arXiv:2602.07143v1 [cs.SD] 6 Feb 2026

大规模声音嵌入基准（MSEB）

Georg Heigold Google Research 慕尼黑，德国
heigold@google.com

Ehsan Variani Google Research 加利福尼亚州，美国
variani@google.com

Tom Bagby Google Research 加利福尼亚州，美国
tombagby@google.com

Cyril Allauzen Google Research 纽约，美国
allauzen@google.com

Ji Ma Google Research 伦敦，英国
maji@google.com

Shankar Kumar
Google Research
纽约 , USA
shankarkumar@google.com

Michael Riley
Google Research
纽约 , USA
riley@google.com

摘要

音频是多模态感知的关键组成部分，任何真正智能的系统都必须展现广泛的听觉能力。这些能力包括转录、分类、检索、推理、分割、聚类、重排序和重建。从根本上说，每项任务都涉及将原始音频信号转换为有意义的“嵌入”——无论是单个向量、连续或离散表示的序列，还是其他结构化形式——并以此作为生成该任务最终响应的基础。为加速迈向稳健的机器听觉智能，我们提出大规模声音嵌入基准（MSEB）：一个可扩展的框架，旨在评估任何多模态系统的听觉组件。在首次发布中，MSEB提供了一套全面的八项核心任务，未来还将规划更多任务，并由多样化的数据集提供支持，其中包括全新的大规模简单语音问题（SVQ）数据集。我们的初步实验确立了明确的性能提升空间，凸显了在以音频为核心信号的真实世界多模态体验中进行改进的重大机遇。我们鼓励研究社区使用MSEB来评估其算法，并为其发展做出贡献。该库公开托管于 https://github.com/google-research/mseb。

1 引言

要实现稳健的机器感知，光靠视觉还不够；还需要能够聆听、理解并根据周围环境中丰富的声学信号采取行动。全面的听觉智能必须涵盖多种能力，从转录和分割等基础感知任务，到检索、推理和聚类等更高层次的认知功能。传统上，这些任务由不同的研究群体分别处理，且往往按声音领域划分（例如语音与非语音）。然而，它们都面临着一个共同的根本挑战：将高维原始音频波形转换为有意义的中间表示——也就是“嵌入”——以支持下游推理和生成。认识到这一共同基础，是摆脱狭窄、特定任务解决方案，迈向通用多模态系统的关键。

从历史上看，不同应用领域发展出了各自专门的嵌入形式。涉及人类语音的技术，如语音搜索 [1, 2] 或语音助手 [3], ，侧重于通过自动语音识别（ASR）生成离散的文本嵌入。相比之下，说话人验证 [4, 5] 或通用声音分类 [6] 等监测技术，则专注于学习稠密、连续的向量表示。尽管这些方法在各自的用途上效果显著，但这种碎片化使得人们难以评估通用机器听觉智能的发展进展。

要在这一方向开展深入研究，不能只依靠新模型，还需要一种标准化的方法来衡量它们的整体影响。具体而言，我们需要确定：1）在特定压缩率和计算复杂度下，端到端性能的现状；2）可实现的最大性能提升空间；以及 3）是否存在能够捕捉与多种任务相关的丰富声学信息的通用音频嵌入技术。为推动这类研究，我们需要一个覆盖多种任务的现代基准测试，其作用类似于面向文本的大规模文本嵌入基准（MTEB） [7] 以及类似的图像基准测试 [8, 9]。我们推出大规模声音嵌入基准（MSEB），以填补这一关键空白，为研究社区提供一个全面的平台，供大家比较方法，并在有意义的现实场景中推动声音理解的发展。

我们的工作做出了几项重要贡献。首先，MSEB 推出了一套涵盖八项关键任务的基准，覆盖广泛的声音技术领域，包括检索、推理和重排序；这些任务在以往的基准测试中鲜有涉及。其次，我们提供了一个可扩展的开源库，旨在支持与多模态系统相关的多种建模范式，让研究人员能够无缝评估自己的算法。第三，我们推出简单语音问题（SVQ）数据集。这是一项新颖的大规模资源，包含来自 26 个地区、使用 17 种语言的口语查询，专为支持基于单一来源开展多项任务的评估而设计。最后，我们的初步实验明确揭示了性能提升空间，凸显了改善现实世界多模态体验的巨大机遇，其中音频是核心信号。

MSEB 与 HARES [10],NOSS [11],和 SUPERB [12] 等以往基准有四个关键区别：1）优先考虑可直接应用于现实世界技术的任务；2）着重通过压缩比和计算复杂度衡量嵌入效率；3）纳入检索和推理等关键的、以用户为中心的任务；4）其评估方法不依赖针对特定任务的微调，能够真正检验嵌入的泛化能力。我们设想 MSEB 将成为一个动态平台，并鼓励社区贡献数据和任务，以推动声音技术的协作发展。

2 大规模声音嵌入基准（MSEB）

大规模声音嵌入基准（MSEB）旨在成为评估任何智能系统听觉能力的核心平台，无论该系统是将声音作为单一模态处理，还是将其作为更广泛多模态上下文的一部分。MSEB 通过明确定义的任务，将这些能力转化为可操作的评估内容；这些任务经过精心挑选，作为广泛应用中核心挑战的可验证替代指标。对于每项任务，受评模型都必须针对给定输入对象提供预测；该对象包含声音信号，也可能包含其他模态，并从我们精选的数据集中抽取。可验证性是 MSEB 的基本原则：每项任务都以公认的真实值和标准化评估方法为依据。为促进广泛采用，该基准配备了灵活的库，旨在简化对各种建模方法的评估，涵盖传统的单模态模型、在模态间进行转换的级联系统，以及大语言模型（LLM）等原生多模态模型。MSEB 的长期愿景涵盖听觉智能的完整范围，而首个版本则优先纳入八项已证明具有现实应用价值的任务，涉及语音搜索、智能助手和生物声学监测等快速发展的技术领域。

2.1 数据集

MSEB 的核心原则是，稳健评估需要高质量且易获取数据。我们依据五项严格标准整理数据集：(a) 公开可用，并采用便于研究使用的许可协议；(b) 数据复杂度较高，确保衍生任务并非微不足道，也不容易通过简单基线解决；(c) 规模庞大，足以在较高统计置信度下报告性能指标；

(d) 高质量标注，提供可验证的真实值；以及 (e) 覆盖多样，涵盖声音这一广阔频谱中相当大的一部分。
MSEB 在首次发布时包含四个符合这些标准的数据集，目前均包含声音和文本模态。我们正积极致力于扩展这一覆盖范围，纳入图像等其他模态。

简单语音问题 (SVQ)：我们推出了 SVQ，这是专为 MSEB 收集的新型大规模数据集。该数据集包含 17.7 万多条简短的口语查询，涵盖 26 个地区和 17 种语言，录制环境各不相同，共有四种（干净环境、背景语音、交通噪声和媒体噪声）。每条查询都配有丰富的文本元数据，这些元数据取自 XTREME-UP 基准 [13]（包括经过对齐的维基百科页面和段落），还包含全面的说话人信息和细粒度的时间标注。SVQ 的独特之处在于以单一、未拆分的数据集形式发布。有关数据收集、具体数据拆分定义和统计分布的详细信息，请参阅附录 A。

Speech-MASSIVE [14]：这个多语言口语理解 (SLU) 数据集扩展了基于文本的 MASSIVE 语料库 [15]，纳入了涵盖 12 种语言的口语话语。该数据集的标注粒度很高，涵盖 18 个领域、60 种意图和 55 个槽位。我们采用其既定的测试集进行标准基准测试。

FSD50K [16]：为覆盖一般声音事件，我们纳入了 FSD50K，其中包含超过 5.1 万个片段，音频总时长超过 100 小时。它设有经过精心设计的既定开发与评估划分，以防止数据泄漏，并直接采用来自 AudioSet 本体的 200 个声音类别 [6]。我们对所有 MSEB 任务均使用标准评估划分。

BirdSet [17]：在生物声学方面，BirdSet 提供了大规模数据集，训练数据超过 6,800 小时，涵盖近 10,000 种鸟类。它提供七个不同的测试集，包含来自被动声学监测设备的超过 400 小时声景录音。这些录音带来了现实挑战，例如高背景噪声以及重叠的叫声。至关重要的是，它们附带可靠且时间精确的标签，其中包含丰富的元数据，如鸟类类型、叫声类型、性别和地理空间坐标。MSEB 使用了全部七个测试集。

2.2 任务

MSEB 的初始版本引入了八项基础超级任务（图 1），旨在全面覆盖现代智能系统所应具备的听觉能力。为确保细粒度且稳健的评估，每个超级任务由众多具体的子变体组成，我们将其简称为任务。在结构上，每项任务都定义了严格的输入——包括主要音频信号以及可能来自其他模态的辅助信息——和所需的输出格式。这些任务的选择与已大规模部署的现实、高实用价值应用直接对齐。

以下按逻辑顺序列出超级任务：从以用户为中心的信息访问（检索、重排序、推理），到基础核心感知（分类、转录、分割），最后是更高层次的组织与生成（聚类、重建）。

检索（查找信息）：这项超级任务模拟人类通过查阅海量知识来回答问题的过程。它为语音搜索提供了一种可验证的替代评估方式：模型接收一个口语查询，并可访问“知识索引”——一个以文档形式存储信息的集合。任务是为查询找到最相关的信息。目标信息可能存在于索引中，也可能需要模型正确判断没有可用答案。我们根据搜索粒度以及查询与索引之间的语言是否匹配，定义了六项子任务。如果要在与查询相同的语言中查找整篇文档（例如网页），该任务称为同语言文档检索。与之对应的还有同语言段落检索（查找段落）和同语言词跨度检索（查找一段词语）。此外，模型也可能听到一种语言的问题，却要在另一种语言的语料库中搜索，这便构成了跨语言文档检索、跨语言段落检索和跨语言词跨度检索任务。SVQ 数据集支持在全部六项子任务、26 种语言区域设置和四种不同录音环境下评估这项超级任务。它提供了一个知识索引（维基百科的一个子集），并为每条音频查询提供对应的真实标注页面、段落和词跨度。

重排序（优化结果）：智能系统与人类相似，常会遇到歧义，即初步感知会产生多种合理解读——例如语音受到噪声干扰，或包含语音上容易混淆的词语时。重排序超级任务旨在作为解决这类歧义的替代评估方式，要求模型根据候选假设与原始音频信号的实际相关性，对给定列表中的候选假设重新排序。这项能力是许多实用系统的核心组成部分，可用于优化搜索引擎中的候选文档，或在自动语音识别中从 N-best列表中选出最佳转录。在当前 MSEB 版本中，我们聚焦于声学假设重排序。借助 SVQ 数据集，每条音频查询都配有一组听起来与真实话语相似的文本候选项。任务是利用音频信号，根据这些假设与实际所说内容的真实相关性进行排序。

推理（提取精准答案）：深刻理解不仅需要定位相关文档，还要求综合信息以推导出精准答案。这项超级任务可作为下一代智能助手的代理任务。假设相关知识已经被检索出来（例如一个段落或文档），推理任务要求模型分析所提供的多模态上下文，以响应口语查询，并准确指出满足用户信息需求的文本片段。至关重要的是，这还要求模型具备稳健的能力，能够正确判断所提供的上下文中是否没有答案。在当前的 MSEB 版本中，我们利用 SVQ 数据集，根据语言对齐情况定义了两个子任务：SpanInLang，即口语查询与文本上下文使用相同语言；以及 SpanCrossLang，即二者使用不同语言。

分类（识别声音）：听到声音后，智能系统可能需要立即识别其来源、说话人的特征，甚至是其背后的意图。这项超级任务评估模型根据不同的声学特征，将音频信号归入预定义类别的能力。尽管潜在范围十分广泛，MSEB 的首个版本聚焦于四个关键领域：说话人属性、用户意图、环境声音和生物声学。对于说话人分类（身份、性别、年龄）和录音环境分类，我们使用 SVQ 数据集，该数据集提供了涵盖 26 个语言区域的丰富标签。意图分类使用 Speech-MASSIVE 数据集进行评估，该数据集为 12 种语言提供了标准基准。对于通用环境声音分类，我们采用 FSD50K，其中包含 200 个源自 AudioSet 的类别。最后，生物声学分类使用大规模 BirdSet 基准进行评估。

转录（将声音转换为文本）：听到声音后，自然会想到将其转录为文本表示。虽然这通常与口语的自动语音识别（ASR）相关，但这一概念也适用于用文本描述任何听觉事件。这项超级任务旨在评估智能系统将音频输入转换为逐字或描述性文本输出的能力。在当前的 MSEB 版本中，我们聚焦于自动语音识别，并使用 SVQ 数据集评估 26 种语言区域的转录质量。此外，SVQ 的四种不同录音环境有助于精确比较模型对不同声学条件（如背景噪声或媒体声音）的鲁棒性。

分割（定位时间上的关键信息）：通常，我们不必听完整段音频录音，只需获取其中的关键信息。这项超级任务模拟了这种需求，评估模型识别音频信号中最显著片段并确定其精确时间戳的能力。在当前的 MSEB 版本中，这项任务基于 SVQ 数据集，涵盖其全部 26 种语言区域。对于每个音频查询，我们都标注了最显著的三个术语及其对应的开始和结束时间。模型的目标是正确预测这些关键术语及其精确的时间位置。

聚类（整理未知声音）：除了识别已知类别，真正智能的系统还必须能够将未知、无结构的声音数据整理成连贯的结构。这个超级任务用于评估模型在不依赖预定义标签的情况下，能否通过将具有相似属性的音频片段分组成同质簇来发现潜在结构。在 MSEB 的初始版本中，我们定义了涵盖不同领域的子任务：使用 SVQ 数据集进行说话者、性别和年龄聚类；使用 FSD50K 进行环境声音聚类；以及使用 BirdSet 进行生物声学聚类。

重建（生成声音）：真正掌握一种模态，意味着不仅能够理解它，还能生成它。这项综合任务要求系统仅凭其内部嵌入表示重建原始音频信号，以此评估系统的生成能力。高保真重建是一项严格的测试，能够检验嵌入是否保留了关键声学细节，而不仅仅是高层语义；这些细节对于神经音频压缩等应用至关重要



图 1：MSEB 的八项超级任务，涵盖信息访问（检索、推理、重排序）、核心感知（转录、分类、分割）以及组织与生成（聚类、重建）。

和语音合成。在首个版本中，我们着重于使用 SVQ 数据集重建多种噪声环境中的语音，同时利用 FSD50K 数据集进行通用环境声音重建。

2.3 评估器

MSEB 评估器旨在完全不依赖特定模型。我们仅假设模型会为每个输入样例生成某种形式的表示。评估器的主要作用是通过将这些表示（如果它们尚未处于任务的目标空间中，则可选择进行映射）映射到该空间，以计算指标，从而实现标准化评估。虽然会计算多项指标以进行细致分析，但每项任务都会指定一项主要指标，用于排行榜报告。除特定任务的性能外，MSEB 还会通过两项关键指标，统一衡量所有提交项的效率：压缩比（CR），用于量化表示相对于原始音频信号的内存占用；以及计算复杂度，以每秒浮点运算次数（FLOPS）衡量。以下各节将详细介绍每项超级任务的具体评估方案。

检索：评估器根据所有候选项在表示空间中与查询的接近程度，对知识索引中的候选项进行排序。主要指标是平均倒数排名（MRR），它会奖励将正确目标排在更靠前位置的模型。我们还报告完全匹配（EM），用于衡量正确目标成为排名第一候选项的频率。

重排序：评估器要求模型为给定候选列表中的每个项目分配相关性分数，然后根据这些分数进行排序。主要指标是平均精度均值（mAP），用于评估排序列表的整体质量。我们还报告平均倒数排名（MRR），以衡量最相关候选项的排名。此外，为了评估重排序对下游任务（例如，选择最佳自动语音识别转录文本）的效用，我们使用词错误率（WER）和候选错误率（CER）衡量排名第一候选项的内容质量。

推理：模型需要执行抽取式问答任务，找出所提供上下文中的精确答案片段，或者在不存在答案时正确指出无答案。主要性能指标是几何平均 F1 分数。该指标将预测的词跨度和真实答案片段视为词元袋，并据此衡量两者的重叠程度，从而提供均衡的评估，既奖励完全匹配，也认可部分重叠。关键在于，几何平均 F1 是通过计算数据中可回答子集和不可回答子集的 F1 分数的几何平均数得出的，而非

算术平均数（如标准 F1）。这种计算方式可避免采用对所有查询一律回答“无答案”这一投机策略，同时在两个子集上的表现相近时，仍能近似标准 F1 分数。

分类：评估策略会根据任务标签空间的性质进行调整。对于标准的多分类问题（每个样本只有一个标签），主要指标是准确率。对于多标签问题，一个样本可能同时具有多个真实标签，其主要指标是平均准确率均值（mAP）（宏平均）。为确保评估全面，我们还报告一系列不同的次要指标：多分类任务包括 Top-k 准确率、平衡准确率和加权 F1 分数；多标签场景则包括微观/宏观 F1 分数、汉明损失和子集准确率。

转录：评估器会衡量模型生成的文本与真实值逐字内容之间的差异。主要指标是词错误率（WER），这是一种标准指标，用于评估相对于参考文本的词级错误（替换、删除、插入）比例。为提供更细致的见解，尤其是针对形态复杂的语言，或用于评估拼写准确性，我们还报告字符错误率（CER）。
分割：评估器会评估模型识别并定位关键事件发生时间的能力。主要指标为归一化折损累计增益（NDCG），它通过奖励在正确时间顺序中找到的正确术语，来评估预测序列的质量。为详细呈现性能表现，我们报告三项具体的准确率指标：内容准确率（无论时间如何，只要标签匹配即可）、时间精确率（开始/结束时间戳在指定容差范围内匹配）以及要求内容和时间对齐均匹配的最严格指标——总体准确率。此外，我们还报告平均准确率均值（mAP），根据置信度分数评估检测性能；并报告词错误率（WER），衡量整体序列差异。

聚类：评估器会评估嵌入空间的内在结构。根据真实值中的簇数量，我们采用标准的 MiniBatch K-Means 算法对所提供的表示进行分组。主要指标为 V-measure [18]，该分数由同质性和完整性的调和平均数得出，用于评估嵌入能否成功地将不同的真实类别（例如说话人、声学环境）分隔到纯净且完整的簇中。
重建：评估器通过将生成音频的频谱图与原始参考信号的频谱图进行比较，来评估其保真度。主要指标是弗雷歇音频距离（FAD）[19]，，用于衡量原始信号嵌入与重建信号嵌入的分布之间的距离。为了全面展示生成质量，我们还报告核音频距离（KAD） [20], 这一基于分布的替代指标，它使用最大均值差异（MMD）；此外还报告嵌入均方误差，用于衡量嵌入帧之间按样本计算的平均均方误差。

2.4 库设计

MSEB 被设计为一个灵活的库，能够在多种类型的编码器上运行各类基准测试。其核心设计原则包括：

- **模块化任务与编码器**：该库支持模块化的任务和编码器，便于添加和配置。
- **批量编码器推理**：为了适应计算开销较大的模型和大型数据集，该库支持批量推理。声音数据集通常涉及大量数据和高开销的预处理（例如音频重采样），因此高效的批量处理是一项关键功能。
- **多样化任务类型**：MSEB 支持多种复杂程度的任务，从设置阶段成本高昂的任务（例如检索语料库生成）到推理步骤复杂的任务（例如聚类）。

图2展示了该库的主要组件和执行流程。Task 表示特定基准数据集的实例，并与评估器配对，用于对编码器模型的输出进行评分。该任务会将数据集（例如 Simple Voice Questions 数据集）映射为系统所需的多模态编码器输入（例如音频、文本或标签对）。

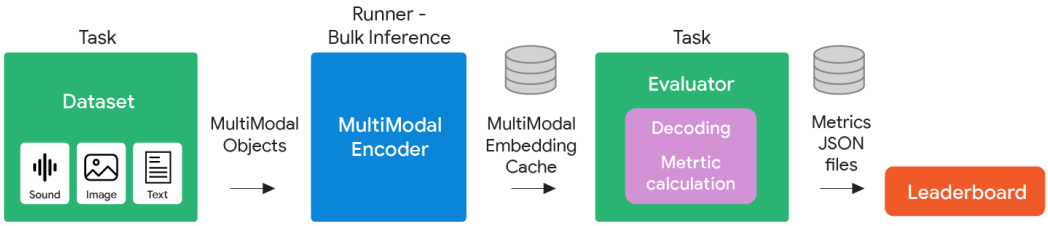


图2：MSEB 库的架构，展示了从 TaskDataset 到排行榜的流程，其中包括批量推理和评估。

对于批量推理，任务会明确列出需要编码的输入集合。Runner 接口可通过多种方式对该集合运行编码器，例如使用 Beam 流水线。Runner 的输出是一个键值存储（嵌入缓存），将任务使用的唯一标识符映射到编码器的输出。此设计允许将批量编码步骤拆分为多个阶段，并为各阶段配置不同的资源需求。
编码器模型实现了 MultiModalEncoder 接口，该接口支持可扩展的输入和输出类型集合，并在运行时进行检查。目前，该接口主要面向声音和文本模态，但设计上也支持在未来迭代中扩展到图像等其他模态。编码器输出的定义涵盖多种形式的“嵌入”：固定大小向量、嵌入序列或离散词元。
编码器也可以组合使用；目前我们支持三种主要类型：

> 1. **端到端编码器**：接收多模态输入并直接生成嵌入。2. **级联**：由一系列首尾相连的编码器组成（例如，自动语音识别 → 文本嵌入）。3. **集合**：多塔编码器，或针对不同模态组合使用的编码器。

评估器负责正确使用这些多样化的编码器输出以计算指标。例如，在分类中，端到端编码器可能直接输出类别标签文本词元，而固定大小的向量编码器则可能需要与标签向量配对，并使用余弦距离进行评估。

该库的核心功能是运行一个（任务，编码器）对的基准实例。系统提供注册表，用于列出和选择这些实例。每次基准运行都会生成一个包含完整元数据和指标的 JSON 文件。排行榜提交通过提交拉取请求来添加这些输出的 JSON 指标文件，随后这些文件会被汇总到最终展示中。

3 实验

我们利用 MSEB 框架建立了初始标准基线，并量化了全部八项超级任务的性能“提升空间”。这些广泛的实验提供了机器听觉能力当前状态的统一视图，既凸显了现有优势，也指出了显著的改进机会。

图 1 全面概述了这些结果。在该图中，每根条形表示特定任务的主要指标，任务名称旁的括号中标明了该指标。条形范围从 0 到该指标的最大值。对于有界指标（例如准确率、F1），条形止于 1.0。对于无界指标（例如词错误率、FAD），末端的虚线表示其开放性质，每个内部区间代表所显示范围的十分之一（这些指标的刻度标注在条形旁边）。此外，每根条形还可视化了性能变异性：标签标明了取得最低分和最高分的特定声音类型（例如语言区域或领域），而两者之间的渐变颜色则表示与该任务相关的所有其他声音类型的分布。

为量化这一性能差距，我们在适用的情况下采用了统一比较。绿色条形表示以声音作为任务主要输入时的性能，代表当前典型系统的表现。蓝色条形表示改用相应的真实文本转写时的性能，

代表一种感知完美的“预言机”情境。这些条形之间的差距量化了提升听觉智能的直接提升空间，尤其是对于主要依赖声音语义内容、而非其他声学特征的任务。

从这一概述中可以得出两个主要观察结果：a）在所有任务上，仍存在显著的提升空间，以达到理论上的最高分；b）当前基于声音的方法明显落后于使用文本模态的方法，这表明仍需开展更多研究，以设计出更优的声音表征，从而实现稳健的听觉理解。

接下来，我们将进一步介绍针对每项超级任务评估的具体基线。关于所有子任务的全面比较，包括次要指标和最新提交结果，请参阅实时 MSEB 排行榜及详细附录。

检索：对于这项超级任务，我们采用了级联编码方法，该方法仍是当前工业语音搜索系统的最先进方案 [21]。我们的标准基线（声音输入）使用 Whisper Large v3 [22] 进行自动语音识别（自动语音识别），随后使用文本嵌入模型对生成的转录文本进行编码。我们评估了两种不同的文本编码器：GeminiEmbedding [23]，因其在我们实验期间的 MTEB 排行榜上表现最先进而选用；以及 Gecko [24]，用于提供采用规模更小、能力较弱的文本模型的对比基线。
我们在所有 SVQ 语言区域进行了评估。对于每个示例，先将口语查询转录，再转换为嵌入。同时，使用同一个文本嵌入模型对文档索引进行显式编码。通过计算查询嵌入与文档嵌入之间的点积执行检索，并以平均倒数排名（MRR）衡量质量。

为确定性能提升空间（图 1 中的文本输入），我们进行了一项“预言机”实验：将人工真实转录文本直接输入文本编码器，绕过自动语音识别步骤。

我们的主要观察结果有三点。第一，如图 1 所示，在所有语言中，基于声音的模型表现始终落后于纯文本模型，这表明端到端音频模型未来仍有明确的追赶空间。第二，不同语言的自动语音识别质量（词错误率）与检索性能（平均倒数排名）并非完全相关（详见附录 B）。这可能是因为标准自动语音识别目标对所有词一视同仁，而有效检索对语义显著词项的依赖程度却并不相同。第三，对比同语言和跨语言任务可以发现，声音和文本模态都存在明显的性能差距。这表明，当前的嵌入向量尚未充分实现通用语义表征，仍受语言限制。有趣的是，与整页检索相比，段落检索中的这一差距相对较小，这意味着较短的文档上下文或许有助于实现更好的跨语言对齐。
重排序：对于这项任务，我们使用 SVQ 数据集，重点研究声学假设重排序。我们的基线（声音输入）采用了与检索中相同的级联方法：使用 Whisper Large v3 转录音频查询，并使用 GeminiEmbedding 对这些转录文本和给定的候选假设进行嵌入。随后，根据候选项与查询嵌入之间的余弦相似度重新排序。我们使用真实转录文本作为重排序查询，确定了文本输入所能达到的提升空间。
结果显示，各地区的表现差异极大，且与底层自动语音识别质量密切相关。英语地区表现出色（例如，澳大利亚英语的平均精度均值为 0.86），而马拉雅拉姆语（马拉雅拉姆语（印度））等具有挑战性的语言则表现严重退化（平均精度均值低至 0.11）。这证实了当前基于文本的重排序严重受制于初始的一最佳自动语音识别输出。利用转录过程中丢失的声学细节直接进行重排序，有望绕过这一限制。

推理：在此任务中，我们使用 Gemma 3 基础模型 [25]，提示它从提供的文本上下文中提取精确的答案片段。我们比较了声音输入基线（将 Whisper 自动语音识别转录文本输入 Gemma）与文本输入预言基线（输入真实转录文本）。
结果表明仍有很大的提升空间，性能差距在很大程度上取决于语言和任务复杂度。某些语言（如阿拉伯语）表现出较强的韧性，而其他语言则因自动语音识别错误而遭受显著的性能损失。例如，在马拉雅拉姆语（ml-IN）的跨语言推理中，F1 分数从 0.54（文本）骤降至仅 0.12（声音）。即使在较简单的语言内场景中，差距依然明显（例如，孟加拉语的 F1 分数从 0.65 降至 0.54），这表明当前的自动语音识别系统往往无法保留稳健下游推理所需的精确语义细节。

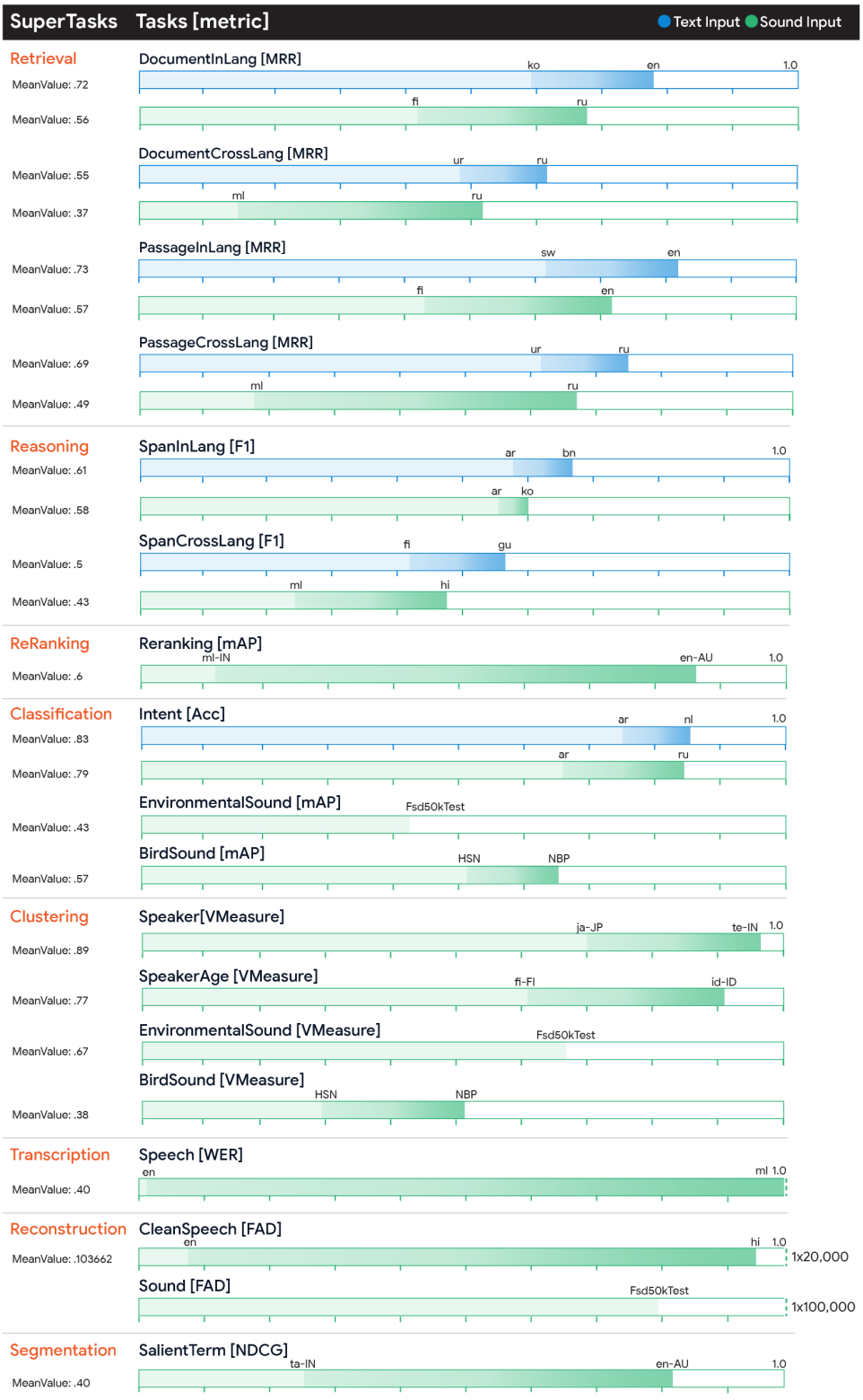


图 3：MSEB 八项超级任务的性能概览。绿色条形表示以声音作为任务输入时的性能，蓝色条形表示使用对应真实文本转写时的性能。范围标记表示不同地区或领域之间的变异性。

分类：我们针对这项任务，在不同领域评估了多种建模策略。对于 Speech-MASSIVE 上的意图分类，我们采用了与检索相同的级联基线（Whisper 自动语音识别 → GeminiEmbedding），通过查找与输入嵌入余弦距离最小的类别标签嵌入来进行预测。比较声音输入（级联）与文本输入（先验转录文本）后可以看出，提升空间因语言而异：德语的准确率仅从 0.84 降至 0.83，损失很小；而阿拉伯语的准确率则从 0.77 大幅降至 0.67，这凸显了自动语音识别对下游理解的影响因语言而异。相反，对于非语音任务，我们使用直接音频编码器来展现 MSEB 的广度。对于 FSD50K 上的环境声音分类，我们评估了 CLAP 编码器 [26], ，利用其双编码器架构，通过直接匹配音频嵌入与类别标签文本嵌入来进行零样本分类（平均精度均值达到 0.43）。对于 BirdSet 上的生物声学任务，我们评估了 Perch [27], ，这是一种专门的野生动物音频模型，并使用其直接 logit 输出（例如，在 NBP 测试集上平均精度均值达到 0.66）。

转录：我们使用 Whisper Large v3 转录全部 26 个 SVQ 语言区域，评估了最先进的自动语音识别的鲁棒性。结果显示，不同语言之间的词错误率（WER）差异显著：在干净条件下，阿拉伯语的词错误率低至 7.7%，而马拉雅拉姆语则超过 100%，这表明即使使用强大的模型，一些长尾语言仍面临严峻挑战。此外，环境噪声的影响虽可测量，但因语言而异；例如，英语在干净条件下的词错误率为 1.6%，在背景语音条件下则升至 3.8%，而其他语言对特定噪声类型的敏感程度各不相同。

分割：我们使用级联方法，在 SVQ 数据集上为这项任务建立了基线。我们使用 Whisper Large v3 转录音频，并利用预定义的 IDF 表识别前三个显著术语。随后，我们利用 Whisper 自带的词级时间戳定位这些术语。以归一化折损累计增益（NDCG）衡量的性能差异很大，且与自动语音识别质量密不可分。英语地区的精确率较高（例如，澳大利亚英语为 0.83），而孟加拉语（0.05）和马拉雅拉姆语（0.002）等较具挑战性的语言表现则急剧下降，这说明完全依赖显式文本转录来定位语义内容的方法十分脆弱。

聚类：我们使用适用于各领域的编码器，评估了在不同领域中发现潜在结构的能力。对于生物声学（BirdSet），Perch 嵌入 [27] 表现出极高的效果，在复杂声景（NBP）上取得最高达 0.53 的 V-measure，显著优于通用音频基线。对于环境声音（FSD50K），CLAP 编码器 [26]表现强劲（V-measure 为 0.68），无需显式标签便成功组织了多样的音频事件。相比之下，在 SVQ 上进行说话人聚类时，简单的原始声谱图特征表现出人意料地与 HuBERT [28] 和 Wav2Vec2 [29] 等复杂预训练模型相当，甚至在许多地区设置中更胜一筹（例如，在 te-IN 中，说话人识别的 V-measure 达到 0.97），这表明在干净条件下，基本声学属性往往足以实现基本的说话人区分。

重建：为该任务建立基线，我们评估了广泛采用的神经音频编解码器 EnCodec [30], 。我们使用弗雷歇音频距离（FAD），在 SVQ 和 FSD50K 上测量重建音频与原始信号之间的保真度。结果表明，当前模型针对特定语言和条件进行了高度优化，但缺乏通用鲁棒性。在 SVQ 上，即使是干净语音，性能也会因地区设置而大幅波动：英语的 FAD 为 ∼15k，而印地语则超过 190k。在嘈杂环境中，音质会进一步下降（例如，交通噪声中的英语 FAD 升至 ∼490k）。FSD50K 上的通用环境声音重建更具挑战性，FAD 得分最高（∼778k），凸显了改进通用音频生成模型的迫切需求。

4 结论

大规模声音嵌入基准（MSEB）旨在成为评估智能系统声音能力的权威平台。其初始版本包含八项多样化的超级任务和四个大规模数据集，涵盖了声音频谱的重要部分以及现实世界中的关键功能。MSEB 灵活且与模型无关的库，可对任何架构进行无缝评估，从传统模型到大语言模型均适用。我们的初步实验揭示，当前基于声音的方法与基于文本的预言模型之间仍有巨大的性能提升空间；此外，不同声音类型、语言和噪声条件下的表现也存在显著差异。这凸显了对稳健、通用声音表征的迫切需求。我们邀请研究社区为 MSEB 作出贡献，共同推动机器声音智能迈向真正的通用化。

参考资料

[1] Charles T Hemphill 和 Philip R Thrift。语音网页浏览。载于第三届 *ACM* 多媒体国际会议论文集，第 215–222 页，1995 年。[2] Johan Schalkwyk、Doug Beeferman、Françoise Beaufays、Bill Byrne、Ciprian Chelba、Mike Cohen、Maryam Kamvar 和 Brian Strope。“你的话就是我的命令”：Google 语音搜索：案例研究。载于语音识别进展：移动环境、呼叫中心与诊所，第 61–90 页。Springer，2010 年。[3] Matthew B Hoy。Alexa、Siri、Cortana 等：语音助手简介。《医学参考服务季刊》，37(1)：81–88，2018 年。[4] Ehsan Variani、Xin Lei、Erik McDermott、Ignacio Lopez Moreno 和 Javier Gonzalez-Dominguez。用于小型文本相关说话人验证的深度神经网络。载于*2014 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 4052–4056 页。IEEE，2014 年。[5] Georg Heigold、Ignacio Moreno、Samy Bengio 和 Noam Shazeer。端到端文本相关说话人验证。载于*2016 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 5115–5119 页。IEEE，2016 年。[6] Aren Jansen、Manoj Plakal、Ratheet Pandya、Daniel PW Ellis、Shawn Hershey、Jiayang Liu、R Channing Moore 和 Rif A Saurous。语义音频表征的无监督学习。载于*2018 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 126–130 页。IEEE，2018 年。[7] Niklas Muennighoff、Nouamane Tazi、Loïc Magne 和 Nils Reimers。MTEB：大规模文本嵌入基准。*arXiv* 预印本 *arXiv:2210.07316*，2022 年。[8] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于计算机视觉—*ECCV 2014*：第 *13* 届欧洲会议，苏黎世，瑞士，*2014* 年 *9* 月 *6—12* 日，论文集，第 *V* 卷，第 740–755 页。Springer，2014 年。[9] Chenghao Xiao、Isaac Chung、Imene Kerboua、Jamie Stirling、Xin Zhang、Márton Kardos、Roman Solomatin、Noura Al Moubayed、Kenneth Enevoldsen 和 Niklas Muennighoff。MIEB：大规模图像嵌入基准。*arXiv* 预印本 *arXiv:2504.10471*，2025 年。[10] Luyu Wang、Pauline Luc、Yan Wu、Adria Recasens、Lucas Smaira、Andrew Brock、Andrew Jaegle、Jean-Baptiste Alayrac、Sander Dieleman、Joao Carreira 等。迈向通用音频表征学习。载于 *ICASSP 2022—2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 4593–4597 页。IEEE，2022 年。[11] Joel Shor、Aren Jansen、Ronnie Maor、Oran Lang、Omry Tuval、Félix De Chaumont Quitry、Marco Tagliasacchi、Ira Shavitt、Dotan Emanuel 和 Yinnon Haviv。迈向通用非语义语音表征学习。*arXiv* 预印本 *arXiv:2002.12764*，2020 年。[12] Shu-wen Yang、Po-Han Chi、Yung-Sung Chuang、Cheng-I Jeff Lai、Kushal Lakhotia、Yist Y Lin、Andy T Liu、Jiatong Shi、Xuankai Chang、Guan-Ting Lin 等。SUPERB：语音处理通用性能基准。*arXiv* 预印本 *arXiv:2105.01051*，2021 年。[13] Sebastian Ruder、Jonathan H Clark、Alexander Gutkin、Mihir Kale、Min Ma、Massimo Nicosia、Shruti Rijhwani、Parker Riley、Jean-Michel A Sarr、Xinyi Wang 等。XTREME-UP：面向代表性不足语言的以用户为中心的稀缺数据基准。*arXiv* 预印本 *arXiv:2305.11938*，2023 年。[14] Beomseok Lee、Ioan Calapodescu、Marco Gaido、Matteo Negri 和 Laurent Besacier。Speech-MASSIVE：用于口语理解及其他任务的多语言语音数据集。*arXiv* 预印本 *arXiv:2408.03900*，2024 年。[15] Jack FitzGerald、Christopher Hench、Charith Peris、Scott Mackie、Kay Rottmann、Ana Sanchez、Aaron Nash、Liam Urbach、Vishesh Kakarala、Richa Singh 等。MASSIVE：一个包含 51*08582*、2022*2204*种类型多样语言的百万示例多语言自然语言理解数据集。*arXiv* 预印本 *arXiv:* 。[16] Eduardo Fonseca、Xavier Favory、Jordi Pons、Frederic Font 和 Xavier Serra。FSD50K：人工标注声音事件的开放数据集。*IEEE/ACM*音频、语音与语言处理汇刊，30：829–852，2021 年。

[17] Lukas Rauch、Raphael Schwinger、Moritz Wirth、René Heinrich、Denis Huseljic、Marek Herde、Jonas Lange、Stefan Kahl、Bernhard Sick、Sven Tomforde 等人。Birdset：用于鸟类生物声学音频分类的大规模数据集。*arXiv* 预印本 *arXiv:2403.10380*，2024年。[18] Andrew Rosenberg 和 Julia Hirschberg。V-measure：一种基于条件熵的外部聚类评估指标。载于*2007*年自然语言处理实证方法与计算自然语言学习联合会议论文集（*EMNLP-CoNLL*），第410–420页，2007年。[19] Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi。Fréchet 音频距离：一种用于评估音乐增强算法的无参考指标。载于*INTERSPEECH 2019* 会议录，第2350–2354页，2019年。[20] Yoonjin Chung、Pilsun Eu、Junwon Lee、Keunwoo Choi、Juhan Nam 和 Ben Sangbae Chon。KAD：不再使用 FAD！一种用于音频生成的有效且高效的评估指标。*arXiv* 预印本 *arXiv:2502 15602*，2025。。[21] Jeff Johnson、Matthijs Douze 和 Hervé Jégou。使用图形处理器进行十亿级相似性搜索。*IEEE*大数据汇刊，7(3):535–547，2019年。[22] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议，第28492–28518页。PMLR，2023年。[23] Jinhyuk Lee、Feiyang Chen、Sahil Dua、Daniel Cer、Madhuri Shanbhogue、Iftekhar Naim、Gustavo Hernán- dez Ábrego、Zhe Li、Kaifeng Chen、Henrique Schechter Vera、Xiaoqi Ren、Shanfeng Zhang、Daniel Salz、Michael Boratko、Jay Han、Blair Chen、Shuo Huang、Vikram Rao、Paul Suganthan、Feng Han、Andreas Doumanoglou、Nithi Gupta、Fedor Moiseev、Cathy Yip、Aashi Jain、Simon Baumgartner、Shahrokh Shahi、Frank Palma Gomez、Sandeep Mariserla、Min Choi、Parashar Shah、Sonam Goenka、Ke Chen、Ye Xia、Koert Chen、Sai Meher Karthik Duddu、Yichang Chen、Trevor Walker、Wenlei Zhou、Rakesh Ghiya、Zach Gleicher、Karan Gill、Zhe Dong、Mojtaba Seyedhosseini、Yunhsuan Sung、Raphael Hoffmann 和 Tom Duerig。Gemini Embedding：来自 Gemini 的通用嵌入，2025年。[24] Jinhyuk Lee、Zhuyun Dai、Xiaoqi Ren、Blair Chen、Daniel Cer、Jeremy R. Cole、Kai Hui、Michael Boratko、Rajvi Kapadia、Wen Ding、Yi Luan、Sai Meher Karthik Duddu、Gustavo Hernandez Abrego、Weiqiang Shi、Nithi Gupta、Aditya Kusupati、Prateek Jain、Siddhartha Reddy Jonnalagadda、Ming-Wei Chang 和 Iftekhar Naim。Gecko：从大语言模型中蒸馏出的多功能文本嵌入，2024年。[25] Gemma Team、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 等人。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*，2025年。[26] 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*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2023年。[27] Bhuvanesh Ghani、Tom Denton、Stefan Kahl 和 Holger Klinck。全球鸟鸣嵌入可实现更优的生物声学分类迁移学习。《科学报告》，13(1):22876，2023年。[28] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdel-rahman Mohamed。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊，29:3451–3460，2021年。[29] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：自监督语音表征学习框架。神经信息处理系统进展，33:12449–12460，2020年。[30] Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。高保真神经音频压缩。机器学习研究汇刊，2023年。

一个 SVQ 数据集

SVQ 数据集是一个开源集合，可通过 https://huggingface.co/datasets/ google/svq 访问。该数据集经过专门整理，旨在支持 MSEB 基准测试中的多种任务。本节将从整体上介绍该数据集的统计概况和描述性细节。

划分：本次发布的音频数据以一个完整的大型集合形式呈现，而非预先划分为训练集、验证集或测试集。这一决定直接源于数据采集过程的设计。具体而言，文本提示和录音环境是在说话人群体中随机分配的。虽然这种做法丰富了条件的多样性，却也给传统的数据划分带来了复杂性：若要创建互不重叠的子集，确保不同子集之间的说话人和文本材料均不重叠（这是常见的最佳实践），就会导致大量数据被舍弃，预计约占录音总量的 40%。

制定这一发布策略的首要目标，是最大限度地提高用户可用数据的效用和数量。因此，为避免如此严重的数据损失并尽可能完整地提供数据集，所有数据均作为一个未划分的评估集整体发布。打算使用这些数据训练模型的用户，需要自行制定并实施划分策略；如果希望复现此类条件，则应考虑数据量与严格的说话人与文本互斥要求之间固有的权衡。

语言构成：该数据集包含 26 个语言区域，代表 17 种不同的语言。与 XTREME-UP 基准 [13], 对应收集的语言分别涵盖 1 至 5 个区域，以确保能够体现各种地区方言和口音。所有已识别区域的录音均已纳入最终数据集。包含多个区域的语言有：

- 英语（en）：en-AU、en-GB、en-IN、en-PH、en-US
- 阿拉伯语（ar）：ar-EG、ar-x-gulf、ar-x-levant、ar-x-maghrebi
- 孟加拉语（bn）：bn-BD、bn-IN
- 乌尔都语（ur）：ur-IN、ur-PK

注意：为与语言代码保持一致，原文中的 uk-PK 推定为 ur-PK e.

数据集规模与音频特征：平均录音时长：5.1 秒
最短录音时长：1 秒 最长录音时长：
62.4 秒 录音总数：171,434 条 不同的文
本转录（提示语）：25,549 条 唯一说话
人总数：700

表 1 展示了 SVQ 数据集主要统计信息的概览。

说话人信息：SVQ 数据集为 700 名唯一参与者提供了有价值的说话人特定元数据，包括匿名标识符、自我报告性别和年龄。对性别分布的考察显示，该数据集具有多元的组成：人数最多的是自我认同为女性的说话人，共 376 人（约占 53.7%）；其次是自我认同为男性的说话人，共 308 人（约占 44.0%）。此外，数据集中还有少数自我认同为非二元性别的参与者，共 2 人（约占 0.3%）；另有 14 名说话人（2.0%）选择不回答性别问题。总体而言，这些数据反映出两个最大性别类别的分布较为均衡，不过女性参与者略多。有关该性别分布更详细的地区特定细分，请参阅表 2。

如表3所详述，SVQ 数据集的整体年龄概况显示，参与者主要由成年人组成。在所有地区，说话人的年龄中位数为29岁。数据集中涵盖的年龄范围相当广，从最年轻的18岁到最年长的71岁。这一范围表明，尽管年龄分布的中心倾向于年轻至

表 1：SVQ 数据集：总体统计数据。

语言	录音	说话人	区域设置	录音	说话人
ar	52453	229	埃及阿拉伯语	13811	59
			海湾阿拉伯语	13452	58
			阿拉伯语（黎凡特）	12600	55
			阿拉伯语（马格里布）	12590	57
bn	8845	42	孟加拉语（孟加拉国）	4119	22
			孟加拉语（印度）	4726	20
en	28004	121	澳大利亚英语	5506	25
			英语（英国）	5586	25
			英语（印度）	5632	23
			英语（菲律宾）	5647	24
			英语（美国）	5633	24
fi	10365	51	芬兰芬兰语	10365	51
gu	3689	16	古吉拉特语（印度）	3689	16
hi	3805	18	印地语（印度）	3805	18
id	5726	28	印度尼西亚语（印度尼西亚）	5726	28
ja	2833	13	日语（日本）	2833	13
kn	3453	15	卡纳达语（印度）	3453	15
ko	6520	28	韩语（韩国）	6520	28
ml	3353	16	马拉雅拉姆语（印度）	3353	16
mr	3699	16	马拉地语（印度）	3699	16
ru	11912	48	俄罗斯俄语	11912	48
sw	5878	25	sw	5878	25
ta	3308	16	泰米尔语（印度）	3308	16
te	10410	45	泰卢固语（印度）	10410	45
ur	7181	32	乌尔都语（印度）	3688	16
			乌尔都语（巴基斯坦）	3493	16

中年成年人，但也包括刚步入成年期的年轻成年人以及年龄较大的老年人。

考察特定语言区域内的年龄分布，可以发现其变异性相当大，这凸显了该数据集人口构成的多样性。各区域层面的年龄中位数有所不同：芬兰芬兰语（fi-FI）区域的说话人年龄中位数较低，为 24 岁；而澳大利亚英语（en-AU）区域的年龄中位数则明显较高，为 39 岁，俄罗斯俄语（ru-RU）和日语（日本）（ja-JP）区域的年龄中位数也达到 37 岁。这表明，不同地区或语言群体的参与者，其典型年龄可能存在显著差异。

此外，各区域内的年龄跨度（从最低年龄到最高年龄）也各不相同。例如，俄罗斯俄语（ru-RU）区域的说话人年龄最高可达 71 岁，与该数据集的总体最高年龄相同；芬兰芬兰语（fi-FI）区域的说话人年龄最高可达 67 岁。相比之下，一些区域的年龄范围较窄，例如古吉拉特语（印度）（gu-IN）区域的参与者最高年龄为 34 岁，孟加拉语（孟加拉国）（bn-BD）区域的最高年龄为 36 岁。这种差异表明，有些区域涵盖了年龄跨度非常广的人群，而另一些区域则侧重于范围更窄、通常更年轻的成年群体。

表 2：SVQ 数据集：说话人性别统计。

语言	区域设置	女性	Male	无答案	非二元性别
ar	埃及阿拉伯语	28	31		
	海湾阿拉伯语	38	16	4	
	阿拉伯语（黎凡特）	36	16	3	
	阿拉伯语（马格里布）	28	29		
en	澳大利亚英语	16	9		
	英语（英国）	15	9		1
	英语（印度）	12	8	3	
	英语（菲律宾）	20	4		
	英语（美国）	10	12	1	1
bn	孟加拉语（孟加拉国）	4	18		
	孟加拉语（印度）	4	16		
fi	芬兰芬兰语	12	38	1	
gu	古吉拉特语（印度）	8	8		
hi	印地语（印度）	9	8	1	
id	印度尼西亚语（印度尼西亚）	23	5		
ja	日语（日本）	8	4	1	
kn	卡纳达语（印度）	11	3	1	
ko	韩语（韩国）	19	9		
ml	马拉雅拉姆语（印度）	9	7		
mr	马拉地语（印度）	8	7	1	
ru	俄罗斯俄语	31	16		1
sw	sw	12	13		
ta	泰米尔语（印度）	10	6		
te	泰卢固语（印度）	18	27		
ur	乌尔都语（印度）	6	10		
	乌尔都语（巴基斯坦）	9	7		
总计		376	308	14	2

表3：SVQ 数据集：说话人年龄统计。

语言	区域设置	年龄中位数	最小年龄	最大年龄
ar	埃及阿拉伯语	32	20	56
	海湾阿拉伯语	30	19	48
	阿拉伯语（黎凡特）	30	20	52
	阿拉伯语（马格里布）	27	18	46
en	澳大利亚英语	39	22	65
	英语（英国）	27	20	49
	英语（印度）	27	21	46
	英语（菲律宾）	31	20	39
	英语（美国）	28	20	62
bn	孟加拉语（孟加拉国）	28	21	36
	孟加拉语（印度）	30	24	51
fi	芬兰芬兰语	24	20	67
gu	古吉拉特语（印度）	28	23	34
hi	印地语（印度）	31	19	47
id	印度尼西亚语（印度尼西亚）	28	22	46
ja	日语（日本）	37	28	52
kn	卡纳达语（印度）	28	22	49
ko	韩语（韩国）	35	20	55
ml	马拉雅拉姆语（印度）	28	22	41
mr	马拉地语（印度）	28	22	45
ru	俄罗斯俄语	37	21	71
sw	sw	27	22	40
ta	泰米尔语（印度）	32	22	48
te	泰卢固语（印度）	26	19	46
ur	乌尔都语（印度）	31	21	4
	乌尔都语（巴基斯坦）	32	22	43
总计		29	18	71

B 检索

我们将检索任务划分为三个不同的粒度层级：页面检索、段落检索和片段检索。每项任务又分为两种形式：语言内检索和跨语言检索。

对于段落检索，我们使用 XTREME-UP 提供的索引。一个包含 271,711 个段落的单一索引涵盖了 9 种语言的段落及额外的困难负例，可支持所有语言内检索任务。对于跨语言检索，我们使用一个单独的索引，其中包含来自英文维基百科的 112,426 个段落。

对于页面检索，我们构建了两种规模的索引："small"和"full"。"small"语言内和跨语言页面索引的构建方式是：将前文所述 XTREME-UP 段落索引中的段落映射到相应的页面标题，然后进行去重。这些标题所指向的页面来自 wikipedia/20190301.[ar|bn|en|fi|id|ko|ru|sw|te] TensorFlow Datasets (TFDS) 快照[1]。这一流程为"small"语言内索引（汇总这九种语言）生成了 8,568 个不同页面，为"small"跨语言索引（源自英语页面标题）生成了 4,559 个不同页面。

相比之下，"full"索引使用 wikipedia/20190301.<lang>TFDS 快照中每个可用语言版本的所有页面。与多语言"small"同语言页面索引不同，"full"同语言页面索引按语言区分。此外，英语（en）"full"索引充当"full"跨语言检索索引。这些"full"索引的大小差异显著，每种语言的页面数从几千到几百万不等，跨度达三个数量级，见表 4。

表 4：MSEB 页面索引大小（full）。

ar	bn	en	fi	id	ko	ru	sw	te
5,824,596	1,272,226	87,566	619,207	947,627	980,493	2,449,364	48,434	91,857

页面的文本内容远多于段落，因此可能超过所用嵌入模型的最大上下文长度。例如，Gecko 和 Gemini 模型的限制为 8,000 个词元。图 4 展示了部分语言的页面文本长度分布。如图所示，虽然大多数页面（例如 99%）都短于这一限制，但仍有一小部分页面长得多。当页面长度超出这一容量时，我们会将文本分割成多个块，并取各块嵌入的均值，作为该页面的最终表示。

对于片段检索，索引基于相应的段落级索引构建，所用方法与前文详述的"小型"页面检索索引类似。我们建议将词跨度限制为最多十个词元。

图 6（Gecko）和图 7（GeminiEmbedding）展示了级联模型按语言划分的全面结果，其中包括使用真实转录文本进行的性能提升空间分析。我们观察到，检索任务的性能提升空间各不相同，具体取决于语言、单语或跨语言条件以及索引大小等因素。

我们重点介绍以下关键观察结果：

- Whisper 在不同语言上的表现差异显著（见图 5），词错误率（WER）从 10%（英语）到 100%（马拉雅拉姆语）不等。
- Gecko 和 Gemini Embedding 在（大多数）语言上的表现都较为稳定。
- Gemini Embedding 的表现始终优于 Gecko。
- 值得注意的是，即便使用近乎"完美"的文本嵌入器，提升空间依然存在，因为答案并非总能从含噪文本中可靠推断出来，使用"small"索引的页面检索任务即为例证。
- 相反，当文本嵌入器的性能较低时，提升空间往往有限；在使用"full"索引的跨语言页面检索任务中，这一点尤为明显。

[1] https://www.tensorflow.org/datasets/catalog/wikipedia

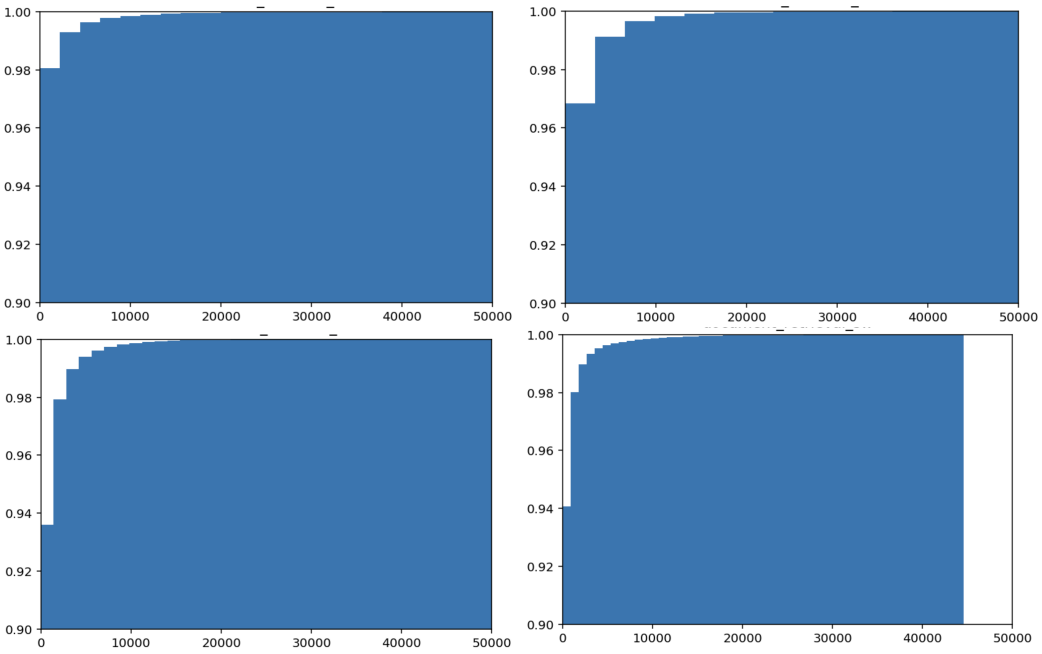


图 4：典型页面长度分布（以 Gemini Embedding 词元计）。

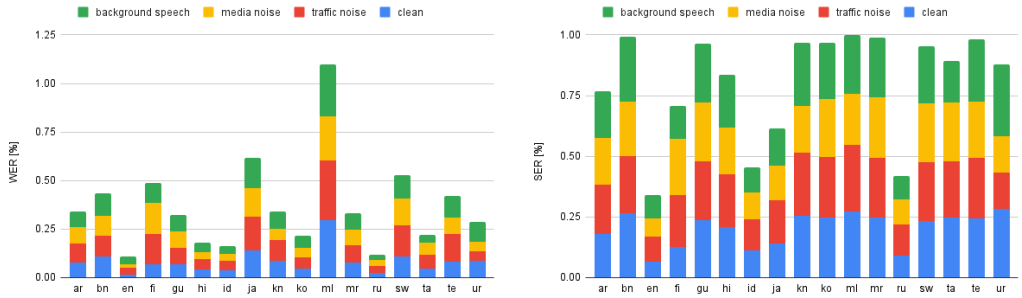


图 5：不同语言和环境下的 Whisper 词错误率（左）以及句子/查询错误率（右）。

总体而言，基于自动语音识别的模型与真实值之间存在很大的提升空间，表明模型改进潜力巨大。

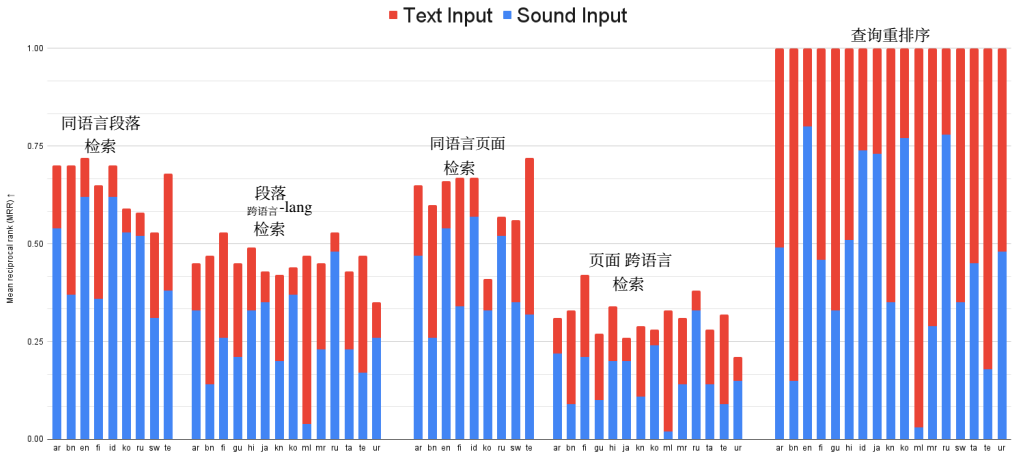


图 6：MSEB 检索与重排序任务：Gecko 的性能提升空间分析。

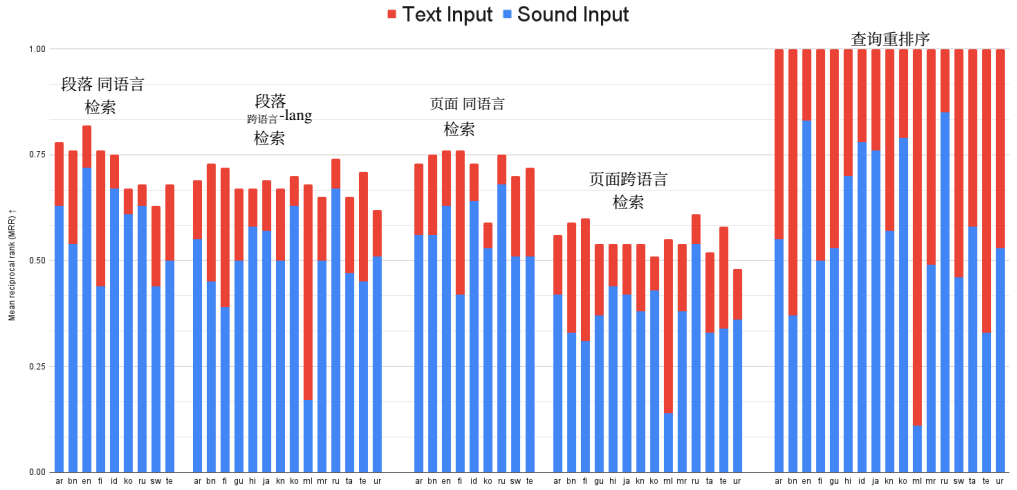


图 7：MSEB 检索与重排序任务：Gemini Embedding 的性能提升空间分析。

C 重排序

对于查询重排序任务，我们的目标是生成一组具有挑战性的候选查询。这组查询经过设计，其中包含与真实查询在语音上相似的查询，以测试声学区分能力；同时还加入语义相似的候选项，作为合理的干扰项，从而增加猜出答案的难度。这些候选项最初由大语言模型（LLM）根据提示生成，提示中包含了这些语音和语义约束，见表 5。

随后，对生成的列表进行若干后处理步骤：添加真实查询、删除重复项、按相对于真实查询的词错误率（WER）从低到高排序，并截取列表，仅保留最多排名前 250 的候选项。由于一个真实查询可能对应多条录制的语音（噪声条件或地区不同），因此所有此类语音都会分配到这组一致的候选查询。表 6 展示了这些生成候选项的示例。

表 5：查询重排序任务中用于生成语音和语义相似候选项的系统指令模板。

给定句子“{text}”，请用 {language} 生成 {num_candidates} 个听起来 与其在语音上相似的句子。 重点关注发音相似的词和节奏。不要考虑句子的语义， 只关注声音。 将结果输出为列表，每行一个句子，不要提供推理过程或评论。
给定句子 ’{文本}’，生成 {num_ 个} 句子，这些句子在 {语言} 中与其语义 相似。 重点关注同义词和相近的语义。不要考虑句子的声音， 只考虑语义。 将结果以列表形式输出，每行一个句子，不要提供推理过程或评论。

表 6：MSEB 查询重排序候选项，查询为“哪些艺术家属于 Clamp？”

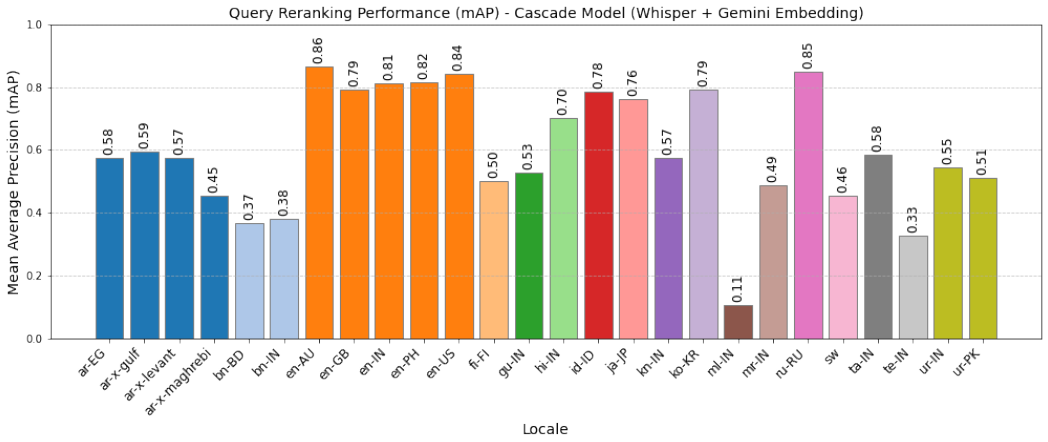
1. “哪些艺术家属于 clamb？”	9. “哪些艺术家与 Clamp 有关联？”
2. “哪些艺术家是 Clamp 的成员？”	10. “哪些屁股是衰退的酸味？”
3. “什么开始是手臂的一部分？”	11. “哪些演员是特朗普的开始？”
4. “哪些艺术家是飞机的组成部分？”	12. “哪些弧形船只是总和的一部分？”
5. “哪些设计师属于 Clamp？”	13. “哪些艺术家对 Clamp 有贡献？”
6. “哪些设计师是 Clamp 的一部分？”	14. “哪些屁属于李子？”
7. “哪些动画师是 Clamp 的成员？”	15. “哪些艺术家属于卡牌群？”
8. “哪些人是 Clamp 的成员？”	. . .

图 8 展示了我们使用 Gemini Embedding 的级联模型在各语言上的结果。在本次评估中，基于文本的基线天然能够达到最佳性能（例如，平均倒数排名=1），因为将真实查询与自身进行比较时，它始终会获得最高的相似度分数。

我们观察到，不同语言的提升空间各不相同。主要观察结果包括：

- Whisper 在不同语言中的表现差异显著（见图 5），词错误率（WERs）从 10%（英语）到 100%（马拉雅拉姆语）不等。
- Gemini Embedding 始终优于 Gecko。

总体而言，基于自动语音识别的模型与真实值之间仍存在显著的提升空间，这表明模型性能有相当大的改进潜力。



Fig图 8：使用级联基线时，SVQ 语言区域的查询重排序性能（平均精度均值）(W耳语自动语音识别 +Gemini Embedding）。 e

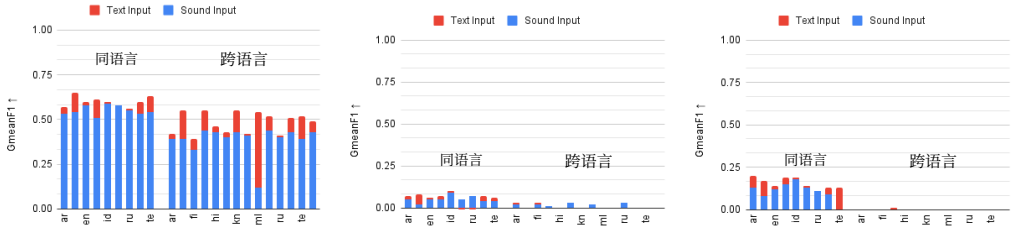


图 9：MSEB 跨度推理任务：生成式（Gemma）与检索（Gemini Embedding 和 Gecko）方法。

D 推理

本文将推理任务定义如下：给定一个查询和上下文，目标是从上下文中预测对应的答案片段；如果找不到答案，则指出这一点。我们还根据所提供的上下文区分两种推理任务：段落级和页面级。在段落级推理中，上下文由标题和维基百科中的特定段落组成，我们复用了原始 XTREME-UP 数据集中的段落。在页面级推理中，上下文包括标题以及该标题所指向的维基百科完整页面。这些维基页面信息来源于 wikipedia/20190301 集合。

我们提供了生成式（Gemma 3）和检索（Gemini embedding）方法的基线结果。在检索方法中，我们使用 0.8 的阈值来区分真实答案和“无答案”。这两种推理方法按语言划分的结果见图 9。检索方法效果不佳，可能是因为 Gemini embedding 和 Gecko 都未针对这类任务进行微调。

表 7：推理任务的系统指令。

任务：可回答性判定和精确答案提取

目标： 判断能否根据提供的上下文和标题回答问题；如果可以，则从文本中逐字提取**确切**答案。

输入： 你将收到一个问题、标题和上下文，每项各占一行。

* “question”：所提出的问题（字符串）。* “title”：上下文所属文档的标题（字符串）。
* “context”：一段文本，可以是维基页面或维基段落，可能包含也可能不包含答案。

输出： 你将以纯文本字符串（无标记）的形式生成单个 JSON 对象。其结构取决于问题是否可回答：
如果问题可以回答：

* "理由": (string) 对所提供答案为何正确的简明解释。请具体说明，并引用句子或短语。
* "answer": (string) 问题的答案，须从标题或上下文中**原样**复制。不要改写或总结。优先使用简洁答案（在完整的前提下尽可能短）。

如果问题无法回答：

* "无答案": (string) 清晰简明地解释为何无法回答该问题。
  请说明缺少哪些信息。

重要注意事项：

* **代码变更**：问题的语言可能与上下文和标题不同。在这种情况下，
  请使用与上下文和标题相同的语言回答问题。
* **精确匹配**：优先使用所提供文本中的原词。不要改写或
  总结。不要翻译成英语。
* **具体性**：在理由和无法_回答的解释中尽可能具体。
* **标题和上下文**：两者都要考虑。
* **直接回答**：仅使用标题或上下文中的文本。不要推断或下结论。
* **歧义**：如果一个问题基于上下文可能有多个不同的答案，则该答案
  应视为“无答案”，因为上下文不够具体。如果有多个答案
  同样有依据且同样正确，请选择“无答案”。
* **纯文本 JSON 输出**：输出必须是有效的 JSON 字符串，但必须是纯文本
  字符串——不得包含任何类型的标记。

E 分类

我们评估了跨不同领域的分类性能，涵盖从人类语音意图到环境声音和生物声学声音。

表 8 展示了 Speech-MASSIVE 上意图分类以及 BirdSet 上鸟鸣分类的参考最先进基准。

在实验中，我们还评估了这些基线之外的零样本方法和专用编码器：

- 意图分类（Speech-MASSIVE）：我们使用 Gemini embedding 评估了一种零样本方法。如图 10 所示，该方法无需特定任务训练，便在许多语言中取得了出色表现。
- 生物声学（BirdSet）：我们评估了 Perch，其在 NBP 测试集上的平均精度均值为 0.66，在 HSN 上为 0.54，在 POW 上为 0.53。
- 环境声音（FSD50K）：我们在零样本设置下评估了 CLAP 编码器，取得了 0.43 的平均精度均值。

表 8：参考最先进分类基准。

name			嵌入		指标		
type	数据集	划分	模型	type	type	值	参考
意图	Speech-MASSIVE	ar de es fr hu ko nl pl pt ru tr vi	Whisper	text	准确率 [%]	61.22 78.64 80.59 85.93 63.93 74.09 77.49 76.88 80.02 79.51 71.14 68.71	[14]
鸟鸣	BirdSet	PER NES UHH HSN NBP SSW SNE	wav2vec	向量	cmAP	0.18 0.39 0.27 0.45 0.63 0.28 0.29	[16]

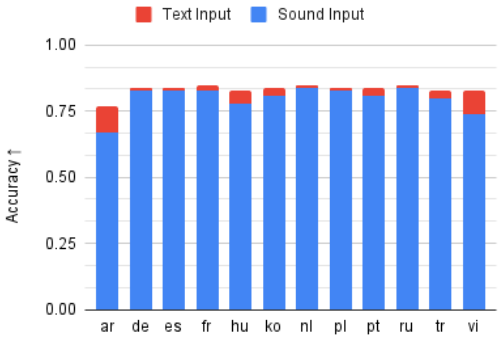


图 10：Speech-MASSIVE：使用 Gemini embedding 进行跨语言零样本意图分类的准确率。

F 转录

对于转录任务，我们在 SVQ 数据集提供的全部 26 个语言区域中评估了 Whisper Large v3 模型 [22] 。评估涵盖了四种声学环境——干净环境、背景语音、交通噪声和媒体噪声——以评估模型的鲁棒性。性能采用两项标准指标进行量化：词错误率（WER）和句子错误率（SER）。

图 11 展示了各语言在所有声学环境下的平均词错误率（WER）和句子错误率（SER）总体情况。结果凸显了不同语言之间显著的性能差异。英语（en）和俄语（ru）等资源丰富的语言词错误率较低，表现出色；而马拉雅拉姆语（ml）和孟加拉语（bn）等其他语言的错误率则极高。这凸显了实现真正通用的自动语音识别能力仍有很大的提升空间。

为进一步分析声学条件的影响，图 12 和图 13 按环境展示了累计错误率。正如预期，干净条件下的错误率始终最低。不过，对特定噪声类型的敏感程度各不相同；例如，交通_噪声会导致许多地区的性能显著_下降，而另一些地区则受背景语音的影响更大。

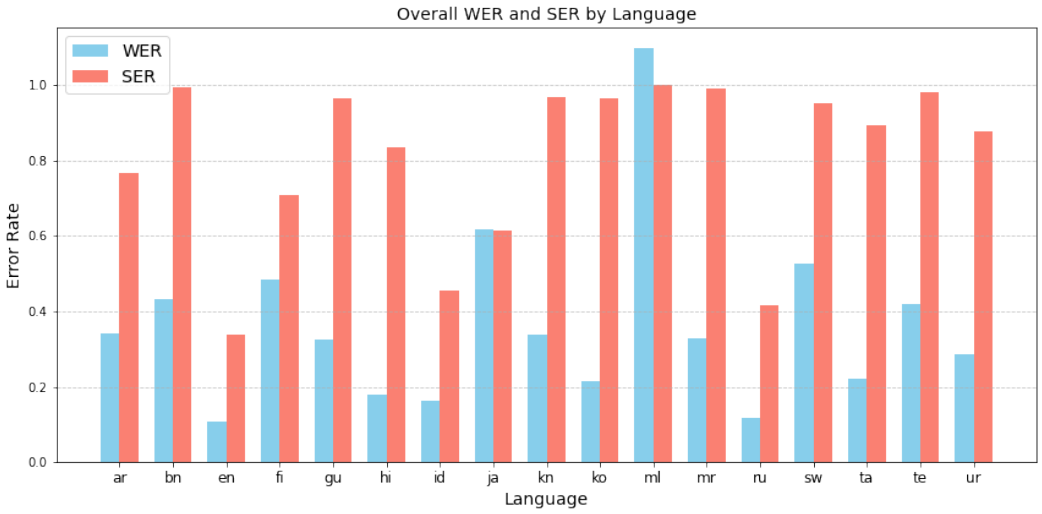


图 11：所有 26 个 SVQ 语言区域在全部声学环境下的平均总体词错误率（WER）和句子错误率（SER）。

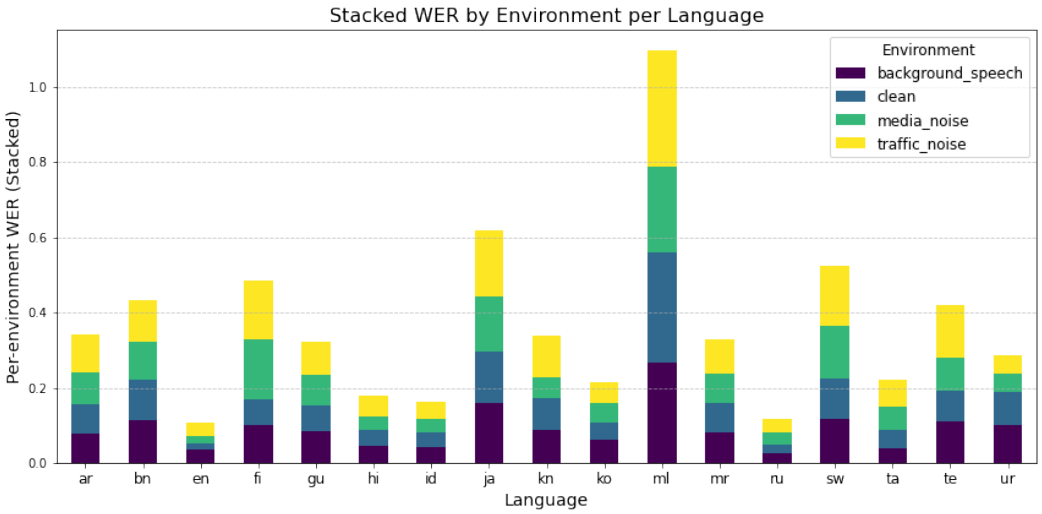
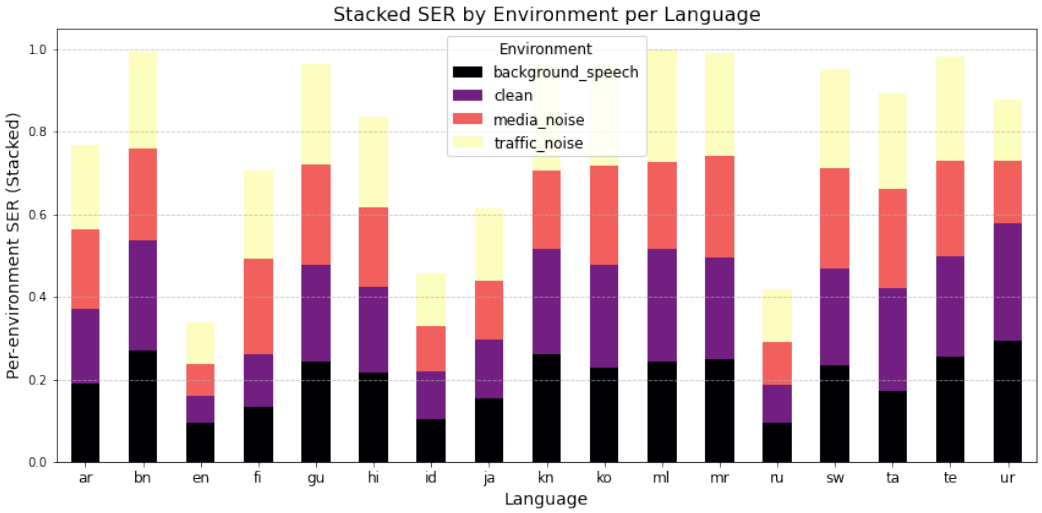


图 12：各语言按环境条件统计的累积词错误率（WER） e.



Fig图 13：各语言按环境条件统计的累积句子错误率（SER） .

G 分割

对于分割任务，我们采用了旨在定位关键信息时间位置的级联基线。首先，使用 **Whisper Large v3** 转写音频查询。接着，我们使用基于维基百科预先计算的逆文档频率（IDF）表，从转写结果中找出最显著的三个术语。最后，我们利用 Whisper 模型提供的词级时间戳，确定这些显著术语的开始和结束时间。

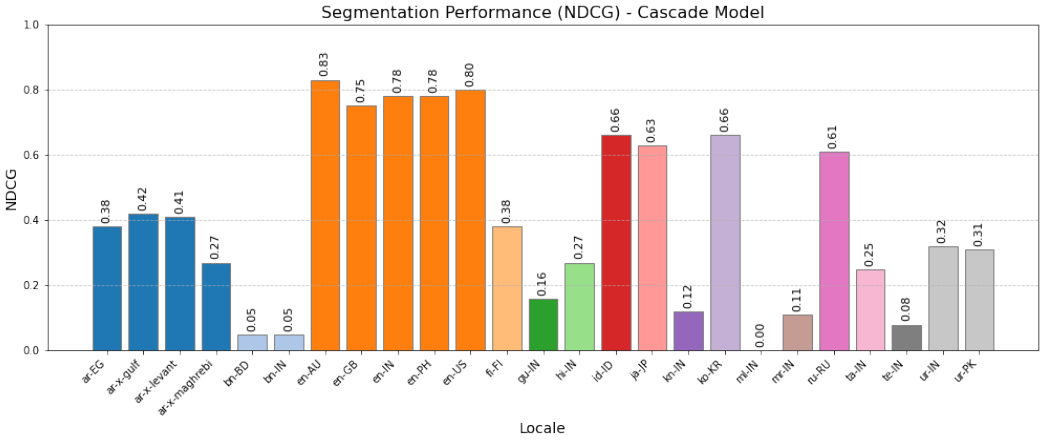
此任务的主要指标是**归一化折损累计增益（NDCG）**，用于评估预测的显著术语序列及其时间顺序的质量。如图 14 所示，各地区的表现差异很大。拥有强大自动语音识别模型的高资源语言（如英语（澳大利亚英语、英语（美国））取得了较高的 NDCG 分数（~0.80）。相反，自动语音识别表现较差的语言（如马拉雅拉姆语（印度）和孟加拉语（孟加拉国））的分数接近于零。

为了解误差来源，我们将表现分解为三项准确率指标（图 15）：

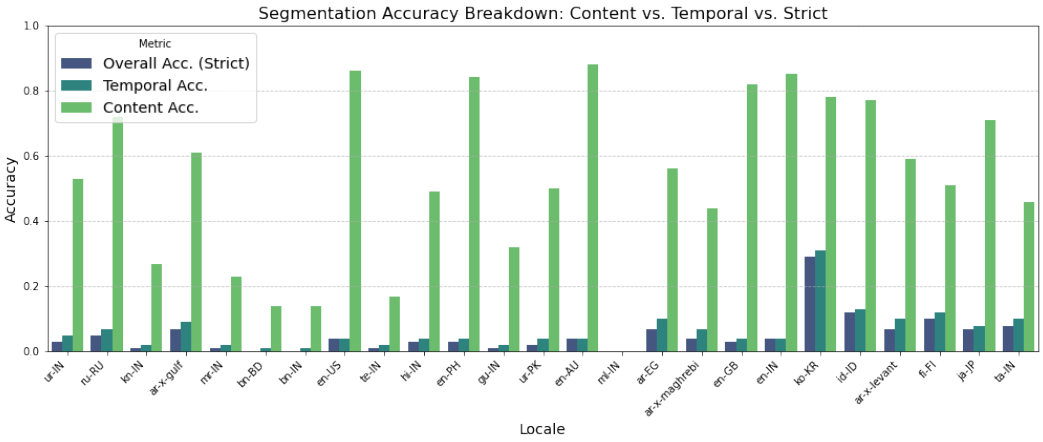
- **内容准确率**：无论时间如何，都要识别出正确的显著词语。
- **时间准确率**：无论词语内容如何，都要识别出正确的时间区间。
- **总体准确率（严格）**：同时正确识别词语及其精确的时间区间。

图 15 显示，主要瓶颈往往在于**内容准确率**。如果自动语音识别未能正确转录显著词语（在难度较高的语言中很常见），严格准确率自然会降至零。有趣的是，对于韩语（韩语（韩国））等某些语言，即使识别出了内容（蓝色柱较高），时间对齐仍然具有挑战性，导致严格总体准确率（绿色柱）明显下降。

图 16 证实了这种对自动语音识别质量的强依赖，其中显示词错误率（WER）与归一化折损累计增益（NDCG）之间存在明显的负相关。



Fig图 14：使用级联基线时，所有 SVQ 语言区域的分割表现（归一化折损累计增益） .



Fig图 15：分割准确率细分。“总体准确率（严格）”要求内容正确 t
ide识别和精确的时间定位（容差范围为 0.1 秒）。

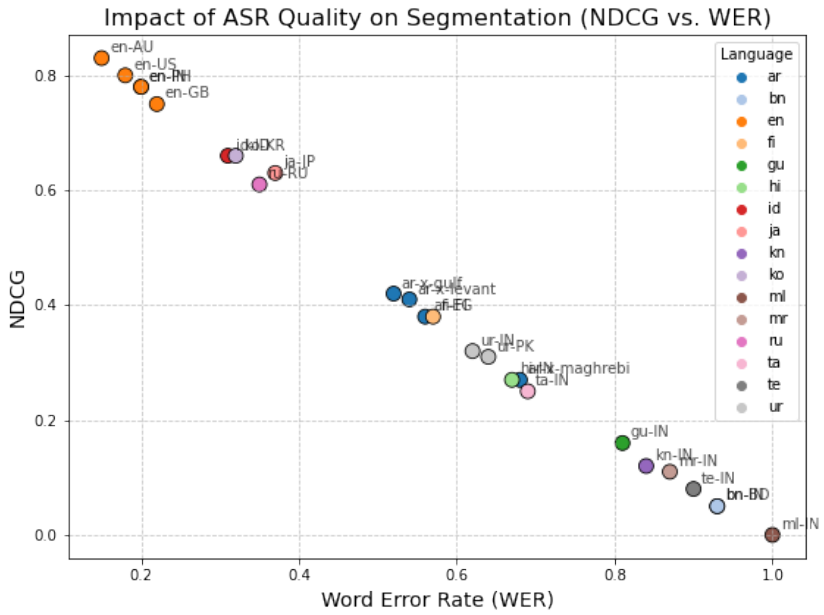


图 16：自动语音识别质量（词错误率）与下游分割性能（归一化折损累计增益）之间的相关性，凸显了级联模型的瓶颈。

H 聚类

对于聚类超级任务，我们评估了各种嵌入在无监督条件下发现潜在声学结构的能力。我们使用 V-measure 对性能进行衡量；在给定真实簇数量的情况下，这是一项平衡同质性与完整性的标准指标。

我们在三个不同领域开展了实验，采用了通用编码器和适用于领域的编码器：

- **生物声学**：使用 Perch（专为生物声学设计）、CLAP（通用音频）和原始声谱图基线，在 BirdSet 的三个测试拆分（HSN、NBP、POW）上进行评估。
- **环境声音**：使用 CLAP 在 FSD50K 上进行评估。
- **人类语音**：使用自监督语音模型（HuBERT Large、Wav2Vec2 Large）和原始声谱图基线，在涵盖 26 个地区的 SVQ 上评估三个子任务（说话人识别、性别、年龄）。

图17 展示了非语音领域的性能。正如预期，专用的 Perch 模型在生物声学领域表现出色，在复杂声景（NBP）上取得了高于 0.5 的 V-measure。不过，CLAP 作为通用编码器也展现出很强的价值，在 BirdSet 上表现不俗，同时在 FSD50K 环境声音上取得了较高的 V-measure （ 0.68 ）。

图18展示了语音领域的一项显著发现。尽管大型预训练自监督学习模型（HuBERT、Wav2Vec2）通常表现良好，但简单的声谱图基线却极具竞争力，尤其是在说话人识别方面。在许多地区，它的表现与复杂模型相当，甚至超过复杂模型。这表明，对于语音搜索中典型的干净近讲查询（如 SVQ 所代表的情况），基本声学特征往往足以实现稳健的无监督说话人区分。

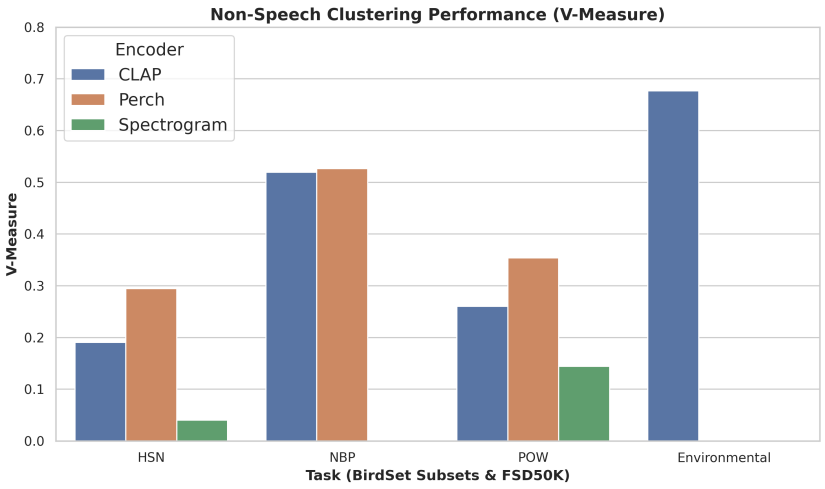


图17：非语音领域的聚类表现。Perch 在其专精领域生物声学（BirdSet）中表现出色，而 CLAP 在环境声音和生物声学任务中均展现出强大的通用能力。

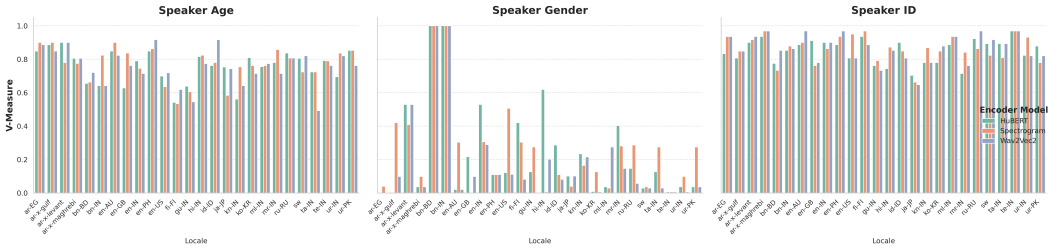


图18：在 SVQ 各语言区域中，针对说话人年龄、性别和识别的说话人聚类表现。令人意外的是，原始声谱图基线（橙色）的表现仍极具竞争力，可与大型预训练模型（HuBERT、Wav2Vec2）媲美，尤其是在说话人识别方面。

I 重建

对于重建任务，我们使用 EnCodec 模型 [30], 一种广泛采用的神经音频编解码器建立了基线。目标是评估该模型从其压缩表示中忠实重建输入音频信号的能力。我们使用弗雷歇音频距离（FAD） [19]指标评估重建质量，该指标通过比较原始音频与重建音频的声谱图计算得出。声谱图采用 1024 个采样点的帧长和 320 个采样点的帧移进行计算。

实验针对两个不同领域开展：使用 SVQ 数据集研究各种条件下的语音，以及使用 FSD50K 数据集研究一般环境声音。表9汇总的结果显示，当前生成式音频模型面临着重大挑战。

如表9所示，EnCodec 在干净语音上的表现最佳，但即便如此，不同语言之间的表现仍存在显著差异，其中印地语（hi）的重建误差明显高于英语（en）。存在背景噪声时，性能会大幅下降，其中交通噪声的影响最为严重。对 FSD50K 中一般环境声音的重建所得 FAD 分数远高于其他情况，表明生成复杂的非语音声学场景对当前模型而言仍是一项特别困难的任务。这些结果确立了明确的基线，并凸显出开发生成式模型的必要性：提升其对噪声的鲁棒性，并增强其对多种声音类型的普适性。

表9：使用 EnCodec 的重建性能（FAD，越低越好）。

数据集	条件	平均FAD	最小FAD（示例）	最大FAD（示例）
SVQ	干净语音	103,662	15,532（英语）	193,799（印地语）
SVQ	背景语音	163,984	20,094（英语）	310,327（印地语）
SVQ	媒体噪声	211,774	125,050（韩语）	346,577（印地语）
SVQ	交通噪声	348,885	133,411（韩语）	596,491（印地语）
FSD50K	环境声音	778,232	N/A	