arXiv:2601.02900v1 [cs.SD] 6 Jan 2026

SPO-CLAPSCORE：通过标准化偏好优化增强基于CLAP的对齐预测系统，参加首届XACLE挑战赛

Taisei Takano，Ryoya Yoshida

东京大学，日本{takano-taisei953,
336921950}@g.ecc.u-tokyo.ac.jp

摘要

首届XACLE挑战赛（文本到音频对齐挑战赛）旨在解决一个关键需求：开发与人类感知的音频—文本语义对齐相关的自动评估指标。本文介绍了提交至XACLE挑战赛的“Takano UTokyo 03”系统。我们的方法采用基于CLAPScore的架构，并结合一种名为标准化偏好优化（SPO）的新型训练方法。SPO对每位听者提供的原始对齐分数进行标准化，使模型能够学习相对偏好，并减轻个体评分偏差的影响。此外，我们采用听者筛选机制，排除评分不一致的听者。实验评估表明，SPO和听者筛选均能有效提升与人类判断的相关性。我们的系统以0.6142的斯皮尔曼秩相关系数（SRCC）在挑战赛中获得第6名，展现出具有竞争力的性能，与排名靠前的系统之间的差距很小。代码可在 https://github.com/ttakano398/SPO-CLAPScore 获取。

索引词— XACLE 挑战赛、平均意见分预测、文本到音频生成、CLAPScore

1. 引言

文本到音频（TTA）生成能够根据文本提示合成音频样本，因此已成为一个重要的研究领域 [1]。由于 TTA 模型根据文本输入生成音频，评估这类模型的关键在于观察文本提示与生成音频之间的语义对齐情况。

目前，人类主观评估仍是评估这一方面的黄金标准，但在时间和资源方面成本高昂。尽管名为 CLAPScore [2] 的客观评估指标在 TTA领域中被广泛使用，但据报道，它与人类主观评估的相关性较低 [3]。已有多项研究致力于开发一种自动评估方法，使其与人类对音频-文本语义对齐的主观评估结果相关 [3, 4, 5, 6, 7]。然而，目前用于统一评估各种评估方法的平台仍寥寥无几。

XACLE 挑战赛 [8] 于今年启动，旨在开发一种音频和文本自动评估模型，使其评估结果与人类主观评价高度相关。该挑战赛提供了 XACLE数据集以及基于LSTM分数预测器的基线模型 [4]。

本文介绍了我们提交给 XACLE 挑战赛的分数预测系统。我们的系统采用基于CLAPScore的架构，利用音频和文本嵌入之间的余弦相似度，并结合一种名为标准化偏好优化（SPO）的新型优化方法。在 XACLE测试数据集上的实验评估表明，SPO 能够使模型的预测结果与人类判断更加一致。此外，通过集成在不同条件下训练的模型，我们在预测音频–文本对齐分数方面取得了稳健的表现。

2. XACLE 挑战赛

XACLE 挑战赛（文本到音频对齐挑战赛）是一项竞赛，旨在开发一种自动评分评估系统，用于评估音频和文本之间的语义对齐度 [8]。该系统的性能通过预测分数与平均语义对齐度之间的相关系数和评分误差进行衡量。所采用的指标包括线性相关系数（LCC）、斯皮尔曼等级相关系数（SRCC）、肯德尔秩相关系数（KTAU）和均方误差（MSE）。在这四项指标中，SRCC 被视为挑战赛期间的主要性能指标。

所提供的数据集包含音频-文本对，每对均由四位听者标注，并获得 0（最低）至 10（最高）的 11 分制语义对齐评分。在训练数据和验证数据中，每个独特的文本提示都配有三个不同的音频样本。值得注意的是，每位听者都对与同一文本对应的全部三个音频样本进行了评分，

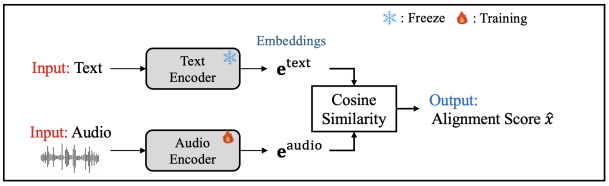


图1。所提出方法的总体架构

这些听者与负责标注训练数据和验证数据的听者不同。该数据集包含训练数据中的 7,500 个音频-文本对、验证数据中的 3,000 个音频-文本对，以及测试数据中的 3,000 个音频-文本对。

3. 所提出的方法

我们提出了SPO-CLAPScore，它结合了基于CLAPScore的分数预测模型和一种称为标准偏好优化（SPO）的偏好优化方法。我们将介绍所提出方法的基本架构、数据处理方法以及所使用的损失函数。

3.1. 基于CLAPScore的架构

我们的SPO-CLAPScore基本架构遵循Human-CLAP [3] 和CLAPScore [2]的框架。图1展示了所提出方法的整体架构。

CLAPScore是TTA领域使用的一种指标，通过计算音频和文本的CLAP [9] 嵌入之间的余弦相似度来衡量二者的对齐程度。类似地，我们根据音频和文本嵌入之间的余弦相似度计算了对齐分数$\hat{x}$:

$$\hat{x} = \frac{\mathbf{e}^{\text{audio}} \cdot \mathbf{e}^{\text{text}}}{\|\mathbf{e}^{\text{audio}}\| \|\mathbf{e}^{\text{text}}\|} \times 10, \tag{1}$$

其中，$\mathbf{e}^{\text{text}}$和$\mathbf{e}^{\text{audio}}$分别表示文本编码器和音频编码器的输出嵌入。由于目标分数是一个范围为0到10的11分制评分，我们将分数乘以10以调整量表。

3.2. 数据处理

为应对噪声评分的影响以及数据中的个体差异，我们对XACLE数据集进行了两种数据处理：听者筛选和标准偏好分数。

3.2.1. 听者筛选

我们发现，XACLE数据集中有一些分数与其他听众标注的分数不一致。由于真实对齐分数是通过计算

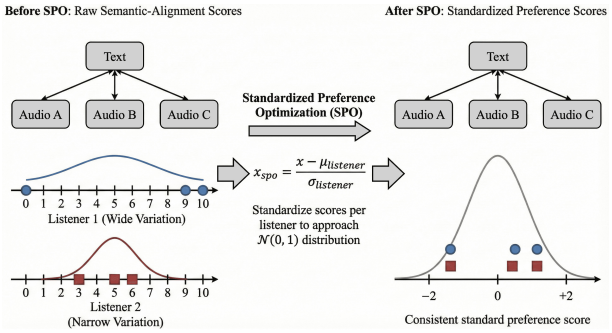


图2。标准偏好优化（SPO）概览

仅四位听者的分数平均值得出的，这些分数可能会产生很大的影响。例如，XACLE 训练数据集中的音频文件“01200.wav”获得了一组分数：（0，8，9，10）。尽管四位听者中有三位给出的分数为 8 分或更高，但由于其中一位给出了 0 分，平均分被拉低至 6。75。

为解决这一问题，我们根据下文所述的算法 $\Pi(\tau,\ r)$筛选听者。

1. 若同一数据的其他分数均不在区间 $[x-\tau, x+\tau]$ 内，则将单个原始分数 x 归类为“NG 分数”。

2. 排除“NG 分数”占比超过阈值 r 的听众。

我们预计，这种听众筛选方法能够排除噪声评分，使模型得以学习对齐分数中更为根本的模式。

3.2.2. 标准偏好分数优化（*SPO*）

在自动预测人工评估分数时，处理听众属性的影响是关键。以往的 MOS 预测模型 [10, 11]，会将“听众 ID”或“听众贡献”纳入主模型；与此不同，我们提出一种名为标准偏好优化（SPO）的基于偏好的优化方法，以减轻模型训练过程中听众属性的影响。

图2展示了 SPO 的概览。如第 2 节所述，XACLE 数据集的训练数据和验证数据中，每段文本对应三个不同的音频样本，且每位听者都会对这三个音频样本全部评分。利用这种成对结构，我们将听者的分数标准化为“标准偏好分数”，以减轻个体评分行为的影响。

人工标注者在评分时往往存在固有偏差。例如，有些听者可能倾向于给出极端分数（如频繁打 0 分或 10 分），而另一些听者则会给出较为保守的分数，将分数限制在较窄的中间范围内（如保持在 3 到 8 分之间）。直接学习

这些原始语义对齐结果会让这类个人差异干扰模型。标准偏好分数（SPO）通过将原始语义对齐结果转换为相对指标来解决这一问题。正的 x_{spo} 表示该样本的评分高于听者的个人平均分，而不受原始数值影响。这种转换使模型能够忽略原始数据中嘈杂的评分尺度，专注于捕捉人类偏好模式。

我们使用下式对每位听者的原始语义对齐分数进行标准化，确保所得分数分布符合 $\mathcal{N}(0,1)$：

$$x_{\text{spo}} = \frac{x - \mu_{\text{listener}}}{\sigma_{\text{listener}}}, \tag{2}$$

其中，μ_{listener} 和 σ_{listener} 分别表示同一听者在整个数据集中给出的所有分数的均值和标准差。x_{spo} 表示“标准偏好分数”，我们在损失计算中使用该分数来优化模型，而非使用原始语义对齐分数。

3.3. 损失函数

我们的模型通过最小化一个结合回归损失 L_{reg} 和对比损失 L_{con}[12]的损失函数进行训练。我们采用预测分数与目标分数之间的均方误差（MSE）作为回归损失 L_{reg}。

由于我们使用 SPO 的标准化偏好分数来优化模型，因此在损失计算过程中，我们使用训练数据集中原始语义对齐分数的全局均值 μ_{train} 和标准差 σ_{train} 对预测分数进行归一化。总体损失函数定义如下：

$$\begin{aligned} L =& L_{\text{reg}}\left(x_{\text{spo}}, \frac{\hat{x} - \mu_{\text{train}}}{\sigma_{\text{train}}}\right) \\ &+ \lambda L_{\text{con}}\left(x_{\text{spo}}, \frac{\hat{x} - \mu_{\text{train}}}{\sigma_{\text{train}}}\right), \end{aligned} \tag{3}$$

其中，x_{spo} 表示目标“标准偏好分数”，$\hat{x}$ 表示预测分数。λ 是用于加权对比损失的超参数。

4. 评估

本节将详细介绍我们提交至 XACLE 挑战赛的系统“Takano UTokyo 03”的实验结果和训练配置。- -

4.1. 实验条件

所提交的系统通过模型集成构建，这些模型分别在下述特定配置（设置 A、设置 B 和设置 C）下进行训练：

- 设置 A：不进行筛选；不使用对比学习损失。
- 设置 B：进行筛选；使用对比学习损失。
- 设置 C：进行筛选；不使用对比学习损失。

对于三种设置中的每一种，我们都使用三个不同的随机种子，分别训练了启用和未启用预热的模型。因此，每种设置共得到六个模型。通过结合在筛选后和未筛选的数据集上训练的模型，我们旨在增强模型在各种条件下的鲁棒性。最终预测分数通过对各个模型的预测结果取平均值获得。

所有模型共用的训练配置汇总于表 1。按照表 1 所示，采用 $\Pi(\tau = 5,\ r = 0.2)$ 进行听者筛选后，XACLE 训练数据集中的原始音频-文本对齐分数从 30,000 个样本减少至 29,308 个样本，XACLE 验证数据集中的样本则从 12,000 个减少至 11,747 个。

在评估了多种骨干编码器 [15, 16] 后，我们采用 M2D-CLAP 2025 [13] 作为音频编码器，并采用 BERT [14] 作为文本编码器。为防止过拟合，我们冻结了文本编码器，仅微调音频编码器。

表 1。常见配置

配置	
音频编码器	M2D-CLAP 2025
文本编码器	BERT（base）
优化器	Adam
总轮次	50
初始学习率（若采用预热）	0.0
峰值学习率（若采用预热）	0.0001
峰值轮次（若采用预热）	5
λ（若采用对比学习）	0.5
τ（如果进行筛选）	5
r（如果进行筛选）	0.2

4.2. 总体结果

我们提交至 XACLE 挑战赛的 SPO-CLAPScore 系统整体性能结果见表 2。这些结果来自官方 XACLE 挑战赛排行榜[1]。我们确认，我们的模型在所有指标上均优于基线模型。值得注意的是，我们的主要 SPO-CLAPScore 模型在所有提交的变体中表现最佳，斯皮尔曼等级相关系数比基线高出 0.27 以上。关于我们的模型集成策略，结果表明，对在不同条件下训练的模型进行集成，例如采用听者筛选

[1]https://xacle.org/results.html

表 2。提交模型在 XA-CLE 测试数据集上的评估结果

	斯皮尔曼等级相关系数 ↑	LCC ↑	KTAU ↑	均方误差 ↓
基线	0.3345	0.3420	0.229	4.811
SPO-CLAPS	**0.6142**	**0.6542**	**0.4407**	2.985
不使用设置 B 和 C	0.6118	0.6510	0.4391	3.072
不使用设置 A	0.6138	0.6542	0.4401	**2.963**

表3。在 XACLE 验证数据上对标准化偏好优化进行消融实验（不使用集成）

	斯皮尔曼等级相关系数 ↑	LCC ↑	KTAU ↑	均方误差 ↓
SPO-CLAPS	**0.6367**	**0.6572**	**0.4606**	**3.256**
不使用 SPO	0.5408	0.5514	0.3837	6.709

和模型预热，使集成能够捕获多样化的特征并增强其鲁棒性。

尽管测试数据集未采用我们的听者筛选方法，但“SPO-CLAPScore 不使用设置 A”（仅使用经过听者筛选的数据训练）在所有指标上的表现均优于“SPO-CLAPScore 不使用设置B和C”（仅使用未经听者筛选的数据训练），尤其是在 LCC 和均方误差方面差异显著。由于 LCC 和均方误差对离群值较为敏感，这些结果表明，我们的听者筛选方法能够帮助模型生成更加稳定、稳健的分数。

4.3. 我们在首届 XACLEChallenge 中的结果

我们的系统“Takano UTokyo 03”在 XACLE 挑战赛官方结果中获得第 6 名。值得注意的是，我们的模型与排名第 5 的系统差距微乎其微，斯皮尔曼等级相关系数仅相差 0.0001。相比之下，我们在斯皮尔曼等级相关系数上仍比排名靠后的系统高出 0.04 分以上。

4.4. 消融研究

我们进行了消融研究，以验证标准偏好分数（SPO）的有效性。在该实验中，我们使用了一个在设置 B 下训练、未进行集成的 SPO-CLAPScore模型，并比较了在 XACLE验证数据上使用和不使用 SPO 时的性能。由于目前无法获取 XACLE测试数据的真实分数，因此本实验改用 XACLE数据集的验证数据。

如表3所示，我们确认，采用SPO的模型在所有指标上都优于未采用SPO的模型。这表明，我们的SPO能够有效地使模型学会更接近人类判断的音频-文本语义对齐偏好。

5. 结论

我们介绍了提交至XACLE 挑战赛的“Takano UTokyo 03”系统。我们的系统基于CLAPScore架构，利用音频和文本嵌入之间的余弦相似度，并结合一种名为标准化偏好优化（SPO）的新型优化方法。通过对每位听者的目标对齐分数进行标准化，SPO使模型能够学会更接近人类判断的偏好，同时减轻个体听者属性的影响。此外，通过集成在不同条件下训练的模型，我们成功地实现了稳健的音频-文本对齐分数，使其接近人类感知。

在 XACLE数据集上进行的实验评估表明，我们的方法成功提高了预测的音频-文本语义对齐分数与人类评估之间的相关性。未来的工作包括收集更多样化的数据，以构建更通用、更稳健的分数预测系统。

6. 参考文献

[1] Navonil Majumder、Chia-Yu Hung、Deepanway Ghosal、Wei-Ning Hsu、Rada Mihalcea 和 Soujanya Poria，“Tango 2：通过直接偏好优化对齐基于扩散的文本到音频生成”，载于第*32*届 *ACM*国际多媒体会议论文集，2024年，第564–572页。[2] Rongjie Huang、Jiawei Huang、Dongchao Yang、Yi Ren、Luping Liu、Mingze Li、Zhenhui Ye、Jinglin Liu、Xiang Yin 和 Zhou Zhao，“Make-an-audio：利用提示增强扩散模型进行文本到音频生成”，载于国际机器学习会议。PMLR，2023年，第13916–13932页。

[3]Taisei Takano、Yuki Okamoto、Yusuke Kanamori、Yuki Saito、Ryotaro Nagase 和 Hiroshi Saruwatari，“Human-CLAP：基于人类感知的语言—音频对比预训练”，亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）论文集，第131–136页，2025年。

[4] Yusuke Kanamori、Yuki Okamoto、Taisei Takano、Shin- nosuke Takamichi、Yuki Saito 和 Hiroshi Saruwatari，《Relate：用于自动评估文本与音频相关性的主观评估数据集》，发表于*Interspeech* 论文集，2025年8月。[5] Minoru Kishi、Ryosuke Sakai、Shinnosuke Takamichi、Yusuke Kanamori 和 Yuki Okamoto，《Audiobertscore：基于音频嵌入序列相似度的环境声音合成客观评估》，发表于以现实世界多模态推理和应用用例为目标的音频中心人工智能（*Audio-AAAI*）论文集，2026年1月。

[6] SohamDeshmukh、DareenAlharthi、Benjamin Elizalde、HannesGamper、MahmoudAl Ismail、Rita Singh、Bhik- sha Raj 和 Huaming Wang，《PAM：通过提示音频语言模型评估音频质量》，发表于*Interspeech 2024*，2024年，第3320–3324页。

[7] HuiWang、JinghuaZhao、ChengLiu、Yuhang Jia、Hao-qin Sun、Jiaming Zhou 和 Yong Qin，“ Audioeval：文本到音频生成的自动双视角、多维度评估”，*arXiv* 预印本 *arXiv:2510.14570*，2025年。

[8] Yuki Okamoto、Riki Takizawa、Minoru Kishi、Yusuke Kanamori、Noriyuki Tonami、Ryotaro Nagase、Shinnosuke Takamichi 和 Keisuke Imoto，“XACLE 挑战赛 2026：首届文本到音频对齐挑战赛”，2025年。

[9] Benjamin Elizalde、Soham Deshmukh、MahmoudAl-Ismail 和 Huaming Wang，“通过自然语言监督学习音频概念的 CLAP”，载于*ICASSP 2023-2023 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2023年，第 1–5 页。

[10] Kaito Baba、Wataru Nakata、Yuki Saito 和 Hiroshi Saruwatari，“t05 系统用于 2024 年 VoiceMOS 挑战赛：从深度图像分类器迁移学习到高质量合成语音的自然度 MOS 预测”，载于*2024 IEEE* 语音语言技术研讨会（*SLT*）。IEEE，2024 年，第 818–824 页。

[11]Yichong Leng、Xu Tan、Sheng Zhao、Frank Soong、Xiang-Yang Li 和 Tao Qin，“Mbnet：利用均值偏差网络预测合成语音的 MOS”，载于*ICASSP 2021—2021 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021 年，第 391–395 页。

[12] Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari，“UTMOS：UTokyo-SaruLab 语音 MOS 挑战赛 2022 系统”，载于*Interspeech 2022*，2022 年，第 4521–4525 页。

[13] Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi 和 Noboru Harada，《M2d-clap：探索超越 clap 的通用音频表征》，*IEEE Access*，第 13 卷，第 163313–163330 页，2025年。

[14]Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：面向语言理解的深度双向 Transformer 预训练》，*CoRR*，第 abs/1810.04805 卷，2018 年。[15] 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》，发表于论文集 第 *32* 届 *ACM*国际多媒体会议，2024 年，第 7356–7365 页。

[16]Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumb- ley, Yuexian Zou 和 Wenwu Wang，《WavCaps：一种由 ChatGPT 辅助生成的弱标注音频字幕数据集，用于音频-语言多模态研究》，*IEEE/ACM Transactions on Audio, Speech, andLan-guageProcessing*，第 1–15 页，2024。