arXiv:2601.04233v1 [cs.SD] 4 Jan 2026

LEMAS：一个包含 15 万小时音频的大规模、可扩展多语言套件，采用生成式语音模型

赵志远†、**林立坚**、**朱烨**、**Kai Xie**、**刘云飞**、**李宇**

国际数字经济研究院（IDEA）
†项目负责人

生成式语音模型在英语和中文等高资源语言中已取得出色表现，但在多语言环境中的能力仍然有限。这一差距主要源于数据方面的挑战：收集并整理大规模、质量一致且带有细粒度时间标注的多语言语音数据要困难得多，而现有的网络爬取语料库在多语言环境下无法充分保证数据质量和标注可靠性。为解决这一瓶颈，我们推出了**LEMAS-Dataset**。据我们所知，它是目前规模最大的、带有词级时间戳的开源多语言语音语料库。LEMAS-Dataset 涵盖 10 种主要语言，总时长超过 15 万小时，采用高效的数据处理流程构建，以确保数据和标注的高质量。为验证 LEMAS-Dataset 在不同生成范式中的有效性，我们在该数据集上训练了两个架构和任务专门化各不相同的基准模型。**LEMAS-TTS**基于非自回归流匹配框架，利用数据集的庞大规模和语言多样性，实现了稳健的零样本多语言合成。我们提出的口音对抗训练和 CTC 损失缓解了跨语言口音问题，提升了合成稳定性。作为互补方案，**LEMAS-Edit** 采用自回归仅解码器架构，将语音编辑构建为掩码词元填充任务。通过利用精确的词级对齐来构建训练掩码，并采用自适应解码策略，该模型实现了无缝、边界平滑且过渡自然的语音编辑。实验结果表明，在 LEMAS-Dataset 上训练的模型能够实现高质量的语音合成和编辑，证实了该数据集的质量。我们相信，这一带有丰富时间戳标注的细粒度多语言语料库将推动基于提示的语音生成系统在未来取得进展。

联系邮箱：赵志远zhaozhiyuan@idea.edu.cn**演示页面：**
https://lemas-project.github.io/LEMAS-Project**数据集：**
https://huggingface.co/LEMAS-Project**代码：**
https://github.com/LEMAS-Project

1. 引言

随着生成式基础模型的兴起，语音合成领域进入了变革性阶段。近期系统利用海量未标注音频进行大规模预训练，在零样本任务中取得了令人瞩目的表现，尤其是在英语和中文等高资源语言中 [11, 18, 39, 9, 38, 44]。尽管取得了这些进展，将生成式语音模型扩展到真正的多语言场景仍然面临挑战。与英语和中文相比，开源多语言数据集和基础模型在规模、覆盖范围和标注质量方面仍然有限，导致模型在不同语言上的表现明显下降。随着模型容量不断增大，主要瓶颈已日益从架构设计转向数据，具体而言，是缺乏大规模、高质量且细粒度的多语言语音语料库。

现有语音数据集的局限性。 尽管近期的工作大幅扩充了公开可用的语音资源，但仍存在若干根本性局限。Libri-Light [20],、GigaSpeech [7], 和 WenetSpeech [43]等大规模语料库提供了丰富的音频-文本对，助力训练出强大的自动语音识别和文本转语音模型。然而，这些数据集主要是单语种的，侧重于英语或中文，未能解决大规模多语言资源匮乏的问题。相比之下，多语言

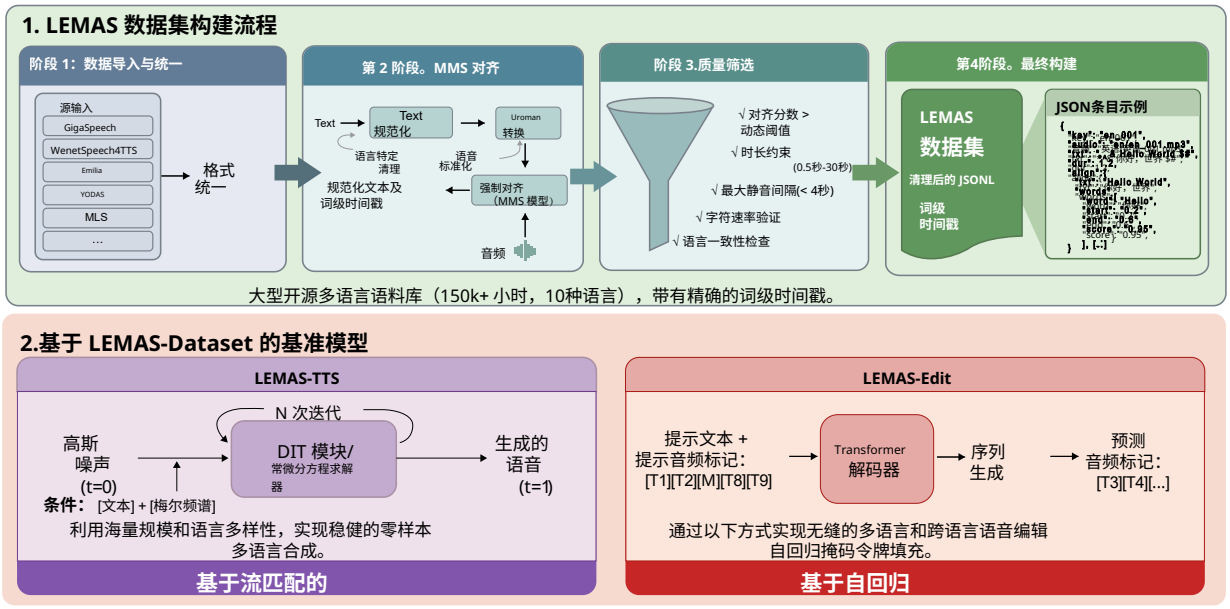

图 1 - LEMAS-Dataset 包含超过**15 万**小时的多说话人语音，并在**10** 种主要语言中提供强制词级对齐。基于 LEMAS-Dataset，我们训练了两个模型。**LEMAS-TTS** 实现了大规模、基于流的神经文本转语音，可根据文本和一段简短的参考音频流式生成高保真语音；而**LEMAS-Edit** 则实现了基于编解码器的词级语音编辑。

数据集，如 TEDx [32]和 MLS [29]，则受限于规模有限、领域覆盖面狭窄。具体而言，TEDx 主要由公开演讲组成，而 MLS 则局限于有声书领域。近期，YODAS [24] 等大规模多语言数据集已问世，但它们高度依赖未经充分筛选的网络抓取数据，导致数据质量参差不齐，也缺乏可靠的细粒度标注。

我们还认可 WenetSpeech4TTS 等富有潜力的工作 [25], ，该工作对 WenetSpeech 进行了严格清洗，并引入了针对语音合成定制的时间戳级标注。尽管这类数据集效果显著，但仍局限于单一语言。相比之下，我们的目标是将这种数据质量和细粒度时间监督水平扩展到真正的多语言环境中，这促成了 LEMAS-Dataset 的构建。此外，我们基于 LEMAS-Dataset 设计了两个基础模型：LEMAS-TTS 和 LEMAS-Edit（图1）。

LEMAS-Dataset。为弥补这一差距，我们推出了 LEMAS-Dataset。据我们所知，LEMAS-Dataset 是规模最大的开源多语言语音语料库，包含严格标注的词级时间戳，涵盖超过**150,000 小时**的语音，涉及**10 种主要语言**，包括中文、英语、俄语、西班牙语、葡萄牙语、德语、法语、意大利语、印度尼西亚语和越南语。表1展示了与现有语料库相比，LEMAS-Dataset 在规模和时间对齐精度方面的优势。

利用真实场景音频构建此类数据集面临着独特的挑战，因为传统的基于 HMM 的强制对齐工具（例如 MFA [26]）往往难以处理含噪且未经切分的语音。为应对这一挑战，我们采用了稳健的多阶段处理流程，利用 MMS 对齐器 [30] 提取词级时间戳；更重要的是，该流程会为每个对齐词分配一个置信度分数。这一设计支持灵活且考虑可靠性的筛选，使研究人员能够根据下游需求动态权衡数据规模与对齐精度，并确保数据对精度要求极高的应用具有很高的实用价值。

为展示 LEMAS-Dataset 在多种生成范式中的适用性，我们分别利用数据集的规模与精度，开发了两个并行的基础模型：

表 1 - LEMAS-Dataset 与现有大规模处理语音数据集的比较。有声读物数据指在洁净声学条件下专业录制、与脚本对齐的语音；而野外数据则是在真实场景中收集的自然语音，通常伴有背景噪声、自发言语风格以及不完美的文本-音频对齐。在涵盖多种数据来源且包含话语级时间戳的数据集中，LEMAS-Dataset 提供了更长的总时长和更广泛的语言覆盖。

数据集	总时长（小时）	语言	时间戳	数据来源
Libri-Light [20]	60k	1	×	有声书
GigaSpeech [7]	10k	1	×	有声书/真实场景
GigaSpeech2 [42]	28k	3	×	有声书/真实场景
WenetSpeech [43]	22k	1	×	真实场景
WenetSpeech4TTS [25]	13k	1	✓	真实场景
MLS [29]	51k	8	×	有声书
Emilia [19]	102k	6	×	真实场景
Voxbox [38]	103k	2	×	有声书/真实场景
LEMAS-Dataset（我们的）	**150k**	**10**	✓	**有声书/真实场景**

LEMAS-TTS。 若没有特定语言对齐正则化，流匹配模型会遭遇对齐漂移，导致低资源语言的可懂度下降；还会出现口音泄漏，使目标韵律被高资源语言的韵律取代 [9], ，最终损害零样本跨语言保真度。为应对这些挑战，我们在 F5-TTS [9] 架构的基础上加入语言感知稳定机制。具体而言，我们将不同的书写系统统一映射到共享语音空间，并引入两个互补目标：连接主义时序分类（CTC）损失 [17] ，以提升语音可懂度；以及口音对抗损失 [16] ，以抑制跨语言口音干扰。这些改进使模型能够有效利用完整的 15 万小时语料库，在零样本跨语言合成中同时实现较高的说话人相似度和自然度。

LEMAS-Edit。 为进一步验证 LEMAS-Dataset 对自回归语音生成与编辑的有效性，我们开发了 LEMAS-Edit，这是一种基于仅解码器架构构建的专用编辑模型。VoiceCraft [28] 将语音编辑建模为掩码令牌填充任务，其生成质量对边界精度和解码稳定性高度敏感。编辑边界未对齐可能会导致明显的伪影，例如静音时间过长或出现不自然的声学不连续现象 [28]。为减轻自回归解码中常见的重复伪影，我们在推理时引入重复惩罚，以抑制令牌循环。我们还通过加入系统级组件扩展了编辑流程，包括去噪、使用 MMS 重新对齐以及一致性验证，从而实现精准、可靠且由用户控制的编辑。正如演示所示，LEMAS-Edit 能够实现无缝、边界平滑的编辑，在带有环境噪声或背景声音的真实场景音频中也表现出良好效果，凸显了 LEMAS-Dataset 所带来的强大泛化能力。

总而言之，我们的贡献主要有三点：

- **LEMAS-Dataset：** 我们发布了目前规模最大的开源多语言语音语料库，包含超过 150,000 小时、涵盖 10种语言的语音数据，并提供严格的词级时间戳。该数据集通过可扩展的对齐流程构建，并采用词级置信度估计，从而为对精度要求极高的语音生成任务提供可靠、细粒度的时间监督。

- **LEMAS-TTS：** 我们开发了一种用于大规模多语言训练的非自回归流匹配文本转语音模型。通过将不同文字系统统一到共享语音空间，并引入基于CTC的对齐正则化和口音对抗目标，LEMAS-TTS 实现了稳健的零样本跨语言合成，并提升了可懂度和口音一致性。

- **LEMAS-Edit：** 我们开发了一种仅解码器自回归语音编辑系统，通过重复惩罚和基于语速的异常检测实现稳定化推理，并由包含去噪、重新对齐和验证的稳健处理流程提供支持。LEMAS-Edit 支持无缝、边界平滑的多语言编辑，并能很好地泛化到真实环境中的噪声音频。

2. 相关工作

2.1 大规模语音语料库

语音数据集的发展历程，体现为从经过精心整理的朗读语音，转向从网络收集的大规模、多领域数据集。这一演进满足了端到端模型对数据规模和多样性日益增长的需求。

精心整理与众包构建的基础。 早期的多语言语料库，如 GlobalPhone [33]和 IARPA BABEL [15]，为高质量、精心整理的录音树立了标准，但规模有限。Mozilla Common Voice [2] 随后通过众包实现数据收集的普及，并优先考虑说话人多样性。在有声读物领域，LibriSpeech [27] 和多语言 LibriSpeech（MLS） [29] 成为自动语音识别的事实标准基准测试集，提供了适用于可复现研究的清晰朗读语音。

超大规模多领域语料库。 为克服有声读物的领域局限，研究人员转向了大规模互联网资源。GigaSpeech [7] 推出了一个不断扩展的多领域英语语料库，素材来自有声读物、YouTube 和播客，包含 10,000 小时的高质量标注音频。这一转变标志着研究方向转向对多样化“真实场景”录音条件进行建模。在中文语音领域，WenetSpeech [43] （及其近期衍生版本 WenetSpeech4TTS [19]）汇集了数万小时的多领域普通话语音，进一步验证了网络挖掘数据在训练工业级自动语音识别和文本转语音模型方面的有效性。

多语言大规模无监督语音语料库。 近期的工作大幅扩展了多语言语音数据集的规模和语言覆盖范围，重点关注弱监督或无标签数据。VoxPopuli [37]提供了来自欧洲议会录音的23种语言、共40万小时的无标签语音。YODAS [24] 从 YouTube 收集了 100+ 种语言的50万+ 小时语音，并使用质量参差不齐的自动字幕。这些数据集以标签质量换取庞大规模，为自监督和半监督语音研究提供了助力。

2.2 生成式语音模型

神经文本转语音已从特定任务模型转向统一的生成式框架。它从 Tacotron [40]等自回归模型，发展到非自回归的 FastSpeech [31], ，再到 VITS [22]等端到端变分模型。如今的前沿方向是大规模基础模型，将语音合成、跨语言克隆和编辑整合于一体。

生成式基础模型。 语音生成分为离散、连续和混合范式，它们在表示方式和生成方法上各不相同。

- **离散方法**将语音建模为编解码器令牌序列，并采用自回归或掩码填充的方式生成，类似于语言模型。VALL-E [8] 率先使用 EnCodec 令牌实现零样本文本转语音；AudioLM [4]则加入分层语义与声学令牌预测。XTTS [5] 将这一方法扩展到多种语言。近期研究聚焦于提升效率和控制能力，例如 MaskGCT [39] 和 FishAudio [36]。

- **连续方法**通过扩散或流匹配，直接从噪声生成梅尔频谱图或波形。NaturalSpeech 2 [35] 展示了高保真潜在扩散；VoiceBox [23] 通过省去外部编码器简化处理流程。F5-TTS [9]采用带有中间填充的 DiT 结构，以精简的设计实现了有竞争力的零样本合成。

- **混合方法**将离散语义建模与连续声学细化相结合，实现多样且忠实的合成。CosyVoice 1–3 [10, 12, 13] 将基于大语言模型的语义 token 与条件流匹配相结合，用于多语言指令跟随文本转语音；Fire Red-TTS [41] 级联离散模型与基于流的模型，以实现韵律控制。IndexTTS 2 [44] 则进一步探索了高效的语义压缩与风格操控。

多语言扩展。 随着训练语料库不断增长，文本转语音从单语/双语发展到多语言。XTTS-v2 [5] 等早期开源模型采用17种语言联合训练。OpenAudio-S1-mini [36] 等社区模型支持13种语言，拥有100k+ 小时的数据，增强了跨语言泛化能力。规模更大的 CosyVoice 3 [13] 采用大型音频理解模型中的语音分词器，并在

涵盖9种语言和18种汉语方言的逾百万小时数据上进行训练。这反映了数据规模的快速增长，以及对大规模多语言数据集的需求。

基于文本的语音编辑。 语音编辑会在保留音色的同时修改语音内容，如今被视为文本转语音的一项子任务。VoiceBox [23] 展现了灵活的上下文学习能力，可用于零样本文本转语音、语音编辑、风格转换和多样化生成。非自回归模型 MaskGCT [39] 采用掩码预测学习；F5-TTS [9] 使用扩散变换器（DiT），同时支持文本转语音和语音编辑。自回归模型 VoiceCraft [28]结合因果掩码和延迟堆叠，实现灵活生成并支持序列内的多样化编辑。

3. LEMAS-Dataset

我们推出了 LEMAS-Dataset，这是一个大型开源多语言语音语料库，涵盖中文（Zh）、英语（En）、俄语（Ru）、西班牙语（Es）、印度尼西亚语（Id）、德语（De）、葡萄牙语（Pt）、越南语（Vi）、法语（Fr）和意大利语（It）这 10 种主要语言，总时长超过 150,000 小时。LEMAS-Dataset 提供词级对齐，并为每个词附有置信度分数。这种细粒度标注有助于广泛的应用，包括基于置信度分数进行数据质量筛选、训练时长预测模型，以及生成详细提示来指导语音合成与理解。

3.1 数据集处理流水线

构建用于生成式建模的大规模多语言语音语料库，不仅仅是合并现有数据集，因为这些数据集在标注、转录质量和声学特征方面存在很大差异。LEMAS-Dataset 处理流程的关键洞见是将语言多样性 与标注脆弱性分离，仅保留时间结构具有可靠依据的样本。为此，我们开发了一个多阶段处理流程，用于统一异构来源、执行稳健的多语言对齐，并实施可扩展的质量控制。

数据导入与结构统一。 我们首先从多种公开可用的语音语料库中汇总音频-文本对，包括 GigaSpeech（英语）[7], GigaSpeech2（印度尼西亚语、越南语） [42],WenetSpeech4TTS（中文，基于 VoiceBox） [25, 38], Emilia（中文和英语，基于 VoiceBox） [19],MLS（英语、德语、法语、葡萄牙语、西班牙语、意大利语） [29], 多语言 TEDx（德语、法语、葡萄牙语、西班牙语、意大利语） [32], Alcaim（葡萄牙语） [6], Golos（俄语） [21], 以及 Yodas（德语、法语、西班牙语、葡萄牙语、俄语、意大利语、印度尼西亚语、越南语） [24]。这些来源涵盖广泛的录音场景，包括有声读物、播客和真实场景中的对话语音。我们不保留数据集特定的模式，而是将所有输入标准化为统一的数据表示，从而支持与语言无关的下游处理和大规模自动化。

通过罗马化文本进行多语言MMS 对齐。 可靠的词级时间戳是文本转语音训练和语音编辑任务的前提，但传统的强制对齐流水线通常依赖特定语言的词典和字素转音素系统，在多语言、噪声和语码转换条件下往往不够稳健。为避免这些局限，我们采用了多语言MMS强制对齐器 [1]，这是一种由 torchaudio [2]提供的基于 wav2vec 的 CTC对齐模型 [3, 17] 。该模型作为“将语音技术扩展至 1,000+ 种语言”计划的一部分，使用涵盖超过 1,100 种语言的 23,000 多小时语音进行训练 [30]。

在进行对齐之前，转录文本会针对不同语言进行规范化处理，然后使用 Uroman [3] 等工具进行罗马化；Uroman 可将多种文字系统（例如汉字、西里尔文字）映射为统一的拉丁字母表示。这种设计消除了对特定语言发音词典的依赖，使对齐模型能够稳健地处理生僻词、命名实体和句内语码转换。随后，MMS 对齐器会生成词级边界以及词元级后验概率，为后续质量控制奠定基础。

置信度驱动的质量筛选。 我们的处理流程有一项核心原则：仅仅成功完成对齐还不够；只有时间可靠性高的对齐才适用于生成式语音建模。

[1]https://docs.pytorch.org/audio/2.3/tutorials/forced_对齐_for_多语言_data_tutorial.html[2] https://github.com/pytorch/audio[3]https://github.com/isi-nlp/uroman

因此，我们仅保留能够提取词级对齐的样本，并应用一系列基于对齐置信度和时间一致性的基于规则的筛选器：

- **对齐置信度筛选**。对于每条话语，我们根据对齐词元的后验概率计算平均对齐置信度分数。为适应不同语言和源语料库之间的差异，我们采用针对各数据集设定的阈值 $x \in [0.2, 0.5]$，仅保留置信度超过相应阈值的样本。

- **时长与停顿约束**。为确保与文本转语音训练兼容，我们会移除时长短于 0.5 秒或长于 30 秒的话语。此外，对于包含较长未对齐区域或静音区间超过 4 秒的样本，也会予以舍弃，因为这类情况通常表明存在分段或转录错误。

- **语速归一化**。为排除语速异常过快或过慢的语音，我们对文本长度与音频时长之比施加按语言设定的约束。对于每种语言 L，基于字符数的比率 len(text)/duration 必须落在预设区间 $[\text{min_ratio}_L, \text{max_ratio}_L]$ 内，该区间根据语料库统计数据通过经验确定。

- **语言与字符验证**。我们将数据集限定为十种目标语言：中文 (zh)、英语 (en)、俄语 (ru)、西班牙语 (es)、印度尼西亚语 (id)、德语 (de)、葡萄牙语 (pt)、越南语 (vi)、法语 (fr) 和意大利语 (it)。我们会验证所有转录文本均属于上述语言之一，并根据各语言的字符集，移除包含不支持的文字系统、表情符号或过多非语言符号的样本。

所有阈值均通过经验确定，以平衡数据规模与对齐精度。值得注意的是，在预处理过程中，我们没有进行明确的语音增强或背景噪声去除。尽管许多文本转语音系统仍容易受到声学干扰的影响，我们仍有意保留真实环境特征，并期望更先进的生成模型能够更好地利用这类变化。与这一设计选择一致，我们演示页面中的示例表明，所提出的编辑模型在含噪录音上仍然有效，并能在一定程度上重现与参考音频相符的声学环境，凸显了保留真实背景条件的实际益处。

最终表示与存储。 最终数据集采用统一的 JSON 格式存储，包含唯一的样本键、音频文件路径、原始和规范化后的转录文本，以及带有相应置信度分数的词级时间戳。为便于存储和分发，所有音频文件均转换为 MP3 格式，同时保留其原始采样率。

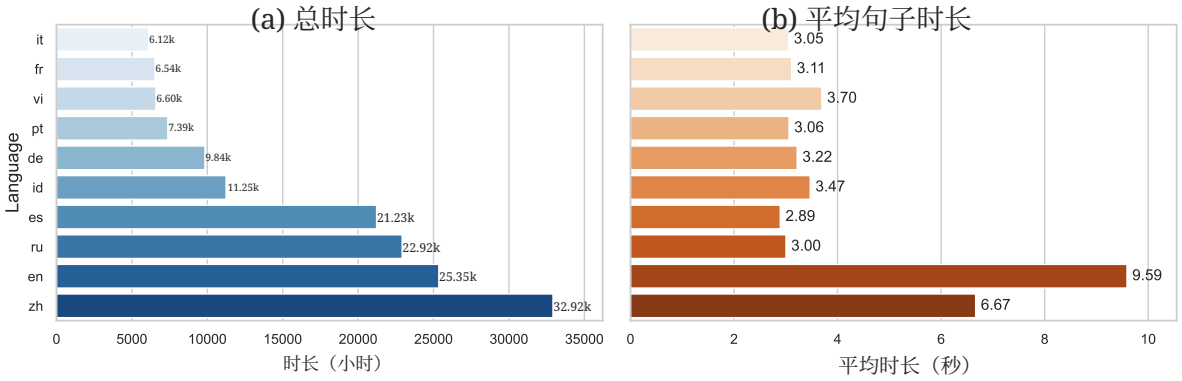

图2 - LEMAS-Dataset 的数据集统计。(a) 各语言的时长；(b) 平均句子时长（秒）。该数据集在不同语言的数据量和句子长度方面均呈现显著差异。

3.2 数据集统计

LEMAS-Dataset 包含约 150,144 小时的音频数据，涵盖 10 种主要语言。如图 2所示，我们对数据集进行了专门整理，以平衡各语言的语音时长。尽管源数据的可用量有所不同，各语言的数据分布仍相对均匀。高资源语言

例如中文（Zh）和英语（En）的数据量仍略多，但所有语言的数据规模都相当可观。具体而言，即使是资源较少的子集（如意大利语和越南语），时长也超过 6,000 小时，足以支持有效的单语训练和联合多语建模。为促进研究的可复现性并确保基准测试的一致性，该数据集划分为训练（train）和评估（eval）子集。

训练集。 训练集划分的详细统计数据见表 2。训练集包含总计超过 150,000 小时的音频，旨在支持大规模声学模型的训练。需要指出的是，由于源语料库不同，各语言的句子级统计数据也有所差异。具体而言，中文和英语数据主要来自 GigaSpeech、Emilia 和 WenetSpeech4TTS，其中包含较长的话语；而其他语言的数据主要来自 YODAS，其话语较短且经过预先切分，未进行额外合并或重组。因此，中文和英语的平均话语长度明显长于其他语言。

表2 - LEMAS-Dataset **训练**集的详细统计数据。单位 **h**、**s** 和 **M** 分别表示小时、秒和百万 (10^6)。平均时长和平均词数分别表示每个话语的平均时长和平均词数。

语言	总时长（小时）	平均时长（秒）	话语数（百万）	总词数（百万）	平均词数
it	6.12k	3.05	7.21	48.86	6.78
fr	6.54k	3.11	7.56	65.84	8.71
vi	6.60k	3.70	6.43	89.90	13.99
pt	7.38k	3.06	8.68	60.32	6.95
de	9.84k	3.22	11.00	80.12	7.28
id	11.25k	3.47	11.66	85.97	7.37
es	21.22k	2.89	26.41	183.67	6.96
ru	22.92k	3.00	27.47	163.02	5.93
en	25.35k	9.59	9.52	268.68	28.24
zh	32.92k	6.67	17.78	496.96	27.96
总计	150.14k	4.18	133.71	1,543.34	12.02

表3 - LEMAS-Dataset 的**评估**集详细统计信息。单位**分钟**和**秒**分别表示分钟和秒。平均时长和平均词数分别表示每条话语的平均时长和平均词数。

语言	总时长（分钟）	平均时长（秒）	话语数	总词数	平均词数
it	44.22	5.31	500	6,599	13.20
fr	38.17	4.58	500	6,546	13.09
vi	36.74	4.41	500	6,727	13.45
pt	41.69	5.00	500	5,812	11.62
de	38.65	4.64	500	5,599	11.20
id	47.20	5.67	500	6,133	12.27
es	40.52	4.86	500	6,216	12.43
ru	40.24	4.82	500	5,138	10.28
en	67.46	8.10	500	11,325	22.65
zh	75.84	9.10	500	18,627	37.25
总计	470.73	5.65	5,000	78,722	15.74

评估集。为建立一致的基准测试集，我们整理了一个独立的评估集，详情见表 3。与训练集不同，评估集通过严格的质量标准筛选每种语言的 500 条话语，以确保数据均衡。具体而言，我们筛选平均词级对齐分数高于 0.9、词数超过 5 且时长介于 3 至 15 秒之间的样本。此外，将句末静音裁剪至最长 0.2 秒，以统一声学特征。在满足上述条件的样本中，我们进一步计算各语言特定的字符时长比统计数据，并选取最接近各语言均值的 500 条话语。这一步有助于排除唱歌或夸张说话风格等非典型或异常样本，同时保留各语言具有代表性的说话模式。最后，为进一步确保数据质量，我们对约 20% 的评估样本进行了人工审核，以确认测试集仅包含自然语音。通过这一整理流程，不同语言的评估指标（例如 WER 或平均意见分）能够进行比较，且不会受到数据量差异或质量变化的偏倚。

4. LEMAS-TTS

我们设计了一种文本转语音方法 LEMAS-TTS，其架构基于 F5-TTS [9]，，采用其非自回归流匹配框架和扩散变换器（DiT）骨干网络。尽管 F5-TTS 在零样本任务上表现出色，但要将其扩展为稳健的多语言模型，仍面临诸多挑战，例如对齐稳定性以及跨语言的有效韵律迁移。为克服这些挑战，我们引入了架构改进，以统一语言表示并稳定训练过程。

4.1 多语言训练框架

为实现高保真多语言合成，我们调整了训练范式，纳入统一语音输入空间、辅助监督目标和精细的韵律建模。

统一语音表示。 为弥合不同语言之间的语言差异，我们不再采用原系统混合使用的字符/词级分词方式，而是改用统一的语音前端。对于中文，我们采用带声调的声母—韵母拼音表示，以捕捉音节结构；对于其他语言，则使用通过 eSpeak-NG [14]生成的国际音标（IPA）序列。选择这些语言的 IPA，是因为 eSpeak-NG 的前端能够自动处理语言切换，而且使用 IPA 也有助于验证发音是否正确。所有语音标记均添加显式的语言标识符（例如 <zh>、 <en>），将不同语言映射到共享的潜在空间，同时保留各语言特有的音系规则。此外，我们利用词级时间戳在音素序列中插入显式停顿标签（#1–#4），分别表示短、中、长和异常停顿，从而实现简单而有效的停顿控制。

辅助稳定目标。标准流匹配目标在处理复杂的多语言场景时，有时会遇到收敛困难。为此，我们引入两个辅助损失函数，以提高对齐准确性和语音可懂度：

- **CTC 对齐损失**。为确保声学轨迹与语言内容之间严格单调，我们在解码器输出上应用连接主义时序分类（CTC）损失。具体而言，我们通过轻量级投影头 [17] 将预测的梅尔频谱图映射为音素序列，从而明确地促进准确的时长建模。
- **口音对抗解耦**。受 IndexTTS2 [44], 启发，我们旨在将说话人音色与口音信息解耦。我们在条件通路中接入口音分类器，并通过梯度反转层（GRL）对其进行优化。利用现成的语言识别模型生成的伪标签监督该分类器，可促使编码器学习口音不变表示，从而减少跨语言口音泄漏。

跨语言韵律建模。为捕捉细粒度的韵律变化，我们集成了一个改编自 Seamless Expressive 架构 [34]的韵律编码器。具体而言，我们从分段后的子语句中提取 80 维滤波器组特征，并使用预训练的 ECAPA-TDNN 主干网络将其编码为 512 维韵律嵌入。这些嵌入与梅尔帧和文本标记对齐，并通过线性投影层注入 DiT。该机制提供了明确的韵律监督，显著提升了跨语言韵律迁移的自然度。

设计分析与排除项。在开发 LEMAS-TTS 期间，我们还探索了显式说话人对比学习（InfoNCE）和裁剪与打乱数据增强。然而，经验观察表明，通过 InfoNCE 强制实现严格的说话人不变性会与流匹配目标相冲突，导致稳定性下降。同样，打乱分段梅尔块会破坏 DiT 学习一致延续内容所需的时间连贯性。因此，最终架构未采用这些组件，而是选择了上述对齐与韵律目标。

4.2 推理流程与采样动态

为确保多语言生成的稳健性，我们的推理框架严格遵循训练范式，以尽量减少分布偏移；同时采用可调节的采样策略，灵活控制采样动态，从而提升系统在不同声学条件和语音风格下的稳定性。

统一多语言前端。我们采用标准化的文本处理流程，使其与训练阶段保持一致，从而确保输入的一致性。首先，系统利用自动语言识别应用特定语言归一化规则（例如数字展开和标点过滤）。随后，文本将转换为训练策略中定义的共享语音空间：中文转换为带声调拼音，并进行声母韵母分解；其他语言则转换为国际音标序列。重要的是，系统支持在文本中显式插入空格和停顿标记（#1-#3），从而精确控制停顿时长。

动态引导与 Sway 采样。 F5-TTS 中原始的 Sway 采样策略会在生成初期分配更多步数，以提高粗略结构的准确性。然而，该策略在多语言和跨语言场景中的表现并不理想。为此，我们重新设计了时间扭曲函数和无分类器引导（CFG）调度，将其构造为单参数族，如图3所示，从而能够在非线性行为与线性行为之间平滑插值。通过增加早期阶段的步数和 CFG 强度，我们提高了发音准确度；而在后期降低 CFG 强度，则能提升模型灵活性，从而改善说话人相似度和音频质量。

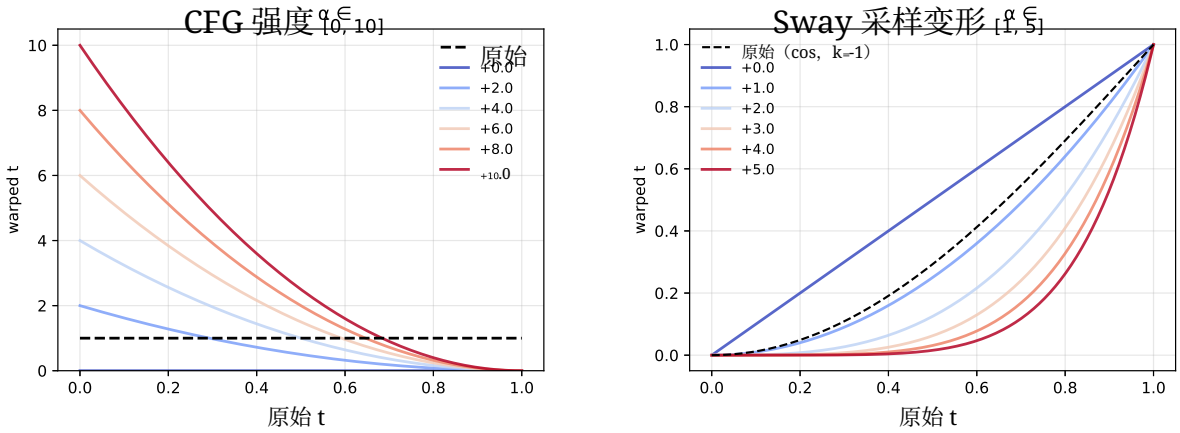

图 3 -采样策略中使用的时间依赖控制调度。左：具有不同最大引导水平的 CFG 强度调度，在采样过程中着重强调早期时间步，并逐渐衰减。右：不同强度的 Sway 采样时间扭曲，与基于余弦的基线（虚线）进行对比；扭曲越强，分配给后期时间步的步数越多。这些调度共同展示了如何随时间调整引导强度和采样步数分配，从而影响生成行为。

显著提升了长篇多语言语句的感知稳定性。

首先，我们改进无分类器引导（CFG）机制。令 $f_\theta^{\text{cond}}(x,\ t)$ 和 $f_\theta^{\text{uncond}}(x,\ t)$ 分别表示条件流和无条件流。我们将引导流定义为：

$$f_\theta^{\text{cfg}}(x,t) \ = \ f_\theta^{\text{cond}}(x,t) \ + \ g(t)\left(f_\theta^{\text{cond}}(x,t) - f_\theta^{\text{uncond}}(x,t)\right), \tag{1}$$

其中，时间相关调度 $g(t)$ 定义为按二次函数衰减：

$$g(t) \ = \ \lambda\,(1-t)^2, \tag{2}$$

其中 λ 为缩放因子（例如 $\lambda = 5$）。如公式（$t \approx 0$）所示，该调度会在初始高噪声阶段施加更强的结构校正，并在接近 $t = 1$ 时逐渐减弱引导，以保持声学自然度，如公式（1）和（2）所示。

其次，为了进一步优化流轨迹的离散化，我们更新了 Sway 采样策略。不同于线性或基于余弦的网格，我们采用公式（3）所示的幂律重新参数化。给定均匀网格 $s_k \in [0,1]$，其中 $k = 0,\ \ldots\ldots,\ K$，有效时间步定义为：

$$t_k \ = \ s_k^{1+\gamma}, \tag{3}$$

其中，$\gamma \geq 0$ 是一个可调节的曲率参数。增大 γ 实际上会拉长前期采样步，并压缩后期采样步。经验表明，这种计算资源的重新分配有助于从噪声中恢复信号，显著提升长篇多语言语句的感知稳定性。

多样化条件控制模式。 推理接口经过扩展，支持灵活控制韵律和风格。启用韵律编码器后，用户可以从参考波形中提取全局韵律嵌入。该嵌入会注入梅尔频谱和文本条件流，以实现精准的风格迁移。此外，我们还引入了无参考模式，不再使用参考音频作为条件，使模型能够生成带有随机采样说话人特征且发音更自然的语音。这促使模型依赖关于音色和能量的内部先验，从而实现纯文本驱动合成。

5. LEMAS-Edit

我们通过大幅扩展 VoiceCraft 的架构来开发 LEMAS-Edit [28], 将其从单语英语模型转变为多语言语音编辑框架。我们的贡献包括扩充训练数据、重新设计分布式训练框架，以及引入自适应解码策略以确保生成稳定性。

5.1 多语言适配与训练

关键在于，当句子中的语言上下文发生切换时，会动态插入新的语言标签。这种结构使自回归模型能够在话语中途切换其内部语言上下文，从而自然支持混合语言编辑任务。

我们重新设计了训练基础设施，以支持大规模分布式学习。该框架现在能够处理多个异构数据集和 GPU 秩之间的数据分片，确保数据均衡消费。一项重要的优化是简化优化步骤。在原始实现中，使用梯度累积来模拟更大的批次，这增加了训练动态的复杂性。我们完全移除此机制，改为对完整批次进行单步更新。这项改动不仅简化了代码库，还使批次大小与有效梯度更新之间的关系更加清晰，并增强了在大型 GPU 集群上处理长多语言序列时的训练稳定性和抵御内存碎片化的能力。

5.2 推理与解码策略

历史感知重复控制。 自回归编解码器模型常会出现重复循环，尤其是在长篇生成中。为缓解这一问题，我们在 top-k/top-p采样过程中引入了一种动态历史感知惩罚机制。与静态惩罚不同，我们的方法会随生成长度变化。我们根据先前生成词元的出现频率，按照以下公式调整其对数几率：

$$\text{penalty} = \frac{\text{repetition_penalty}}{100} \cdot \text{num_gen} + 1. \tag{4}$$

其中，num_gen 表示当前生成的词元数量。对于历史记录中出现的任意词元，其正对数几率会除以该惩罚值，而负对数几率则会乘以该惩罚值。随着序列增长，这会逐渐提高重复词元的代价，从而有效抑制静音循环和内容卡顿，同时不降低对初始提示的遵循度。

自适应重新生成机制。 为确保生成的音频时长符合自然语音模式，我们采用多轮自适应推理方案。首先，通过分析参考音频，确定基准语速：

$$\texttt{avg_speed} = \frac{\#\text{encodec frames}}{\#\text{raw text tokens}}. \tag{5}$$

这一估算出的语速使我们能够对解码器施加动态 max_rate 约束，防止生成失控。

然而，简单的约束不足以应对复杂编辑。因此，我们引入了反馈循环。在生成过程中，如果模型触发内部异常标志（RE_GEN），或者输出长度异常短（具体而言，$< 0.5 \times$ avg_speed），系统就会自动启动重新生成轮次。在每次

后续轮次中，我们通过扩大编辑掩码边界（提供更多声学上下文）并增加重复_penalty，迭代放宽约束。这个迭代过程会持续进行，直到输出符合预期的韵律结构，或达到重试上限。

5.3 多语言编辑前端

我们升级了推理前端，以支持专业级多语言工作流程，重点提升准确性和灵活性：

稳健识别与对齐。 精确的文本与音频对齐对于准确的语音编辑至关重要。为提升转录的稳健性，我们用多语言 Whisper 大模型取代标准的英语自动语音识别。此外，我们采用 MMS 强制对齐器，以获取更精确的词级时间戳。这些细粒度的时间信息对于构建精确的编辑掩码至关重要，可确保修改严格局限于目标区域，不影响相邻词语。

信号增强。 为处理真实场景中的录音，处理流程集成了两种不同的降噪后端。用户可以选择 UVR5[4] 进行轻度降噪，以保留背景环境声（这对于维持编辑后语音的自然度至关重要）；也可以选择 DeepFilterNet[5]，以强力抑制低质量输入中的噪声。

长篇处理。 一次性处理较长录音往往会降低注意力性能。我们的系统会自动将长音频文件分割成易于处理的分块，分别处理编辑内容，再通过在过零点处进行平滑交叉淡化，将结果拼接起来。这确保了整段录音的感知连续性。

统一界面。 系统在单一界面中提供两种不同的操作模式：*TTS* 模式，使用参考文本和音频，从句末开始向前自回归地生成语音；以及编辑模式，用于修改现有音频。后者借助 LEMAS-Dataset 精确的对齐能力，支持对起始和结束时间进行细粒度的人在回路调整，从而弥合自动化模型与专业编辑需求之间的差距。

6. 实验

我们基于 LEMAS-Dataset 构建了一个统一的多语言基准，用于评估文本转语音合成和基于文本的语音编辑。该基准着重评估发音稳定性、跨语言泛化能力和词级编辑自然度，并利用数据集中细粒度的词级时间戳。对于多语言 TTS 评估，我们从 10 种语言中每种随机抽取 10 条话语，并使用所有目标文本合成语音，得到 10,000 个样本，涵盖语言内和跨语言设置。客观评估采用词错误率（WER）和说话人相似度指标。对于语音编辑，我们在 7 种语言的对齐语句中随机替换连续词语，构建词级编辑案例，并通过主观 A/B 测试评估自然度，重点关注局部编辑后的边界平滑度和感知连贯性。所有评估均遵循统一的数据构建和推理协议，以确保不同模型和任务之间的公平比较。为便于复现，我们将发布评估列表和生成脚本。

6.1 LEMAS-TTS 评估

我们在10种语言上对 LEMAS-TTS 进行评估，并将其与同期的开源基线模型 OpenAudio-S1-mini[6]，进行比较，后者支持相同的语言集合。对于中文，使用 FunASR Paraformer-zh[7] 评估词错误率（WER）；对于其他所有语言，则使用 Whisper-large-v3[8] 进行评估。同时，遵循 Seed-TTS 评估协议 [1]，使用 WavLM-large[9]，计算说话人相似度（SIM）。由于缺乏中文和英语以外语言的母语评估者，因此未进行主观评估。

[4]https://ultimatevocalremover.com [5] https://github.com/Rikorose/DeepFilterNet [6] https://huggingface.co/spaces/fishaudio/openaudio-s1-mini [7] https://huggingface.co/funasr/paraformer-zh [8] https://huggingface.co/openai/whisper-large-v3 [9]https://huggingface.co/microsoft/wavlm-large

表 4 - 多语言测试集上文本转语音模型的客观评估。词错误率（WER）用于衡量转录准确率（越低越好），而相似度（SIM）用于评估与目标说话人的感知接近度（越高越好）。结果比较了 OpenAudio-S1-Mini 和 LEMAS-TTS 模型在有无韵律控制时的表现，涵盖 10 种语言。

语言	模型	词错误率↓(%)	SIM↑
de	OpenAudio-S1-Mini	3.95	0.448
	LEMAS-TTS（无韵律控制）	1.20	0.533
	LEMAS-TTS（含韵律）	**0.44**	**0.535**
en	OpenAudio-S1-Mini	4.98	0.451
	LEMAS-TTS（不含韵律）	2.21	**0.557**
	LEMAS-TTS（含韵律）	**1.07**	0.556
es	OpenAudio-S1-Mini	7.66	0.488
	LEMAS-TTS（不含韵律）	4.93	**0.560**
	LEMAS-TTS（带韵律）	**4.28**	0.552
fr	OpenAudio-S1-Mini	13.41	0.458
	LEMAS-TTS（不带韵律）	10.45	0.517
	LEMAS-TTS（带韵律）	**9.87**	**0.518**
id	OpenAudio-S1-Mini	32.77	0.518
	LEMAS-TTS（不带韵律）	7.05	**0.550**
	LEMAS-TTS（含韵律）	**6.38**	0.534
it	OpenAudio-S1-Mini	9.27	0.489
	LEMAS-TTS（不含韵律）	5.57	**0.544**
	LEMAS-TTS（含韵律）	**4.43**	0.537
pt	OpenAudio-S1-Mini	8.05	0.476
	LEMAS-TTS（不含韵律）	3.12	**0.555**
	LEMAS-TTS（含韵律）	**2.09**	0.546
ru	OpenAudio-S1-Mini	20.75	0.475
	LEMAS-TTS（不含韵律）	12.29	**0.545**
	LEMAS-TTS（含韵律）	**11.84**	0.538
vi	OpenAudio-S1-Mini	-	0.451
	LEMAS-TTS（不含韵律）	22.14	**0.541**
	LEMAS-TTS（含韵律）	**15.38**	0.531
zh	OpenAudio-S1-Mini	9.59	0.550
	LEMAS-TTS（不含韵律）	4.64	**0.570**
	LEMAS-TTS（含韵律）	**3.00**	0.546
Avg	OpenAudio-S1-Mini	12.27	0.480
	LEMAS-TTS（不含韵律）	8.06	**0.547**
	带韵律的 LEMAS-TTS	**6.39**	0.539

如表 4所示，LEMAS-TTS 在所有语言上的表现均持续优于 OpenAudio-S1-mini，取得了更低的词错误率和更高的说话人相似度。我们进一步比较了 LEMAS-TTS 的两个版本，即带有和不带有韵律编码器的版本。具备韵律感知能力的模型始终能取得更低的词错误率，表明其发音稳定性有所提升，但说话人相似度略有降低。主观听感表明，韵律条件控制以表达丰富性为代价换取了更高的发音清晰度稳定性。因此，我们同时发布这两个版本，以满足不同的应用偏好。对于越南语，OpenAudio-S1-mini 的词错误率在多次测量中始终异常偏高；因此，我们将该结果排除在平均词错误率计算之外，以避免其对汇总结果产生不成比例的影响。

6.2 LEMAS-Edit 的评估

VoiceCraft [28] 最初仅使用英语 GigaSpeech 语料库进行训练。相比之下，LEMAS-Edit 使用从 LEMAS-Dataset、WenetSpeech4TTS、GigaSpeech 和 MLS 中精选的子集进行训练。由此形成的合并训练语料库涵盖七种主要语言（中文、英语、德语、法语、葡萄牙语、西班牙语、意大利语），使

模型能够学习跨语言音系规则和韵律迁移。为了高效适应多语言环境，我们采用热启动初始化策略，使用预训练的 3.3 亿参数 VoiceCraft 检查点初始化 LEMAS-Edit。这一设计使 LEMAS-Edit 能够在不改变模型架构的情况下，将 VoiceCraft 扩展至多语言语音编辑。

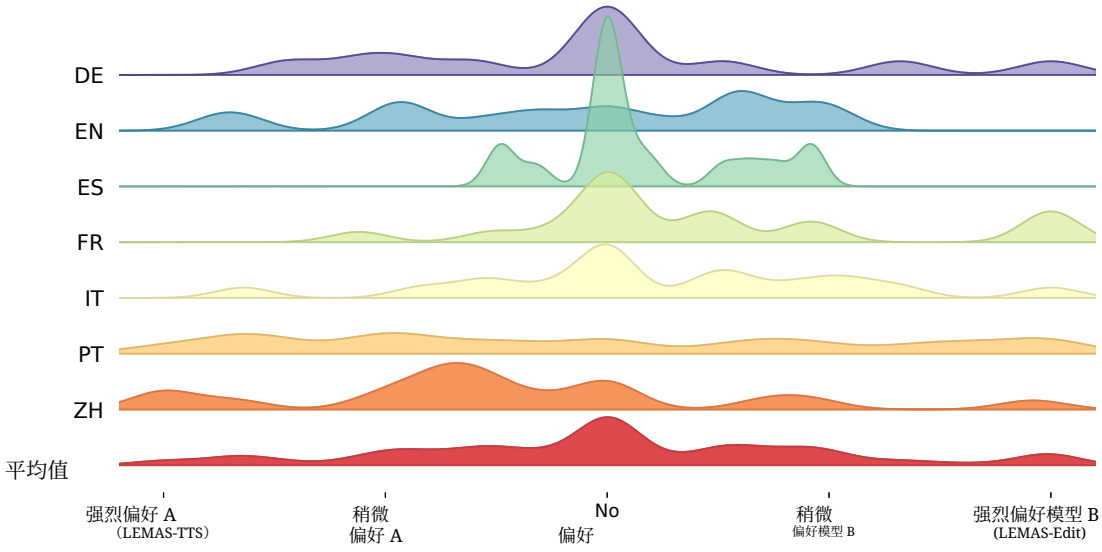

图 4 - 不同语言下音频自然度的偏好分布。山脊线图展示了 A/B 偏好测试的主观结果，对比了*LEMAS-TTS*（模型 A）和*LEMAS-Edit*（模型 B）。各项得分均归一化至 0–100 的范围，其中 0 表示强烈偏好模型 A，100 表示强烈偏好模型 B。每种语言的垂直分布（核密度估计）揭示了用户的整体意见，“平均值”行则呈现了所有受测语言的汇总表现。

我们进一步评估了 LEMAS-TTS 和 LEMAS-Edit 在基于文本的语音编辑任务上的表现。从 LEMAS-Dataset 评估集中随机选取了 20 条话语，涵盖 7 种语言（每种语言 2–3 条话语）。对于每条话语，均提供精确的词级对齐，并使用 ChatGPT 随机替换单个词或短语，以生成编辑案例。六位人类听众通过 A/B 偏好测试评估编辑后音频的自然度（图 4）。尽管不同语言之间存在细微的偏好差异，但听众的整体判断较为均衡，表明 LEMAS 为多语言语音编辑评估提供了稳健且与架构无关的基准测试集。所有评估样本均可通过演示页面公开访问。

7. 结论

在这项工作中，我们介绍了**LEMAS**，一个配备生成式语音模型、拥有 15 万小时音频的大规模可扩展多语言音频套件，为大规模多语言生成式语音研究提供统一且可扩展的基础。LEMAS-Dataset 涵盖 10 种语言、超过 15 万小时的音频，并配有高质量的词级时间戳和置信度分数，为训练和评估新一代生成式语音模型奠定了基础。我们从两个互补的角度展示了**LEMAS-Dataset** 的价值。通过**LEMAS-TTS**，我们表明，精心设计的对齐与解耦目标对于稳定大规模多语言流模型至关重要。通过**LEMAS-Edit**，我们验证了精准的时间监督是实现稳健零样本语音编辑的关键。总的来说，这些结果表明，生成式语音建模的进步不仅日益取决于模型架构，也取决于底层数据的规模、多样性和粒度。我们希望 LEMAS 能成为社区共享的基准和实用资源，推动多语言语音合成、语音编辑和可控音频生成领域的进一步研究。

致谢

我们谨向以下人士致以诚挚谢意，感谢他们提供的宝贵帮助和富有见地的讨论。以下姓名按姓氏字母顺序排列：Xie Chen、Chen Cheng、Zhen Guo、Jiaxin Huang、Rui Li、Fei Peng、Daniel van Strien、Tiezhen Wang、Baogang Yao、Fei Yu、Xiaoling Zhang、Changyin Zhou。

参考文献

[1] Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao、Mingqing Gong、Peisong Huang、Qingqing Huang、Zhiying Huang、Yuanyuan Huo、Dongya Jia、Chumin Li、Feiya Li、Hui Li、Jiaxin Li、Xiaoyang Li、Xingxing Li、Lin Liu、Shouda Liu、Sichao Liu、Xudong Liu、Yuchen Liu、Zhengxi Liu、Lu Lu、Junjie Pan、Xin Wang、Yuping Wang、Yuxuan Wang、Zhengnan Wei、Jian Wu、Chao Yao、Yifeng Yang、Yuan-Qiu-Qiang Yi、Junteng Zhang、Qidi Zhang、Shuo Zhang、WenJie Zhang、Yang Zhang、Zilin Zhao、Dejian Zhong 和 Xiaobin Zhuang。《Seed-TTS：一系列高质量、通用的语音生成模型》。*arXiv* 预印本 *arXiv:2406.02430*，2024年。

[2] Rosana Ardila、Megan Branson、Kelly Davis、Michael Henretty、Michael Kohler、Josh Meyer、Reuben Morais、Lindsay Saunders、Francis M Tyers 和 Gregor Weber。Common Voice：一个超大规模多语言语音语料库。载于第十二届语言资源与评测会议论文集，2020年。

[3] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：语音表征自监督学习框架。载于神经信息处理系统进展，第33卷，第12449–12460页，2020年。

[4] Zal´an Borsos、Rapha¨el Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matthew Sharifi、Dominik Roblek、Olivier Teboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour。AudioLM：一种音频生成的语言建模方法。*IEEE ACM* 音频、语音与语言处理汇刊，31：2523–2533，2023年。

[5] Edresson Casanova、Kelly Davis、Eren G¨olge、G¨orkem G¨oknar、Iulian Gulea、Logan Hart、Aya Aljafari、Joshua Meyer、Reuben Morais、Samuel Olayemi 和 Julian Weber。Xtts：一款支持海量语言的零样本文本转语音模型。*arXiv* 预印本 *arXiv:2406.04904*，2024年。

[6] CETUC。CETUC Alcaim：欧洲葡萄牙语语音数据集。网址 http://www02.smt.ufrj.br/~igor.quintanilha/alcaim.tar.gz。访问日期：2025-12-28。

[7] Guoguo Chen、Shuzhou Chai、Guanbo Wang、Jiayu Du、Wei-Qiang Zhang、Chao Weng、Dan Su、Daniel Povey、Jan Trmal、Junbo Zhang 等。Gigaspeech：一个不断发展的多领域自动语音识别语料库，包含 10,000 小时的转录音频。*arXiv* 预印本 *arXiv:2106.06909*，2021年。

[8] Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li 等。神经编解码器语言模型是零样本文本转语音合成器。*IEEE*音频、语音与语言处理汇刊，2025年。

[9] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu 和 Xie Chen。F5-TTS：一种通过流匹配生成流畅且忠实语音的讲故事者。*arXiv* 预印本 *arXiv:2410.06885*，2024年。

[10] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma 等。CosyVoice：一种基于监督式语义令牌的可扩展多语言零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2407.05407*，2024年。

[11] Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma 等。CosyVoice：一种基于监督式语义令牌的可扩展多语言零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2407.05407*，2024年。

[12] Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等。CosyVoice 2：基于大语言模型的可扩展流式语音合成。*arXiv* 预印本 *arXiv:2412.10117*，2024年。

[13] Zhihao Du、Changfeng Gao、Yuxuan Wang、Fan Yu、Tianyu Zhao、Hao Wang、Xiang Lv、Hui Wang、Chongjia Ni、Xian Shi、Keyu An、Guanrou Yang、Yabin Li、Yanni Chen、Zhifu Gao、Qian Chen、Yue Gu、Mengzhe Chen、Yafeng Chen、Shi-Min Zhang、Wen Wang 和 Jieping Ye。CosyVoice 3：通过扩大规模和后训练，迈向真实场景语音生成。*arXiv* 预印本 *arXiv:2505.17589*，2025年。

[14] eSpeak NG。eSpeak NG（版本 1.51）。https://github.com/espeak-ng/espeak-ng/tree/1.51，2022年。访问日期：2025年12月25日。[15] Mark JF Gales、Shakti P Rath 等。低资源语言的语音识别与关键词检测：剑桥大学工程系 Babel 项目研究。资源匮乏语言口语技术，2014年。[16] Yaroslav Ganin 和 Victor Lempitsky。通过反向传播实现无监督领域自适应。载于《国际机器学习会议》，第 1180–1189 页。PMLR，2015年。[17] Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber。连接主义时序分类：利用循环神经网络标注未分段序列数据。载于《第23届国际机器学习会议论文集》，第 369–376 页，2006年。[18] Hao-Han Guo、Yao Hu、Kun Liu、Fei-Yu Shen、Xu Tang、Yi-Chen Wu、Feng-Long Xie、Kun Xie 和 Kai-Tuo Xu。FireRedTTS：面向工业级生成式语音应用的基础文本转语音框架。arXiv 预印本 arXiv:2409.03283，2024年。[19] Haorui He、Zengqiang Shang、Chaoren Wang、Xuyuan Li、Yicheng Gu、Hua Hua、Liwei Liu、Chen Yang、Jiaqi Li、Peiyang Shi 等。Emilia：用于大规模语音生成的广泛、多语言且多样化的语音数据集。载于《2024 IEEE口语技术研讨会（SLT）》，第 885–890 页。IEEE，2024年。[20] Jacob Kahn、Morgane Riviere、Weiyi Zheng、Eugene Kharitonov、Qiantong Xu、Pierre-Emmanuel Mazar´e、Julien Karadayi、Vitaliy Likhomanenko、Vineel Pratap、Awni Kahn 等。Libri-Light：一个面向有限监督或无监督自动语音识别的基准测试集。载于《IEEE国际声学、语音与信号处理会议（ICASSP）》，第 7669–7673 页，2020。[21] Nikolay Karpov、Alexander Denisenko 和 Fedor Minkin。Golos：用于语音研究的俄语数据集。arXiv 预印本 arXiv:2106.10161，2021年。[22] Jaehyeon Kim、Jungil Kong 和 Juhee Son。结合对抗学习的条件变分自编码器，用于端到端文本转语音。载于《国际机器学习会议》，第 5530–5540 页。PMLR，2021年。[23] Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 等。VoiceBox：由文本引导的规模化多语言通用语音生成。神经信息处理系统进展，36：14005–14034，2023年。[24] Xinjian Li、Shinnosuke Takamichi、Takaaki Saeki、William Chen、Sayaka Shiota 和 Shinji Watanabe。YODAS：面向YouTube的音频与语音数据集。载于《2023 IEEE自动语音识别与理解研讨会（ASRU）》，第 1–8 页。IEEE，2023年。[25] Linhan Ma、Dake Guo、Kun Song、Yuepeng Jiang、Shuai Wang、Liumeng Xue、Weiming Xu、Huan Zhao、Binbin Zhang 和 Lei Xie。WenetSpeech4TTS：一个用于大型语音生成模型基准测试的 12,800 小时普通话文本转语音语料库。arXiv 预印本 arXiv:2406.05763，2024年。[26] M. McAuliffe、M. Socolof、S. Mihuc、M. Wagner 和 M. Sonderegger。蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本与语音对齐。载于 Interspeech 2017 论文集，第 498–502 页，2017年。[27] Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。载于《IEEE国际声学、语音与信号处理会议（ICASSP）》，2015年。[28] Puyuan Peng、Po-Yao Huang、Shang-Wen Li、Abdelrahman Mohamed 和 David Harwath。VoiceCraft：真实场景中的零样本语音编辑与文本转语音。arXiv 预印本 arXiv:2403.16973，2024年。[29] Vineel Pratap、Qiantong Xu、Anuroop Sriram、Gabriel Synnaeve 和 Ronan Collobert。MLS：用于语音研究的大规模多语言数据集。载于 Interspeech，2020。[30] Vineel Pratap、Andros Tjandra、Bowen Shi、Paden Tomasello、Arun Babu、Sravya Kundu、Ali Elkahky、Zhaoheng Ni、Apoorv Vyas 等。将语音技术扩展至 1,000+ 种语言。《机器学习研究期刊》，24(368)，2023年。[31] 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。第 32 卷，2019年。[32] Elizabeth Salesky、Matthew Wiesner、Jacob Bremerman、Roldano Cattoni、Matteo Negri、Marco Turchi、Douglas W. Oard 和 Matt Post。用于语音识别与翻译的多语言 TEDx 语料库。arXiv 预印本 arXiv:2102.01757，2021年。

[33] Tanja Schultz 和 Tim Schlippe。GlobalPhone：涵盖 20 种语言的多语言文本与语音数据库。载于 IEEE国际声学、语音与信号处理会议，2013年。[34] Seamless Communication，Lo¨ıc Barrault、Yu-An Chung、Mariano Cora Meglioli、David Dale、Ning Dong、Paul- Ambroise Duquenne、Hady Elsahar、Hongyu Gong、Kevin Heffernan、John Hoffman、Christopher Klaiber、Pengwei Li、Daniel Licht、Jean Maillard、Alice Rakotoarison、Kaushik Ram Sadagopan、Guillaume Wenzek、Ethan Ye、Bapi Akula、Peng-Jen Chen、Naji El Hachem、Brian Ellis、Gabriel Mejia Gonzalez、Justin Haaheim、Prangthip Hansanti、Russ Howes、Bernie Huang、Min-Jae Hwang、Hirofumi Inaguma、Somya Jain、Elahe Kalbassi、Amanda Kallet、Ilia Kulikov、Janice Lam、Daniel Li、Xutai Ma、Ruslan Mavlyutov、Benjamin Peloquin、Mohamed Ramadan、Abinesh Ramakrishnan、Anna Sun、Kevin Tran、Tuan Tran、Igor Tufanov、Vish Vogeti、Carleigh Wood、Yilin Yang、Bokai Yu、Pierre Andrews、Can Balioglu、Marta R. Costa-juss`a、Onur Celebi、Maha Elbayad、Cynthia Gao、Francisco Guzm´an、Justine Kao、Ann Lee、Alexandre Mourachko、Juan Pino、Sravya Popuri、Christophe Ropers、Safiyyah Saleem、Holger Schwenk、Paden Tomasello、Changhan Wang、Jeff Wang 和 Skyler Wang。《SeamlessM4T：大规模多语言、多模态机器翻译》，2023年。[35] Kai Shen、Zeqian Ju、Xu Tan、Eric Liu、Yichong Leng、Lei He、Tao Qin、Sheng Zhao 和 Jiang Bian。NaturalSpeech 2：潜在扩散模型实现自然的零样本语音与歌声合成器。载于第十二届国际学习表征会议，2024年。[36] Fish Audio团队。Fish Speech：OpenAudio S1 Mini。https://huggingface.co/fishaudio/openaudio-s1-mini，2024年。[37] Changhan Wang、Morgane Riviere、Ann Lee、Anne Wu、Chaitanya Talwalkar、Elizabeth Xiao、Yatharth Saraf、Piyush Puvvada 和 Emmanuel Dupoux。VoxPopuli：用于表征学习、半监督学习和解释的大规模多语言语音语料库。载于计算语言学协会（ACL），2021年。[38] Xinsheng Wang、Mingqi Jiang、Ziyang Ma、Ziyu Zhang、Songxiang Liu、Linqin Li、Zheng Liang、Qixi Zheng、Rui Wang、Xiaoqin Feng 等。Spark-TTS：一种采用单流解耦语音标记的高效基于大语言模型的文本转语音模型。arXiv预印本 arXiv:2503.01710，2025年。[39] Yuancheng Wang、Haoyue Zhan、Liwei Liu、Ruihong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu。MaskGCT：采用掩码生成式编解码器变换器的零样本文本转语音。arXiv预印本 arXiv:2409.00750，2024年。[40] Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio 等。Tacotron：迈向端到端语音合成。2017年。[41] Kun Xie、Feiyu Shen、Junjie Li、Fenglong Xie、Xu Tang 和 Yao Hu。FireRedTTS-2：迈向播客和聊天机器人的长篇对话语音生成。arXiv预印本 arXiv:2509.02020，2025年。[42] Yifan Yang、Zheshu Song、Jianheng Zhuo、Mingyu Cui、Jinpeng Li、Bo Yang、Yexing Du、Ziyang Ma、Xunying Liu、Ziyuan Wang 等。GigaSpeech 2：一种不断演进的大规模多领域自动语音识别语料库，面向低资源语言，并采用自动化爬取、转录和精炼。载于《第63届计算语言学协会年会论文集》（第1卷：长篇论文），第2673–2686页，2025年。[43] Binbin Zhang、Hang Lv、Pengcheng Guo、Qijie Shao、Chao Yang、Lei Xie、Xin Xu、Hui Bu、Xiaoyu Chen、Chenchen Zeng 等。WenetSpeech：一个用于语音识别的 $10000+$ 小时多领域普通话语料库。载于ICASSP 2022-2022 IEEE国际声学、语音与信号处理会议（ICASSP），第6182–6186页。IEEE，2022年。[44] Siyi Zhou、Yiquan Zhou、Yi He、Xun Zhou、Jinchao Wang、Wei Deng 和 Jingchen Shu。IndexTTS2：富有情感表现力、时长可控的自回归零样本文本转语音领域的一项突破。arXiv预印本 arXiv:2506.21619，2025年。