S2SCORE：面向语音到语音翻译系统的无文本统一评估指标

Lalaram Arya* Mrinmoy Bhattacharjee† S. R. Mahadeva Prasanna**印度理工学院达尔瓦德分校电气、电子与计算机工程系，达尔瓦德，印度
202021004@iitdh.ac.in, prasanna@iitdh.ac.in
†印度理工学院查谟分校计算机科学与工程系，查谟，印度mrinmoy.bhattacharjee@iitjammu.ac.in

arXiv:2611.29557v2 [eess.AS] 24 Nov 2026

摘要——**评估语音到语音翻译（S2ST）系统，需要考察翻译语音的多个维度，包括语义充分性、说话人相似度和声学自然度。现有方法大多基于文本，依赖自动语音识别转录文本，且主要评估语义充分性；其他维度则需要使用单独的指标。我们提出 S2SCORE，这是一种统一的无文本评估框架，通过多任务架构联合建模这些维度，并生成经人工校准的总体质量分数。S2SCORE 使用源自成熟评估模型的监督信号进行训练，无需直接进行人工标注即可实现可扩展学习。在五种语言对上进行评估后，S2SCORE 与人类判断、独立参考指标及训练监督指标均表现出高度一致性，表明其具有稳健的泛化能力。无参考版本 S2SCORE-QE 始终优于 S2SCORE-REF，在语义充分性预测方面也优于 BLASER** 2。0。

索引词——**语音质量评估、语音到语音翻译、自动评估、统一评估指标。**

I. 引言

语音到语音翻译（S2ST） [1] 通过将源语言语音转换为目标语言语音，实现跨语言的口语交流。近年来，S2ST [2] 逐渐转向端到端架构，避免使用中间的自动语音识别（ASR） [3],、机器翻译（MT） [4], 和文本转语音（TTS） [5]组件，从而降低延迟并减少错误传播。然而，移除中间文本表示也带来了新的评估挑战。

传统上，翻译质量通过 BLEU [6], 、METEOR [7],、BLEURT [8],、COMET [9], 和 MetricX-24 [10]等基于文本的指标进行评估。现代自动语音识别系统 [11] 使基于自动语音识别的评估成为语义评估的有力基准。然而，由于这些方法仅依赖文本，它们会舍弃与翻译质量评估相关的声学信息。此外，这些方法还需要额外的自动语音识别系统，增加了计算开销，并可能引入转录错误。

BLASER [12] 推出了首批基于语音、通过人工判断监督训练的语义充分性指标之一，而 BLASER 2.0 [13] 则将其扩展到了大规模多语言和无参考值评估。与此同时，有研究提出了一种无文本语音到语音翻译指标 [14], 无需依赖转录文本即可近似基于文本的指标；而 SpeechQE [15]最近提出了一个基于大语言模型（LLM）的语音翻译（ST）质量估计框架。

尽管取得了这些进展，现有的无文本语音到语音翻译评估方法仍主要关注语义充分性。对于现代语音到语音翻译系统 [16], ，输出质量是多维的：译文不仅应保留原意，还应保持说话人身份和自然度。因此，评估依赖针对各个维度的独立指标，导致评估流程碎片化并增加了评估开销 [17]。据我们所知，目前尚无无文本评估框架能够联合建模这些维度并提供统一的总体质量分数。

为解决这些局限，我们提出 S2SCORE，这是一种统一的无文本评估框架，采用多任务架构联合建模语义充分性、说话人相似度和声学自然度，同时生成一个经人工判断校准的学习式总体质量分数。由于收集大规模人工判断成本高且耗时，该框架使用来自成熟评估指标的监督信号进行训练：使用 xCOMET-XXL [18]评估语义充分性，使用 WavLM-SV [19] 评估说话人相似度，使用 UTMOSv2 [20] 评估声学自然度。S2SCORE 利用 SONAR [21], 、WavLM-SV [19],和 log-Mel 频谱图 [22] 特征，实现可扩展的多语言评估。训练和评估使用源自 CVSS [23] 和 BhasaAnuvaad [24] 语料库的多语言语音到语音翻译数据，涵盖法语（fr）、西班牙语（es）、德语（de）、中文（zh）和印地语（hi）语音译为英语（en）的任务；数据由级联系统和直接式语音到语音翻译系统生成，包含多种翻译质量条件。实验结果表明，S2SCORE 能够有效评估语音的多维质量，与人工判断高度一致，并独立的参考指标。主要贡献如下：

- 我们提出了 S2SCORE，这是一种统一的无文本语音到语音翻译评估框架，联合建模语义充分性、说话人相似度和声学自然度，并生成经人工校准的总体质量分数。
- 我们提出了一种可扩展的监督策略，利用成熟的指标训练多维度语音评估模型，而无需大规模人工标注。
- 我们在五种语言对上评估了 S2SCORE，结果表明，它与人工判断和独立参考指标高度一致。

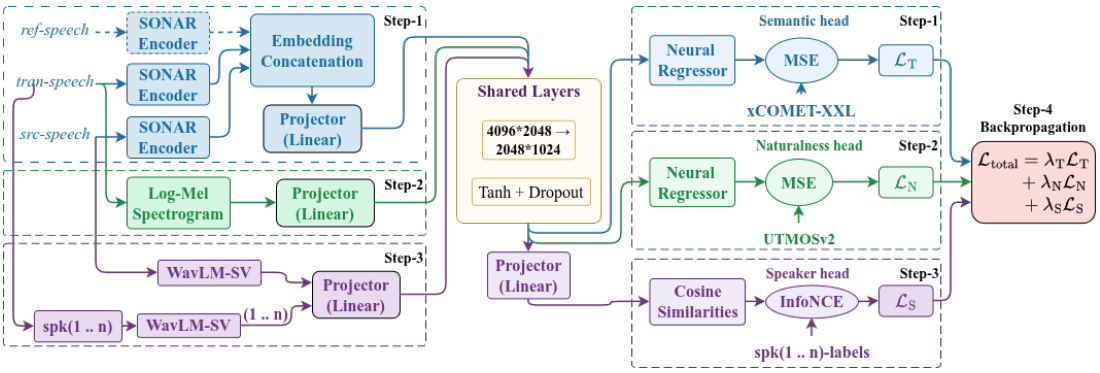

图 1. S2SCORE 的架构。语义、声学和说话人表示被投影到共享的多层感知机中，并由特定任务的预测头进行处理，这些预测头分别由 xCOMET-XXL [18], UTMOSv2 [20],和WavLM-SV [19]监督。虚线箭头表示仅在基于参考的变体中使用的参考语音。

II. S2SCORE 框架

S2SCORE采用统一的多头架构，对语音到语音翻译系统中的语义充分性、说话人相似度和声学自然度进行评估，同时生成统一的总体质量分数。如图1所示，该框架由任务专属的特征提取器、共享的多层感知机模块（S-MLP）和三个预测头组成。输入语音被编码为任务专属表示，投影到共同的嵌入空间，并经由共享的S-MLP处理，随后路由至相应的预测头。最终得到的语义充分性、说话人相似度和自然度分数，随后会汇总为统一的总体质量分数。下文将介绍各个组成部分。

A. 语音特征表示

S2SCORE 从三个维度评估翻译语音，每个维度均采用特定任务的特征提取器进行表示。语义表示由 SONAR 编码器提取；该编码器提供多语言语音和文本嵌入，训练数据涵盖 57 种语音语言和 200 多种文本语言，因此非常适合跨语言评估。声学自然度由对数梅尔频谱图表示，它能捕捉与感知相关的时频特征。说话人相似度则通过预训练 WavLM-SV 嵌入建模；WavLM-SV 是一种说话人验证模型，使用约 9.4 万小时的语音进行训练，可提供稳健的说话人表示。

B. 共享多层感知机层

按照硬参数共享方法 [25]，语义、说话人和自然度头共享一个通用的 S-MLP 模块。S-MLP 将 4096 维输入映射为共享表示 $\mathbf{h} \in \mathbb{R}^{1024}$，该过程通过两个全连接层实现，层的单元数分别为 2048 和 1024，且每层后均接 *tanh*激活函数和 Dropout（$p = 0.1$）。所得表示随后被传递至相应的预测头，S-MLP 的参数则通过三个任务损失的加权和进行联合优化。

C. 评估头

评估头以 S-MLP 的输出（第 II-B 节）为输入，预测特定于任务的得分。

1) 翻译评估头：翻译头使用 SONAR 嵌入 [21]估计语义充分性。根据是否有参考语音，构建了 S2SCORE 的两种变体，即 S2SCORE-QE 和 S2SCORE-REF。遵循 BLASER 2.0 [13]，利用源嵌入与译文嵌入之间的逐元素乘积和绝对差构建交互特征；在 REF变体中，还额外加入参考语嵌入。所得特征经投影后由 S-MLP 处理，以生成 $\mathbf{h}_{\mathrm{T}} \in \mathbb{R}^{1024}$。随后，回归头 $f_{\mathrm{T}}(\cdot)$ 预测翻译质量分数，公式如下：

$$\hat{S2SCORE}_{\mathrm{T}} = f_{\mathrm{T}}(\mathbf{h}_{\mathrm{T}}) \tag{1}$$

使用 xCOMET-XXL [18] 分数 $\hat{y}_{\mathrm{xCOMET-XXL}}$ 作为监督信号，计算均方误差（MSE）损失，公式如下：

$$\mathcal{L}_{\mathrm{T}} = \|\hat{S2SCORE}_{\mathrm{T}} - \hat{y}_{\mathrm{xCOMET-XXL}}\|^2 \tag{2}$$

2) 声学自然度头：自然度头遵循 UTMOSv2 [20]，根据对数梅尔频谱图特征 [22]，估计翻译语音的感知质量。从翻译信号中提取对数梅尔频谱图 $\mathbf{M} \in \mathbb{R}^{T\times F}$，将其展平后，通过任务特定的投影器映射到共同的 4096 维空间。随后，投影后的表示由 S-MLP 处理以生成$\mathbf{h}_{\mathrm{N}} \in \mathbb{R}^{1024}$，并传递给回归头 $f_{\mathrm{N}}(\cdot)$：

$$\hat{S2SCORE}_{\mathrm{N}} = f_{\mathrm{N}}(\mathbf{h}_{\mathrm{N}}) \tag{3}$$

该头部通过与 UTMOSv2 [20]分数 y_{UTMOSv2} 计算均方误差进行优化：

$$\mathcal{L}_{\mathrm{N}} = \|\hat{S2SCORE}_{\mathrm{N}} - \hat{y}_{\mathrm{UTMOSv2}}\|^2 \tag{4}$$

3）说话人相似度头：为评估说话人保持情况，将源语句与 N 个翻译候选配对，其中包含一个正匹配和 $N-1$ 个负样本。使用以下方法提取说话人嵌入 $\mathbf{e}^{\mathrm{sv}}_{\mathrm{src}}$ 和 $\mathbf{e}^{\mathrm{sv}}_{\mathrm{tran},\ i}$

WavLM-SV [19], , 并投影到共享的4096维特征空间，再由 S-MLP 处理以获得 $\mathbf{h}_{\mathrm{S}} \in \mathbb{R}^{1024}$。最终投影将这些表示映射为128维嵌入， $\tilde{\mathbf{e}}_{\mathrm{src}}^{\mathrm{sv}}$ 和 $\tilde{\mathbf{e}}_{\mathrm{tran},\,i}^{\mathrm{sv}}$，并使用余弦相似度计算对比得分，定义如下：

$$S2\hat{SCORE}_{\mathrm{S}}^{i} = \frac{\tilde{\mathbf{e}}_{\mathrm{src}}^{\mathrm{sv}} \cdot \tilde{\mathbf{e}}_{\mathrm{tran},i}^{\mathrm{sv}}}{\|\tilde{\mathbf{e}}_{\mathrm{src}}^{\mathrm{sv}}\| \|\tilde{\mathbf{e}}_{\mathrm{tran},i}^{\mathrm{sv}}\|} \tag{5}$$

该头部使用 InfoNCE损失进行训练，如下所示： [26]

$$\mathcal{L}_{\mathrm{S}} = -\log \frac{\exp(S2\hat{SCORE}_{\mathrm{S}}^{+}/\tau)}{\sum_{i=1}^{N} \exp(S2\hat{SCORE}_{\mathrm{S}}^{i}/\tau)} \tag{6}$$

其中，$\hat{}S2SCORE^{+}$ s 表示匹配的源语音–翻译语音对的相似度分数，$\hat{}S2SCORE^{i}$ s 表示i^{th} 候选项的分数， τ 表示温度参数。

4）统一训练目标：整体训练目标是三个特定任务损失的加权和：

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{T}}\mathcal{L}_{\mathrm{T}} + \lambda_{\mathrm{N}}\mathcal{L}_{\mathrm{N}} + \lambda_{\mathrm{S}}\mathcal{L}_{\mathrm{S}} \tag{7}$$

权重 λ_{T}、 λ_{N} 和 λ_{S} 在验证集上选取，以便在联合优化过程中平衡这三个损失。

5）统一语音质量分数：除各项单独分数外，S2SCORE 还会生成统一语音质量分数，将语义充分性、声学自然度和说话人相似度整合为一个易于解释的数值。由于这三个头部的取值范围不同（$S2SCORE_{\mathrm{T}} \in [0,1]$、$\hat{}\hat{}$ $S2SCORE_{\mathrm{N}} \in [1,5]$ 和 $S2SCORE_{\mathrm{S}} \in [-1,1]$），因此会将分数归一化为 $[0,1]$。统一分数通过对归一化分数（∵）进行加权组合计算得到：

$$\begin{aligned} S2SCORE = \alpha_{\mathrm{T}}\, S2\tilde{SCORE}_{\mathrm{T}} + \alpha_{\mathrm{N}}\, S2\tilde{SCORE}_{\mathrm{N}} \\ + \alpha_{\mathrm{S}}\, S2\tilde{SCORE}_{\mathrm{S}} \end{aligned} \tag{8}$$

其中 $\alpha_{\mathrm{T}} + \alpha_{\mathrm{N}} + \alpha_{\mathrm{S}} = 1$ 和 $S2SCORE \in [0,1]$。聚合权重通过留出验证集上的整体质量人工评分进行学习，从而得到一个经过校准、与人类对语音到语音翻译质量的判断相符的统一分数。

III. 实验设置

本节介绍用于训练和评估 S2SCORE 的数据集、训练配置和评估指标。

A. 数据集

训练 S2SCORE 需要源语音、翻译语音和参考语音，以及相应的转录文本，以生成监督标签（仅 REF 变体需要参考语音）。现有语音到语音翻译语料库提供了源语音和参考语音，但缺少足够多样的翻译语音，无法涵盖广泛的语义充分性、说话人相似度和声学自然度。

为解决这一局限，我们基于 CVSS [23] 和 BhasaAnuvaad [24]构建了一个多语言数据集。我们使用 CVSS 中的法语–英语、德语–英语、西班牙语–英语和中英 语言对，以及 BhasaAnuvaad 中的印地语–英语语言对。CVSS 提供源语音、转录文本、英语译文和合成参考语音

表 I 数据集统计信息：按语言对汇总训练集、验证集和测试集中的句子数量（千，K）及语音时长（小时，H）。

语言	训练集	验证集	Test
法语–英语	11.8k | 18.4小时	1.5k | 2.3小时	1.5k | 2.2小时
西班牙语–英语	10.6k | 9.8小时	1.3k | 1.2小时	1.3k | 1.2小时
德语–英语	10.8k | 9.7小时	1.4k | 1.2小时	1.4k | 1.1小时
中英	5.6k | 7.4小时	0.7k | 1.1小时	0.7k | 1.0小时
印地语–英语	9.6k | 11.8小时	1.2k | 1.6小时	1.2k | 1.5小时

表 II 各语言对数据集质量评估所用的语料级 ASR-BLEU [6] 和 xCOMET-XXL [18] 分数。

指标	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语
ASR-BLEU	26.57	23.91	20.12	15.16	23.48
xCOMET-XXL	0.587	0.571	0.554	0.567	0.591

（CVSS-T），而 BhasaAnuvaad 提供带有对齐转录文本的源语音和参考语音。我们使用多个语音到语音翻译系统（第 III-A1 节）生成翻译语音，从而引入翻译质量差异。最终数据集按 80%/10%/10% 的比例划分为训练集、验证集和测试集，且句子和说话人均无重叠。数据集统计信息汇总于表 I。

1）翻译语音生成：为生成涵盖广泛质量条件的翻译语音，我们采用了三种语音到语音翻译系统：

- 一个基于 Whisperlarge-v2 [11] 的级联流程，用于语音翻译（ST），随后使用 XTTS-v2 [5] 进行语音合成。
- SeamlessM4T-v2 [27], 同时用于级联模式（自动语音识别+机器翻译+文本转语音）和 ST 模式，并使用 XTTS-v2 进行语音合成。
- DS2ST-LM [28], 是一种直接语音到语音翻译模型。

对于每个系统，每条源语句都会使用六种目标语音（男性三种、女性三种）合成翻译语音，从而为说话人相似度头的对比训练提供匹配与不匹配的说话人对。

2) 数据质量：如表 II 所示，ASR-BLEU 在不同语言对之间差异显著（15.16–26.57），反映了翻译难度的差异；而 xCOMET-XXL [18] 则相对稳定（0.554–0.591），表明语义充分性一致。图 2(c) 中句子级 xCOMET-XXL 的分布覆盖了整个 $[0,1]$ 范围，且没有明显偏斜，说明翻译质量丰富多样。

3) 监督标签生成：跨多种语言开展大规模人工标注成本高昂且耗费人力。因此，使用成熟的评估指标，大规模自动生成各个评估维度的监督标签。利用根据翻译语音的 Whisper large-v2 [11]自动语音识别转录文本计算得到的 xCOMET-XXL [18]分数，对语义充分性进行监督；利用 UTMOSv2 [20],评估声学自然度；并使用第 III-A1 节所述的匹配与不匹配的说话人对，以对比方式评估说话人保留。

聚合权重学习和评估所需的人工评分由五名具备语音与语言处理研究生学历、英语水平熟练但非英语母语的说话人完成。为学习聚合权重（α_{T}，α_{N}，α_{S}），从训练集和验证集中随机抽取了 2500 条翻译语句（每个语言对 500 条）。标注者根据翻译语音、目标说话人的参考语音样本以及英语参考转写，综合考虑语义充分性、声学自然度和说话人保留情况，以五分制对整体质量进行评分。随后对评分取平均，并线性归一化至 $[0,1]$，用于权重学习。在评估阶段，另外从留出测试集中抽取了 2500 条语句（每个语言对 500 条），按照相同流程，分别对语义充分性、声学自然度、说话人保留和整体质量进行独立评分。五名标注者的 Krippendorff'sα 为 0.79，表明标注者间具有较高的一致性。

B. 训练设置

SONAR [21]和 WavLM-SV [19] 编码器保持冻结。对于自然度分支，对数梅尔频谱图 [22]（含 $F = 64$ 个梅尔频率 bins）会截断或补零至 $T = 256$ 帧，然后展平为一个 16，384 维向量。语义头和自然度头均使用两个隐藏层，维度为 $[512, 64]$,，并采用 ReLU 激活和 Dropout（$p = 0.1$）；说话人分支则通过温度为 $\tau = 0.07$ 的 InfoNCE [26] 目标进行训练。模型使用 AdamW [29] 训练 250 个 epoch，学习率为 5×10^{-4}，批次大小为 64，并采用线性预热。多任务损失权重设为 $\lambda_{\mathrm{T}} = 1.0$、$\lambda_{\mathrm{N}} = 0.5$ 和 $\lambda_{\mathrm{S}} = 0.1$，并通过验证集选定。随后根据人工总体质量评分学习聚合权重（α_{T}，α_{N}，α_{S}）。所有实验均在单张 NVIDIA Quadro RTX 6000（24 GB）GPU 上进行。

C. 评估指标

我们使用斯皮尔曼秩相关系数 [30] (ρ) 作为主要评估指标。报告 S2SCORE 预测结果与人工评分在语义充分性、声学自然度、说话人保留和整体质量方面的相关性。为评估泛化能力和一致性，我们报告其与独立指标（MetricX-24 [10],NISQA [31],和 ECAPA-TDNN [32]）以及训练信号（xCOMET-XXL [18],UTMOSv2 [20], 和 WavLM-SV [19]）的相关性。使用配对自助法重采样 [33]评估统计显著性。

IV. 实验结果与分析 IS

本节介绍 S2SCORE 的实验评估、结果分析和消融研究。

A. 语义充分性评估

我们使用三种互补的参考信号来评估语义头：人工充分性评分、MetricX-24 [10],和 xCOMET-XXL [18]。人工评分是评估语义预测质量的主要依据，而 MetricX-24 和 xCOMET-XXL 分别用于评估其与独立语义指标和训练监督信号的一致性。

表 III 语义质量预测结果与人工充分性评分、MetricX-24 [34], 和 XCOMET-XXL [18] 分数在测试集上的斯皮尔曼相关系数（ρ）。平均值根据五种语言对计算。每个参考值类别中的最佳结果以粗体显示。

模型	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语	Mean
人工充分性评分						
BLASER 2.0-参考版本	0.615	0.608	0.583	0.578	0.603	0.597
BLASER 2.0-QE	0.619	0.613	0.597	0.591	0.610	0.606
S2SCORE$_T$-REF	0.635	0.629	0.617	0.604	0.619	0.621
S2SCORE$_T$-QE	**0.649**	**0.637**	**0.628**	**0.607**	**0.633**	**0.631**
MetricX-24 （独立语义指标）						
S2SCORE$_T$-参考	0.648	0.637	0.625	0.610	0.634	0.631
S2SCORE$_T$-质量估计	**0.659**	**0.653**	**0.644**	**0.629**	**0.650**	**0.647**
xCOMET-XXL（训练监督）						
BLASER 2.0-参考	0.633	0.624	0.615	0.610	0.621	0.621
BLASER 2.0-质量估计	0.657	0.649	0.620	0.619	0.642	0.637
S2SCORE$_T$-参考	0.671	0.662	0.655	0.635	0.659	0.656
S2SCORE$_T$-QE	**0.686**	**0.673**	**0.663**	**0.641**	**0.670**	**0.667**

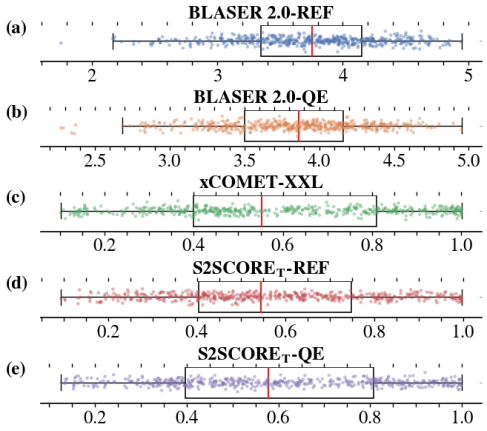

图 2。法语–英语测试集的语义指标分数分布。各面板：(a) BLASER 2.0-REF，(b) BLASER 2.0-QE，(c) xCOMET-XXL 监督，(d) S2SCORE$_T$-REF，以及 (e) S2SCORE$_T$-QE。箱线图显示中位数（红色）、四分位距（箱体）以及延伸至 1.5 × 四分位距的须线。

如表 III 所示，S2SCORE$_T$ 始终比 BLASER 2.0 与人工评分表现出更强的一致性。S2SCORE$_T$-QE 在五种语言对上的 Spearman相关系数为 0.607–0.649，平均 $\rho = 0.631$；而 S2SCORE$_T$-REF 的相关系数为 0.604–0.635（平均 $\rho = 0.621$）。相比之下，BLASER 2.0-QE 和 BLASER 2.0-REF 的平均相关系数分别为 0.606 和 0.597。相较于对应的 BLASER 2.0 基线模型，改进具有统计显著性（$p < 0.05$）。由于训练过程中未使用人工评分，这些结果表明，所提出的基于语音的指标能够更好地反映人类对语义充分性的感知，并且能够在包括中英在内的多种语言对之间实现良好的泛化。

S2SCORE$_T$ 与 MetricX- 24 也表现出很强的一致性，相关系数为 0.629–0.659（平均值 $\rho = 0.647$），这表明学习到的表征能够泛化到一种

表 IV S2SCORE$_S$ 预测结果与测试集上的人类说话人相似度评分、ECAPA-TDNN [32] 余弦相似度 [32], 和 WAVLM-SV [19] 余弦相似度之间的 Spearman相关系数（ρ）。平均值按五种语言对计算。

参考值	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语	Mean
人类	0.857	0.864	0.859	0.862	0.870	0.862
ECAPA-TDNN	0.861	0.873	0.868	0.867	0.893	0.872
WavLM-SV	0.872	0.891	0.893	0.881	0.904	0.888

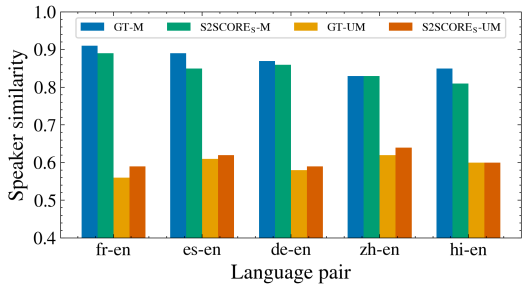

图 3. 各语言对的平均说话人相似度。GT-M/GT-UM 表示匹配对/不匹配对的真实 WavLM-SV [19] 余弦相似度；S2SCORE$_S$-M/S2SCORE$_S$-UM 表示相应的预测结果。

表 V S2SCORE$_N$ 预测结果与人工自然度评分的 Spearman相关系数（ρ）、测试集上的 NISQA [31], 和 UTMOSV2 [20] 分数。平均值根据五种语言对计算得出。

参考值	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语	Mean
人类	0.890	0.901	0.883	0.851	0.871	0.879
NISQA	0.899	0.918	0.892	0.875	0.882	0.893
UTMOSv2	0.914	0.925	0.902	0.887	0.895	0.905

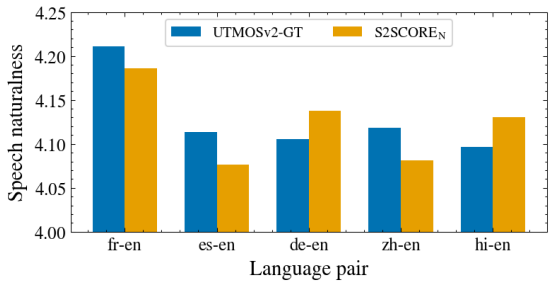

图 4. 各语言对的平均自然度分数。蓝色柱表示 UTMOSv2 参考分数，橙色柱表示 S2SCORE$_N$ 在 $[1, 5]$ MOS 量表上的对应预测结果。

独立的语义指标。与 xCOMET-XXL 的一致性仍然很高，相关系数为 0.641–0.686（平均值 ρ =0.667），表明其与训练目标保持一致。

除秩相关外，图 2 还显示，在法语–英语测试集上，S2SCORE$_T$ 产生的分数分布明显宽于 BLASER 2。0。BLASER 2。0 的归一化标准差为 0。119–0。144，而 S2SCORE$_T$ 的归一化标准差增至 0。228–0。237，接近 xCOMET-XXL 的水平（0。251）。更宽的分布表明其具有更强的判别能力，能够在直接处理语音的同时，更细致地区分语义质量不同的翻译输出。

B. 说话人相似度评估

我们使用三种互补的参考信号评估说话人头：人类说话人相似度评分、ECAPA-TDNN [32]余弦相似度和 WavLM-SV [19]余弦相似度。人类评分提供对说话人保留情况的主要评估，而 ECAPA-TDNN 和 WavLM-SV 分别衡量与独立说话人验证模型及训练监督信号的一致性。如表 IV 所示，S2SCORE$_S$ 在所有语言对上都取得了与人类说话人相似度评分的高度一致性，Spearman相关系数范围为 0。857 至 0。870（平均值 $\rho = 0.862$）。持续较高的相关系数表明，学习到的表征能够有效捕捉说话人保留情况，并且能够跨语言对泛化。

S2SCORE$_S$ 还与 ECAPA-TDNN 表现出高度一致性，相关系数达到 0.861–0.893（平均值 ρ =0.872）。由于训练过程中未使用 ECAPA-TDNN，这些结果表明，学习到的表征也与独立的说话人验证模型具有良好的一致性。与 WavLM-SV 监督信号的一致性仍然很高，相关系数为 0.872–0.904（平均值 $\rho = 0.888$），表明其与训练目标具有高度一致性。

除了秩相关外，图 3 还显示，S2SCORE$_S$在不同语言间保留了说话人相似度的区分能力。真实 WavLM-SV 余弦相似度将匹配和不匹配的说话人对区分开来，前者范围为 0.831 至 0.910，后者范围为 0.562 至 0.623。S2SCORE$_S$ 复现了这一模式：匹配对的预测值为 0.810–0.890，不匹配对的预测值为 0.590–0.640。真实相似度（0.208–0.348）与预测值（0.170–0.300）的区分幅度相近，表明该指标保留了不同语言间说话人相似度的相对结构。

C. 声学自然度评估

我们将自然度头与三种互补的参考信号进行比较评估：人工自然度评分、NISQA [31]预测结果和UTMOSv2 [20]预测结果。人工评分是感知自然度的主要评估依据，而 NISQA和UTMOSv2分别用于衡量与独立语音质量模型及训练监督信号的一致性。如表 V 所示，S2SCORE$_N$在所有语言对上的人工自然度评分一致性均较强，Spearman相关系数介于0.851至0.901之间（平均值 $\rho = 0.879$）。这些结果表明，该模型能够有效捕捉感知自然度，并在不同语言间实现泛化。

S2SCORE$_N$ 还显示出与 NISQA（$\rho = 0.875$–0.918，平均值为 $\rho = 0.893$）的高度一致性，表明其泛化能力并不局限于单一语音质量模型。与 UTMOSv2 的一致性同样很高（$\rho = 0.887$–0.925，平均值为 $\rho = 0.905$），说明从监督信号中学习到的特征在很大程度上得以保留。

表 VI 统一 S2SCORE 预测结果与测试集上的人工总体质量评分之间的 Spearman相关系数（ρ）。平均值根据五种语言对计算得出。

模型	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语	Mean
S2SCORE-REF	0.702	0.713	0.697	0.681	0.706	0.700
S2SCORE-QE	**0.718**	**0.726**	**0.711**	**0.694**	**0.719**	**0.714**

表 VII 匹配（M）与不匹配（UM）说话人条件下各语言对的统一 S2SCORE 值。平均值根据五种语言对计算得出。

模型	法语–英语	西班牙语–英语	德语–英语	中英	印地语–英语	Mean
REF-M	0.787	0.760	0.767	0.749	0.753	0.763
REF-UM	0.687	0.684	0.677	0.685	0.683	0.683
QE-M	0.792	0.764	0.771	0.752	0.758	0.767
QE-UM	0.692	0.687	0.681	0.689	0.688	0.687

除秩相关外，图 4 显示，S2SCORE$_N$ 保留了底层 UTMOSv2 参考值中的跨语言自然度模式。UTMOSv2平均分在 $[1, 5]$ MOS 量表上介于 4.10–4.21 之间，而 S2SCORE$_N$ 预测结果与这些数值的差异均在 ± 0.04 MOS 以内。S2SCORE$_N$ 略微低估了法语–英语、西班牙语–英语和中英的自然度，并高估了德语–英语和印地语–英语的自然度。尽管存在这些轻微偏差，S2SCORE$_N$ 仍保留了自然度信号的动态范围和跨语言变化。

D. 统一的 *S2ST* 质量评分

我们使用跨语言对和说话人条件下的人工总体质量评分来评估统一 S2SCORE。如表 VI 所示，S2SCORE-QE 的 Spearman 相关系数为 0.694–0.726（平均值 $\rho = 0.714$），而 S2SCORE-REF 的表现相当（0.681–0.713，平均值$\rho = 0.700$）。这些结果表明，该统一分数能够反映不同翻译和声学条件下的人工判断。QE 变体的表现略好，这与语义评估结果一致。这可能是因为源语音和翻译语音包含了大部分与质量相关的信息，而相较于增加的建模复杂度，参考语音提供的额外信息有限。

表 VII 进一步显示，说话人匹配对的统一分数始终高于不匹配对；REF 和 QE 变体的平均差距均约为 0.08，凸显了说话人保留的重要性。学习得到的权重表明，语义充分性的贡献最大（$\alpha_{\mathrm{T}} = 0.47$），其次是说话人保留（$\alpha_{\mathrm{S}} = 0.28$）和声学自然度（$\alpha_{\mathrm{N}} = 0.25$）。这些权重反映了各个维度对人工总体质量的相对贡献；较高的语义权重意味着翻译正确性的贡献更大，而说话人保留和自然度则提供互补信息。

E. 消融研究

为评估参数共享的影响，表 VIII 将所提出的多任务框架与独立训练的单任务模型进行了比较，采用 Spearman相关系数评估其与测试集上人工评分的相关性。多任务框架在所有维度上均 consistently 优于单任务模型：语义充分性、声学自然度和说话人保留度分别提升了 $+0.014$、$+0.020$ 和 $+0.013$，平均增益为 0.016 的 Spearman相关系数。尽管提升幅度不大，但所有维度上的一致性增益表明，联合多任务训练能够改善统一评估。

表 VIII 多任务训练的影响。测试集上预测结果与人工评分之间的 Spearman相关系数(ρ)，在五种语言对上取平均。

方法	语义	自然度	说话人	Mean
单任务	0.617	0.859	0.849	0.775
多任务	**0.631**	**0.879**	**0.862**	**0.791**

表 IX 聚合函数比较。测试集上与人工总体质量评分的 Spearman相关系数(ρ)，取五种语言对的平均值。

指标	非加权平均值	学习得到的线性函数（我们的方法）	MLP
Spearman ρ	0.671	0.714	**0.720**

为评估聚合函数，表 IX 比较了统一分数的三种策略：等权平均值（$\alpha_{\mathrm{T}} = \alpha_{\mathrm{N}} = \alpha_{\mathrm{S}} = 1/3$）、学习式线性加权（其中使用人工评分优化聚合系数），以及非线性双层 MLP回归器。与等权平均值相比，学习式加权使 Spearman相关系数提高了 0.043，证明了根据人工判断学习聚合权重的优势。尽管双层 MLP 增加了参数量和复杂度，但其带来的提升仅为 0.006，十分有限。

V. 结论

我们提出了 S2SCORE，这是一种面向语音到语音翻译的统一评估框架，可在单一架构中联合建模语义充分性、声学自然度和说话人相似度。S2SCORE 直接在语音域中运行，无需中间文本表示即可对翻译语音进行整体评估。在五种语言对上，所提出的框架与人工判断及成熟的参考指标均表现出较强的一致性；其统一分数还能可靠地预测整体语音到语音翻译质量，并推广至所评估的各语言对。本研究聚焦于语音到语音翻译质量的三个关键维度，并评估了五种语言对中译为英语的翻译。未来工作将扩展 S2SCORE 框架，纳入口语交流的其他方面，例如韵律、情感表现力和说话风格，并评估其在更多语言、翻译方向以及类型学更加多样化的环境中的稳健性。

参考文献

[1] S. Dhawan，“语音到语音翻译：挑战与未来”，IJCATR，第 11 卷，第 3 期，第 36–55 页，2022年。[2] T. Etchegoyhen、H. Arzelus、H. Gete、A. Alvarez、I. G. Torre、J. M. Mart´ín-Do˜nas、A. Gonz´ález-Docasal 和 E. B. Fernandez，“级联式还是直接语音翻译？一个案例研究”，《应用科学》，第 12 卷，第 3 期，2022年。[3] S. Alharbi、M. Alrazgan、A. Alrashed、T. Alnomasi、R. Almojel、R. Al-harbi、S. Alharbi、S. Alturki、F. Alshehri 和 M. Almojil，“自动语音识别：系统性文献综述”，《IEEE Access》，第 9 卷，第 131 858–131 876 页，2021年。[4] Z. Tan、S. Wang、Z. Yang、G. Chen、X. Huang、M. Sun 和 Y. Liu，“神经机器翻译：方法、资源与工具综述”，《AI Open》，第 1 卷，第 5–21 页，2020年。[5] E. Casanova、K. Davis、E. G¨olge、G. G¨oknar、I. Gulea、L. Hart、A. Aljafari、J. Meyer、R. Morais、S. Olayemi 和 J. Weber，“XTTS：一种大规模多语言零样本文本转语音模型”，载于 Interspeech 论文集，2024年，第 4978–4982 页。[6] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，“BLEU：一种用于机器翻译自动评估的方法”，载于 ACL 论文集，2002年，第 311–318 页。[7] S. Banerjee 和 A. Lavie，“METEOR：一种能够提高与人工判断相关性的机器翻译评估自动指标”，载于 ACL研讨会论文集，2005年，第 65–72 页。[8] T. Sellam、D. Das 和 A. Parikh，“BLEURT：学习用于文本生成的稳健指标”，载于 ACL 论文集，2020年，第 7881–7892 页。[9] R. Rei、C. Stewart、A. C. Farinha 和 A. Lavie，“COMET：用于机器翻译评估的神经框架”，载于 EMNLP 论文集，2020年，第 2685–2702 页。[10] J. Juraska、D. Deutsch、M. Finkelstein 和 M. Freitag，“MetricX-24：Google 在 WMT 2024 指标共享任务中的提交”，载于 WMT 论文集，2024年，第 492–504 页。[11] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现稳健的语音识别”，载于 ICML 论文集，2023年，第 28 492–28 518 页。[12] M. Chen、P.-A. Duquenne、P. Andrews、J. Kao、A. Mourachko、H. Schwenk 和 M. R. Costa-juss`a，“BLASER：一种无文本的语音到语音翻译评估指标”，载于 ACL 论文集，2023 年，第 9064–9079 页。[13] D. Dale 和 M. R. Costa-juss`a，“BLASER 2.0：一种用于大规模多语言语音与文本翻译评估及质量估计的指标”，载于 EMNLP论文集，2024年，第 16 075–16 085 页。[14] L. Besacier、S. Ribeiro、O. Galibert 和 I. Calapodescu，“一种用于语音到语音比较的无文本指标”，2022年，arXiv:2210.11835。[15] H. Han、K. Duh 和 M. Carpuat，“SpeechQE：估计直接语音翻译的质量”，载于 EMNLP 论文集，2024年，第 21 852–21 867 页。[16] C. Le、Y. Qian、D. Wang、L. Zhou、S. LIU、X. Wang、M. Yousefi、Y. Qian、J. Li 和 M. Zeng，“TransVIP：一种保留语音特征和等时性的语音到语音翻译系统”，载于 NeurIPS 论文集，2024年，第 89 682–89 705 页。[17] Y. Jia、R. J. Weiss、F. Biadsy、W. Macherey、M. Johnson、Z. Chen 和 Y. Wu，“使用序列到序列模型进行直接语音到语音翻译”，载于 Interspeech 论文集，2019年，第 1123–1127 页。[18] N. M. Guerreiro、R. Rei、D. van Stigt、L. Coheur、P. Colombo 和 A. F. T. Martins，“xCOMET：通过细粒度错误检测实现透明的机器翻译评估”，TACL，第 12 卷，第 979–995 页，2024年。[19] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao、J. Wu、L. Zhou、S. Ren、Y. Qian、Y. Qian、J. Wu、M. Zeng、X. Yu 和 F. Wei，“WavLM：面向全栈语音处理的大规模自监督预训练”，《IEEE信号处理专题期刊》，第 16 卷，第 6 期，第 1505–1518 页，2022年。[20] K. Baba、W. Nakata、Y. Saito 和 H. Saruwatari，“VoiceMOS挑战赛2024中的T05系统：将深度图像分类器迁移学习用于高质量合成语音的自然度MOS预测”，载于 SLT 论文集，2024年。[21] P.-A. Duquenne、H. Schwenk 和 B. Sagot，“SONAR：句子级多模态、语言无关表示”，2023年，arXiv:2308.11466。[22] S. B. Davis 和 P. Mermelstein，“连续语句中单音节词识别的参数化表示比较”，《IEEE声学、语音与信号处理汇刊》，第 28 卷，第 4 期，第 357–366 页，1980年。[23] Y. Jia、M. Tadmor Ramanovich、Q. Wang 和 H. Zen，“CVSS 语料库与大规模多语言语音到语音翻译”，载于*LREC* 论文集，2022年，第 6691–6703 页。[24] A. Sankar、S. Jain、N. Narasimhan、D. Choudhary、D. Suman、M. S. U. R. Khan、A. Kunchukuttan、M. M. Khapra 和 R. Dabre，“面向构建大规模数据集及适用于14种印度语言的先进自动语音翻译系统”，载于*ACL* 论文集，2025年，第 32 945–32 966 页。[25] J. Yu、X. Liu、C. Luo、R. Zhou、Y. Liu、J. Hu、J. Chen、K. Zhang、D. Zhang、Y. Shen、E. Adhikarla、Y. Dai、K. Zhang、Z. Kong、W. Ye、Y. Yin、V. Namboodiri、B. Davison、J. Moore 和 Y. Chen，“ 多任务学习 1997–2024：第一部分 —— 基础 ”，《哈佛数据科学评论》，第 7 卷，第 3 期，2025年。[26] A. van den Oord、Y. Li 和 O. Vinyals，“基于对比预测编码的表征学习”，2019年，arXiv:1807.03748。[27] S. Communication、L. Barrault、Y.-A. Chung、M. C. Meglioli、D. Dale、N. Dong、P.-A. Duquenne、H. Elsahar、H. Gong、K. Heffernan、J. Hoffman、C. Klaiber、P. Li、D. Licht、J. Maillard、A. Rakotoarison、K. R. Sadagopan、G. Wenzek、E. Ye、B. Akula、P.-J. Chen、N. E. Hachem、B. Ellis、G. M. Gonzalez、J. Haaheim、P. Hansanti、R. Howes、B. Huang、M.-J. Hwang、H. Inaguma、S. Jain、E. Kalbassi、A. Kallet、I. Kulikov、J. Lam、D. Li、X. Ma、R. Mavlyutov、B. Peloquin、M. Ramadan、A. Ramakrishnan、A. Sun、K. Tran、T. Tran、I. Tufanov、V. Vogeti、C. Wood、Y. Yang、B. Yu、P. Andrews、C. Balioglu、M. R. Costa-juss`a、O. Celebi、M. Elbayad、C. Gao、F. Guzm´an、J. Kao、A. Lee、A. Mourachko、J. Pino、S. Popuri、C. Ropers、S. Saleem、H. Schwenk、P. Tomasello、C. Wang、J. Wang 和 S. Wang，“SeamlessM4T：大规模多语言、多模态机器翻译”，2023年，arXiv:2308.11596。[28] L. Arya、M. Bhattacharjee、A. C. R. 和 S. R. M. Prasanna，“可扩展至多种语言对的音色感知型、基于大语言模型的直接语音到语音翻译”，2026年，arXiv:2601.16023。[29] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，载于*ICLR* 论文集，2019年。[30] C. Spearman，“两种事物之间关联的证明与测量”，《美国心理学杂志》，第 15 卷，101，1904 1，72号，第 – 页。[31] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，“NISQA：一种基于深度卷积神经网络自注意力模型的多维语音质量预测方法，使用众包数据集”，载于*Interspeech* 论文集，2021年，第 2127–2131 页。[32] B. Desplanques、J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：用于基于 TDNN 的说话人验证的强调通道注意力、传播与聚合”，载于*Interspeech* 论文集，2020年，第 3830–3834 页。[33] P. Koehn，“机器翻译评测的统计显著性检验”，载于*EMNLP* 论文集，2004年，第 388–395 页。[34] J. Juraska、M. Finkelstein、D. Deutsch、A. Siddhant、M. Mirzazadeh 和 M. Freitag，“MetricX-23：Google 在 WMT 2023 评测指标共享任务中的投稿”，载于*WMT* 论文集，2023年，第 756–767 页。