神经编解码器能否泛化？针对未见语言和非语音任务的受控研究

Shih-Heng Wang[1]*，Jiatong Shi[2],Jinchuan Tian[2],Haibin Wu[3],Shinji Watanabe[2],[1]南加州大学，[2]卡内基梅隆大学，[3]Meta

arXiv:2601.12205v1 [cs.SD] 18 Jan 2026

摘要

本文探讨了神经音频编解码器（NAC）的泛化能力中三个关键但尚未得到充分研究的方面：（i）NAC 能否在预训练期间泛化到未见语言；（ii）仅用语音预训练的神经音频编解码器能否有效泛化到非语音（例如环境声音、音乐、动物发声等）应用；（iii）纳入非语音预训练数据能否提升语音和非语音任务的性能？现有研究依赖现成的神经音频编解码器进行比较，但由于 NAC 实现方式各异，所得见解有限。在本研究中，我们采用严格受控的配置从头训练 NAC，并精心筛选预训练数据覆盖范围，以进行公平比较。我们使用 11 项指标，对 NAC 的信号重建质量和下游应用性能进行全面评估。研究发现，NAC 能在预训练期间泛化到未见语言；仅用语音预训练的神经音频编解码器在非语音任务上的性能会下降；而在预训练期间纳入非语音数据，可以提升非语音任务的性能，同时保持语音任务上的性能相当。

1 引言

神经音频编解码器（NAC）（Zeghidour 等人，2022年；Défossez 等人，2023；Kumar 等人，2023；Zhang 等人，2024年；Yang 等人，2023；Wu 等人，2024c；Mousavi 等人，2025a；Arora 等人，2025）已成为一种流行的信号压缩技术，这主要得益于它们在压缩后重建信号方面的强大能力。通常，NAC 在推理过程中由三个主要组件构成：（i）编码器，将原始波形转换为连续表征；（ii）量化器，将这些连续表征转换为离散表示；（iii）解码器，根据这些离散表示重建原始波形。压缩后的离散令牌可实现高效传输，并能准确重建原始信号。此外，这些离散表示适用于语音与音频语言建模任务（Wang 等人，2023；Chen 等人，2024b；Zhang 等人，2023；Li 等人，2024年），进一步凸显了它们的重要性。

尽管神经音频编解码器日益普及，但关于它们仍有若干基本问题尚未得到解答。本文旨在探讨有关其泛化能力的三个关键问题（见图 1）：

1. **神经音频编解码器能否泛化到预训练期间未见的语言？**
2. **仅用语音预训练的神经音频编解码器能否有效泛化到非语音（例如环境声音、音乐、动物发声等）应用？**
3. **纳入非语音预训练数据是否会提升语音和非语音任务的表现？**

先前的研究（Shi 等人、2024a；Wu 等人，2024a；Mousavi 等人，2024年；Chang 等人，2024年；Yang 等人，2024年；Mousavi 等人，2025b；Park 等人，2025）主要侧重于评估现成的神经音频编解码器，对这些具体问题的见解有限。然而，现有的神经音频编解码器在架构、训练目标和预训练数据方面各不相同，这使得直接比较变得复杂，也阻碍了详细分析。明确回答这些问题对于语音领域至关重要，因为这有助于指导神经音频编解码器的预训练数据整理，并明确哪些神经音频编解码器适用于特定应用。值得注意的是，近期的综述（Mousavi 等人，2025b）对神经音频编解码器进行了广泛概述。

*本研究是在2024年7月至2024年11月期间于卡内基梅隆大学访问期间开展的。

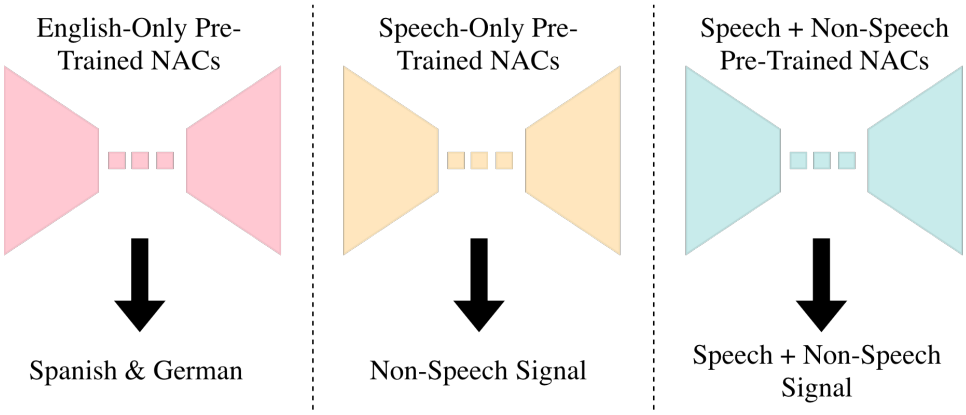

图1：我们提出的三个关键研究问题概述。

在概述我们的方法之前，我们首先详细阐述这三个问题背后的动机：

问题一：神经音频编解码器能否在预训练过程中泛化到未见语言？理想情况下，神经音频编解码器的性能应不受语言影响，从而适用于多语言应用。然而，现有与神经音频编解码器相关的研究（Shi 等人，2024a；Wu 等人，2024a，b；Mousavi 等人，2024年；Chang 等人，2024年；Yang 等人，2024年；Mousavi 等人，2025b）主要评估单一语言（英语），因此神经音频编解码器的语言泛化能力仍有待探究。因此，我们旨在研究神经音频编解码器是否存在语言依赖问题。填补这一研究空白对于确定神经音频编解码器能否可靠地应用于其预训练过程中未包含的语言至关重要（Zhang 等人，2023）。

问题二：仅用语音预训练的神经音频编解码器能否有效泛化到非语音应用？神经音频编解码器越来越多地应用于非语音场景，例如环境声音检测（Wu 等人，2024a）和音乐生成（Xu 等人，2025；Chen 等人，2024a）。然而，一些现有的神经音频编解码器（Zhang 等人，2024年；Yang 等人，2023；Ju 等人，2024年）仅使用语音数据训练，从未接触过非语音信号。鉴于这两个领域之间存在显著差异，这类模型在处理非语音输入时预计会出现性能下降。尽管有这样的预期，目前仍缺乏系统性证据。因此，我们开展实验，以实证检验这一假设。

问题三：纳入非语音预训练数据能否提升语音和非语音任务的表现？基于问题二，我们研究在神经音频编解码器的预训练中纳入非语音数据，能否提升其在语音和非语音任务上的表现。具体而言，我们旨在考察纳入非语音数据能否缩小非语音任务中观察到的领域差距。

为回答这些问题，我们在受控条件下开展实验。由于实现方式各异，直接比较现有神经音频编解码器所得的见解有限。因此，我们从头预训练神经音频编解码器（第2.1节），并精心设计数据覆盖范围，以便进行公平评估。具体而言，我们考虑三种预训练数据覆盖范围（第2.2节），并从信号重建质量和下游性能两方面评估由此得到的神经音频编解码器（第2.3节）。我们采用 11 项指标（第3.4节），全面评估不同设置下的神经音频编解码器。结果（第4节）表明：**(i)**神经音频编解码器能够泛化到预训练期间未见过的语言；**(ii)** 仅语音神经音频编解码器在非语音任务上的性能会下降；**(iii)** 预训练期间加入非语音数据，可提升非语音任务上的性能，同时保持语音任务上的性能相当。

2 方法论

我们的首要目标是在受控条件下进行比较，以确保评估公平且一致。为增强研究结果的可靠性和实用性，我们将所有结果完全开源并确保可复现。本节概述我们的方法。首先，在第2.1节中，我们介绍神经音频编解码器的预训练过程和架构。接着，在第2.2节中，我们介绍三种神经音频编解码器预训练数据覆盖范围设置。最后，在第2.3节中，我们说明神经音频编解码器的评估标准。

2.1 NAC 预训练

这里，我们将简要介绍神经音频编解码器的预训练方法。

总体框架。神经音频编解码器（Zeghidour 等，2022；Défossez 等，2023；Kumar 等，2023；Zhang 等，2024；Yang 等，2023；Wu 等，2024c；Mousavi 等，2025a；Arora 等，2025）的预训练采用重建与生成对抗网络（GAN）相结合的框架。在预训练期间，涉及两个主要模块：生成器 $\mathcal{G}$（编解码器模型）和判别器。生成器 $\mathcal{G}$ 由编码器$\mathcal{E}$、量化器 $\mathcal{Q}$ 和码本 $\mathcal{B}$以及解码器 $\mathcal{D}$组成。预训练完成后，仅使用生成器 $\mathcal{G}$。

具体而言，给定输入波形 $Y = [y_1, y_2, \ldots, y_T]$，其中每个 $y_t \in \mathbb{R}$ 和 T表示其长度，我们首先将 Y输入 $\mathcal{E}$ 和 $\mathcal{Q}$，以获得离散表示$\mathbf{C}$：

$$\mathbf{C} = \mathcal{Q}(\mathcal{E}(Y), \mathcal{B}). \tag{1}$$

其中，$\mathbf{C} \in \mathbb{Z}^{T \times L}$ 是一个二维矩阵，T 表示压缩序列长度，L 表示所使用的码本数量。接下来，我们通过将 $\mathcal{D}$与 $\mathcal{B}$ 结合应用于$\mathbf{C}$，来重建波形 Y：

$$\hat{Y} = \mathcal{D}(\mathbf{C}, \mathcal{B}). \tag{2}$$

判别器模块 $\mathcal{D}$ 由基于短时傅里叶变换（STFT）的判别器和基于波形的判别器组成。这些判别器旨在区分 Y 与 $\hat{Y}$，从而提升 Y 的感知质量。

训练损失。我们使用生成对抗网络框架训练模型，总训练损失由三个部分组成：重建损失、对抗损失和量化损失（Zeghi- dour 等人，2022年；Défossez 等人，2023；Kumar 等人，2023；Zhang 等人，2024年；Yang 等人，2023；Wu 等人，2024c；Mousavi 等人，2025a；Arora 等人，2025）。具体而言，重建损失结合了时域损失和多尺度梅尔频谱图损失。对抗损失包括生成器损失、判别器损失和特征匹配损失。对于量化损失，我们在完整量化器输出和各个量化级别上均采用承诺损失。这些损失的详细公式和数学表达式可见于 ESPnet-Codec 论文（Shi 等人，2024a）。我们将在终稿版本中发布预训练的神经音频编解码器，并补充实现细节。

模型架构。 为确保开源可复现性，我们采用 ESPnet-Codec 工具包（Shi 等人，2024a），使用 Sound-Stream 架构（Zeghidour 等人，2022年）及其默认配置。具体而言，$\mathcal{E}$ 和 $\mathcal{D}$ 基于 SEANet（Li 等人，2021年），主要由卷积模块组成，以提升计算效率。我们的量化器 $\mathcal{Q}$ 采用残差向量量化（RVQ）（Razavi 等人，2019年），由 32 个向量量化模块组成。为提高效率，我们遵循以往的实践（Wang 等人，2023；Chen 等人，2024b），在实验中仅使用前八个模块的输出。我们使用基于卷积波形的判别器和基于 STFT 的判别器。

2.2 预训练数据覆盖范围

为了解答第1节中提出的三个问题，我们使用三种不同的预训练数据覆盖范围对神经音频编解码器进行了预训练：英语、多语言和多语言+音频。具体而言，英语数据覆盖范围仅包含英语语音数据；多语言数据覆盖范围包含多种语言的语音数据；而多语言+音频数据覆盖范围则进一步纳入多语言语音数据和非语音音频数据。请注意，这些配置所用训练数据的总量各不相同（第3.1节）。

2.3 评估标准

我们使用两个主要标准评估所有神经音频编解码器：信号重建质量和下游任务应用。请注意，在所有评估实验中，我们始终保持神经音频编解码器参数不变。

信号重建的表述。神经音频编解码器的信号重建能力反映了其在离散化后准确还原原始波形的能力。我们评估语音和非语音领域中重建波形的质量。此重建过程遵循公式（1）和公式（2）中的表述。具体而言，我们按照公式（2）评估其重建保真度。

下游任务应用。 除了信号级评估外，评估这些神经音频编解码器在真实任务中的表现也至关重要。因此，我们在文本转语音（TTS）任务中评估每种编解码器，该任务是应用最广泛的任务之一——

应用（Wang等，2023；Chen等，2024b；Zhang等，2023）中的一种神经音频编解码器。

在文本转语音（TTS）中，每个语音波形 Y都与相应的音素序列 P及其从神经音频编解码器提取的真实表示 $\mathbf{C}$配对。形式化地，令 $P = [p_1, p_2, \ldots, p_M], \quad p_t \in \mathbb{P}$，其中$M$是音素序列的长度，$\mathbb{P}$是所有音素的集合。TTS 模型表示为

$$\hat{\mathbf{C}} = \text{TTS}(P), \tag{3}$$

ˆ其中 $\mathbf{C}$ 应近似于 $\mathbf{C}$。接下来，我们将ˆ解码器 $\mathcal{D}$和码本 $\mathcal{B}$应用于 $\mathbf{C}$（见式 2）ˆ，以合成波形 Y_{tts}:

$$\hat{Y}_{\text{tts}} = \mathcal{D}(\hat{\mathbf{C}}, \mathcal{B}). \tag{4}$$

ˆ Y_{tts} 的质量反映了 TTS 模型的性能。

3 实验设置

3.1 数据集

预训练数据覆盖范围。 对于第2.2节提到的预训练数据覆盖范围，我们使用了 MLS 数据集（Pratap 等人，2020年）和 AudioSet（Gemmeke 等人，2017）语料库。英语 覆盖范围仅包含 MLS 的英语子集，而 多语言 覆盖范围则包含 MLS 中涵盖 8 种语言的全部语音数据。最后，多语言+音频覆盖范围将所有 MLS 数据与 AudioSet 相结合。

语音信号重建与文本转语音。 在语音信号重建和文本转语音实验中，我们使用 LJSpeech（Ito 等，2017年）和 CSS10（Park 等，2019年）数据集，这两个数据集都是常用的单说话人文本转语音数据集。LJSpeech 是一个干净的英语数据集，包含约 24 小时的语音；CSS10 则是一个涵盖十种不同语言的多语言数据集。在本研究中，我们仅使用了 CSS10 的德语和西班牙语子集，分别提供约 16 小时和 24 小时的数据。所有波形文件均降采样至 16 kHz，以适配我们的神经音频编解码器。我们报告相应测试划分上的性能。

非语音信号重建。 在非语音信号重建实验中，我们使用 AudioSet（Gemmeke 等人，2017）测试集，其中包含 17,142 个最长 10 秒的音频片段，涵盖环境声音、音乐、动物发声等。请注意，此测试集未用于我们的 NAC 预训练。所有文件均降采样至 16 kHz，多声道录音则转换为单声道录音。

3.2 信号重建设置

我们使用第 2.3 节所述流程评估语音和非语音数据集上的信号重建性能（见第 3.1节）。由于我们的神经音频编解码器是从头开始训练的，因此我们在语音重建实验中提供成熟声码器基线的结果作为参考。具体而言，这些基线会将每个波形转换为梅尔频谱图，然后应用声码器还原波形。声码器可作为上限，因为这一过程不涉及离散化。此外，我们还纳入真实值（GT）评估结果作为参考。

对于声码器基线，我们采用 HiFi- GAN（Kong 等人，2020年），因为它在现有声码器中表现突出。我们采用其官方配置，但调整跳跃长度，以适应 16 千赫的采样率。我们分别使用第 3.1 节所述的英语、德语和西班牙语数据集训练 HiFi-GAN 模型。

3.3 下游应用设置

我们在文本转语音任务上评估每种神经音频编解码器（参见第 2.3 节）。在文本转语音模型架构方面，我们采用 FastSpeech2（Ren 等人，2021年），这是一种广为人知的非自回归端到端文本转语音框架。我们对 FastSpeech2 进行修改，使其预测神经音频编解码器，而非梅尔频谱图。具体而言，我们的实现分别预测八个编解码器流。我们的设置遵循 ESPnet-Codec 中的非自回归文本转语音实现，并采用其结果作为参考。请注意，我们在文本转语音训练中不包含说话人嵌入，因为我们已在语音重建实验中验证了神经音频编解码器的说话人保留能力（参见第 3.2 节）。

作为基线，我们训练了一个连续表征的文本转语音模型，由一个 FastSpeech2 文本到梅尔频谱图模型和一个 HiFi-GAN 梅尔频谱图到波形声码器组成。我们采用 ESPnet（Watanabe 等人，2018）中 FastSpeech2 的实现，使用其官方配置，并采用第3.2节所述的声码器。这些连续文本转语音基线被视为上限，因为它们不涉及任何量化过程。

表1：使用不同数据覆盖范围进行预训练的神经音频编解码器的语音重建结果。粗体数字表示神经音频编解码器中的最佳结果。

数据覆盖范围	侵入式				非侵入式			感知		最佳/最差
	MCD (↓)	基频相关系数 (↑)	PESQ (↑)	S-BERT (↑)	DNSMOS (↑)	SHEET-SSQA (↑)	UTMOS (↑)	SPK-SIM (↑)	词错误率 (%) (↓)	
英语										
英语	**4.43**	0.68	**2.85**	**0.97**	3.14	4.21	**4.05**	**0.91**	3.36	5 / 0
多语言	4.95	0.67	2.63	0.96	**3.16**	4.15	3.97	0.86	**3.21**	2 / 6
多语言+音频	4.63	**0.70**	2.77	0.96	3.13	**4.24**	3.91	0.90	3.38	2 / 4
声码器	3.76	0.71	3.31	0.98	3.15	4.43	4.30	0.93	3.33	–
GT	–	–	–	–	3.17	4.52	4.44	–	3.31	–
西班牙语										
英语	4.83	**0.44**	**2.84**	**0.98**	**3.27**	**4.14**	**2.68**	**0.93**	**1.94**	8 / 0
多语言	5.31	0.42	2.51	0.91	**3.27**	3.75	2.48	0.87	2.34	1 / 8
多语言+音频	**4.65**	0.43	2.81	0.93	3.25	4.07	2.63	0.92	2.31	1 / 1
声码器	3.24	0.49	2.98	0.97	3.29	4.44	2.99	0.90	1.01	–
GT	–	–	–	–	3.30	4.51	3.19	–	2.66	–
德语										
英语	**4.60**	0.63	2.71	**0.93**	**3.30**	**4.11**	**3.02**	**0.90**	**4.71**	7 / 0
多语言	4.93	0.62	2.49	**0.93**	3.21	3.93	2.79	0.88	6.17	1 / 8
多语言+音频	4.66	**0.64**	**2.72**	**0.93**	3.18	4.06	2.95	0.89	4.83	3 / 1
声码器	3.88	0.70	3.08	0.96	3.24	4.29	3.31	0.98	3.30	–
GT	–	–	–	–	3.26	4.48	3.60	–	4.60	–

3.4 指标

为全面评估神经音频编解码器，我们使用 VERSA 工具包（Shi 等人，2024年c），该工具包提供了丰富多样的评估指标。遵循 ESPnet-Codec 的做法，我们纳入以下指标：

侵入式（全参考）。对于侵入式指标，我们采用六种基于参考信号的评估方法：用于衡量频谱失真的梅尔倒谱距离（MCD）（Kubichek，1993），用于衡量音高一致性的 F0 皮尔逊相关系数（F0-CORR）（Hayashi 等人，2021年），用于评估语音感知质量的感知语音质量评估（PESQ）（Rix 等人，2001年），用于评估语音语义相似度的 S-BERT（Saeki 等人，2024年），用于衡量信号保真度的尺度不变信号失真比（CI-SDR）（Boeddeker 等人，2021年），以及用于感知质量建模的 VISQOL（Chinen 等人，2020年）。所有这些指标在评估期间都需要获取真实参考信号。

非侵入式（无参考）。对于非侵入式指标，我们报告深度噪声抑制平均意见分（DNSMOS）（Reddy 等，2021年）、东京大学 SaruLab 面向语音质量挑战赛的系统（UTMOS）（Saeki 等，2022年），以及 SHEET 工具包中的主观语音质量评估（SHEET-SSQA）（Huang 等，2024年）。这些指标无需参考信号即可估计主观质量分数，反映噪声抑制程度或语音感知整体质量。

感知。对于感知指标，我们使用 Whisper-Large（Radford 等，2022年）报告词错误率（词错误率），并使用来自 ESPnet-SPK 的说话人嵌入提取器 [1]（weon Jung 等，2024年）。这些指标评估波形的可懂度和说话人保持能力。

我们尽可能纳入更多指标，以便从语音质量和可懂度等多个方面全面评估神经音频编解码器的性能。

4 结果与分析

4.1 结果

信号重建——语音。表1展示了语音数据集测试集上的信号重建结果（参见第3.1节）。表格分为三个部分，分别对应评估所用数据的语言。每一行都用颜色标示神经音频编解码器的预训练覆盖范围：红色表示英语，橙色表示多语言，蓝色表示多语言+音频；声码器和真实值则不使用任何颜色。为便于比较，我们为每个神经音频编解码器计算“最佳 / 最差”指标，表示其在每种语言中表现最佳 / 最差的指标数量。不同语言的结果不可比较。

[1] espnet/voxcelebs12_rawnet3

表2：使用不同数据覆盖范围进行预训练的神经音频编解码器的非语音重建结果。粗体数字表示神经音频编解码器中的最佳结果。

数据覆盖范围	梅尔倒谱距离（↓）	尺度不变信号失真比（↑）	VISQOL（↑）	最佳 / 最差
英语	7.32	-10.25	4.21	0 / 1
多语言	7.56	-10.15	4.19	0 / 2
多语言+音频	**5.53**	**-7.31**	**4.37**	3 / 0

表3：使用不同数据覆盖范围预训练的神经音频编解码器的文本转语音结果。粗体数字表示神经音频编解码器中的最佳结果。

数据覆盖范围	词错误率 (%) (↓)	UTMOS(↑)	DNSMOS (↑)	SHEET-SSQA (↑)	最佳 / 最差
英语					
英语	**4.33**	2.36	2.36	3.06	1 / 3
多语言	4.83	**2.84**	**2.72**	**3.26**	3 / 1
多语言+音频	4.71	2.47	2.44	3.11	0 / 0
连续文本转语音	3.63	4.23	3.06	4.15	–
ESPnet-Codec (Shi 等人，2024a)	4.70	1.84	–	–	–
西班牙语					
英语	**4.04**	1.75	2.95	2.65	1 / 1
多语言	6.89	**1.95**	**3.07**	**2.88**	3 / 1
多语言+音频	6.30	1.64	2.77	2.67	0 / 2
连续文本转语音	4.72	3.40	3.18	4.58	–
德语					
英语	**13.30**	**1.51**	2.16	2.29	2 / 0
多语言	16.13	1.49	**2.36**	2.17	1 / 3
多语言+音频	13.55	**1.51**	2.05	**2.38**	2 / 1
连续文本转语音	11.21	3.35	2.60	4.04	–

表1 表明，使用英语数据覆盖范围进行预训练的神经音频编解码器，在语音重建任务中的表现通常优于使用其他覆盖范围进行预训练的神经音频编解码器。具体而言，使用英语数据覆盖范围进行预训练的神经音频编解码器，在英语、西班牙语和德语上分别有 5、8 和 7 项指标（共 9 项）取得最佳表现。

信号重建——非语音。在表 2 中，我们展示了 AudioSet（Gemmeke 等人，2017）测试集上的信号重建结果（参见第 3.1 节）。遵循 ESPnet-Codec（Shi 等人，2024a），我们报告三项信号级指标：梅尔倒谱距离、尺度不变信号失真比和 VISQOL。

从表 2 中可以看出，使用 多语言+音频数据训练的 NAC，在三项指标上均优于未使用非语音数据训练的模型。请注意，重建质量较差时，尺度不变信号失真比可能为负值；不过，相对提升（例如从 -10 dB 提升至 -7 dB）仍表明质量有明显改善。

下游应用。 在表3中，我们展示了各神经音频编解码器的文本转语音结果。报告的指标包括词错误率（用于评估可懂度）以及三种用于评估合成波形质量的非侵入式指标。

就词错误率（WER）而言，采用英语数据覆盖范围进行预训练的神经音频编解码器取得了最佳结果，多语言+音频紧随其后。对于非侵入式质量指标，采用多语言数据覆盖范围进行预训练的神经音频编解码器总体表现最佳。值得注意的是，采用英语进行预训练的神经音频编解码器在西班牙语和德语中排名第二，而这两种语言在其预训练期间均属于未见语言。需要注意的是，我们结果中的 UTMOS 得分仍然相对较低，尤其是在西班牙语和德语中。我们认为，这一现象源于：在小型数据集上使用神经音频编解码器训练文本转语音（TTS）模型存在固有限制，以及 UTMOS 模型的跨语言泛化能力有限。先前的研究（Shi 等人，2024a；Mousavi 等人，2024年）也报告了类似的较低 UTMOS 得分，表明这一趋势与现有研究结果一致，并非异常现象。

4.2 分析

问题一： 根据表1 和表3，神经音频编解码器对预训练期间未见的语言表现出很强的泛化能力。对于西班牙语和德语，采用英语数据进行预训练的模型（最佳 / 最差：表1中的 15 / 0；表3中的 3 / 1）总体上取得了比多语言模型更多的最佳次数（最佳 / 最差：表1中的 2 / 16；表3中的 4 / 4）。这与语音自监督学习（SSL）模型（Shi 等人，2023，2024b；Chou 等人，2023；Wang 等人，2025）形成对比，后者通常会在预训练期间未见的语言上性能下降。这些结果凸显了神经音频编解码器的语言鲁棒性及其在更广泛语言学应用中的潜力。

问题二. 如表 2所示，仅用语音预训练的神经音频编解码器（最佳 / 最差：0 / 2）的表现不如用语音和非语音数据预训练的神经音频编解码器（最佳 / 最差：3 / 0）。这符合预期，因为语音与非语音信号在信号层面存在显著差异。

问题三. 如表 1、表 2和表3所示，在预训练期间纳入非语音数据可以提升非语音任务的表现（见上文问题二），同时保持语音任务上的表现处于可接受水平（表1中的最佳 / 最差：6 / 6；表 3中的最佳 / 最差：2 / 3）。这一发现有些出乎意料，因为我们最初假设纳入非语音数据可能会带来取舍。基于这些结果，我们建议在神经音频编解码器预训练中纳入非语音数据，以增强模型在语音与非语音领域之间的泛化能力。

5 结论

本文探讨了关于神经音频编解码器（NACs）的三个关键但尚未充分研究的问题。首先，我们使用对齐的模型配置和三种经过精心筛选的预训练数据覆盖范围对 NACs 进行预训练，以确保比较公平。随后，我们使用 11 项互补指标评估其在信号重建和下游任务中的表现，进行全面评估。结果表明：（i）NACs 能够泛化到预训练期间未见过的语言；（ii）仅使用语音数据进行预训练的 NACs 在非语音任务中表现不佳；（iii）在预训练期间纳入非语音数据，能够提升其在非语音任务中的表现，同时保持在语音任务中的表现相当。这些发现表明，NACs 可以可靠地应用于多语言设置，且面向非语音的预训练对于非语音应用至关重要，而同时使用语音和非语音数据进行联合预训练，则能带来最稳健的整体表现。

6 局限性

神经音频编解码器架构。 在本研究中，我们采用 SoundStream 架构（Zeghidour 等人，2022年），并使用其默认配置。未来，我们计划将研究拓展至其他神经音频编解码器架构，例如 EnCodec（Défossez 等人，2023）、DAC（Kumar 等人，2023）和 HiFi-Codec（Yang 等人，2023），以进一步验证研究结果的普适性。

评估标准。 我们从两个角度评估神经音频编解码器。对于信号重建，我们同时考察语音与非语音信号。然而，对于下游应用，我们仅关注文本转语音，因为这本质上是一项仅语音任务。未来，我们计划纳入更多非语音下游任务，例如音乐生成以及与环境声音相关的任务，从而提供更全面的评估。

文本转语音架构。 我们计划将文本转语音实验扩展到更强大的自回归模型，例如 VALL-E（Wang 等人，2023；Chen 等人，2024b；Zhang 等人，2023）。虽然这些模型的计算开销更大，但它们可能具备更强的生成能力，并有助于进一步了解神经音频编解码器与先进文本转语音架构之间的交互。

参考文献

Siddhant Arora 和另外 1 位作者。2025年。《口语语言模型的发展图景：全面综述》。*arXiv* 预印本 *arXiv:2504.08528*。

Christoph Boeddeker 等人。2021年。面向多通道混响语音分离的卷积传递函数不变 SDR 训练准则。收录于*ICASSP 2021 - 2021*年 *IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，第 8428–8432 页。

Xuankai Chang 等人。2024年。使用离散语音单元探索语音识别、翻译与理解：一项比较研究。收录于*ICASSP 2024 - 2024*年 *IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*，第 11481–11485 页。

Jianyi Chen 等人。2024a。PyramidCodec：用于长篇音乐生成的分层编解码器

音频领域。收录于 计算语言学协会 *EMNLP 2024* 论文集，第 4253–4263 页。计算语言学协会。

Sanyuan Chen 和其他 1 位作者。2024年。 VALL-E 2：神经编解码器语言模型是达到人类水平的零样本文本转语音合成器。 预印本，arXiv:2406.05370。

Michael Chinen 和其他 1 位作者。2020年。 ViSQOL v3：一种开源、可用于生产的客观语音与音频指标。载于*2020*年第十二届多媒体体验质量国际会议（*QoMEX*），第 1–6 页。

Yi-Hui Chou 和其他 1 位作者。2023年。 在台湾闽南语语料库上评估自监督语音模型。载于 *2023*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–7 页。

Alexandre Défossez 和其他 1 位作者。2023年。 高保真神经音频压缩。 《机器学习研究汇刊》。特色认证、可复现性认证。

Jort F. Gemmeke 和其他 1 位作者。2017年。 Audio Set：用于音频事件的本体和人工标注数据集。载于 *2017*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 776–780 页。

Tomoki Hayashi 等人。2021年。 ESPnet2-TTS：拓展文本转语音研究的边界。 预印本，arXiv:2110.07840。Wen-Chin Huang 等人。2024年。 MOS-Bench：评测主观语音质量评估模型的泛化能力。 预印本，arXiv:2411.03715。Keith Ito 等人。2017年。LJ 语音数据集。 https://keithito.com/LJ-Speech-Dataset/。

Zeqian Ju 等人。2024年。NaturalSpeech 3：采用因子分解编解码器和扩散模型的零样本语音合成。载于 第*41*届国际机器学习会议论文集，ICML'24。JMLR.org。Jungil Kong 等人。2020年。 HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。载于 神经信息处理系统进展，第33卷，第17022–17033页。Curran Associates, Inc.

Robert F. Kubichek。1993年。用于客观语音质量评估的梅尔倒谱距离度量。 *IEEE*太平洋地区通信、计算机与信号处理会议论文集，第1卷：125–128页。Rithesh Kumar 等人。2023年。采用改进型 RVQGAN 的高保真音频压缩。载于 神经信息处理系统进展，第36卷，第27980–27993页。Curran Associates, Inc.

李佳琦 等人。2024年。 研究神经音频编解码器在基于语音语言模型的语音生成中的应用。载于 *2024*年 *IEEE* 口语技术研讨会（*SLT*），第554–561页。

李云鹏等。2021年。实时语音频带扩展。载于 *ICASSP 2021 - 2021*年 *IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 691–695 页。

Pooneh Mousavi 等。2024年。 DASB - 离散音频与语音基准。预印本，arXiv:2406.14294。

Pooneh Mousavi 等。2025a。离散音频令牌：不止是一篇综述！*arXiv* 预印本 *arXiv:2506.10274*。

Pooneh Mousavi 等。2025b。离散音频令牌：不止是一篇综述！预印本，arXiv:2506.10274。

Joonyong Park 等。2025年。分析神经音频编解码器的语言。预印本，arXiv:2509.01390。

Kyubyong Park 等 1 人。2019年。CSS10：一个涵盖 10 种语言的单说话人语音数据集合集。载于*Interspeech 2019*，第 1566–1570 页。

Vineel Pratap 等人。2020年。 MLS：用于语音研究的大规模多语言数据集。载于 *Interspeech 2020*，第 2757–2761 页。Alec Radford 等人。2022年。 通过大规模弱监督实现稳健的语音识别。预印本，arXiv:2212.04356。Ali Razavi 等人。2019年。使用 VQ-VAE-2 生成多样化的高保真图像。载于 神经信息处理系统进展，第 32 卷。Curran Associates, Inc. Chandan K A Reddy 等人。2021年。 DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标。载于 *ICASSP 2021 - 2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6493–6497 页。任怡等人。2021年。 FastSpeech 2：快速且高质量的端到端文本转语音。载于 国际学习表征会议。A.W. Rix 等人。2001年。 语音质量感知评估（PESQ）——一种用于电话网络和编解码器语音质量评估的新方法。载于 *2001*年 *IEEE* 国际声学、语音与信号处理会议。会议论文集（目录编号：*01CH37221*），第 2 卷，第 749–752 页，第 2 卷。Takaaki Saeki 等人。2022年。 UTMOS：UTokyo-SaruLab 的 VoiceMOS挑战赛2022 系统。载于*Interspeech 2022*，第 4521–4525 页。

Takaaki Saeki 等人。2024年。 SpeechBERTScore：利用自然语言处理评估指标进行参考感知的语音生成自动评估。载于*Interspeech2024*，第 4943–4947 页。

Jiatong Shi 和另外 1 位作者。2023年。 ML-SUPERB：多语言语音通用性能基准。载于*Interspeech2023*，第 884–888 页。

Jiatong Shi 和另外 1 位作者。2024a。 ESPnet-Codec：面向音频、音乐和语音的神经编解码器综合训练与评估。 预印本，arXiv:2409.15897。

Jiatong Shi 和另外 1 位作者。2024b。 ML-SUPERB 2.0：在建模约束、语言和数据集方面对多语言语音模型进行基准测试。载于 *Interspeech2024*，第 1230–1234 页。

Jiatong Shi 等人。2024c。 VERSA：用于语音、音频和音乐的多功能评估工具包。预印本，arXiv:2412.17667。Chengyi Wang 等人。2023。 神经编解码器语言模型是零样本文本转语音合成器。 预印本，arXiv:2301.02111。Shih-Heng Wang 等人。2025。 如何学习一门新语言？针对低资源场景下自监督学习模型适配未见语言的高效方案。 预印本，arXiv:2411.18217。Shinji Watanabe 等人。2018。 ESPnet：端到端语音处理工具包。载于*Interspeech* 论文集，第 2207–2211 页。Jee weon Jung 等人。2024。 ESPnet-SPK：完整流程的说话人嵌入工具包，包含可复现方案、自监督前端和现成模型。载于 *Interspeech2024*，第 4278–4282 页。Haibin Wu 等人。2024a。 “codec-SUPERB：对声音编解码器模型的深入分析”。载于计算语言学协会 *ACL 2024* 论文集，第 10330–10348 页，泰国曼谷及线上会议。计算语言学协会。Haibin Wu 等人。2024b。Codec-superb@ SLT 2024：神经音频编解码器模型的轻量级基准。载于 *2024 IEEE* 语音语言技术研讨会（*SLT*），第 570–577 页。IEEE。Haibin Wu 等人。2024c。Ts3-codec：基于 Transformer 的简单流式单编解码器。*arXiv* 预印本 *arXiv:2411.18803*。Yaoxun Xu 等人。2025。 Mucodec：用于音乐生成的超低码率音乐编解码器。载于 第 *33* 届 *ACM* 国际多媒体会议论文集，MM ’25，第 689–698 页。计算机协会。Dongchao Yang 等人。2023。HiFi-Codec：用于高保真音频编解码器的组残差向量量化。*ArXiv*，abs/2305.02765。

Yifan Yang 等人。2024。 迈向通用语音离散标记：自动语音识别和文本转语音的案例研究。载于*ICASSP2024*，第 10401–10405 页。

Neil Zeghidour 和另外 1 位作者。2022年。 SoundStream：端到端神经音频编解码器。*IEEE/ACM* 音频、语音与语言处理汇刊，30:495–507。

Xin Zhang 和另外 1 位作者。2024年。 SpeechTokenizer：面向语音语言模型的统一语音分词器。载于 第十二届国际学习表征会议。

Ziqiang Zhang 等人，2023 年。用自己的声音说外语：跨语言神经编解码器语言建模。预印本，arXiv:2303.03926。