arXiv:2601.06829v1 [cs.SD] 11 Jan 2026

MOESCORE：基于混合专家的文本-音频相关性评分预测，用于文本到音频系统评估

Bochao Sun[1] *Yang Xiao*[2] *Han Yin*[3*]

[1]海洋科学与技术学院，西北工业大学，中国西安[2]计算与信息系统学院，墨尔本大学，澳大利亚墨尔本[3]电气工程学院，韩国科学技术院，大田，大韩民国

*通讯作者。电子邮箱：*hanyin@kaist.ac.kr*

摘要

近年来，生成模型的进步使现代文本到音频（TTA）系统能够合成具有高感知质量的音频。然而，TTA 系统通常难以保持与输入文本的语义一致性，从而导致声音事件、时间结构或上下文关系不匹配。如何评估 TTA 中的语义保真度仍是一项重大挑战。传统方法主要依赖主观的人工听音测试，耗时较长。为解决这一问题，我们提出了一种基于混合专家（MoE）架构和序列交叉注意力（SeqCoAttn）的客观评估器。我们的模型在 XACLE 挑战赛中取得第一名，在测试数据集上的斯皮尔曼等级相关系数（SRCC）为 0.6402（**较挑战赛基线提升了 30.6%**）。代码地址：https://github.com/S-Orion/MOESCORE。

索引词— 文本到音频、文本-音频相关性评估、混合专家、跨模态特征融合

1. 引言

文本到音频（TTA） [1, 2, 3] 旨在根据文本描述合成音频样本，已广泛应用于各种场景，例如语音助手 [4], 、媒体内容制作 [5, 6],以及虚拟现实声学环境模拟 [7, 8]。

评估 TTA 系统 [9] 通常涉及两个核心维度：音频质量和文本-音频相关性。尽管现代 TTA 模型在生成高质量音频方面取得了显著进展，但它们往往无法充分体现输入文本的内容，导致文本-音频相关性表现存在关键差距 [10]。传统上，文本-音频相关性通过人工主观评估进行衡量，但这种方式成本高、结果不一致且难以扩展，因此亟需客观评估指标。

为应对这一挑战，XACLE 挑战赛 2026[1] 旨在开发能够预测 TTA 系统文本-音频相关性评分（TARS）的模型。基线方法依赖单一跨模态模型，难以同时捕捉全局语义对齐、细粒度语义和时间对应关系。为克服这些局限，我们提出一种基于混合专家（MoE） [11] 架构并采用序列交叉注意力（SeqCoAttn）增强的 TARS 预测器。

我们发现，与依赖单一模型相比，CLAP 系列模型在文本-音频相关性预测方面具有互补的归纳偏置，这源于它们在粒度建模、时间敏感性和预训练目标方面的差异。受这一发现启发，我们通过特征感知门控网络整合了多个专门化专家，包括三个基于 CLAP 的变体和一个经 SeqCoAttn 增强的专家模型。基于 SeqCoAttn 的专家模型对音频与文本之间细粒度的时间对应关系进行建模，补充了基于嵌入的全局对齐。在 XACLE 2026 挑战赛上的实验表明，所提出的混合专家框架达到了最先进的性能。

[1]XACLE 2026 挑战赛：https://xacle.org/

2. 提出的方法

2.1. 模型架构

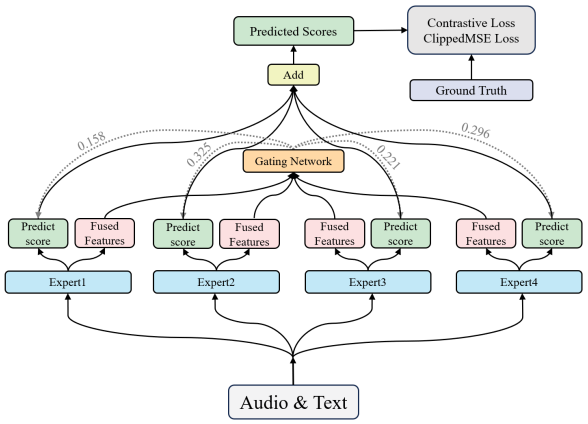
如图1所示，我们采用混合专家（MoE）架构，利用多个专用模型的互补优势，预测合成音频文本对的 TARS。混合专家框架通过门控机制动态组合多个专家模型的预测结果，从而实现更细致、更准确的估计。具体而言，我们的系统包含**四个采用不同架构和训练策略的专用专家模型**。对于**专家1–3**，我们采用预训练的、基于 CLAP 的 [12] 模型，并通过计算音频和文本特征之间的相似度来生成 TARS。对于**专家4**，我们引入 SeqCoAttn 模块，以实现更细粒度的音频文本对齐。

专家1采用 LAION-CLAP [12]骨干网络，这是一个通过大规模音频文本对比学习训练的双编码器模型。该预训练模型将音频和文本映射到共享潜在空间，在零样本音频分类和文本音频检索方面表现出色。在混合专家框架中，它提供可泛化的跨模态语义线索，以补充其他专家模型。

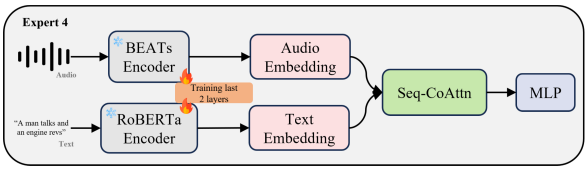
专家2基于 MGA-CLAP [13], 这一 CLAP 扩展，通过引导式注意力机制改进多粒度音频–文本对齐。MGA-CLAP 通过增强局部音频模式与全局语义表示之间的交互，在复杂音频场景中提供更强的细粒度语义对应关系，从而补充 CLAP 的全局对齐优势。

专家3建立在 M2D-CLAP [14], 这一多阶段框架之上，将自监督音频表征学习与 CLAP 式跨模态对比学习相结合。M2D-CLAP 先学习通用音频特征，再与文本嵌入进行语义对齐，从而生成稳健的音频–语言表示，其泛化能力远超标准 CLAP 设置，并支持可靠的语义评分和回归式评估。

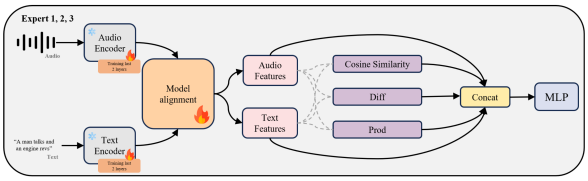
专家4基于基线系统开发，整合了 BEATs 音频编码器 [15], RoBERTa 文本编码器 [16],和 SeqCoAttn 融合模块，以建模跨模态时间对应关系。音频逐层特征通过可学习的



(a) 总体架构



(b) 专家4架构



(c) 微调后的专家1、专家2和专家3架构

图1。所提出的文本-音频相关性预测器的架构。权重并投影至512维；文本嵌入对齐至相同维度。双向SeqCoAttn实现跨模态注意力，随后通过自适应最大池化和两层多层感知机（含随机失活/Tanh裁剪）输出对齐分数。该模型专注于局部时间匹配，弥补了专家1至3的不足。

所有四个专家模型均在官方训练集上分别进行了微调。在混合专家训练期间，专家参数保持冻结，仅优化门控网络，以有效整合各专家模型。

2.2. SeqCoAttn

SeqCoAttn通过两个多头注意力层，在音频序列与文本序列之间执行双向交叉注意力。经过注意力处理的序列通过自适应最大池化进行池化，并拼接形成融合表示。

2.3. 损失函数

我们采用基线的损失函数形式 [17], ，该形式结合了裁剪均方误差（MSE）损失和基于间隔的对比损失。对比损失定义如下：

$$\mathcal{L}_{\text{con}} = \max(0, \left|\hat{d} - d\right| - \varepsilon) \tag{1}$$

ˆ其中 d 和 d 分别表示样本对之间预测分数与真实分数的差值。裁剪后的均方误差损失为：

$$\mathcal{L}_{\text{ClippedMSE}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{I}(|\hat{y} - y_i| > \tau) \cdot (\hat{y} - y_i)^2 \tag{2}$$

表1。我们的方法与基线在官方测试集上的性能对比。

系统	斯皮尔曼等级相关系数 ↑	线性相关系数 ↑	KTAU ↑	均方误差 ↓
Ours	**0.6402**	**0.6873**	**0.4612**	**3.0111**
基线	0.3345	0.3420	0.2290	4.8110

表2。单个专家和混合专家变体在官方验证集上的性能比较。

系统	斯皮尔曼等级相关系数 ↑	线性相关系数 ↑	KTAU ↑	均方误差 ↓
专家1	0.6302	0.6492	0.4542	3.5382
专家2	0.6297	0.6562	0.4550	3.7208
专家3	0.6257	0.6438	0.4496	3.7369
专家4	0.5808	0.5865	0.4147	3.8082
混合专家（3个专家）	0.6480	0.6745	0.4693	3.3592
混合专家（4个专家）	**0.6680**	**0.6845**	**0.4861**	**3.3462**

其中 y_i 和 $\hat{y}_i$ 分别表示第 i 个话语的真实分数和预测分数。最终的损失函数定义为：$\mathcal{L} = \beta \cdot \mathcal{L}_{\text{mse}} + \gamma \cdot \mathcal{L}_{\text{con}}$ 其中 β 和 γ 为超参数。

3. 结果与讨论

模型在官方盲测集上的表现汇总于表1。如表所示，我们提出的方法在所有评估指标上均持续优于基线，斯皮尔曼等级相关系数（SRCC）绝对提升了30.6%。

为进一步分析单个专家的贡献及专家聚合的效果，我们在验证集上进行了详细的消融研究，结果见表2。专家1–3凭借强大的基于CLAP的跨模态表征，取得了相近的性能；而专家4（基于SeqCoAttn的专家）侧重于序列交叉注意力和时间对应关系，其单独使用时的性能较低。尽管如此，通过混合专家框架聚合多个专家，能够持续提升所有评估指标。由三个专家（专家1–3）组成的混合专家模型已优于所有单个专家，表明该模型能够有效融合互补语义线索。纳入全部四个专家后，性能进一步提升；混合专家（4个专家）在斯皮尔曼等级相关系数、线性相关系数和KTAU上均取得最佳表现，同时均方误差最低。这些结果表明，专家多样性和动态门控使模型能够捕捉全局语义对齐和细粒度时间对应关系，验证了所提出的混合专家架构的有效性和可扩展性。

4. 结论

为解决文本到音频（TTA）语义对齐主观评估效率低且结果不一致的问题，我们提出了一种基于混合专家（MoE）和序列交叉注意力（SeqCoAttn）的客观评估器。我们采用四个专门化专家，通过门控网络进行动态融合，并借助混合损失提升鲁棒性和排序准确度。在官方盲测数据集上，我们的模型以斯皮尔曼等级相关系数（SRCC）高出基线 30.6% 的表现位居第一，显示出与人类判断高度一致。未来的工作将进一步分析专家选择行为，并将所提出的框架拓展至挑战赛场景之外。

5. 参考文献

[1] Haohe Liu、Zehua Chen、Yi Yuan、Xinhao Mei、Xubo Liu、Danilo Mandic、Wenwu Wang 和 Mark D Plumbley，《Audioldm：基于潜扩散模型的文本到音频生成》，*arXiv* 预印本 *arXiv:2301.12503*，2023 年。[2] Huadai Liu、Rongjie Huang、Yang Liu、Hengyuan Cao、Jialei Wang、Xize Cheng、Siqi Zheng 和 Zhou Zhao，《Audiolcm：基于潜一致性模型的文本到音频生成》，*arXiv* 预印本 *arXiv:2406.00356*，2024 年。[3] Jiawei Huang、Yi Ren、Rongjie Huang、Dongchao Yang、Zhenhui Ye、Chen Zhang、Jinglin Liu、Xiang Yin、Zejun Ma 和 Zhou Zhao，《Make-an-audio 2：时间增强的文本到音频生成》，*arXiv* 预印本 *arXiv:2305.18474*，2023 年。[4] Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech：快速、稳健且可控的文本转语音》，神经信息处理系统进展，第 32 卷，2019 年。.[5] Sercan O Arık、Mike Chrzanowski、Adam Coates、Gregory Diamos、Andrew Gibiansky、Yongguo Kang、Xian Li、John Miller、Andrew Ng、Jonathan Raiman 等，《Deep Voice：实时神经文本转语音》，载于国际机器学习会议。PMLR，2017 年，第 195–204 页。[6] Rongjie Huang、Jiawei Huang、Dongchao Yang、Yi Ren、Luping Liu、Mingze Li、Zhenhui Ye、Jinglin Liu、Xiang Yin 和 Zhou Zhao，《 Make-an-audio：基于提示增强扩散模型的文本到音频生成》，载于国际机器学习会议。PMLR，2023 年，第 13916–13932 页。[7] Dongchao Yang、Jianwei Yu、Helin Wang、Wen Wang、Chao Weng、Yuexian Zou 和 Dong Yu，《 DiffSound：用于文本到声音生成的离散扩散模型》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 1720–1733 页，2023 年。[8] Han Yin、Yang Xiao、Rohan Kumar Das、Jisheng Bai、Haohe Liu、Wenwu Wang 和 Mark D Plumbley，《EnvSDD：环境声音深度伪造检测基准测试》，*arXiv* 预印本 *arXiv:2505.19203*，2025 年。[9] Hui Wang、Cheng Liu、Junyang Chen、Haoze Liu、Yuhang Jia、Shiwan Zhao、Jiaming Zhou、Haoqin Sun、Hui Bu 和 Yong Qin，“Tta-bench：用于评估文本到音频模型的综合基准”，*arXiv*预印本 *arXiv:2509.02398*，2025。[10] Junwon Lee、Modan Tailleur、Laurie M Heller、Keunwoo Choi、Mathieu Lagrange、Brian McFee、Keisuke Imoto 和 Yuki Okamoto，“声音场景合成挑战：评估文本到音频生成”，*arXiv*预印本 *arXiv:2410 17589*，2024。[11] Saeed Masoudnia 和 Reza Ebrahimpour，“混合专家：文献综述”，《人工智能评论》，第42卷，第2期，第275–293页，2014年。[12] Yusong Wu*、Ke Chen*、Tianyu Zhang*、Yuchen Hui*、Taylor Berg-Kirkpatrick 和 Shlomo Dubnov，“通过特征融合和关键词到描述增强进行大规模对比式语言-音频预训练”，载于*IEEE*国际声学、语音与信号处理会议，*ICASSP*，2023年。[13] Yiming Li、Zhifang Guo、Xiangdong Wang 和 Hong Liu，“推进对比式语言-音频预训练的多粒度对齐”，载于第*32*届*ACM*国际多媒体会议论文集，2024年，第7356–7365页。[14] Daisuke Niizumi、Daiki Takeuchi、Masahiro Yasuda、Binh Thien Nguyen、Yasunori Ohishi 和 Noboru Harada，“M2d-clap：探索超越CLAP的通用音频-语言表征”，《*IEEE*开放获取》，第13卷，文章163330，2025年 163313pp. – 。[15] Sanyuan Chen、Yu Wu、Chengyi Wang、Shujie Liu、Daniel Tompkins、Zhuo Chen 和 Furu Wei，“BEATs：使用声学分词器进行音频预训练”，2022年。[16] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov，“RoBERTa：一种稳健优化的BERT预训练方法”，*arXiv*预印本 *arXiv:1907 11692*，2019。[17] Yusuke Kanamori、Yuki Okamoto、Taisei Takano、Shinnosuke Takamichi、Yuki Saito 和 Hiroshi Saruwatari，“RELATE：用于自动评估文本与音频之间相关性的主观评估数据集”，载于*INTERSPEECH*会议论文集，2025年。