Listen, Attend, Understand：针对高方差标签的稳定端到端语音翻译训练正则化技术

Yacouba Diarra 和 Michael
LeventhalRobotsMali AI4D 实验室，巴马科，马里 diarray 和 mleventhal |@robotsmali.org

arXiv:2601.01121v1 [cs.CL] 3 Jan 2026

摘要

当目标转录文本存在较高方差和语义歧义时，端到端语音翻译往往收敛较慢、性能也较差。我们提出 Listen, Attend, Understand（LAU），这是一种语义正则化技术，可在训练期间约束声学编码器的潜在空间。LAU 利用冻结的文本嵌入提供方向性辅助损失，在不增加推理成本的情况下，为声学表征注入语言学层面的依据。我们在一个班巴拉语到法语数据集上评估了该方法，该数据集包含由非专业人士翻译的 30 小时班巴拉语语音。实验结果表明，与使用多出 100% 数据预训练的端到端语音翻译系统相比，LAU 模型在标准指标上的表现相当，同时在保留语义方面表现更佳。此外，我们提出总参数漂移这一指标，用以量化正则化的结构影响，并证明语义约束会主动重组编码器权重，使其优先考虑意义而非字面语音学特征。研究结果表明，LAU 是事后重评分的一种稳健替代方案，也是端到端语音翻译训练的重要补充，尤其适用于训练数据稀缺和／或噪声较大的情况。

1 引言

继 Listen, Attend and Spell（LAS）（Chan 等人，2015年）取得成功后，能够同时学习语音识别器所有组件的端到端神经网络架构已成为自动语音识别领域的主流范式，并在多个基准测试中达到人类水平的表现（Xiong 等人，2016年）。然而，即使在某些基准测试中已接近人类水平，自动语音识别系统仍会产生不可忽视的错误。对于一些特别敏感的应用，使用外部预训练语言模型对自动语音识别模型的 n-best 候选转录文本（通常称为假设）进行第二轮重打分或事后校正，已成为标准做法（Ogawa 等人，2018年；Futami 等人，2020年；Zhao 等人，2021年；Chen 等人，2023年；Li 等人，2024年）。将现代大语言模型与自动语音识别级联使用，可以通过纠正第一轮自动语音识别假设中的语义和语法错误来提升转录质量，但第二轮校正需要运行大型大语言模型，因而会带来资源开销。

SpecAugment 等训练时鲁棒性方法（Park 等人，2019年）解决了问题的声学层面。针对转录文本和翻译文本歧义的鲁棒性研究相对较少，研究重点逐渐转向推理阶段的事后纠正。标签歧义会减缓收敛并降低训练稳定性。在语音翻译中，这一问题尤为突出：即使结合上下文，一个源语言词语或短语通常也有多种可接受的翻译（Prior 等人，2010）。在低资源语言（低资源语言）中，这一问题通常会进一步加剧，因为有限的数据和较少的标准化标注规范会增加不同标注人员及领域之间的差异，而规模较小的语料库也为模型提供了更少的示例来平均消除这种噪声。

本文介绍一种用于训练端到端语音翻译模型（**LAU：聆听、关注、理解**，简称端到端语音翻译）的新型正则化技术，旨在引导声学编码器输出的表征也存在于由预训练嵌入模型定义的高资源语义空间中，而非仅存在于声学空间中。我们使用 jeli-asr 在班巴拉语到法语的语音翻译任务上进行了实验，所用数据集为 30 小时半专业翻译数据集（Diarra 等人，2022年）。实验结果表明，这一新的训练约束确实发挥了作用，在参数总漂移量（即初始与最终编码器权重之差的 L2 范数）方面表现出明显差异；尽管预训练数据显著更少，该方法在语义导向任务上的表现仍优于传统的端到端语音翻译模型。

我们的两项主要贡献如下 s:

• 我们发现，对于班巴拉语到法语的语音翻译，一个仅使用63小时班巴拉语语音训练的简单编码器—解码器端到端语音翻译模型（其中只有30小时语音提供了法语翻译文本），其表现优于级联式自动语音识别到机器翻译的流程。这一结果为低资源语言设置提供了相关证据：尽管直接端到端学习通常更难，但避免两个低资源组件之间的错误累积可能是有益的。

•我们为端到端语音翻译实现了一种新的约束，通过迁移学习稳定高方差、非专业翻译文本上的训练，迫使编码器的输出在外部预训练嵌入空间中承载有意义的信息。我们使用同一个30小时数据集进行实验，并在4中讨论结果。

2 相关工作

2.1 端到端语音翻译

给定三元组训练语料库 <语音、转录、翻译文本>，高质量端到端语音翻译的传统训练方法是先使用<语音、转录> 配对数据预训练自动语音识别基线，然后利用 <语音、翻译文本>配对数据进一步优化。然而，端到端语音翻译近期的大多数突破都来自双任务方法：通过基于Transformer的架构使用多任务提示词元，或联合训练两个解码器（分别对应一项任务），使神经网络模型同时执行自动语音识别和语音翻译（Radford 等人，2022年；Puvvada等人，2024），Ko等人，2021；Chuang 等人，2020年）。

先前研究表明，在高资源场景下，只要有足够的数据，直接端到端语音翻译就能胜过传统的自动语音识别到机器翻译级联流程（Weiss等人，2017年；Sperber等人，2019年）。在第4.2 节和第4.3节中，我们将展示，对于班巴拉语这样的低资源语言，即使只有30小时的数据，也可能呈现相同的定性趋势。除了比较端到端与级联范式，近期研究还通过在语音编码器与大语言模型之间学习一个可训练的连接模块（例如 Dense、Transformer 或 Q-former），探索了语音模型与大语言模型之间更紧密的集成方式，从而使大语言模型能够将语音衍生的表示作为输入（Šimon Sedlá˘cek等人，2024；Yu等人，2023）。

为了让端到端语音翻译训练更加稳健，Ko等人还在多任务训练中加入了自动语音识别混淆这一因素。他们提出了一种基于后验的自动语音识别损失，将目标词元替换为预训练自动语音识别模型生成的后验分布。这种更柔和的参考编码了相互竞争的假设，并将自动语音识别混淆注入学习信号。由于语音翻译解码器将自动语音识别解码器和编码器的状态都作为输入，这种能够感知不确定性的自动语音识别约束有助于翻译模型应对识别歧义。该方法在他们的Fisher西英数据集上取得了更好的双语评估替补分数（Ko等人，2021年）。

Du 等人认为，传统的端到端语音翻译训练方法在每个阶段都将三元组数据拆分为二元组数据（数据对），因而未能利用三元组数据之间的关联。他们训练单个编码器-解码器模型，通过双路径连续解码方法学习转录和翻译的联合概率。解码器按照自动语音识别-机器翻译的顺序，或反向按照语音翻译-回译的顺序，输出两个任务拼接后的对数概率。他们还在模型训练目标中引入了两个库尔贝克-莱布勒散度正则化项，以减少两条路径输出概率之间的不匹配。与传统的解耦方法相比，该方法在MuST-C基准上的自动语音识别和语音翻译任务中均取得了提升（Du 等人，2022年）。

2.2 在口语理解中引入语义

近期研究指出，自监督学习（SSL）语音编码器（Chung 等人，2019年；Baevski 等人，2020年）广泛用作口语理解任务（如自动语音识别、意图分类和问答）中的通用前端，但往往更注重声学和语音学结构，其表征中的语义信息相对较弱（Pasad 等人，2022年）。

沿着这一思路，Xu 等人利用无监督自动语音识别（ASR-U）作为桥接模块，将语音编码器嵌入转换为大语言模型使用的子词标记输入，从而为 SSL 语音编码器注入来自大语言模型的语义信息。他们的桥接模块沿用了 wav2vec-U 2.0 的做法，使用生成对抗网络（GAN）组件从 SSL 特征中预测音素序列

随后应用基于词典构建的 WFST 解码器；该词典编码了已知的音素到子词的映射，用于将预测的音素转换为大语言模型的子词标记。尽管他们在多个语音语言理解任务上取得了令人瞩目的成果，且无需带标注的自动语音识别数据（Xu 等人，2023a），但其方法较为复杂，需要额外引入自动语音识别-理解桥接模块，而且研究重点在于通用表征框架。相比之下，我们的 LAU 实现以任务为导向，仅添加一项辅助约束，专门用于稳定特定语言对的端到端语音翻译，而非优化旨在广泛迁移到各种任务的通用表征。

Chuang 等人提出了一种语义感知端到端语音翻译方法，通过在双解码器设置中鼓励自动语音识别解码器的隐藏状态与预训练的源语言词嵌入对齐。在我们的实现中，我们采用单解码器架构，并在编码器层级应用嵌入对齐约束：使用一个仅用于训练的语义头，将编码器输出投影到与嵌入模型相同的维度（在我们的设置中为目标语言）。这样可以约束音频输入中信息最丰富的表示，避免其被映射到有限的输出词汇表之后才进行约束。在推理时，语义头会被移除，因此该正则化项不会带来额外开销。

在高维嵌入空间中，余弦相似度是实现此类对齐的常用选择，因为它衡量向量之间的夹角，而夹角通常与语义相关性相关；此外，许多嵌入模型采用能够促进方向一致性的目标进行训练（Mikolov 等人，2013；Grill 等人，2020年；Chen 等人，2020年）。

最后，尽管余弦相似度因具有尺度不变性，且符合许多嵌入模型的训练目标而被广泛使用，但近期研究表明，嵌入幅值也能编码有意义的语义信号，而这些信号会被余弦相似度忽略（You，2025年）。受这一观察启发，在构建编码器层级嵌入对齐正则化项时，我们同时考虑基于余弦相似度和对幅值敏感的对齐损失。

3 实现与实验

我们使用了 `RobotsMali/soloni-114m-tdt-ctc-v0`（soloni-v0），这是基于英伟达 1.14 亿参数 Parakeet 模型针对班巴拉语微调的模型（Rekesh 等人，2023年；Xu 等人，2023b），在 jeli-asr（Diarra等人，2022年）上训练；以及 `RobotsMali/soloni-114m-tdt-ctc-v1`（soloni-v1），在此模型基础上使用 Kunkado（Diarra等人，2025b）进行微调，作为我们的自动语音识别基线模型。

我们还使用 jeli-asr 的 <语音、翻译> 对对 soloni-v1 进行了训练，共训练 12,000 步；采用 AdamW 优化器和 Noam 调度器，学习率缩放因子为 2.0，预热步数为 1k（Vaswani 等人，2017），并将其用作传统端到端语音翻译模型。

我们在所有 LAU 实验中均使用 soloni-v0 作为基线，并将 `dangvantuan/sentence-camembert-base`用作预训练法语嵌入模型（Nils Reimers，2019年；Martin 等人，2020年）；随后，我们添加了一个浅层语义头，其中仅包含两个全连接层，用于将编码器的输出映射到与预训练嵌入相同的维度。我们希望语义头保持轻量，以便语义损失对编码器产生更大的影响。我们使用序列到序列任务（自动语音识别/语音翻译）的损失与语义损失的加权和，对更新后的架构进行训练

$$\mathcal{L}_{\text{LAU}} = \mathcal{L}_{\text{Seq}} + \lambda \times \mathcal{L}_{\text{semantic}} \tag{1}$$

其中 λ 是用于控制正则化强度的超参数。

序列损失。该项是我们要学习的主要序列到序列任务（本实验中的语音翻译）的损失函数，通常为转导器损失。在我们的案例中，由于 Parakeet 模型采用了混合架构，包含两个相互独立的解码器（TDT 和 CTC），因此该项也是多个损失之和：

$$\mathcal{L}_{\text{Seq}} = (1-\alpha) \times \mathcal{L}_{\text{TDT}} + \alpha \times \mathcal{L}_{\text{CTC}} \tag{2}$$

其中 α 也是一个超参数；其中：

$$\mathcal{L}_{\text{TDT}} = -\log P_{\text{TDT}}(\mathbf{y}|\mathbf{x}) \tag{3}$$

其中 $P_{\text{TDT}}(\mathbf{y}|\mathbf{x})$ 是目标序列 $\mathbf{y}$ 的总概率，通过对词元（包括空白符）和时长的所有有效联合对齐进行边缘化计算得出（Xu 等人，2023b）；以及：

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{y})} P(\pi|\mathbf{x}) \tag{4}$$

其中 $\mathcal{B}^{-1}(\mathbf{y})$表示映射到目标序列的所有可能帧级路径的集合

y 在移除空白符和重复项之后（Graves 等人，2006）。

我们的实现基于英伟达的 NeMo 工具包，该工具包已经实现了这些训练目标（Kuchaiev 等人，2019年）。我们对该开源工具包的源代码仅做了少量修改，以集成嵌入模型和语义损失。

语义损失。对于语义正则化项，我们尝试了两种经典的相似度度量：余弦嵌入损失（5）和均方误差损失（6）。

$$\mathcal{L}_{\text{cosine}} = 1 - \frac{Y \cdot \hat{Y}}{\|Y\|\|\hat{Y}\|} \tag{5}$$

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{6}$$

ˆ其中，Y 是语义头的输出，Y 是通过使用我们的法语嵌入模型对参考翻译进行编码而得到的参考嵌入向量。

训练期间，SentenceTransformer 模型保持冻结，仅为计算语义损失提供参考嵌入。语义损失的梯度仅流入浅层语义头和编码器，从而利用预训练嵌入空间中的迁移学习。语音翻译部分独立训练。在推理阶段，语义头会被移除，因此推理图与基线模型完全相同，并且可以使用 CTC 或 TDT 解码器进行解码。图1展示了训练和推理图。

4 结果

我们使用所提出的架构进行了训练实验，并分别测试了以余弦损失和均方误差作为语义损失且不设置权重的情况，作为基线（$\lambda = 1.0$）。我们还探究了不同语义损失权重（$\lambda = 0.2$ 和 $\lambda = 5.0$）下正则化的影响；不过，由于计算资源有限，我们仅针对余弦损失进行了不同权重的实验（相关观察结果可能同样适用于均方误差损失）。所有实验均在 soloni-v0 上运行了 11，800 步，并采用与我们的传统端到端语音翻译模型相同的优化器配置。

4.1 用于正则化的 LAU

我们通过总参数漂移来评估 LAU 正则化的有效性，即比较训练前后编码器权重差值的 L2 范数（7）。该指标反映了正则化损失对编码器变化的约束程度：数值越小，更新幅度越小，也就是说有效正则化越强。我们的实验表明，均方误差损失最适合用于正则化，且 λ 的最优值1.0左右。更低的值（例如 0.2）会导致更大的参数变化，使模型对高方差翻译文本过拟合；而更高的值会使辅助语义目标主导主要任务，从而导致更大幅度的更新，如图2所示。这与近期研究结果一致：在某些情境下，余弦相似度可能并不是最佳的语义相似度度量，其尺度不变性可能导致它忽略重要的语义信息（You，2025年）

$$\text{Drift} = ||\Theta_{final} - \Theta_{initial}||_2 \tag{7}$$

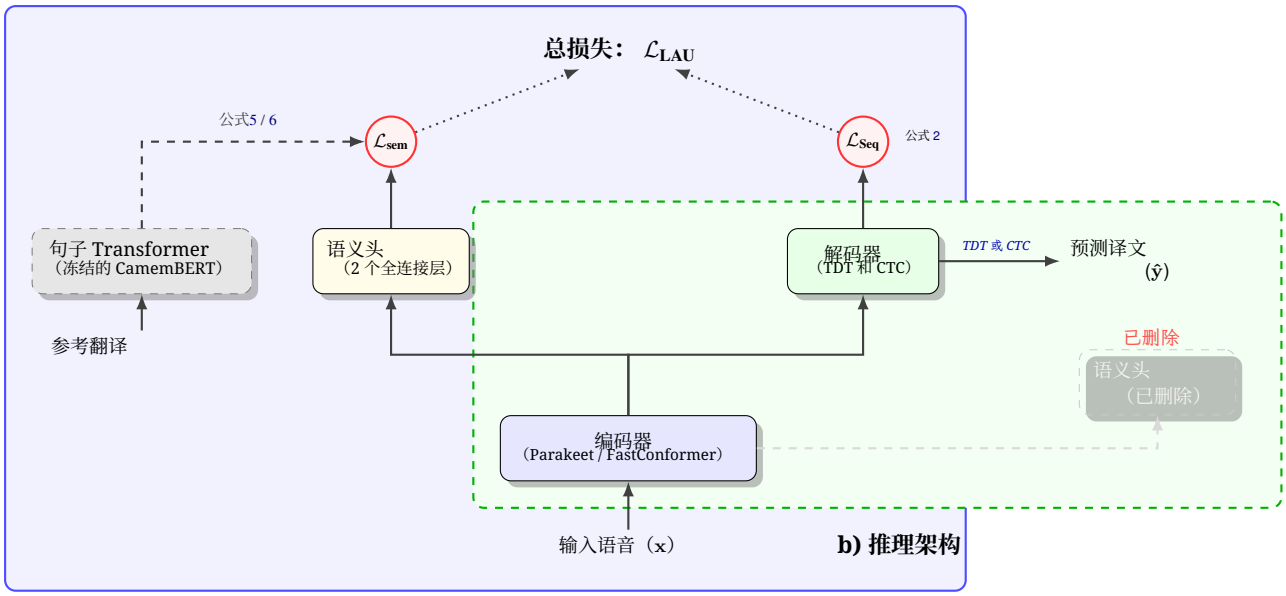
其中 Θ 表示编码器的权重。

4.2 双语评估替补与词错误率评估

双语评估替补（BLEU）、词错误率（WER）和字符错误率（CER）是评估语音翻译系统的标准指标。表1 报告了我们的端到端语音翻译模型（soloni-v1）的 BLEU-4、WER 和 CER 得分。该模型使用 30 小时的 jeli-asr 音频和 33 小时的 Kunkado 音频进行训练；还报告了经 LAU 训练的模型（soloni-v0）的得分，该模型使用 30 小时的 jeli-asr 音频进行训练（Diarra等人，2022年），并将其与采用 soloni-v0 和 Djelia 机器翻译模型的自动语音识别→机器翻译流水线进行比较（Djelia 团队，2025年）。

模型（解码）	词错误率↓	字符错误率↓	双语评估替补↑
~~自动语音识别-机器翻译（TDT）~~	0.9109	0.6884	0.0880
端到端语音翻译（TDT）	**0.7043**	0.5817	**0.2418**
LAU-λ0.2（CTC）	0.7451	0.5742	0.1642
LAU-λ5（CTC）	0.7455	**0.5683**	0.1611
LAU-λ1（CTC）	0.7553	0.5858	0.1592
LAU-均方误差-λ1（~~CTC~~）	0.7608	0.5864	0.1429

表1：模型在词错误率、字符错误率和双语评估替补指标上的性能对比。由于每个模型都有两个解码器，因此报告了能够取得最佳双语评估替补分数的解码策略。粗体数值表示各项指标的最佳分数。



a) 提出的训练架构

图1：**a)** 训练图在编码器输出上添加一个浅层语义头，使其与冻结的参考嵌入对齐（LAU正则化）。**b)** 推理期间会移除语义分支，从而形成标准的编码器-解码器架构。

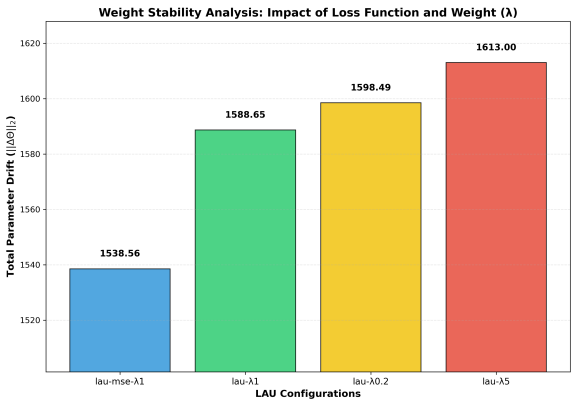


图2：不同语义正则化损失选项及权重下的总参数漂移

尽管 LAU 模型的表现不及完全预训练的端到端语音翻译模型（该模型使用多 33 小时的自动语音识别数据进行预训练，Diarra等人，2025b），但它们的表现已非常接近，甚至在字符错误率方面超越了该模型（−1.3% 左右），这一结果来自采用余弦损失训练的 LAU 模型，并且 $\lambda = 5$。

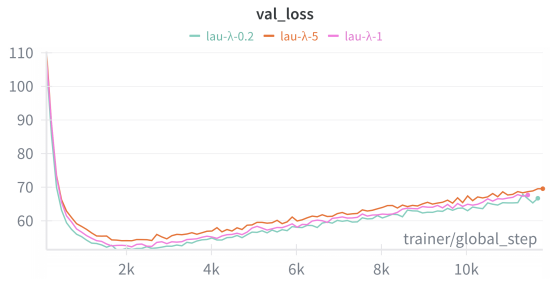
最值得注意的是，完全预训练的端到端语音翻译模型和所有经LAU训练的变体，在全部三项指标上都优于级联的自动语音识别到机器翻译流水线。这一结果拓展了此前在资源更丰富的环境中取得的发现 Weiss等人，2017；Sperber等人，2019年 并支持端到端训练在资源受限场景中的价值，因为两个低资源组件之间的误差累积可能会造成尤为严重的影响。

叠加效应可能尤其有害。

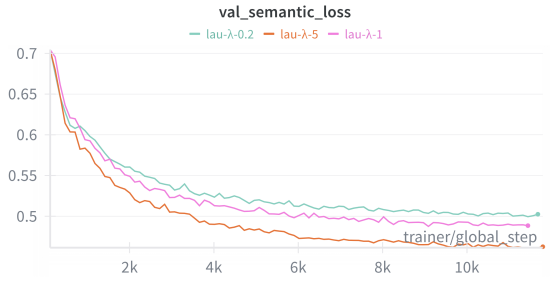
4.3 LAU 在语音语言理解中的潜力

在他们的多任务实现中，Chuang 等人，2020年报告称，集成嵌入模型后，词错误率有所下降，而双语评估替补得分有所提升。在我们的实现中，嵌入模型保持冻结状态，语义损失仅作为辅助正则项使用，这似乎使词错误率和双语评估替补得分更加一致（表1）。尽管如此，我们仍观察到多目标干扰：总验证损失起初下降，但随后上升，而语义损失则持续下降。这种模式表明存在真正的权衡，说明语义损失确实发挥作用，并对共享编码器施加了有意义的约束，而 TDT/CTC目标必须适应这一约束（见图3）。

Chuang 等人，2020年的研究结果还表明，如果我们想衡量模型输出是否具有正确的语义，词错误率和双语评估替补可能并非合适的指标。有鉴于此，我们使用两项语音语言理解风格的任务评估模型：基于大语言模型的问答（问答），以及基于主题的音频聚类（主题音频聚类）。问答任务用于评估模型的输出是否提供了足够的信息，使大语言模型能够根据音频回答问题或执行单次指令



(a) 总加权损失和（1）



(b) 语义损失（5）

图3：多任务冲突表明，新的正则化技术已启动；λ值越高，越会促使编码器进行更多更新，以符合语义约束，但这会损害词汇准确率

内容，而主题音频聚类则评估输出在多大程度上支持按主题或意图对话语进行分组。

对于LLM-QA，我们使用经过指令微调的LLAMA-3.1-8B模型，根据测试集中的音频话语生成了2230个法语问题和答案，并根据参考翻译为每条音频话语分配一个主题（Grattafiori 等人，2024）。我们使用同一模型，仅依据给定模型的输出回答生成的问题，最后还将其用于大语言模型评审流程，以评估大语言模型答案的正确性（Gu 等人，2025）。

该模型将测试集划分为6个主要主题组：家庭与出身、知识传承、社会与文化、冲突与竞争、宗教与灵性以及“其他主题”。对于TAC任务，我们使用k均值聚类，仅根据不同模型的输出，通过最小化其嵌入的余弦距离，在同一测试数据中创建6个簇。表2 展示了不同模型的大语言模型问答准确率（Acc）、簇纯度（Pur）和归一化互信息（NMI）。

簇纯度。簇纯度表示每个 k均值簇中“占主导地位”的真实主题所占的平均比例。纯度为 1。0 这意味着每个簇都只包含一个由大语言模型指定的主题。

$$Purity = \frac{1}{N}\sum_{k=1}^{K}\max_{j}|\omega_k \cap c_j| \tag{8}$$

其中，N：样本总数。ω_k：第 k 个 k均值簇中的点集，c_j 第 j 个真实类别（由大语言模型确定）中的点集。

归一化互信息（NMI）。纯度衡量同质性，而 NMI 通过同时考虑同质性和完整性来反映整体一致性。尤其是，如果某个语义类别（主题）被拆分到多个簇中，NMI 会对这种碎片化进行惩罚。它可以反映 k 均值簇与大语言模型推导出的簇有多吻合。NMI 能揭示聚类模型是否只分离出少数几个主题，却未能捕捉其他主题，而仅靠纯度可能无法识别这种情况。

$$NMI(\Omega, C) = \frac{2 \times I(\Omega; C)}{H(\Omega) + H(C)} \tag{9}$$

其中：$I(\Omega; C)$：衡量 k 均值簇（Ω）与真实类别（C）之间共享的信息量。$H(\Omega)$ 和 $H(C)$ 分别是 k 均值簇和由大语言模型分配的簇的熵。

模型（解码）	准确率 ↑	纯度 ↑	归一化互信息 ↑
自动语音识别-机器翻译 (tdt)	0.3165	0.5947	0.0590
端到端语音翻译 (tdt)	0.3762	0.6177	0.0543
LAU-λ0.2（ctc）	0.3789	0.6184	0.0623
LAU-λ5 (ctc)	0.3721	0.6150	0.0592
LAU-λ1 (tdt)	0.3740	**0.6279**	0.0673
LAU-均方误差-λ1 (ctc)	**0.3834**	0.6218	**0.0705**

表 2：不同模型在 LLM-QA准确率、簇纯度和归一化互信息方面的性能比较。

这项评估表明，经LAU训练的模型保留了更多语义信息，这有利于下游ST，尤其是在与大语言模型结合时，也有利于其他口语语言理解任务。在所有模型中，“家庭与出身”这一主题都能以很高的准确率（65–91%）被识别出来。然而，若干受评模型将同一主题拆分到多个簇中，而每个簇都以“家庭与出身”示例为主，

这种现象可能会因底层的类别不平衡而加剧。因此，纯度仍相对较高，因为各簇内部较为同质；但归一化互信息仍然较低（6–7%），因为聚类未能形成完整且一一对应的主题划分。最后，尽管自动语音识别—机器翻译流程取得了有竞争力的归一化互信息（5.9%），但在准确率和纯度两项指标上都明显不及其他模型[1]。

5 讨论

5.1 低资源语言语音应用的巨大潜力

班巴拉语采用音位文字，这意味着字素与音素之间存在一致的对应关系（Konta 和 Vydrin，2014年；Vydrin，2022年），这本应使其非常适合语音识别；然而，其他因素（标准不稳定、文盲率较高，以及口语表达优先于书面表达）使得为班巴拉语自动语音识别开发标注数据尤其困难。尽管端到端语音翻译是一项比自动语音识别复杂得多的任务，但近期研究表明，与寻找能够准确转写母语的班巴拉语使用者相比，寻找具备足够法语读写能力、能够将班巴拉语语音直接翻译成法语的班巴拉语使用者可能更容易（Diarra等人，2025a）。

我们的研究结果有助于为许多面临类似读写能力问题的低资源语言开发实用的语音应用。创建语音翻译数据集可能是一种成本更低、成效更高的数据策略；由于大多数现代应用都依赖于已经在高资源语言领域发展成熟的大语言模型，端到端语音翻译有望成为与这些应用集成的一条更快速、更实用的路径。

在本文中，我们不仅提出了一种能够稳定在非专业翻译数据集上进行端到端语音翻译训练的技术，还展示了通过在语音编码器中加入语义约束，LAU 在面向大语言模型的应用和语音理解方面优于传统的端到端语音翻译（4.3）。

5.2 自动语音识别的无监督近似

基于主题的音频聚类（TAC）虽然在此仅用于评估，但也可以支持有限的免转录语义索引，从而应用于原本需要依赖自动语音识别的场景。如果数据库中已经包含一部分带有主题或意图标签的项目，则可以利用这些已标注示例为 TAC 识别出的簇命名或提供锚定。随后，可以将新的未标注音频分配到最近的簇，从而无需词级转录即可实现主题级搜索和筛选。

由于翻译文本提供了每个话语的文本表示，这种方法还可以扩展到跨语言聚类：将语音映射到一种高资源语言，从而更可靠地推导类别标签。这是一个很有前景的方向，因为我们的 LAU 结果表明，即使翻译文本在词汇上并非完全准确，只要语义忠实，仍能保留有效无监督聚类所需的信息。

6 结论

在本研究中，我们提出了一种新颖的正则化框架，旨在应对高方差训练标签，稳定端到端语音翻译。通过强制声学编码器与预训练语义嵌入空间之间实现语义对齐，我们证明，即使不进行大规模预训练，模型也能达到有竞争力的词汇准确率（字符错误率），并捕捉更多底层语义信息（LLM-QA准确率、归一化互信息）。此外，我们对受约束语音编码器参数更新幅度的分析证实，语义损失是一种主动的优化力量；较高的权重（$\lambda = 5$）会促使编码器发生有意义的结构性重组。最终，LAU正则化为开发更稳健的语音翻译系统提供了一条计算效率高的途径，尤其适用于低资源语言。

作者贡献

Yacouba Diarra提出了研究方向，设计并开展了实验，并撰写了稿件。Michael Leventhal作为首席研究员负责监督项目，并参与了稿件的编辑。

局限性

尽管取得了令人鼓舞的结果，本研究仍存在若干局限。首先，LAU正则化的有效性取决于目标语言（法语）中是否有高质量的预训练文本嵌入模型。其次，尽管多任务冲突的发生以及参数

[1] jeli-asr 测试集 及其模型生成的翻译文本、端到端语音翻译模型和 LAU-mse 模型，均可在 RobotsMali 的 Hugging Face 个人主页公开获取；页面还提供了训练和评估所用全部代码与提示词的链接。

漂移分析表明该机制确实在发挥作用，但我们尚未确定正则化权重 λ 的理论上限，超过该上限后，自动语音识别/语音翻译的性能可能会崩溃。最后，我们的评估仅限于一个相对较小的语料库，且实验预算有限，因此模型和超参数探索的范围受到限制。同时，这一设置或许也有其益处，因为它反映了许多低资源语言研究所面临的实际条件：在这些条件下，收集平行数据并进行大规模实验往往不可行。还需要在更多样的语言对和更大的数据集上进一步验证，以确认这种“语义牵引”能否推广到资源更丰富的语言。

参考文献

Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。wav2vec 2.0：一种用于语音表示自监督学习的框架。预印本，arXiv:2006.11477。

WilliamChan、Navdeep Jaitly、Quoc V.Le 和 Oriol Vinyals。2015年。Listen, attendandspell。预印本，arXiv:1508.01211。

Chen Chen、Yuchen Hu、Chao-Han Huck Yang、Sabato Marco Siniscalchi、Pin-Yu Chen 和 Eng-Siong Chng。2023年。Hyporadise：一个使用大语言模型进行生成式语音识别的开放基线。神经信息处理系统进展，36:31665–31688。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 GeoffreyHinton。2020 年。一种用于视觉表示对比学习的简单框架。预印本，arXiv:2002.05709。

Shun-Po Chuang、Tzu-Wei Sung、Alexander H. Liu 和 Hung-yi Lee。2020年。词错误率更高，但双语评估替补得分更好？将词嵌入用作多任务端到端语音翻译中的中间表示。载于计算语言学协会第*58*届年会论文集，第5998–6003页，线上。计算语言学协会。

Yu-An Chung、Wei-Ning Hsu、Hao Tang 和 James Glass。2019年。用于语音表示学习的无监督自回归模型。预印本，arXiv:1904.03240。

Sebastien Diarra、Michael Leventhal 和 Allahsera Auguste Tapo。2022年。Robotsmali Griots语音数据集及自动语音识别。https://github.com/robotsmali-ai/jeli-asr/。

Yacouba Diarra、Nouhoum Souleymane Coulibaly 和 Michael Leventhal。2025a。以口语为主的语言中人工校正转录的成本分析。预印本，arXiv:2510.12781。

Yacouba Diarra、Panga Azazia Kamate、Nouhoum Souleymane Coulibaly 和 Michael Leventhal。2025年b。 Kunnafonidilaw ka cadeau：一份关于当代班巴拉语的自动语音识别数据集。 预印本，arXiv:2512.19400。Djelia团队。2025年。Djelia：专为班巴拉语打造的首个人工智能平台。 https://djelia.cloud。面向班巴拉语的先进文本转语音、翻译和转录模型。Yichao Du、Zhirui Zhang、Weizhi Wang、Boxing Chen、Jun Xie 和 Tong Xu。2022年。 通过三角分解一致性正则化端到端语音翻译。 预印本，arXiv:2112.10991。Hayato Futami、Hirofumi Inaguma、Sei Ueno、Masato Mimura、Shinsuke Sakai 和 Tatsuya Kawahara。2020年。 蒸馏 BERT 的知识用于序列到序列自动语音识别。 预印本，arXiv:2008.03822。Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan、Amy Yang、Angela Fan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark 以及其他542位作者。2024年。 Llama 3模型群。 预印本，arXiv:2407.21783。Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。2006年。 连接主义时序分类：使用循环神经网络为未分段序列数据添加标签。载于 *2006*年国际机器学习会议，第2006卷，第369–376页。Jean-Bastien Grill、Florian Strub、Florent Altché、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad Gheshlaghi Azar、Bilal Piot、koray kavukcuoglu、Remi Munos 和 Michal Valko。2020年。自举自身潜变量 —— 一种新的自监督学习方法。载于 神经信息处理系统进展，第33卷，第21271–21284页。Curran Associates, Inc. Jiawei Gu、Xuhui Jiang、Zhichao Shi、Hexiang Tan、Xuehao Zhai、Chengjin Xu、Wei Li、Yinghan Shen、Shengjie Ma、Honghao Liu、Saizhuo Wang、Kun Zhang、Yuanzhuo Wang、Wen Gao、Lionel Ni 和 Jian Guo。2025年。 关于大语言模型评审的综述。预印本，arXiv:2411.15594。Yuka Ko、Katsuhito Sudoh、Sakriani Sakti 和 Satoshi Nakamura。2021年。基于自动语音识别后验的多任务端到端语音翻译损失。载于 *Interspeech*，第2272–2276页。Mamadou Konta 和 Valentin Vydrin。2014年。班巴拉语正字法提案。*Mandenkan*，（52）：3–38。Oleksii Kuchaiev、Jason Li、Huyen Nguyen、Oleksii Hrinchuk、Ryan Leary、Boris Ginsburg、Samuel Kriman、Stanislav Beliaev、Vitaly Lavrukhin、Jack Cook、Patrice Castonguay、Mariya Popova、Jocelyn Huang,

以及 Jonathan M. Cohen。2019年。 NeMo：使用神经模块构建人工智能应用的工具包。预印本，arXiv:1909.09577。

Sheng Li、Chen Chen、Chin Yuen Kwok、Chenhui Chu、Eng Siong Chng 和 Hisashi Kawai。2024年。探究利用大语言模型和多语言一优假设进行自动语音识别错误纠正。收录于《语音处理会议论文集》，第 2024 卷，第 1315–1319 页。

Louis Martin、Benjamin Muller、Pedro Javier Ortiz Suárez、Yoann Dupont、Laurent Romary、Éric Ville- monte de la Clergerie、Djamé Seddah 和 Benoît Sagot。2020年。CamemBERT：一种美味的法语语言模型。计算语言学协会第 *58* 届年会论文集。

Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。2013年。高效估算词表示的向量空间。预印本，arXiv:1301.3781。

Iryna Gurevych Nils Reimers。2019年。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。*https://arxiv.org/abs/1908.10084*。

Atsunori Ogawa、Marc Delcroix、Shigeki Karita 和 Tomohiro Nakatani。2018年。基于编码器分类器模型进行一对一假设比较的语音识别 N-best列表重排序。载于*2018*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 6099–6103 页。

Daniel S. Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D. Cubuk 和 Quoc V. Le。2019年。Specaugment：一种用于自动语音识别的简单数据增强方法。载于*Interspeech 2019* 会议论文集，第 2613–2617 页。ISCA。Ankita Pasad、Ju-Chieh Chou 和 Karen Livescu。2022年。自监督语音表征模型的逐层分析。预印本，arXiv:2107.04734。Anat Prior、Shuly Wintner、Brian MacWhinney 和 Alon Lavie。2010年。语境内外的翻译歧义。应用心理语言学，32(1):93–111。获取地址：https://doi.org/10.1017/S0142716410000305。Krishna C. Puvvada、Piotr ˙Zelasko、He Huang、Oleksii Hrinchuk、Nithin Rao Koluguri、Kunal Dhawan、Somshubra Majumdar、Elena Rastorgueva、Zhehuai Chen、Vitaly Lavrukhin、Jagadeesh Balam 和 Boris Ginsburg。2024年。少即是多：无需网络规模数据即可实现准确的语音识别与翻译。预印本，arXiv:2406.19674。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。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。Dima Rekesh、Nithin Rao Koluguri、Samuel Kriman、Somshubra Majumdar、Vahid Noroozi、He Huang、Oleksii Hrinchuk、Krishna Puvvada、Ankur Kumar、Jagadeesh Balam 和 Boris Ginsburg。2023年。采用线性可扩展注意力机制的快速 Conformer，实现高效语音识别。预印本，arXiv:2305.05084。

Matthias Sperber、Graham Neubig、Jan Niehues 和 Alex Waibel。2019年。用于稳健且数据高效的端到端语音翻译的注意力传递模型。计算语言学协会会刊，7：313–325。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需要的一切。预印本，arXiv:1706.03762。

Valentin Feodosievich Vydrin。2022年。《班巴拉语正字法词典》。*Mandenkan*：《曼德语语言学研究半年刊》，（68）：59–82。

Ron J. Weiss、Jan Chorowski、Navdeep Jaitly、Yonghui Wu 和 Zhifeng Chen。2017年。序列到序列模型可以直接翻译外语语音。预印本，arXiv:1703.08581。

Wayne Xiong、Jasha Droppo、Xuedong Huang、Frank Seide、Michael Seltzer、Andreas Stolcke、Dong Yu 和 Geoffrey Zweig。2016年。在会话语音识别方面达到人类水平。*IEEE/ACM* 音频、语音与语言处理汇刊，PP。

DerekXu、ShuyanDong、ChanghanWang、SuyounKim、Zhaojiang Lin、Bing Liu、Akshat Shrivastava、Shang-Wen Li、Liang-Hsuan Tseng、Guan-Ting Lin、Alexei Baevski、Hung-yi Lee、Yizhou Sun 和 Wei Wang。2023a。将语义引入语音编码器。载于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第11413–11429页，多伦多，加拿大。计算语言学协会。Hainan Xu、Fei Jia、Somshubra Majumdar、He Huang、Shinji Watanabe 和 Boris Ginsburg。2023b。通过联合预测词元和时长实现高效的序列转导。预印本，arXiv:2304.06795。Kisung You。2025年。语义的角度：余弦相似度何时有效，何时失效。预印本，arXiv:2504.16318。Wenyi Yu、Changli Tang、Guangzhi Sun、Xianzhao Chen、Tian Tan、Wei Li、Lu Lu、Zejun Ma 和 Chao Zhang。2023年。连接语音编码器与大语言模型以实现自动语音识别。预印本，arXiv:2309.13963。Yun Zhao、Xuerui Yang、Jinchao Wang、Yongyu Gao、Chao Yan 和 Yuanfu Zhou。2021年。基于 BART 的普通话自动语音识别系统语义纠错。*arXiv* 预印本 *arXiv:2104.05507*。Šimon Sedlá˘cek、Santosh Kesiraju、Alexander Polok、˘and Jan Cernocký。2024年。为口语翻译对齐预训练模型。预印本，arXiv:2411.18294。