音合成、风格迁移
SoundStorm	非自回归(NAR)	并行非自回归语音生成器	极快速度,高自然度	实时TTS、语音生成系统
DiffWave	扩散模型(Diffusion)	扩散声码器	音质高,推理快	高保真语音合成、波形生成
AudioLDM[60]	扩散模型(Diffusion)	文本到音频生成器	训练高效,生成质量高	文本转音频,音频风格迁移

4　训练策略

语音大模型的训练策略主要针对其性能与泛化能力. 通常,训练过程包括三个主要阶段:预训练(Pre-training)、指令微调(Instruction Tuning)和对齐调优(Alignment Tuning)[61]. 以下将详细介绍每个阶段的策略与实施方法.

4.1　预训练策略

预训练是构建语音大模型语义建模能力的第一步,旨在通过大规模语音-文本数据学习模态内表示与模态间对齐关系,为后续任务微调奠定基础. 根据目标任务的差异,预训练策略通常分为三类:从头预训练、继续预训练和多阶段预训练.

从头预训练(Pre-training from Scratch):该策略指在模型参数随机初始化的基础上,直接使用大规模语音-文本数据进行训练,从零构建语义建模能力[62]. 其优势在于结构自由度高,具备较强的知识融合与泛化潜力,适合构建通用的多模态基础模型. 典型代表如OpenAI的Whisper[3]和Google的AudioLM[63].

继续预训练(Continued Pre-training):继续预训练基于已有的预训练语言模型(如LLaMA、Qwen等),在特定领域或多模态数据集上进一步微调以适配任务. 该方法显著提高了多任务性能与零样本泛化能力,尤其适合构建跨语言迁移系统或小样本微调系统. 然而,该策略的性能依赖于语言模型主干的预训练质量,若原始语言模型存在偏差或模态隔阂严重,将影响语音模态的融合质量.

多阶段预训练(Multi-stage Pre-training):该策略以逐步训练的形式构建语音-语言一体化能力. 通常包括:第一阶段在文本上进行语言建模,第二阶段通过语音自监督学习提取模态表示,第三阶段进行跨模态对齐训练. 此策略强调"分阶段建模+跨模态融合",可在控制训练资源消耗的同时逐步提升语音语言协同能力.

4.1.1　训练目标

语音大模型的预训练阶段旨在为模型建立跨模态理解与生成能力,主要围绕以下三大目标展开:跨模态对齐(Cross-modal Alignment)、模态内表征学习(Intra-modal Representation Learning)与多模态融合(Multi-modal Fusion). 三者从不同层面分别解决语音模态"连续性强、结构弱、冗余多"的表征难题,是支撑语音大模型理解、生成、翻译等任务的关键训练支柱.

(1)跨模态对齐:跨模态对齐是多模态语音模型预训练中的核心目标,其本质是建立语音模态与文本模态之间的语义对应关系,确保不同模态的信息可以在共享空间中进行连通与映射.

例如,在自动语音识别(ASR)中,目标是学习从语音序列 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_T\}$ 预测文本序列 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_L\}$,通常建模条件概率 $P(\boldsymbol{y}|\boldsymbol{x})$ 并优化其最大化:

$$\max_{\theta} P(\boldsymbol{y}|\boldsymbol{x};\theta) \tag{7}$$

在文本到语音合成(TTS)中,则反向建模,从文本序列 $\boldsymbol{y}$ 生成语音帧序列 $\boldsymbol{x}$,对应条件概率 $P(\boldsymbol{x}|\boldsymbol{y})$,并优化

其最大化：

$$\max_{\theta} P(\boldsymbol{x}\mid \boldsymbol{y};\theta) \tag{8}$$

近年来更为主流且有效的做法是采用对比学习(Contrastive Learning，CL)，建立语音与文本模态在统一语义空间中的隐式对齐机制[64]. 其核心思想是：(*a*)将配对的语音-文本样本作为“正例”拉近表示；(*b*)将未配对样本作为“负例”进行区分，对表示空间施加对比约束，对比损失定义为

$$\mathcal{L}_{\mathrm{CL}}=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\exp\left(\mathrm{sim}(v_i^{\mathrm{speech}},v_i^{\mathrm{text}})/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}(v_i^{\mathrm{speech}},v_j^{\mathrm{text}})/\tau\right)} \tag{9}$$

其中，$\boldsymbol{v}_i$分别为第i对语音与文本的向量表示；sim(·)是余弦相似度函数；τ是温度系数；N是batch大小. 理论意义上，对比学习提供了一种无监督框架，可广泛适用于多语种、多场景下的预训练.

(2)模态内表征学习：该阶段旨在分别优化语音与文本模态的内部建模能力，提升单模态表示质量. 语音模态侧重于学习音高、音色、说话人特征和韵律等属性，典型模型如HuBERT[12]、WavLM[13]，通过自监督遮蔽预测挖掘语音序列模式. 文本模态则多采用掩码语言建模任务，捕捉上下文依赖与语义关系. SpeechT5[25]等模型通过独立优化语音与文本编码器，在增强单模态建模的同时，为后续多模态任务奠定基础.

(3)多模态融合：该目标旨在实现语音与文本信息的深度融合，构建统一的跨模态表示空间，从而支持语音情感识别、语音翻译、多轮人机对话等复杂跨模态任务. 现有研究中常采用的关键技术包括：跨模态注意力机制(Cross-modal Attention)[65]、联合编码器(Joint Encoder)结构[66]、联合对比学习范式，以及对齐-融合并联训练策略等.

4.1.2 训练细节

为了在预训练阶段高效利用已有的模型能力，语音大模型通常采用参数冻结+轻量优化策略，并辅以多任务损失函数以实现模态内学习、模态对齐与生成优化的多目标协同训练. 这种方法不仅能保留大型语言模型(LLM)和语音编码器中已有的语言/声学知识，还能显著降低训练所需的计算资源.

(1)参数冻结策略：当前主流语音大模型(如Whisper[3]、VALL-E[9]、SpeechT5[25])广泛采用“冻结主干、训练接口”的策略，即冻结LLM与语音编码器权重，仅对输入投影器、输出投影器与对齐模块等轻量结构进行更新.

同时，为提升接口训练的表达能力，常采用如下微调方法. LoRA(Low-Rank Adaptation)：通过低秩矩阵调整主干权重方向，保持参数冻结同时注入新知识[67]. Adapter Tuning：在主模型层间插入小型瓶颈模块，仅更新Adapter参数. Prefix Tuning：对Transformer输入添加可学习的提示向量，引导任务特化[68].

(2)损失函数设计：语音大模型往往涉及多个任务目标，如序列生成、上下文理解、模态对齐等. 为实现这些目标，预训练阶段通常采用多种损失函数的联合优化，具体如表7所示[42,64,69,70].

表7 常见的损失函数

损失类型	适用目标	示例模型	描述
交叉熵损失(CE)[69]	自回归语言建模、ASR	Whisper,VALL-E	用于优化token预测准确性
掩码语言模型(MLM)[70]	模态内语言建模	SpeechT5	屏蔽部分输入,训练上下文恢复能力
对比学习(CL)[64]	跨模态对齐	Kimi Audio,AudioCLIP	强化配对语音-文本间相似度
VAE重建损失[42]	潜变量建模、说话人建模	Voicebox,SpeechT5	生成语音时提升多样性与连贯性

4.1.3 预训练数据

在语音大模型中，预训练数据的多样性、语义层级与模态匹配度是支撑语义理解、语音生成与跨模态对齐的核心基础. 不同类型的数据对应不同的训练目标，因此合理的数据分类与任务映射至关重要. 根据标注粒度与模态结构，常用数据可分为三类：语音-文本配对数据、多模态音频数据和未标注语音数据(表8~表10[71~75]). 其中，若数据集既含文本标签又具备多模态属性(如CoVoST2[74])，则按主要用途归入“多模态”类；若数据集同时支持识别与其他任务(如Uwspeech[76])，则归入对齐数据类.

(1)语音-文本配对数据. 这是最基础的数据类型，用于训练语音识别与转录能力. 代表性数据集包括LibriSpeech[71]和Common Voice[72]. 其中Common Voice是最大规模的多语种开源语音集，覆盖70余种语言，并提供性别、年龄、地区等元信息，有助于模型学习语音-语言-说话人风格的映射关系，从而提升多语种迁移与个性化合成能力(图4).

(2)多模态数据集. 这类数据不一定具备精确对齐标注，但通常提供更丰富的上下文、标签信息(如情感、场景、音效类别等)，适用于模型语义理解与多模态任务迁移能力的训练. 代表数据集包括：AudioSet[77]，YouTube音频事件数据集，标签覆盖环境音、语言、音乐等；VoxCeleb2[75]，用于说话人识别，音频含强身份标签. 这些数据增强了模型对语音情绪、身份、场景等非语言特征的理解能力.

(3)未标注语音数据. 近年来，自监督学习的发展使得大规模未标注语音数据成为模型预训练的重要资

图4　数据示例

表8　语音-文本配对数据(用于语音识别、转录、TTS等任务)

数据集名	规模/h	年份	简述
LibriSpeech[71]	1 000	2015	基于LibriVox有声读物的英语朗读语音识别数据集
Common Voice[72]	4 200+	2020	Mozilla众包多语言语音数据集,包含丰富的元信息
GigaSpeech	10 000	2021	多领域英语语音识别数据集,包含高质量转录音频
WenetSpeech[73]	10 000	2021	多领域中文语音识别数据集,包含强标签和弱标签数据
LibriTTS	585	2019	LibriSpeech的多说话人版本,适用于语音合成任务
Fisher	2 000	2004	英语对话语音数据集,广泛用于语音识别研究

表9　多模态音频数据(用于语音翻译、情感识别、跨模态理解等任务)

数据集名	规模/h	年份	简述
Multilingual TEDx	1 000	2018	多语言TED演讲录音数据集,包含文本转录和语音
CoVoST2[74]	2 800	2020	多语言语音翻译数据集,包含语音和文本对
CVSS	1 900	2022	基于Common Voice构建的语音分割数据集
MUST-C	1 400	2020	多语言电影和电视节目语音翻译数据集
VoxCeleb2[75]	2 000+	2018	包含人脸识别和语音数据,用于说话人识别和语音增强

表10　未标注语音数据(用于自监督学习、特征预训练等任务)

数据集名	规模/h	年份	简述
Libri-light	60 000	2019	包含约60 000 h未标注英语语音,用于自监督预训练
The People's Speech	30 000	2021	多样化的英语语音识别数据集,适用于商业用途
Voxblink2	16 000	2024	包含带说话人标注的多语言音频数据

源. 通过设计如遮蔽预测、对比学习、潜变量建模等任务,模型可在无标签语音中学到深层的结构特征. 代表数据集如表10所示.

一些典型模型,例如HuBERT[12]、VALL-E[13]、SeamlessM4T都大量采用此类数据进行预训练. 这类数据能显著提升模型的声学鲁棒性、多语言泛化能力与低资源任务适应性.

4.2　指令微调

指令微调(Instruction Tuning)是语音大模型训练中的关键环节,旨在通过引导模型学习"任务指令+输入→合理响应"的范式,从而提升其指令理解能力、任务泛化能力以及零样本(zero-shot)[78]和少样本(few-shot)执行能力. 与传统训练不同,指令微调强调"对话式"人机交互场景中的指令理解与跨模态响应建模.

4.2.1　数据格式与任务定义

指令微调的数据设计围绕"指令、输入、输出"三要素展开,主要分为以下两类:

(1)指令跟随格式. 指令跟随格式用于单轮任务建模,结构清晰,常见于以下场景:语音问答(Speech-in-Text-out)和指令生成语音(Text-in-Speech-out). 指令示例,"请根据以下语音回答问题:'这段语音中提到的主要人物是谁?'"输入是一段语音文件,输出为回答问题的文本,如图5所示.

(2)多轮对话格式. 多轮对话格式通常采用结构化的"轮次对话(Turns)"表示方式,每一轮由对话的发起方和对应输入/输出内容构成,如图6所示,具备以下要素:角色标注(Role Tags),标记发言主体,如human(用户)、gpt(模型)、audio(语音输入)等;输入模态,对应发

指令/Instruction: <instruction>
输入/Input: {<audio>, <text>}
输出/Response: <output>

图5 指令跟随格式示意图

言内容，可为文本、音频或其他模态数据；轮次序号（Turn Index），所有轮次组织为数组形式，保留上下文顺序，便于模型理解会话演进过程.

```
{
 "conversations": [
   {
    "from": "human",
    "value": <text>
   },
   {
    "from": "audio",
    "value": <audio>
   },
   {
    "from": "gpt",
    "value": <output>
   }
 ]
}
```

图6 多轮对话格式示意图

4.2.2 训练机制与损失设计

监督微调（Supervised Fine-Tuning，SFT）[79]是指令微调的核心，通过高质量指令数据训练模型响应能力. 在训练过程中，模型的输出是基于指令和输入的预测，假设模型接收指令I和输入模态M（语音或语音-文本对），则目标生成输出响应A可表示为

$$A=\mathrm{LLM}(I,M) \tag{10}$$

其中，训练的目标是通过最小化损失函数来优化模型的参数：

$$\mathcal{L}=-\sum_{i=1}^{N}\ln P(R_i|R_{<i},I,M) \tag{11}$$

其中，N是标响应的长度；R_i为第i个词元；$R_{<i}$为前缀序列. 该损失函数促使模型通过自回归方式生成符合指令的输出. 而对于多轮对话任务，模型不仅需要根据当前指令和输入，还需考虑对话历史C：

$$\mathcal{L}_{\mathrm{dialog}}=-\sum\ln P(R_i|R_{<i},I,M,C) \tag{12}$$

此类训练可用于强化模型在复杂对话中保持语境一致性与任务相关性.

4.3 对齐调优

对齐调优（Alignment Tuning）旨在使语音大模型的生成行为更贴合人类偏好，减少“幻觉”（hallucination）现象，即模型输出不准确、不相关或虚假的内容. 当前主流方法采用结合监督学习与基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）策略，以优化模型在流畅表达、语义契合、情感表现等维度的综合表现.

4.3.1 基于人类反馈的强化学习

RLHF是当前对齐调优的核心方法，流程分为三个阶段.

第1步：监督微调（Supervised Fine-Tuning）. 首先，利用对齐调优数据对模型进行监督微调，使其初步具备满足用户需求的能力. 在这一阶段，模型学会生成更符合预期的输出. 监督微调的目标是最小化模型生成的输出与高质量标注数据之间的差异. 这可以通过最小化交叉熵损失函数来实现：

$$L_{\mathrm{SFT}}=-\frac{1}{N}\sum_{i=1}^{N}\ln P_\theta(y_i|x_i) \tag{13}$$

其中，x_i是输入数据；y_i是目标输出；$P_\theta(y_i|x_i)$是模型根据参数θ生成y_i的概率. 此阶段的优化目标是调整模型参数θ，使得生成的输出更接近标注数据.

第2步：奖励建模训练（Reward Modeling）. 在监督微调的基础上，训练一个奖励模型，用于评价生成内容的优劣. 例如，奖励模型可以基于用户的偏好分数对生成内容进行打分，从而为后续强化学习提供目标函数. 损失函数为

$$L_{\mathrm{RM}}=-\frac{1}{N}\sum_{i=1}^{N}\left[R_\phi(y_i^+)-R_\phi(y_i^-)\right] \tag{14}$$

其中，y_i^+是用户偏好较高的生成内容；y_i^-是用户偏好较低的生成内容；R_ϕ是奖励模型，其参数为ϕ. 奖励模型通过区分y_i^+和y_i^-的得分来学习用户的偏好.

第3步：策略优化（PPO）. 使用强化学习算法[如近端策略优化（Proximal Policy Optimization，PPO）][80]优化模型的生成策略，使其能够在奖励模型的指导下，生成更符合偏好的内容. 通过引入KL惩罚项，可以防止优化过程偏离初始策略过远，从而保持生成内容的多样性和稳定性，优化目标函数为

$$L_{\mathrm{PPO}}=\mathbb{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}A_t\right]-\beta\cdot\mathrm{KL}\left[\pi_\theta\|\pi_{\theta_{\mathrm{old}}}\right] \tag{15}$$

其中，π_θ是当前策略；参数为θ；$\pi_{\theta_{\mathrm{old}}}$是初始策略；$A_t$是优势函数，用于评估动作$a_t$的质量，$\beta$是KL惩罚项的权重. 通过平衡奖励最大化和策略稳定性，模型能够生成更符合用户偏好的内容.

4.3.2 新兴对齐优化方法：DPO与KTO

为解决RLHF流程复杂、训练不稳定等问题，近年来出现了无需强化学习的新范式：直接偏好优化（Direct Preference Optimization，DPO）与卡尼曼-特沃斯基优化（Kahneman-Tversky Optimization，KTO）.

（1）DPO

DPO方法[81]省去奖励模型与策略优化过程，直接以偏好对（正例vs负例）为监督信号，引入对比损失函数优化语言模型参数. 其核心优势在于：结构简洁、训

练稳定、无需强化学习,适合大模型的大规模偏好对齐.

(2)KTO

KTO方法借鉴了行为经济学中的前景理论(Prospect Theory),KTO在偏好对之外引入“参考响应”作为评价基准,构建三元组(偏好响应、非偏好响应、参考响应),通过建模“相对于参考响应的增益/损失”来引导模型学习人类的真实偏好趋势.该方法在模拟复杂偏好行为方面表现出更强的拟合能力,并在存在评分噪声或主观分歧的场景下展现出更强的鲁棒性.

5　语音大模型的典型体系与比较分析

随着语音大模型的快速发展,其在语音识别、语音理解、语音合成、语音转换及跨模态任务等方面展现出显著性能提升和应用潜力.本节从任务体系的角度出发,梳理当前语音大模型的发展路径与代表模型,分析其关键技术特点、适用场景与存在问题,以期为后续研究提供系统性参考.

5.1　语音识别模型体系

当前主流语音识别大模型多采用Transformer结构或自监督预训练策略,在多语言、多场景任务中展现出强大的鲁棒性与泛化能力.代表模型包括Whisper[3]、Coqui STT、XLS-R、HuBERT[12]、WavLM[13]、Paraformer[11]、DeepSpeech[6]等.其中,Whisper基于端到端Transformer架构,具备多语言、长音频处理与强鲁棒性能力,是现阶段通用语音识别的代表.XLS-R和HuBERT强调自监督训练,通过对海量未标注语音的特征学习,降低对人工标注数据的依赖.Paraformer模型则在中文语音识别中表现突出,兼具建模效率与推理速度优势.相较而言,DeepSpeech虽在工程部署上优势明显,但其泛化能力和多语种适应性已显不足.SpeechGPT[28]、讯飞星火语音大模型、星辰超多方言语音识别大模型等国内大模型,也在细分应用上逐步追赶国际一流水平.

5.2　语音情感识别模型体系

在语音情感识别(SER)领域,模型体系经历了从依赖声学特征工程向基于深度大模型建构的演化路径.近年来,多模态驱动的情感大模型逐渐成为研究焦点,代表性模型包括BLSP-Emo[15]、LTU-AS[16]、OSUM及AudioPaLM、SPIRIT-LM、Soul语音大模型等.其中,BLSP-Emo构建了多尺度说话人嵌入与语义感知模块,提升了情感建模的上下文适应能力;LTU-AS与OSUM分别在表征稀疏情绪标签和跨文化情感迁移方面提出了新型结构设计,体现了情感识别任务对泛化性和可解释性的双重诉求.

5.3　语音合成(TTS)模型体系

语音合成技术在近年来取得显著突破,主流模型包括Tacotron 2、FastSpeech 2[18]、VITS、Glow-TTS[19]、Bark、BigVGAN[48]、NaturalSpeech 2[20]等.Tacotron系列采用端到端序列建模,实现了高自然度语音合成;FastSpeech系列则以非自回归方式大幅提升推理效率.VITS整合变分自编码器与GAN,兼顾音质与多样性.Bark、SpeechGPT等大模型通过大规模语音与文本联合预训练,实现了强泛化和多任务适配,能够生成多语言、多说话人、高表达力的语音.国内模型如UniSpeech-SAT和鹏城实验室多语言TTS系统已在多语种支持与音色复杂度方面取得突破[14].

5.4　语音转换(VC)模型体系

语音转换(VC)任务旨在在不改变语义内容的前提下,实现说话人、语音风格、情绪状态等语音属性的灵活转换,是实现个性化语音交互与语音编辑的重要技术.当前主流方法可分为两大类:一类是基于生成对抗网络(GAN)和循环一致性学习的非平行数据建模方法,如StarGAN-VC[21]、CycleGAN-VC等,强调多说话人建模与配对语料依赖的减弱;另一类是结合离散表示学习与大模型框架的最新方案,如VQ-VAE-VC、AudioPaLM、Salmon、Moshi、Qwen-Audio-VC、Voicebox[30]等,具备更强的语音属性建模与可控生成能力.

6　评估

语音大模型的评估是衡量其性能和实用性的重要环节.由于其涉及语音识别(ASR)、语音合成(TTS)、语音理解、多模态推理等多项任务,需从多个维度展开评估.评估方式主要包括:自动评估(Automatic Evaluation)和人工评估(Human Evaluation).本节从这两个维度详细介绍常用的评估方法、基准以及其核心细节.

6.1　自动评估

自动评估利用可复现、统一的指标体系,实现对语音大模型多任务能力的快速、客观评测.常见指标按任务类型分类如下:

在语音识别(ASR)任务中,常用指标包括词错误率(Word Error Rate, WER)、字错误率(Character Error Rate, CER)和句错误率(Sentence Error Rate, SER),分别衡量词、字和句级别的错误比例.

在语音合成(TTS)任务中,常用指标包括语音相似度(Similarity, SIM)、平均意见得分(Subjective MOS, SMOS)以及基于自动识别的TTS-WER.

在语音转换(VC)任务中,主要关注转换后语音的词错误率(VC-WER)、相似度(VC-SIM)和自然度(VC-SMOS).

在跨模态语音任务中,则依赖BLEU、准确率等指

标评估语音到文本、语音到语音及语音-图文对齐的效果. 下文通过表11~14系统对比当前主流语音大模型在上述任务中的性能表现.

在语音识别任务中(表11)[3,13,28],WavLM Large在Content数据集上取得3.44%的最低WER,体现其在自监督预训练上的优势;AudioPaLM-2在VoxPopuli上将WER降至9.8%,展现了语音-文本联合建模的有效性;USM-M与mSLAM等模型在多语种和低资源场景中表现突出,说明其具备较强的迁移能力.

在语音合成任务中(表12)[9,20,28,30,82],VALL-E 2在SIM(0.782)和SMOS(3.947)上达到最佳表现,接近真实语音的自然度与保真度;Voicebox与SoundStorm在低延迟生成和风格控制方面优势明显,适合实时交互应用;SpeechGPT兼具TTS与VC能力,在多任务适配上具有潜力.

在语音转换任务中(表13)[28,82],SpeechGPT在显式与隐式链路下均实现约3.1的VC-WER和0.86的相似度,说明其在保持语义和实现风格迁移方面取得良好平衡;Voicebox则突出在多说话人和多情感的可控转换场景.

在跨模态语音任务中(表14)[27,83],WavLLM在语音问答(SQA/SQQA)中达到67.55%的准确率,展现出较强的复杂语义建模能力;AudioPaLM系列在语音到语音翻译(S2ST)与语音-图文对齐(AST)中表现优异,BLEU分别达到35.4和37.8,体现其三模态联合建模优势. 与此同时,国内在跨模态语音处理方面也已有相关探索,例如兰朝凤等[84]提出基于DCNN与BiLSTM的单通道视听融合语音分离方法,从语音与视觉的联合建模角度为跨模态语音任务提供了重要参考.

表11~表14分别呈现语音大模型在语音识别、合成、转换与跨模态任务中的性能表现. 总体来看,不同任务的评估指标各有侧重:ASR更关注识别准确率,TTS与VC注重语音的自然度和可控性,跨模态任务则强调语义对齐与生成质量. 不同模型在各任务中展现出差异化优势,这也说明语音大模型的性能不能单一依赖某类指标进行评价.

6.2 人工评估

尽管自动化评估指标(如词错误率WER、感知语音质量评估PESQ等)在语音大模型性能评估中发挥着重要作用,但此类指标主要关注可量化的客观特性,难以全面反映语音的主观感知质量. 语音的自然度、情感表达、语调韵律与个性化风格等特征往往具有较强的主观性,需要依赖人类听觉系统的综合判断. 因此,人工评估依然是当前语音处理任务中不可替代的重要手段,尤其在语音合成、语音转换与情感生成等应用中,发挥着关键作用.

表11 语音识别(ASR)任务模型性能对比

模型名称	数据集	WER ↓
Whisper Large-v2[3]	VoxPopuli	13.6
mSLAM-CTC	VoxPopuli	9.1
USM-M	VoxPopuli	9.1
AudioPaLM 8B	VoxPopuli	11.1
AudioPaLM-2 8B	VoxPopuli	9.8
WavLM Base+[13]	Content	5.59
WavLM Large[13]	Content	3.44
SpeechGPT(显式链)[28]	—	3.1
SpeechGPT(隐式链)[28]	—	2.4

表12 语音合成(TTS)任务模型性能对比

模型名称	SIM ↑	SMOS ↑
VALL-E[9]	0.773	3.942
VALL-E 2	0.782	3.947
NaturalSpeech 2[20]	0.62	—
YourTTS[82]	0.34	3.14
AudioLM	—	3.93
SpeechGPT(显式链)[28]	0.63	4.08
Voicebox[30]	0.68	—

表13 语音转换(VC)任务性能比较

模型名称	VC-WER ↓	VC-SIM ↑	VC-SMOS ↑
YourTTS[82]	10.1	0.72	3.25
SpeechGPT(显式链)[28]	3.1	0.86	3.54
SpeechGPT(隐式链)[28]	3.1	0.86	3.72
SoundStorm	7.7	0.81	3.41

表14 跨模态语音(SQA/S2ST/AST)任务性能比较

模型名称	AST BLEU ↑	SQA ↑	SQQA ↑
SALMONN-13B[27]	—	43.35%	43.35%
Qwen-Audio-Chat 7B	—	54.25%	38.0%
WavLLM 7B[83]	—	67.55%	67.55%
AudioPaLM 8B	35.4	—	—
AudioPaLM-2 8B	37.8	—	—
USM-M	30.7	—	—

目前常用的主观评估方法中,平均意见得分(Mean Opinion Score,MOS)是最具代表性的形式[85]. MOS通过邀请多名评估员对模型生成语音的自然度和清晰度进行1~5分打分(其中1分表示“非常差”,5分表示“非常好”),最终以平均分作为主观质量的量化指标. MOS简单直观,适用于语音合成与语音转换等场景,可有效衡量语音输出与真实语音之间的感知相似性. 具体如表15所示.

表15　人工评估方法对比与适用性分析

评估方法	优点	局限	适用任务
MOS(平均意见得分)	操作简单,结果直观,可以量化主观感受	主观性强,波动大	TTS,VC
ABX测试	区分能力强,适用于感知差异判断	操作复杂,样本需求高	VC,语音克隆
比较评分	易用于多模型横向对比,适应性强	易受样本偏差影响	多模型性能排序
SDG评分	差异明显时效果好,适合版本间比较	不适用于改进幅度较小的任务	模型版本升级、微调分析
定性分析	可补充解释定量评分结果,揭示细节问题	难以量化,依赖评估员语言能力和经验	辅助优化设计、情感与语调评估

7　挑战与未来发展方向

尽管语音大模型已在识别、生成、情感建模与多模态交互等任务中展现出卓越能力,并逐渐成为语音智能的核心支柱,但作为快速演进的前沿技术,其在结构设计、数据利用、泛化能力与安全隐私等方面仍存在挑战,制约了实际应用的普及与可信部署. 为此,本文从七个方向归纳关键瓶颈,并提出系统性解决路径.

7.1　架构设计与模块协同优化不足

当前模型多采用模块化组合,例如“编码器+投影器+LLM+声码器”,各模块性能虽不断提升,但整体缺乏统一设计理念,往往依赖浅层耦合,限制了端到端性能优化,尤其在多语言或复杂任务中表现不稳. 未来应从结构重构与协同优化两方面入手:一方面构建统一语义表示空间并引入跨模块动态信息流机制(如交叉注意力、残差耦合),提升语义协同;另一方面结合任务驱动的架构搜索(Task-Aware NAS),在预训练-微调过程中优化效率与适配度.

7.2　安全性、隐私保护与防伪检测技术不足

语音大模型在提升生成质量的同时,也带来了伪造欺诈、数据泄露与情绪操控等风险. 随着深度伪造技术发展,部分模型(如Voicebox、VALL-E)的语音克隆与风格迁移能力已接近真实语音,滥用风险显著增加. 针对这一问题,国际上已形成专门的检测研究体系,如Nes2Net在多类攻击下保持高检测精度,OpenAI Sora(2025版)集成实时伪造检测功能,ASVspoof竞赛则为性能评测和标准制定提供了权威平台. 同时,国内也有研究者提出了基于多特征融合与BiLSTM的语音隐写检测方法,在对抗隐蔽伪造与异常信号识别方面取得了良好效果[86]. 未来应在生成链路中深度融合防伪机制:在生成端嵌入可追溯水印,在接收端部署检测网络,并结合差分隐私与敏感特征屏蔽,构建端到端可信的语音AI防护体系.

7.3　低资源语种与个性语音建模能力不足

目前大多数语音大模型依赖大规模高质量语料进行预训练,导致在低资源语言、地域方言、特殊口音等场景中性能严重下降. 现有迁移学习机制难以充分捕捉语种间语音—语义的结构映射规律,表现出泛化能力弱与误识率高的局限. 未来可通过自监督预训练(如HuBERT、WavLM)结合跨语种建模策略(如多语种共享编码器+语言特征适配器)增强表示通用性. 值得探索的方向包括构建语音领域的大规模跨语言对比学习框架、引入语言条件控制机制、发展低资源语音领域的参数高效微调技术(如Adapter、LoRA),以提升模型对长尾语言的支持能力.

7.4　实时语音生成响应延迟与质量平衡难题

在语音助手、虚拟客服等场景中,低延迟、高自然度的响应是关键指标. 然而现有语音生成流程通常为“文本生成+波形解码”两阶段串行结构,难以满足交互式场景的即时反馈需求. 未来应聚焦于“听说同步”结构设计,如引入具备前瞻预测能力的并行式合成模块,或开发融合LLM与声码器的端到端生成框架(如Prompt-to-Audio结构). 此外,发展基于流式建模的生成策略(如流式TTS+缓冲预测机制),可在提升响应速度的同时兼顾语音质量与语义连贯性.

8　结束语

本文系统回顾了语音大模型的发展脉络,深入分析了其核心架构、训练范式、评估方法及面临的关键技术挑战. 在模型架构方面,当前主流语音大模型多采用模块化设计,集成语音编码器、语义投影器、语言建模模块及语音生成器,具备较强的多模态感知与生成能力. 在训练策略上,预训练—指令微调—偏好对齐三阶段范式逐步完善,显著提升了模型在语音理解、表达生成与人机交互中的综合表现. 在评估体系上,通过结合自动化指标与人工主观评估,实现了对模型识别准确性、语音自然度与交互体验的多维度量化与验证.

尽管语音大模型近年来取得了实质性进展,但仍存在诸多待解问题,包括模块间协同优化不足、长语音建模能力有限、低资源语言覆盖度不足、对齐训练效率偏低以及安全性与隐私保护机制尚不健全等. 特别是在真实应用场景中的可迁移性、鲁棒性与可信性建设方面,仍是当前研究的重点与难点.

面向未来,语音大模型研究可从以下几个方向深入推进:一是加强跨模态联合建模与长时序语音理解能力,提升模型对复杂语音场景的适应能力;二是推动面向低资源语种和多方言的统一语音建模框架,促进语音智能的公平普惠应用;三是构建高效对齐机制与

可解释评估体系,增强系统的稳定性、透明度与可控性.

致谢 感谢中国社会科学院语言学重点实验室(项目编号:2024SYZH001)、河南省自然科学基金面上项目(面向情感与抑郁症联合识别的多视角交互方法研究,项目编号:242300421412)对本文研究成果的支持.

参考文献

[1] LI C, WONG C, ZHANG S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day[J]. Advances in Neural Information Processing Systems, 2023, 36: 28541-28564.

[2] HUANG D W, YAN C, LI Q, et al. From large language models to large multimodal models: A literature review[J]. Applied Sciences, 2024, 14(12): 5068.

[3] CAO N, LIN Y R, SUN X H, et al. Whisper: Tracing the spatiotemporal process of information diffusion in real time[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2649-2658.

[4] HORI T, MORITZ N, HORI C, et al. Transformer-based long-context end-to-end speech recognition[C]//Interspeech 2020. Barcelona: ISCA, 2020: 5011-5015.

[5] STRIK H, CUCCHIARINI C. Modeling pronunciation variation for ASR: A survey of the literature[J]. Speech Communication, 1999, 29(2/3/4): 225-246.

[6] AMODEI D, ANANTHANARAYANAN S, ANUBHAI R, ET AL. Deep speech 2: End-to-end speech recognition in english and mandarin[C]//Proceedings of the International Conference on Machine Learning (PMLR). Cambridge: PMLR, 2016: 173-182.

[7] SHEN J, PANG R M, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2018: 4779-4783.

[8] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023: 28492-28518.

[9] CHEN S Y, WANG C Y, WU Y, et al. Neural codec language models are zero-shot text to speech synthesizers[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 705-718.

[10] ZHANG Z Q, CHEN S Y, ZHOU L, et al. SpeechLM: Enhanced speech pre-training with unpaired textual data[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2177-2187.

[11] LU X Y, YAN Y P, KANG B, et al. ParaFormer: Parallel attention transformer for efficient feature matching[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1853-1860.

[12] HSU W N, BOLTE B, TSAI Y H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[13] CHEN S Y, WANG C Y, CHEN Z Y, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[14] CHEN S Y, WU Y, WANG C Y, et al. Unispeech-sat: Universal speech representation learning with speaker aware pre-training[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 6152-6156.

[15] WANG C, LIAO M P, HUANG Z Q, et al. BLSP-emo: Towards empathetic large speech-language models[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2024: 19186-19199.

[16] GONG Y, LIU A H, LUO H Y, et al. Joint audio and speech understanding[C]//2023 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE, 2024: 1-8.

[17] FITZGERALD J, ANANTHAKRISHNAN S, ARKOUDAS K, et al. Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 2893-2902.

[18] REN Y, HU C X, TAN X, et al. FastSpeech 2: Fast and high-quality end-to-end text to speech[EB/OL]. (2022-08-08)[2025-03-20]. https://arXiv.org/abs/2006.04558.

[19] KIM J, KIM S, KONG J, et al. Glow-tts: A generative flow for text-to-speech via monotonic alignment search[C]//Advances in Neural Information Processing Systems 33. San Diego: NeurIPS, 2020: 8067-8077.

[20] SHEN K, JU Z, TAN X, et al. NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers[EB/OL]. (2023-05-30)[2025-03-20]. https://arXiv.org/abs/2304.09116.

[21] HU H R, SONG Y, ZHANG J T, et al. Stargan-vc based cross-domain data augmentation for speaker verification[C]// ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[22] AKBARI H, YUAN L, QIAN R, et al. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text[C]//Advances in Neural Information Processing Systems 34. San Diego: NeurIPS, 2021: 24206-24221.

[23] GUZHOV A, RAUE F, HEES J, et al. Audioclip: Extending clip to image, text and audio[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 976-980.

[24] CHUANG Y S, LIU C-L, LEE H Y, et al. SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering[C]//Interspeech 2020. Barcelona: ISCA, 2020: 4168-4172.

[25] AO J Y, WANG R, ZHOU L, et al. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 5723-5738.

[26] DESHMUKH S, ELIZALDE B, SINGH R, et al. Pengi: An audio language model for audio tasks[J]. Advances in Neural Information Processing Systems 36. San Diego: NeurIPS, 2023: 18090-18108.

[27] TANG C L, YU W Y, SUN G Z, et al. SALMONN: Towards generic hearing abilities for large language models [EB/OL]. (2024-04-08) [2025-03-20]. https://arXiv. org/abs/2310.13289.

[28] ZHANG D, LI S M, ZHANG X, et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 15757-15773.

[29] KONG Z, GOEL A, BADLANI R, et al. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities[C]//International Conference on Machine Learning. Cambridge: PMLR, 2024: 25125-25148.

[30] LE M, VYAS A, SHI B, et al. Voicebox: Text-guided multilingual universal speech generation at scale[C]//Advances in Neural Information Processing Systems 36. San Diego: NeurIPS, 2023: 14005-14034.

[31] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[32] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2025, 16(5): 1-72.

[33] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[C]// Advances in Neural Information Processing Systems 33. San Diego: NeurIPS, 2020: 17022-17033.

[34] MAITI S, PENG Y F, CHOI S, et al. VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks[C]// ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 13326-13330.

[35] GAIDO M, PAPI S, NEGRI M, et al. Speech translation with speech foundation models and large language models: What is there and what is missing? [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 14760-14778.

[36] DOMINGUEZ-OLMEDO R, HARDT M, MENDLER-DÜNNER C. Questioning the survey responses of large language models[C]//Advances in Neural Information Processing Systems 37. San Diego: NeurIPS, 2024: 45850-45878.

[37] TAUD H, MAS J F. Multilayer perceptron (MLP) [M]// Geomatic Approaches for Modeling Land Change Scenarios. Cham: Springer International Publishing, 2017: 451-455.

[38] FATHULLAH Y, WU C Y, LAKOMKIN E, et al. Prompting large language models with speech recognition abilities[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 13351-13355.

[39] JEON J, LEE S, CHOI S. A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models[J]. Interactive Learning Environments, 2024, 32(8): 4613-4631.

[40] ZEGHIDOUR N, LUEBS A, OMRAN A, et al. SoundStream: An end-to-end neural audio codec[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 495-507.

[41] SATYANARAYAN A, HEER J. Lyra: An interactive visualization design environment[J]. Computer Graphics

Forum, 2014, 33(3): 351-360.

[42] VAN DEN OORD A, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning[C]//Advances in Neural Information Processing Systems 30 (NIPS 2017). San Diego: NeurIPS, 2017: 6309-6318.

[43] RAZAVI A, OORD A V D, VINYALS O. Generating diverse high-fidelity images with VQ-VAE-2[C]//Advances in Neural Information Processing Systems 32. San Diego: NeurIPS, 2019: 11240.

[44] CHEN S, WU Y, WANG C, et al. BEATs: Audio pre-training with acoustic tokenizers[C]//International Conference on Machine Learning. Cambridge: PMLR, 2023: 5178-5193.

[45] AN W D, LI R W, GE H Y, et al. An end-to-end audio transformer with multi-student knowledge distillation algorithm for deepfake speech detection[C]//Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition. New York: ACM, 2024: 366-371.

[46] BARNES C F, RIZVI S A, NASRABADI N M. Advances in residual vector quantization: A review[J]. IEEE Transactions on Image Processing, 1996, 5(2): 226-262.

[47] YAO X, NEWSON A, GOUSSEAU Y, et al. A style-based GAN encoder for high fidelity reconstruction of images and videos[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 581-597.

[48] CHANG K W, WU H B, WANG Y K, et al. SpeechPrompt: Prompting speech language models for speech processing tasks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3730-3744.

[49] CHEN Z H, HUANG H, ANDRUSENKO A, et al. SALM: Speech-augmented language model with in-context learning for speech recognition and translation[C]// ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 13521-13525.

[50] GAJDOŠ M, HUMMER K, KRESSE G, et al. Linear optical properties in the projector-augmented wave methodology[J]. Physical Review B, 2006, 73(4): 045112.

[51] HUANG Z L, WANG X G, HUANG L C, et al. CCNet: Criss-cross attention for semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 603-612.

[52] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. Cambridge: PMLR, 2023: 19730-19742.

[53] LIN J H, JIANG N F, ZHANG Z T, et al. LMQFormer: A Laplace-prior-guided mask query transformer for lightweight snow removal[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6225-6235.

[54] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[55] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.

[56] WYATT S, ELLIOTT D, ARAVAMUDAN A, et al. Environmental sound classification with tiny transformers in noisy edge environments[C]//2021 IEEE 7th World Forum on Internet of Things. Piscataway: IEEE, 2021: 309-314.

[57] GU J X, LI C, ZHANG B C, et al. Projection convolutional neural networks for 1-bit CNNs via discrete back propagation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8344-8351.

[58] GAUTHIER J. Conditional generative adversarial nets for convolutional face generation[J]. Class project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter Semester, 2014, 2014(5): 2.

[59] Kumar K, Kumar R, de Boissiere T, et al. MelGAN: Generative adversarial networks for conditional waveform synthesis[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 14910-14921.

[60] LIUH, CHENZ, YUANY, et al. AudioLDM: Text-to-audio generation with latent diffusion models[C]//Proceedings of the 40th International Conference on Machine Learning. Cambridge: PMLR, 2023: 21450-21474.

[61] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]// Advances in Neural Information Processing Systems 35. San Diego: NeurIPS, 2022: 27730-27744.

[62] JING Y, ZHU X L, LIU X B, et al. Exploring visual pre-training for robot manipulation: Datasets, models and methods[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2023: 11390-11395.

[63] BORSOS Z, MARINIER R, VINCENT D, et al. AudioLM: A language modeling approach to audio generation[J]. IEEE/ACM Transactions on Audio, Speech, and Lan-

guage Processing, 2023, 31: 2523-2533.

[64] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[EB/OL]. (2021-03-10) [2025-03-20]. https://arXiv.org/abs/2004.11362.

[65] XU X, WANG T, YANG Y, et al. Cross-modal attention with semantic consistence for image-text matching[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5412-5425.

[66] GUO C P, WANG S Y, XIE R L, et al. Estimating energy consumption of neural networks with joint Structure-Device encoding[J]. Sustainable Computing: Informatics and Systems, 2025, 45: 101062.

[67] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[EB/OL]. (2019-06-13)[2025-03-20]. https://arXiv.org/abs/1902.00751.

[68] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 4582-4597.

[69] MAO A Q, MOHRI M, ZHONG Y T. Cross-entropy loss functions: Theoretical analysis and applications[EB/OL]. (2023-06-20)[2025-03-20]. https://arXiv.org/abs/2304.07288.

[70] SALAZAR J, LIANG D, NGUYEN T Q, et al. Masked language model scoring[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 2699-2712.

[71] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2015: 5206-5210.

[72] ARDILA R, BRANSON M, DAVIS K, et al. Common voice: A massively-multilingual speech corpus[EB/OL]. (2020-03-05)[2025-03-20]. https://arXiv.org/abs/1912.06670.

[73] ZHANG B B, LV H, GUO P C, et al. WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 6182-6186.

[74] WANG C H, WU A, GU J T, et al. CoVoST 2 and massively multilingual speech translation[C]//Interspeech 2021. New York: ACM, 2021: 2247-2251.

[75] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[C]//Interspeech 2018. Los Angeles: ISCA, 2018: 1086-1090.

[76] ZHANG C, TAN X, REN Y, et al. UWSpeech: Speech to speech translation for unwritten languages[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14319-14327.

[77] GEMMEKE J F, ELLIS D P W, FREEDMAN D, et al. Audio Set: An ontology and human-labeled dataset for audio events[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2017: 776-780.

[78] XIAN Y Q, SCHIELE B, AKATA Z. Zero-shot learning: The good, the bad and the ugly[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3077-3086.

[79] DONG G T, YUAN H Y, LU K M, et al. How abilities in large language models are affected by supervised fine-tuning data composition[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 177-198.

[80] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[EB/OL]. (2022-11-04)[2025-03-20]. https://arXiv.org/abs/2103.01955.

[81] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[EB/OL]. (2024-07-29)[2025-03-20]. https://arXiv.org/abs/2305.18290.

[82] CASANOVA E, WEBER J, SHULBY C, et al. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone[EB/OL]. (2023-04-30)[2025-03-20]. https://arXiv.org/abs/2112.02418.

[83] HU S J, ZHOU L, LIU S J, et al. WavLLM: Towards robust and adaptive speech large language model[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 4552-4572.

[84] 兰朝凤, 王顺博, 郭小霞, 等. 基于DCNN和BiLSTM的单通道视听融合语音分离方法研究[J]. 电子学报, 2023, 51(4): 914-921.
LAN C F, WANG S B, GUO X X, et al. A single channel audio-visual fusion speech separation method based on DCNN and BiLSTM[J]. Acta Electronica Sinica, 2023, 51(4): 914-921. (in Chinese)

[85] STREIJL R C, WINKLER S, HANDS D S. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives[J]. Multimedia Systems, 2016, 22(2): 213-227.

[86] 苏兆品, 张羚, 张国富, 等. 基于多特征融合和BiLSTM的语音隐写检测算法[J]. 电子学报, 2023, 51(5): 1300-1309.

SU Z P, ZHANG L, ZHANG G F, et al. A speech steganalysis algorithm based on multi-feature fusion and BiLSTM[J]. Acta Electronica Sinica, 2023, 51(5): 1300-1309. (in Chinese)

作者简介

张亚洲 男,1990年出生于河南省安阳市. 现为天津大学计算机科学与技术学院副研究员、研究生导师. 主要研究领域为大型语言模型、情感计算、价值观对齐.
E-mail: yzhou_zhang@tju.edu.cn

刘祈蒙 男,2000年出生于河南省安阳市. 现为郑州轻工业大学硕士研究生. 主要研究领域为大型语言模型、价值观对齐.
E-mail: qimengliu2023@outlook.com

戎 璐 1995年出生于山西省忻州市. 现为天津大学博士研究生. 主要研究领域为大型语音模型、学习科学.
E-mail: ronglu@tju.edu.cn

赵 彬 女,1992年出生于山西省忻州市. 现为中国社会科学院语言研究所助理研究员. 主要研究领域为脑认知和计算神经科学.
E-mail: zhaobin@cass.org.cn

李爱军 女,1966年出生于湖北省郧县. 现为中国社会科学院语言研究所纪委书记、副所长,博士生导师,创新工程项目“语言与言语科学重点实验室”首席研究员. 主要研究领域为语言合成、语言分析及语音数据库.
E-mail: liaj@cass.org.cn