RO-N3WS：借助多样化的罗马尼亚语语音基准测试，提升低资源自动语音识别的泛化能力

Alexandra Diaconu、M˘ad˘alina Vînaga、Bogdan Alexe
布加勒斯特大学计算机科学系 {madalina.vinaga,
alexandra.diaconu}@s.unibuc.ro
bogdan.alexe@fmi.unibuc.ro

arXiv:2603.02368v1 [cs.CL] 2 Mar 2026

摘要

我们推出了 RO-N3WS，这是一个罗马尼亚语语音基准数据集，旨在提升自动语音识别（ASR）在低资源和分布外（OOD）条件下的泛化能力。RO-N3WS 包含超过 126 小时的转写音频，素材涵盖广播新闻、文学有声读物、电影对白、儿童故事和播客对话语音。如此丰富的多样性有助于在风格迥异的领域中进行稳健训练和微调。我们在零样本和微调设置下评估了多个先进的自动语音识别系统（Whisper、Wav2Vec 2.0），并利用表现力文本转语音模型生成的合成数据开展受控比较。结果表明，即使仅使用 RO-N3WS 中少量真实语音进行微调，也能显著降低词错误率，优于零样本基线。我们将发布所有模型、脚本和数据划分，以支持多语言自动语音识别、领域适配和轻量化部署方面的可复现研究。

1 引言

自动语音识别（ASR）已成为现代机器学习流水线的核心组成部分，可用于媒体内容转录、无障碍服务和下游语言模型训练 Kanade 等人（2015）；Pelloin 等人（2022）；Kuhn 等人（2025）。尽管语音转文本系统取得了显著进展，但大多数高性能模型都侧重于英语或其他广泛使用的语言，导致许多语言资源匮乏。罗马尼亚语便是其中之一。现有数据集，如 Vox Populi Wang 等人（2021）、Common Voice Ardila 等人（2020）、FLEURS Conneau 等人（2023）以及志愿者收集的语料库 Stan 等人（2017）；Georgescu 等人（2018；2020）；Ungureanu & Dascalu（2024），提供了重要资源，但往往受限于领域（议会演讲）、录音风格（朗读语音）或语言变化有限等因素。虽然 Whisper Radford 等人（2023）和 Wav2Vec 2.0 Baevski 等人（2020）等预训练多语言模型具有出色的零样本性能，但在此类场景下，它们应对领域偏移或富有语言表现力的语音的鲁棒性仍缺乏充分研究。

本文介绍了*RO-N3WS*，这是一个罗马尼亚语语音基准测试数据集，旨在支持对自动语音识别模型进行细粒度评估与适配。该数据集包含*126*小时来自罗马尼亚两大广播新闻来源、经人工筛选和转写的语音。数据集分为：（i）一个域内子集 （105小时），包含演播室新闻报道和现场新闻报道；（ii）一个分布外（*OOD*）子集 （21小时），包含有声读物、罗马尼亚电影、儿童故事和播客中的富有表现力的自发言语。

RO-N3WS 涵盖了丰富的语言和声学特征，包括韵律变化、命名实体、偏离音位拼写的情况、变音符号以及领域专用术语。我们在零样本和微调两种设置下，对包括 Whisper 和 Wav2Vec 2.0 在内的多种先进模型进行了基准测试。实验表明，即使仅使用 RO-N3WS 的小规模子集进行微调，也能显著提升域内和分布外性能。

表1：多语言及罗马尼亚语专用语音数据集概览，重点展示罗马尼亚语覆盖范围、说话人多样性和录音类型。

数据集	总计（小时）	罗马尼亚语（小时）	罗马尼亚语说话人	Type
Common Voice	7,335	47	428	Read, 众包
VoxPopuli	1,791	89	18	脚本化、正式
FLEURS	1,400	12	19	朗读，已对齐
SWARA	21	21	17	朗读，语音平衡
RODigits	38	38	11	朗读，数字
RSC	100	100	164	Read, 孤立词
Echo	378	378	343	Read, 多领域
RO-N3WS（我们的工作）	126	126	100+	朗读、即兴、新闻 + 分布外

我们的贡献主要有三点：(i) 我们推出了 RO-N3WS，这是一个罗马尼亚语自动语音识别基准，包含经过精心对齐的域内和分布外子集，可用于系统评估领域鲁棒性和泛化能力；(ii) 我们对多种性能强劲的预训练自动语音识别模型进行了基准测试，并证明在 RO-N3WS 上进行有针对性的微调，可显著提升模型在域内和分布外场景中的性能；(iii) 我们比较了使用自然语音与合成语音进行模型适配的效果，结果表明，尽管富有表现力的文本转语音（TTS）可能有所助益，但真实录音始终能提供更强的监督信号。

RO-N3WS 为低资源自动语音识别、领域适配和以数据为中心的评估研究提供了一个真实且灵活的测试平台。完整语料库和微调模型将在论文获批后公开发布。

2 相关工作

近年来，多语言自动语音识别的发展很大程度上得益于强大的预训练模型，例如 Whisper（Radford 等人，2023）和 Wav2Vec 2.0（Baevski 等人，2020）；这些模型已展现出在多种语言上的出色零样本性能。然而，它们对领域偏移、风格变化和低资源适配的鲁棒性仍是一个活跃的研究领域（Radford 等人，2023）。在此背景下，使用真实且语言多样的语料库训练和评估模型，对于理解模型局限性和提升泛化能力都至关重要。

对于罗马尼亚语，大多数现有语料库要么是多语言的，要么涵盖范围狭窄（见表 1）。Common Voice Ardila et al. (2020) 提供了 47 小时由用户提交的朗读语音，但质量参差不齐。VoxPopuli Wang et al. (2021) 和 FLEURS Conneau et al. (2023) 包含罗马尼亚语子集（分别为 89 小时和 12 小时），主要用于跨语言评估。然而，这些数据集缺少富有表现力的内容或对话内容，因此限制了它们在研究真实场景中的适应能力方面的用途。

目前已有若干仅包含罗马尼亚语的语料库，包括 SWARA Stan et al. (2017)、RODigits Georgescu et al. (2018) 和 RSC Georgescu et al. (2020)，但它们仅限于录音室条件、朗读语音或狭窄领域（例如数字、孤立词）。近期推出的 Echo 数据集 Ungureanu & Dascalu (2024) 更具雄心，涵盖 14 个领域，共计 378 小时。不过，该数据集仍采用脚本化内容，并通过众包方式收集，且自发语音或情感变化丰富的语音占比不足。

相比之下，RO-N3WS 提供了一个可直接用于基准测试的数据集，包含域内和分布外划分。其域内部分收录了来自电视广播的真实世界罗马尼亚语语音，包括直播访谈和多种说话风格。分布外子集则通过有声读物、儿童故事、电影对白和对话类播客中的富有表现力的语音，展现了风格多样性，这对于评估预训练模型的细粒度泛化能力十分重要。这种结构支持使用 Whisper、Wav2Vec 2.0 及相关模型开展受控的适配实验。

此外，RO-N3WS 涵盖了以往语料库中涉及不足的语言现象，例如命名实体、韵律变化、偏离音位拼写的情况，以及情绪充沛的

表2：罗马尼亚语语音数据集中的命名实体密度（每100个词元包含的实体数）。RO-N3WS的整体密度最高，彰显了其词汇丰富性。

标签	FLEURS	Common Voice	Vox Populi	Echo	RO-N3WS	RO-N3WS OOD
人物	2.06	1.35	1.66	2.89	**3.56**	2.76
LOC	**0.25**	0.07	0.17	0.08	0.12	0.04
NAT_REL_POL	0.70	0.31	**0.94**	0.72	0.42	0.16
产品	1.24	0.28	0.26	1.04	**1.44**	0.87
日期时间	0.93	0.45	0.49	0.83	**1.46**	0.50
数字 –值	0.81	0.45	0.67	1.23	**1.61**	0.46
序数	**0.27**	0.18	0.19	0.23	0.22	0.14
货币	0.08	0.02	0.06	0.09	**0.17**	0.03
GPE	0.75	0.23	**1.31**	0.54	0.90	0.17
设施	0.62	0.13	0.13	0.54	**0.89**	0.33
数量	**0.19**	0.01	0.01	0.05	0.14	0.02
组织	0.26	0.24	**0.54**	0.32	0.36	0.09
语言	**0.08**	0.01	0.01	0.02	0.01	0.02
WORK_OF_ART	**0.02**	0.01	0.01	0.01	0.01	0.01
事件	**0.04**	0.03	0.01	0.02	0.03	0.01
时期	**0.03**	0.01	0.01	0.01	0.01	0.01
总计	8.33	3.78	6.47	8.62	**11.35**	5.61

表达方式；这些对于以数据为中心的自动语音识别研究，以及在低资源场景中开发鲁棒自动语音识别系统至关重要。因此，它是研究领域不匹配、合成与自然监督，以及有针对性的微调策略下模型行为的重要资源。第3节提供了各语料库在命名实体使用情况和韵律表现力方面的定量比较。

3 数据集分析

为展现 RO-N3WS 在语言和韵律方面的丰富性，我们进行了两项互补分析：命名实体密度，反映词汇内容和主题具体程度；以及韵律变化，作为情感表现力的替代指标。我们对五个罗马尼亚语语音语料库开展了分析：我们的数据集（RO-N3WS）以及四个广泛使用的基准测试，即 FLEURS、Common Voice、Vox Populi 和 Echo。

命名实体密度。我们使用 spaCy 的 ro_core_news_md 模型（Honnibal 等（2020））进行了详细的命名实体识别（NER）分析，并将实体密度计算为每 100 个词元中的命名实体数量。结果汇总于表 2。RO-N3WS 的总体命名实体密度最高（每 100 个词元中有 11.35 个实体），高于 Echo（8.62）、FLEURS（8.33）、Vox Populi（6.47）和 Common Voice（3.78）。具体而言，RO-N3WS 显示出：（i）每 100 个词元中有 3.56 个人物实体（Echo 为 2.89）；（ii）1.44 个产品实体（FLEURS 为 1.24）和 1.46 个日期时间提及（FLEURS 为 0.93）；（iii）1.61 个数字值（Echo 为 1.23）和 0.17 个货币提及（Echo 为 0.09）。这些数据反映出 RO-N3WS 中事实性语言和实体丰富语言的高度集中，因此该数据集尤其适用于评估自动语音识别系统处理命名实体、数量和时间表达的能力。即使是 RO-N3WS-OOD 子集，其实体密度也达到 5.61，不容忽视；这一数值高于 Common Voice（3.78），且接近 Vox Populi（6.47）。

韵律变化与情感表现力。 我们使用 Parselmouth 库 Jadoul 等（2018），即 Praat Boersma（2001）的 Python 接口，从每个数据集中提取韵律特征：（1）平均基频（F0）：平均声音基频，反映说话人的总体发声音区；（2）音高标准差：音高变化，与表达性调节相关；（3）基频范围：最高与最低基频值之差；范围越大，表明情绪动态越强；（4）平均强度：平均响度，在情绪强烈的话语中通常更高。

表3：罗马尼亚语自动语音识别数据集的韵律统计数据。

数据集	平均基频（Hz）	基频标准差（Hz）	基频范围（Hz）	平均强度（dB）
FLEURS	172.01	60.00	**394.19**	54.37
Common Voice	151.45	37.21	212.08	49.92
Vox Populi	**197.99**	36.49	304.07	61.45
Echo	175.10	46.67	341.19	50.95
RO-N3WS	182.19	39.33	266.32	**63.65**
RO-N3WS-OOD	197.72	**71.92**	370.07	58.19

这些指标可作为声音表现力的定量代理指标。结果见表 3。由于 RO-N3WS 与 RO-N3WS-OOD 在风格和构成上有所不同，我们分别对其进行分析。从表 3 中可以得出若干关键观察：（i）RO-N3WS-OOD 在所有数据集中表现出最高的音高标准差，表明其表达变化性很高，尤其与电影对白和对话类播客等自发言语相关；（ii）RO-N3WS-OOD 的平均基频几乎最高，非常接近 Vox Populi，进一步印证了其富有表现力且生动活泼的特点；（iii）RO-N3WS（域内部分）显示出最高的平均强度，反映出专业广播语音（如新闻主播和记者）典型的发声力度与清晰度。与 Echo 相比，RO-N3WS 提供了互补的韵律特征：（i）其平均基频更高（182.19 Hz 对 175.10 Hz），与记者和主播所使用的广播语音风格一致；（ii）尽管 Echo 表现出更大的基频范围和变化，RO-N3WS 仍保持较高的清晰度与发声力度，这一点由其显著更高的平均强度（63.65 对 50.95 dB）得以证明。

4 RO-N3WS：构建、标注与数据划分

广播语音采集与切分。我们从两个主要电视新闻来源采集口语罗马尼亚语，构建RO-N3WS 语料库：*ProTV* 和*Antena 1*。我们使用基于 UiPath Studio 构建的定制工作流，分别从在线存档 [1] 和 YouTube 频道[2]抓取录制内容。每段视频都会被下载、转换为 .wav 格式，并通过 Whisper 模型 Radford 等人（2023）进行处理。为提高时间精度，我们使用 whisper-timestamped 获取词级时间戳。我们通过贪心方式将连续的完整句子分组，在确保片段时长少于 10 秒的同时提取片段。这既能确保片段适用于训练，也便于人工标注。通过初始抓取流水线，我们从 ProTV 获取了*42,225* 个片段，从 Antena 1 获取了*78,831*个片段，原始广播音频总计*207* 小时 。

数据清洗与质量控制。为确保转写准确度，我们剔除包含重叠语音、音乐、言语不流畅、外语、Whisper幻觉或切分不佳的文件。尤其是 Antena 1 的片段，需要移除重复的宣传内容。约有 *60%* 的 Antena 1 片段被剔除。最终清洗后的数据集包含 ProTV 的*32,319* 个文件（~54.5 小时）和 Antena 1 的*30,073* 个文件（~50.5 小时），共计*105* 小时 经过整理的广播新闻语音。

高质量转录与标注规范。每个音频文件均由一组 15 名受过培训的标注员手动校正并标注。Whisper 生成的初始转录文本作为第一轮假设。标注员遵循以下准则：(i) 补全罗马尼亚语变音符号（ș, t ˘a、â、î）；(ii) 将数字写成文字（例如，*100* 变为*o sut˘a* - 一百）；(iii) 展开口语缩略语；(iv) 命名实体按其读音保留；(v) 更正拼写错误。每个文件均由两名标注员交叉核对。此流程确保转录文本在正字法上正确，且语音准确。附录 A.2 中列出了一些有代表性的标注结果示例及相应的 Whisper 输出。

数据划分与可复现性策略。我们采用20折策略进行分层划分，将数据分为训练集（85%）、验证集（10%）和测试集（5%）。同一部长视频中的所有片段均划分到同一折，以避免说话人或上下文信息泄漏。因此，共有*53,049* 个训练片段（~89.4小时）、*6,219* 个验证片段（~10.4小时）和*3,124* 个测试片段（~5.3小时）。

[1] [https://stirileprotv.ro/video/[2] [https://www.youtube.com/@ObservatorTV

表4：RO-N3WS 数据集中音频文件的详细统计信息。

数据划分 / 来源	文件数	总时长（秒）	最短时长（秒）	最长时长（秒）	平均时长 (s)	标准差 (s)
域内（新闻）						
ProTV新闻	32,319	196,405	3.00	11.98	6.08	1.57
Observator新闻	30,073	182,423	1.72	12.00	6.07	1.57
所有新闻	62,392	378,828	1.72	12.00	6.07	1.57
分布外（OOD）						
有声书	2,475	15,422	1.20	10.14	6.23	1.37
电影	4,723	29,814	3.00	11.94	6.31	1.32
故事	2,821	17,029	0.66	21.46	6.04	2.11
播客	1,723	12,061	0.52	29.18	7.00	2.76
全部分布外	11,742	74,327	0.52	29.18	6.33	1.83

领域偏移下的评估：分布外语音。除了广播新闻录音外，我们还使用在风格、领域和声学条件方面存在显著差异的语音数据扩充数据集，并将其称为分布外（*OOD*）数据。这些录音用作评估材料，用于测试基于域内（新闻）数据训练的自动语音识别模型的泛化能力。按照前文小节所述的相同流程，我们从三个OOD来源收集、清理并标注音频片段：（1）有声读物[3]，包含富有表现力、叙事风格的朗读；（2）著名的罗马尼亚电影[4]，包含不同声学环境中的对话语音；（3）儿童故事[5]，包含面向年轻听众、富有表现力且韵律丰富的叙述；（4）罗马尼亚播客[6]，，提供涵盖广泛主题和多种说话风格的无脚本即兴对话，因此带来自然的犹豫、轮流发言和非正式措辞，而这些特点在更为脚本化的来源中并不存在。这些子集采用相同的Whisper+标注流程进行转录和整理。它们为评估模型对风格和领域变化的鲁棒性提供了宝贵的测试条件。如表4所示，OOD评估集包含约*4.2*小时的有声读物语音、*8.2*小时的电影对白、*4.4*小时的儿童故事讲述，以及*3.3*小时的播客语音。

域内和分布外子集的更多音频时长统计数据及各来源直方图见附录A.1。

5 实验评估

我们在 RO-N3WS 上的两种关键设置下评估了多种先进的语音转文本模型：零样本（不进行微调）和有监督适配（在 RO-N3WS 上进行微调）。为评估转写准确度和领域鲁棒性，我们报告了域内（ProTV 和 Antena1 新闻）及分布外测试集（有声读物、电影、儿童故事和播客）上的词错误率（WER）。

在我们的实验中，我们分别报告 RO-N3WS 测试集内 ProTV 和 Antena1 子集的 WER 分数。这种区分是有意为之：尽管这两个来源都属于同一种广播新闻体裁，但它们代表着不同的语言和声学分布。差异体现在呈现风格、说话人身份、背景条件和编辑内容等方面。将这些子集合并为单一指标会掩盖这些领域特定差异。分别报告各子集的分数，可以让我们更细致地分析模型性能，并更清楚地凸显不同域内条件下的泛化差异。此外，这种数据划分还支持对域内泛化进行受控实验：例如，在一个来源（如 ProTV）上训练，然后

[3] *Cortina* – 波洛最后一案 [4] *Buletin de București, C˘as˘atorie cu repetiție, Toamna bobocilor, Iarna bobocilor, Prim˘avara bobocilor, Liceenii, Toate pânzele sus, Cu mâinile curate, Duelul, Revanșa, Un comisar acuz˘a, Supraviețuitorul, Pistruiatul* [5] https://www.youtube.com/@RomanianFairyTales [6] https://www.youtube.com/@fainsisimplu, https://www.youtube.com/@unpodcastmisto, https://www.youtube.com/@IlincaVandici./podcasts

在另一个来源（如 Antena1）上测试，反之亦然。即使领域保持不变，这也提供了一个用于评估跨分布稳健性的真实场景。

所有实验均在配备 NVIDIA H100 GPU（80GB）、Intel Xeon Platinum 8480+ CPU（112 核）和 2TB 内存的高性能计算环境中执行。Whisper 和 Wav2Vec 2.0 模型的完整训练超参数列于附录 A.3。

5.1 评估的模型

我们在实验中评估了五个语音转文本系统，分为两类：（i）开源、可微调模型，以及（ii）商业黑盒 API。开源模型既在零样本设置下评估，也在我们的 RO-N3WS 数据集上进行监督微调后评估。商业模型通过 API 访问，且仅在零样本模式下评估。

5.1.1 开源模型

Wav2Vec 2.0（针对罗马尼亚语微调的 VoxPopuli）Baevski 等人（2020）是一种自监督模型，使用对比学习和掩码在原始波形上训练，类似于自然语言处理中的掩码语言建模。在我们的实验中，我们使用可通过 Hugging Face 获取的模型检查点 facebook/wav2vec2-base-10k-voxpopuli-ft-ro。[7]

Whisper Radford 等人（2023）是一种多语言编码器-解码器 Transformer，在 680,000 小时的监督数据上训练，涵盖多种任务，包括自动语音识别、语音翻译和语音活动检测。我们评估了若干 Whisper变体，如 Small 和 Large，以及一个在 Echo 数据集上微调的变体。在 *Echo* 数据集上微调的 *Whisper Small* Ungureanu & Dascalu（2024年）是一个公开发布的模型，在约 356 小时先前的罗马尼亚语音频数据与 378 小时 Echo 数据集的组合上训练，共计超过 730 小时带标注的罗马尼亚语语音。该模型基于 small 的 Whisper 架构，并使用带有数字字面展开的规范化转录文本进行微调。该模型被纳入作为强有力的罗马尼亚语专用基线。

5.1.2 商业模型（黑盒 API）

Vatis[8] 是一项以罗马尼亚语转写准确度著称的罗马尼亚语商业自动语音识别服务。其模型在专有数据上训练，并针对生产用途进行了优化。由于该模型未开源，我们使用其公开 API 获取转录文本。

Microsoft Transcribe[9] 是一个生产级语音转文本系统，集成于 Azure 认知服务，并针对跨通用领域的多语言转录进行了优化。

Google Chirp Zhang 等人（2023）正式称为通用语音模型（*USM*），是一个拥有 20 亿参数的大型多语言编码器-解码器系统。该系统使用基于 Conformer 的编码器架构，在数百万小时的音频和数十亿条文本样本上进行训练。

5.2 评估指标

我们将词错误率（*WER*）作为主要评估指标，其计算方式是将预测序列转换为参考序列所需的最少插入、删除和替换次数，再除以参考长度。使用 WER 面临的一项挑战，是语言模型或后处理启发式方法（如逆文本归一化）造成的文本格式差异。例如，数字短语 *"o mie opt sute șaizeci și trei"*/ *"一千八百六十三"* 可能会被某些系统呈现为*"1.863"* 。表 5 展示了这类非语义差异。为减轻特定格式不匹配带来的影响，我们构建了多个参考标注，以涵盖可接受的格式变体。在这一宽松设置下，我们不会惩罚语义等同但句法形式不同的输出。

[7]https://huggingface.co/facebook/wav2vec2-base-10k-voxpopuli-ft-ro[8] https://vatis.tech[9]https://azure.microsoft.com/en-us/products/cognitive-services/ 语音转文本

表 5：尽管输出与参考序列在语义上等价，但由于格式差异，词错误率较高的示例。在计算词错误率之前，会对商业自动语音识别 API 的输出进行后处理，以移除标点、统一大小写并规范空格。

转录（罗马尼亚语）	**WER**
真实标注： 从一月一日起，最短时长工资将为三千列伊税前，而一名员工在˘方面将会收到一千八百六十三列伊 （“从一月一日起，最短时长工资将为三千列伊税前，而在到手后，员工将会收到 *1,863* 列伊”）	0.000
Vatis 输出： 从 1 月 1 日起，最低工资将为税前 3,000 列伊，而员工到手˘的工资将为˘1,863 列伊。	0.387
Microsoft Transcribe 输出： 从 1 月 1 日起，最低工资将为税前 3,000 罗马尼亚列伊，而员工到手˘的工资将应收到˘a 1.863 罗马尼亚列伊˘a。	0.451

Ta表 6：域内和分布外测试集上的零样本词错误率（WER %）。数值越低越好。

模型	**域内（RO-N3WS）**		**分布外（OOD）**			
	ProTV	**Antena1**	**有声读物**	**电影**	**故事**	**播客**
W2V2	27.7	25.6	40.8	75.4	54.1	36.3
Whisp-S	31.6	26.8	40.0	60.0	41.1	31.9
Whisp-L	12.3	5.9	14.8	27.3	10.9	11.7
Whisp-S +Echo	9.4	10.1	18.8	54.1	21.0	21.6
Microsoft Transcribe	2.9	4.8	10.6	31.1	17.6	11.5
Google Chirp (USM)	12.1	11.3	20.2	37.6	22.4	22.4
Vatis	5.2	4.4	13.0	31.2	16.0	10.2

5.3 零样本评估

在零样本设置下，模型在未针对 RO-N3WS 数据集进行任何适配的情况下接受评估。该设置用于评估各系统对罗马尼亚语广播语音以及风格多样的分布外场景的泛化能力。表 6 分别报告了域内（ProTV、Antena1）和分布外测试集（有声读物、电影、儿童故事、播客）的结果。总体而言，结果表明，不同模型和评估领域的零样本性能存在显著差异。

域内泛化。 在开源模型中，*Whisper Large* 和*Whisper Small + Echo*的表现远胜 Wav2Vec 2.0。Whisper Large 得益于其规模（15.5 亿个参数），而 Whisper Small + Echo 则展现了罗马尼亚语专项微调的有效性，在 ProTV 上的表现甚至超过 Whisper Large（词错误率为 9.4%，而非 12.3%）。不过，域内性能整体最佳的是商业系统：*MicrosoftTranscribe* 在 ProTV 上达到 2.9%，*Vatis*在 Antena1 上达到 4.4%，这表明专有数据和领域调优具有优势。

分布外鲁棒性。 所有模型在分布外数据集上的性能都显著下降，尤其是在电影上，声学变化和重叠对话构成了重大挑战。*Whisper Large*是跨域外领域表现最稳健的开源模型，在儿童故事上的词错误率为 10.9%，在有声读物上为 14.8%。令人意外的是，*WhisperSmall + Echo* 在富有表现力的语音上表现不俗（例如，在故事上的词错误率为 21.0%），但在电影音频上表现不佳，这可能是因为 Echo 的朗读风格与电影自然语音领域不匹配。

尽管 Whisper 等多语言模型展现出强大的零样本能力，领域不匹配仍是一项重大障碍。罗马尼亚语专项微调（例如通过 Echo）能够提升域内性能，但要实现稳健的领域泛化，就需要更丰富的风格多样性，而这正是 RO-N3WS 所提供的。

表7：在 RO-N3WS 及其子集上微调的自动语音识别模型的词错误率（%）。数值越低越好 .

模型	域内（RO-N3WS）		分布外（OOD）			
	ProTV	Antena1	有声读物	电影	故事	播客
W2V2 + RO-N3WS	9.8	12.1	25.6	62.4	33.6	22.8
W2V2 + ProTV	10.5	17.5	31.8	69.6	40.5	30.6
W2V2 + Antena1	14.5	13.4	28.2	62.3	36.2	23.8
Whisp-S + RO-N3WS	4.1	6.4	18.9	43.3	21.1	14.1
Whisp-S + ProTV	5.8	11.1	27.5	59.6	33.6	30.6
Whisp-S + Antena1	9.6	7.7	21.3	46.3	25.8	16.0
Whisp-S + Echo + RO-N3WS	4.0	6.1	17.6	42.3	20.0	14.8
Whisp-L + RO-N3WS	2.9	4.4	13.6	31.7	14.0	11.1
Whisp-S + ProTV（4 次运行）	6.0 ± 0.2	11.0 ± 0.1	25.2 ± 1.9	55.8 ± 2.9	30.8 ± 2.8	22.9 ± 1.0
Whisp-S + Antena1（4 次运行）	9.7 ± 0.1	7.9 ± 0.2	21.2 ± 0.3	47.8 ± 1.7	24.8 ± 1.7	16.7 ± 1.0

5.4 监督微调场景

我们使用 RO-N3WS 训练集和验证集对 Whisper 和 Wav2Vec 2.0 模型进行微调。表7报告了域内和分布外测试集上的词错误率（WER）。微调相较于零样本基线提升了性能，证实了针对性适配对罗马尼亚语自动语音识别的作用。

RO-N3WS 微调的影响。 在 RO-N3WS 上进行微调，能让不同模型在多个领域取得显著提升。例如，*Whisper Small* 在 ProTV 上的词错误率从零样本时的 31.6% 降至 4.1%，在儿童故事上的词错误率从 41.1% 降至 21.1%。Wav2Vec 2.0 在 ProTV 上的词错误率也从 27.7% 降至 9.8%。即使数据有限，这些提升也证实了域内监督的重要性。

最佳域内性能。 *Whisper Large +RO-N3WS* 在 ProTV（2.9%）和 Antena1（4.4%）上取得了最低的词错误率。然而，在分布外数据集上的性能较零样本设置略有下降（例如，故事数据上的词错误率为 14.0%，而非 10.9%），这表明模型对新闻领域存在轻微过拟合。

特定来源微调的影响。 仅使用一个新闻来源进行训练会损害泛化能力。*Whisper Small + ProTV* 在 ProTV 上表现良好（5.8%），但在 Antena1 上降至 11.1%。同样，*WhisperSmall + Antena1* 在 ProTV 上进行测试时性能也会下降。这证实了尽管二者都是广播新闻，但它们在风格和声学特征上的差异值得分别考量。随着训练数据增加而分析模型性能的学习曲线见附录 A.4。

Echo 预训练提升了分布外鲁棒性。 在 Echo 上进行预训练可带来 modest 但一致的分布外性能提升。*Whisper Small +Echo + RO-N3WS* 在所有分布外子集上的词错误率均优于仅使用 Whisper Small + RO-N3WS（在故事上的词错误率为 20.0%，相比之下为 21.1%），尽管 Echo 预训练涉及 730 小时数据，而 RO-N3WS 仅有100小时。这表明 RO-N3WS 具有较高的监督效率，可能得益于其语言复杂性和真实的语音表达。

Wav2Vec 2.0：提升适中。 Wav2Vec 2.0 受益于 RO-N3WS 微调，但在域外领域的鲁棒性仍较差（电影上的词错误率为 62.4%），这反映出其容量有限且采用了单语预训练。与 Whisper 模型相比，它在分布偏移下的适应效率较低。

多次运行的变异性与错误分析。 为评估资源有限的设置下微调结果的稳定性，我们使用不同的随机种子，对*Whisper Small +ProTV* 和*Whisper Small + Antena1*分别进行了四次独立运行。表7的最后两行报告了所有测试子集上词错误率的均值和标准差。我们观察到，域内数据集上的变异性较低（标准差通常低于0.2–0.3%），表明微调结果稳定。然而，分布外测试集上的偏差较高，尤其是在电影和儿童故事上，这表明在声学特征和风格更多样的条件下，结果对初始化更加敏感。

Tab表 8：在自然、合成和混合子集上微调的模型的词错误率（%）。所有模型均使用 Whispe r Small.

模型	域内（RO-N3WS）		分布外（OOD）			
	ProTV	Antena1	有声读物	电影	故事	播客
零样本 Whisp-S	31.6	26.8	40.0	60.0	41.1	31.9
仅合成（文本转语音）	24.3	22.5	34.2	58.5	39.5	33.0
混合（50% 文本转语音）	19.0	20.5	33.8	**55.7**	36.7	31.5
仅自然语音（RO-N3WS）	**16.6**	**18.1**	**31.0**	58.1	**33.3**	**27.1**

5.5 自然监督与合成监督

为评估高质量合成语音能否替代或补充自然监督，我们使用 ElevenLabs 的富有表现力的语音，开展了一项受控的文本转语音微调实验。[10]我们合成了约4小时的 RO-N3WS 转录文本音频，并使用 Whisper Small 创建了三种训练配置：

- **仅自然语音：** 3小时真实语音（训练） + 1小时真实语音（验证）；
- **仅合成语音：** 3小时文本转语音音频 + 1小时文本转语音验证集；
- **混合：** 1.5小时真实语音 + 1.5小时文本转语音（训练），0.5小时真实语音 + 0.5小时文本转语音（验证）。

表 8 报告了 RO-N3WS 数据集域内和分布外测试集上的词错误率。这些结果凸显了几个关键发现：(1) 仅使用自然语音进行微调的模型在所有领域都优于合成语音模型和零样本基线，凸显了人类录音中韵律和语境线索的价值；(2) 合成语音训练的效果显著优于零样本 Whisper，表明设计良好的表现力语音合成在低资源场景中可以成为可行的资源。不过，其表现仍不及真实数据，尤其是在情感丰富的语音（如儿童故事）上；(3) 自然语音与合成语音的混合训练缩小了与仅使用自然语音训练的模型之间的差距，甚至在声学多样性较高的领域（如电影）中胜过后者。这表明，高质量的文本转语音增强能够提升鲁棒性，尤其是在发生领域偏移时。

这些结果表明，尽管表现力合成语音对低资源自动语音识别有所帮助，自然录音仍能提供无可匹敌的监督保真度。混合流程是一种很有前景的折中方案，尤其是在扩充高质量数据集成本高昂或耗时较长时。未来的工作可以探索课程学习或自适应混合策略，以进一步发挥合成数据增强的作用。

6 结论与未来工作

我们介绍了 RO-N3WS，这是一个高质量的罗马尼亚语自动语音识别数据集，源自广播新闻，并通过涵盖有声读物、电影、儿童故事和播客的分布外子集加以扩充。我们的基准测试表明，即使是中等规模、语言内容丰富的监督数据，也能显著提升模型在域内和域偏移条件下的性能。

展望未来，RO-N3WS 开辟了多个颇具前景的方向。若将数据从广播语音扩展至更多说话人和更丰富的方言多样性，并通过噪声内容或用户生成内容扩大分布外覆盖，将进一步增强模型鲁棒性。我们在表现力语音合成和罗马尼亚语专用预训练方面的初步成果，也表明合成数据增强、课程学习和多语言迁移具有发展潜力。

我们希望 RO-N3WS 能成为推动罗马尼亚语鲁棒自动语音识别发展的基础，并激发低资源建模、零样本适配和以数据为中心的评估领域的更广泛研究。

[10]https://www.elevenlabs.io

参考文献

Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis M Tyers 和 Gregor Weber。Common Voice：一个大规模多语言语音语料库。*arXiv* 预印本 *arXiv:1912.06670*，2020。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：用于语音表示自监督学习的框架。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第 33 卷，第 12449–12460 页。Curran Associates, Inc.，2020。网址：https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf。

Paul Boersma。Praat，一种通过计算机进行语音学研究的系统。*Glot. Int.*，5(9):341–345，200 1.

Alexis Conneau、Patrick von Platen Ma、AmanpreetSingh 等。Fleurs：用于多语言语音识别的少样本学习基准测试。*arXiv* 预印本 *arXiv:2205.12446*，2023。

Alexandru Lucian Georgescu、Alexandru Caranica、Horia Cucu 和 Corneliu Burileanu。RODigits——用于自动语音和说话人识别的罗马尼亚语连接数字语音语料库。布加勒斯特理工大学科学通报，*C*辑：电气工程与计算机科学，80(3):45–62，2018。

Alexandru-Lucian Georgescu、Horia Cucu、Andi Buzo 和 Corneliu Burileanu。RSC：用于自动语音识别的罗马尼亚语朗读语音语料库。载于第十二届语言资源与评估会议论文集，第6606–6612页，法国马赛，2020年5月。欧洲语言资源协会。网址：https://aclanthology.org/2020.lrec-1.814/。

Matthew Honnibal、Ines Montani、Sofie Van Landeghem、Adriane Boyd 等。spaCy：基于 Python 的工业级自然语言处理。2020。

Yannick Jadoul、Bill Thompson 和 Bart De Boer。介绍 Parselmouth：面向 Praat 的 Python 接口。语音学杂志，71：1–15，2018。

Akhil Kanade、Sourabh Gune、Shubham Dharamkar 和 Rohan Gokhale。视频自动字幕生成。国际工程研究与通用科学期刊，3(6)，2015。

Korbinian Kuhn、Verena Kersken 和 Gottfried Zimmermann。通过协作式纠正自动语音识别，实现交流无障碍实时翻译。在*CHI* 扩展摘要中，2025年。

Valentin Pelloin、Franck Dary、Nicolas Herve 等。用于语音任务的语言模型预训练之 ASR生成文本。在*Interspeech*中，2022年。

Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。在国际机器学习会议中，第 28492–28518 页。PMLR，2023。

Adriana Stan、Florina Dinescu、Cristina Tiple、Serban Meza、Bogdan Orza、Magdalena Chirila 和 Mircea Giurgiu。《SWARA语音语料库：大规模罗马尼亚语平行朗读语音数据集》。载于第九届语音技术与人机对话会议（*SpeD*）论文集，罗马尼亚布加勒斯特，2017年7月6日至9日。

Remus-Dan Ungureanu 和 Mihai Dascalu。《Echo：众包罗马尼亚语语音数据集》。交互设计与建筑学期刊（*IxD&A*），62：141–152，2024年。DOI：10.55612/s- 5002-062-009。

Changhan Wang、Paul Mineiro、Kyu J Sim 等。《Voxpopuli：用于表示学习、半监督学习和解释的大规模多语种语音语料库》。载于*NAACL-HLT*论文集，2021年。

Yanping Zhang、Wei Han、Jing Qin 等。Google USM：扩展超过100种语言的自动语音识别。*arXiv* 预印本 *arXiv:2303.01037*，2023。

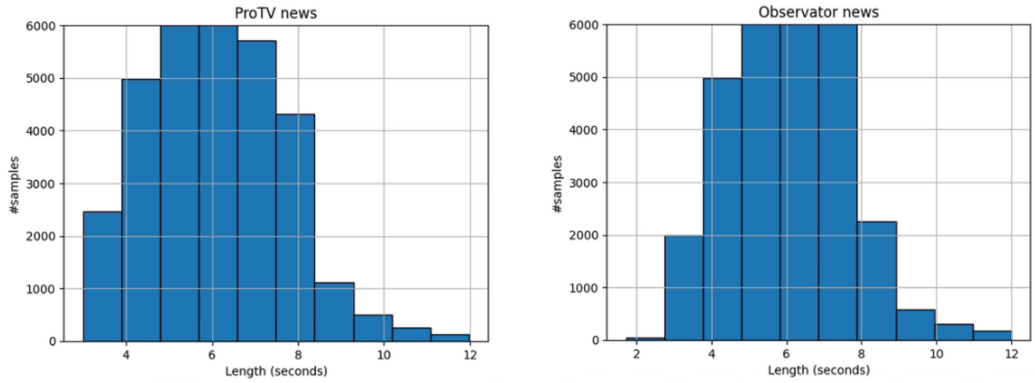

图1：所采集音频文件的录音时长直方图（以秒为单位），分别来自*ProTV*新闻（左）和*Observator*新闻（右）。

A 附录

本附录提供了补充论文的更多细节。具体而言，我们将介绍以下四个主题：

1. **数据集分析**。 我们提供了RO–N3WS语料库域内与域外数据划分的录音时长直方图。
2. **标注示例：** 我们纳入了人工标注流程中的代表性样本，将Whisper生成的转录文本与最终校正后的真实标注进行对比。3. **训练配置**。 汇总表格列出了用于微调Wav2Vec 2.0以及两种Whisper变体（*Small* 和*Large*）的关键超参数。4. **学习曲线**。我们绘制了随着监督数据量增加（5、10以及全部17个可用数据块）时，*ProTV*新闻和*Observator*新闻 测试集上的词错误率。

A.1 数据集分析

图1展示了两档新闻简报（*ProTV* 与*Antena 1* 新闻）的录音时长分布（以秒为单位）。两幅直方图形态相近，表明我们的采集与清洗流程从两个来源得到了长度相当的录音。

图 2 展示了分布外（OOD）数据划分中各类录音时长（秒）的分布：有声读物、电影、儿童故事 和播客。在电影 子集中，将相关影片标题归入系列后，每个系列的录音数量如下：

- *Buletin de Bucure. ti*，*C˘as˘atorie cu repetit. ie*：377；
- *Toamna bobocilor*，*Iarna bobocilor*，*Prim˘avara bobocilor*：442；
- *Liceenii*：641；
- *Toate pânzele sus*：1 297；
- *Cu mâinile curate*，*Duelul*，*Revans. a*，*Un comisar acuz˘a*，*Supraviet. uitorul*：1239；
- *Pistruiatul*：727。

A.2 数据集标注示例

如第4节所述，我们的标注流程确保转录文本在正字法上准确，且在语音上忠实于原始语音。表9列出了一些具有代表性的示例，突显了 Whisper 生成的输出与相应修正后的标注之间的差异。

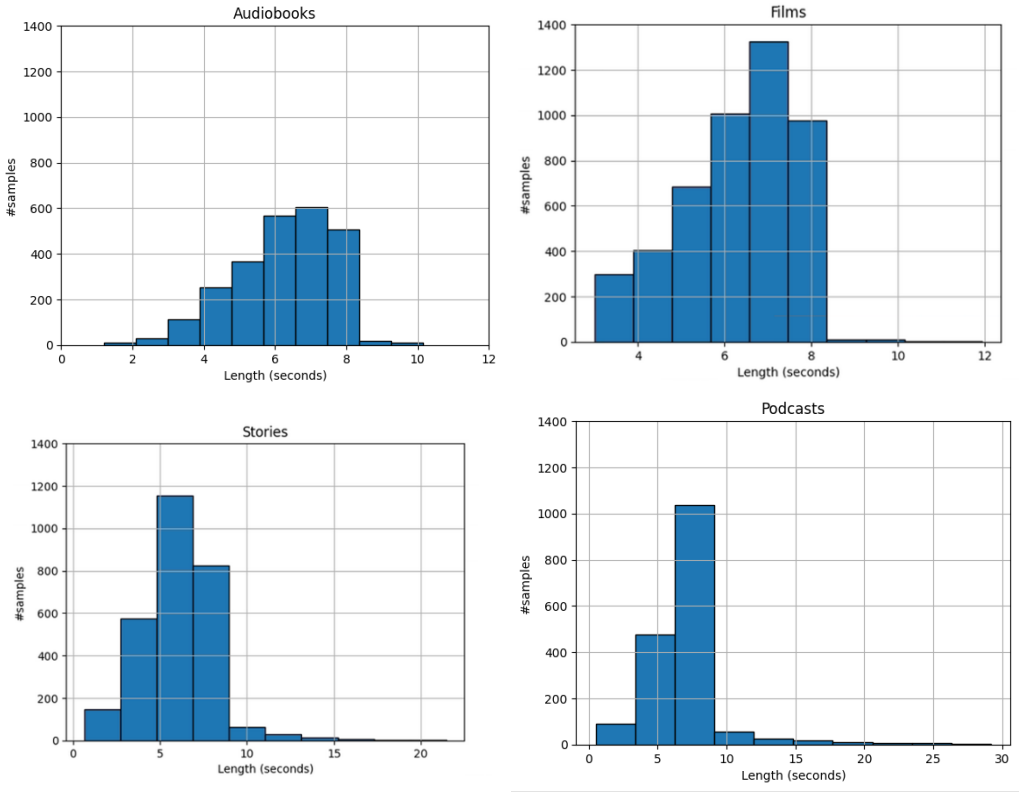

图 2： 分布外子集的录音时长直方图（秒）：有声读物、罗马尼亚电影、儿童故事和播客。

Tab表 9：Whisper 生成的转录文本及其修正后的标注示例。英文 h tran译文对应真实标注版本。

Whisper 标注（罗马尼亚语）	**真实标注** **（罗马尼亚语）**	**英文翻译** **（标准答案）**
克卢日的政治家追查到了 一个月前偷窃的人 来自科学大学的资金办公室 90,000列伊	克卢日警方追查到了 那些在ˇ一个月ˇ前偷走 农业科学ˇ大学财务室里的 农业九万列伊	克卢日警方追查到了那些人 他们一个月前偷走了九十 农业科学大学的数千列伊 农业科学。
在绑好并挟持了 两位年长的业主后，窃贼 sase la numˇar	在挟持并绑好他们之后 两位年长的房主 六名劫匪	在绑住两位年长的 房主后，六名劫匪.....
意大利南部一座城市的居民 有机会目睹 最新一部电影的拍摄 詹姆斯·邦德系列	来自南部某座城市的人们 意大利有机会ˇ观看 拍ˇ摄最新ˇ一部电影 詹姆斯·邦德系列电ˇ影	意大利南部一座小镇上的人们 有机会亲眼见证 最新一部詹姆斯·邦德电影的 拍摄。

A.3 训练配置

表10列出了正文表6、表7和表8中报告的 Whisper 实验的主要超参数。它涵盖三种训练配置：

- *Whisper-small* 在一个仅包含*ProTV* 或*Antena 1* 新闻简报的狭窄子集上微调；
- *Whisper-small* 在完整的 RO-N3WS 语料库上微调；
- *Whisper-large* 在相同的完整 RO-N3WS 数据集上进行微调。

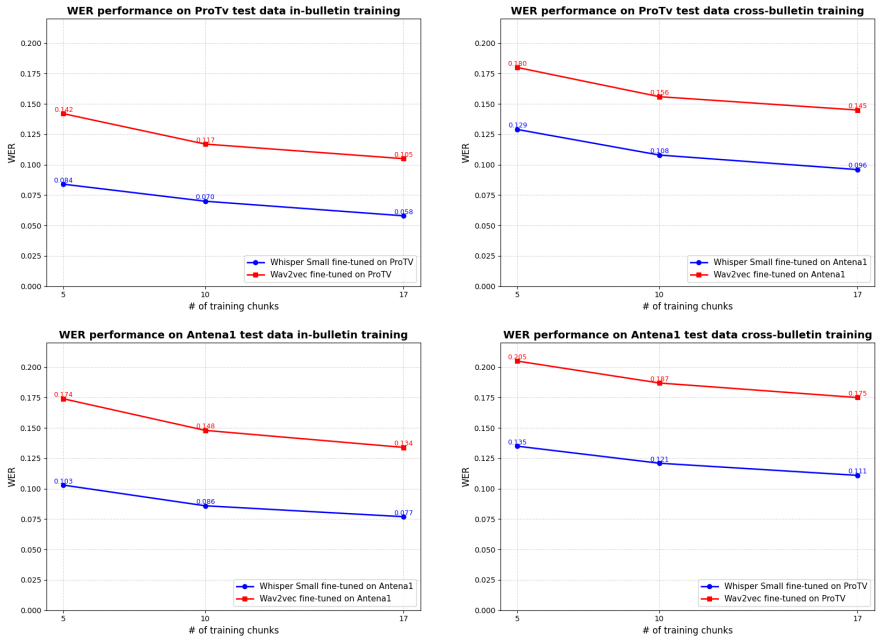

图 3： *ProTV* 和 *Antena 1* 测试集上的学习曲线。展示了在同一新闻简报或另一新闻简报的 *5*、*10* 个训练块以及全部 *17* 个训练块上，对 *Wav2Vec 2.0* 和 *Whisper Small* 进行微调后的词错误率。彩色显示效果最佳。

表 10：Whisper 配置的关键超参数。

超参数	**Whisper-small（ProTV / Antena 1）**	**Whisper-small（RO-N3WS）**	**Whisper-large（RO-N3WS）**
每设备批次大小	16	32	32
梯度累积步数	1	2	2
初始学习率	1×10^{-5}	5×10^{-5}	2×10^{-5}
学习率调度器	线性	线性	线性
预热步数	300	500	3 000
总训练步数	5 000	20 000	20 000

表 11：Wav2Vec 2.0 微调的关键超参数

超参数	**值**
每设备批次大小	16
梯度累积步数	2
初始学习率	5×10^{-5}
权重衰减	0.01
学习率调度器	线性（默认）
预热步数	500
训练轮数	15

表 11 详细列出了用于在完整 RO-N3WS 语料库以及范围较窄的*ProTV* 或*Antena 1* 子集上微调 Wav2Vec 2.0 模型的主要超参数。

A.4 学习曲线

图 3 展示了随着监督数据量的增加，*ProTV* 和*Antena1* 测试集上的词错误率（WER）。对比了两种系统：Wav2Vec 2.0 和 Whisper Small。对于每种

模型，我们绘制了：（i）同新闻简报内 曲线，表示系统在其所评估的同一新闻简报上进行微调；（ii）跨新闻简报 曲线，表示系统在另一份新闻简报上进行微调。
RO-N3WS 语料库中 *ProTV* 和 *Antena 1* 的子集均划分为 17 个训练块、2 个验证块和 1 个测试块。因此，横轴报告的是在 5、10 个训练块以及全部 17 个训练块上微调后的词错误率。

在所有数据设置下，Whisper Small 的词错误率始终低于 Wav2Vec 2.0。正如预期，同一公告内的微调（ProTV→ProTV、Antena 1→Antena 1）优于跨公告设置（Antena 1→ProTV、ProTV→Antena 1），凸显了训练数据与评估数据之间封闭分布匹配的价值。