一个批改全部的低声

Nhan Phan, Anusha Porwal, Yaroslav Getman, Ekaterina Voskoboinik, Tamás Grósz, Mikko Kurimo

Department of Information and Communications Engineering, Aalto University, Finland

{ firstname.lastname } @aalto.fi

Abstract

我们提出了一种高效的端到端方法，用于全面的多部分二语测试自动口语评估 (ASA)，这是为 2025 年的 Speak & Improve Challenge 开发的。我们系统的主要创新是能够通过一个单一的 Whisper-small 编码器处理所有四个口语回答，通过轻量化的聚合器结合所有信息，并预测最终分数。这种架构消除了转录和每个部分模型的需要，减少了推理时间，使得 ASA 在大型计算机辅助语言学习系统中实际可行。我们的系统实现了 0.384 的均方根误差 (RMSE)，优于文本为基础的基线 (0.44)，同时最多使用 168M 个参数（约占 Whisper-small 的 70 %）。此外，我们提出了一种数据采样策略，使模型只需在语料库中 44.8 % 的说话人上训练，仍可以达到 0.383 的 RMSE，这表明在不平衡类别上的性能提升以及强大的数据效率。

Index Terms: automatic speaking assessment, computer-assisted language learning, L2 proficiency, Whisper

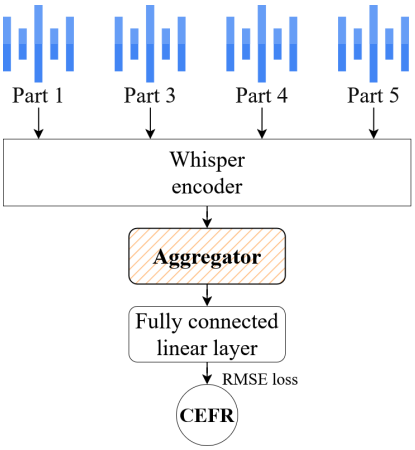

Figure 1: 我们模型架构的概述。

1. 引言

口语是交际能力的核心维度，因此必须进行评估，以保持任何语言资格的有效性 [1] 。然而，人工作答的评分既耗时又昂贵。此外，由于疲劳 [2] 、培训不足 [3] 或对二语使用者口音的接触有限 [4] ，评分者的一致性可能会受到影响。自动口语评估 (ASA) 可以通过提供可扩展、客观的分数来缓解这些缺点，同时大大降低成本。在计算机辅助语言学习 (CALL) 应用中，ASA 还可以通过提供即时的、无偏见的反馈来支持学习，这可以减轻口语焦虑并促进自主练习 [5] 。

早期的 ASA 系统通过手动设计的声学或文本特征预测整体熟练程度 [6, 7] 。由于这些特征是手工制作的，它们的实用性很大程度上取决于开发者选择包含的内容，并且它们可能忽略数据驱动方法可能捕捉到的线索。因此，最近的研究探索了从音频信号中自动提取特征的使用，使得 ASA 系统可以作为端到端评分系统，仅以学习者的语音为输入 [8–12] 。虽然这些方法可以超越基于手工特征的模型，但它们也引入了与计算复杂性相关的新挑战。大多数端到端模型被设计用于预测单个问题或任务的整体评分 [8–10] ；对于多部分测试，它们通常需要多个模型或模型集群 [11, 12] 。

多模型计算多部分口语考试的综合评分的需求限制了现有 ASA 系统的实用性，特别是在速度和计算成本方面。尽管每个模型的推理时间可能相似，但依次加载它们会影响延迟，而并行运行它们又会增加内存需求。这些限制使得大规模实时部署更加困难。为了解决这个问题，我们提出了一种更高效的 CALL 应用解决方案：利用单个 Whisper small [1] 编码器 [13] ，可以顺序处理来自多部分考试的所有口语回答，并以最小的性能损失预测演讲者的综合得分（图 1 ）。这种设计减少了推理时间和资源使用，使得能够同时向大量学习者提供即时反馈，这对于可扩展的 ASA 系统来说是一个关键优势。

我们在 Speak & Improve Challenge 2025 的口语语言评估（SLA）赛道 [12] 中展示了我们提出的解决方案，其取得的均方根误差（RMSE）为 0.384。这个结果明显优于官方基线，该基线结合了 Whisper-small 用于自动语音识别（ASR）与四个独立的基于 BERT 的评分模型，达到了 0.44 的 RMSE。我们的架构和详细配置已作为开源发布 [2] 。

2. 数据

我们的数据全部来自 2025 年说 & 改进语料库 (S & I) [14] ，作为 SLA 挑战的一部分。任务是开发模型，通过使用部分 1、3、4 和 5（不包括第 2 部分）的口语回答来评估 L2 学习者的语言水平。第 1 部分和第 5 部分是简短回答 ($\leq$ 20 秒)，而第 3 部分和第 4 部分是一个长回答 ($\leq$ 1 分钟)。每个样本包括四个部分：第 1 部分和第 5 部分（短回答）首先分别连接，和第 3 部分及第 4 部分一起，每个说话者大约形成四分钟的音频。

我们遵循官方的数据集划分：用于训练（TRAIN）的语音为 244 小时，开发（DEV）和评估（EVAL）各为 35 小时。然而，我们在实验中并未使用整个数据集（详情见表 1 ）。每个说话者的最终整体评分遵循欧洲共同语言参考框架（CEFR）[15] ，分数按 0.125 分的增量映射到从 2.0（A2）到 5.5（C1+）的数值。完整语料库的详细信息请见 [14] 。

数据集中面临的一个挑战是在整体水平的可用训练样

[1] https://huggingface.co/openai/whisper-small

[2] https://github.com/aalto-speech/slate-2025

Table 1: 用于训练和开发集的唯一说话人 *ID* 与完整的 *SMATHXADI* 语料库（*TRAIN* + *DEV*）相比。

Dataset version	Unique speaker IDs	% of corpus
Standard	1684	77.2 %
4x	1871	85.8 %
OE	1134	52.0 %
OE+OE $_d$	976	44.8 %

本的不平衡。在这里，我们将 CEFR 水平为 A2+ 或更低（≤ 2.5 ）和 C1 或更高（≥ 5.0 ）的说话者定义为边缘案例。虽然这些水平占可能分数范围的三分之一以上，但它们仅代表大约 13 % 的数据。边缘案例样本的稀缺对我们这样的模型带来了额外的挑战，因为这些模型需要在所有部分都有完整的回应。此外，Bannò 等人的早期工作 [11] 显示，基于音频的模型在高水平能力的说话者上可能表现不佳。这进一步放大了在 C1 及以上水平的数据不平衡的影响。

2.1. 交换采样策略

与为每个部分单独训练的模型相反，我们的方法需要对于所有四个部分（1、3、4 和 5）的完整响应。然而，语料库中的许多说话者缺少响应，这将可用的训练集缩减到只有 77.2 % 的可用说话者 ID（见表 1 ）。这种限制对于已经代表性不足的边缘案例尤其严重。

为了解决这个问题，我们引入了一种简单但有效的增强方法，称为交换采样。其核心思想是通过结合不同说话者的回答合成地构建四部分样本。例如，如果说话者 A 缺少第 5 部分，而说话者 B 有，我们可以使用 A 的第 1、3 和 4 部分与 B 的第 5 部分形成一个新的完整样本。如果 A 和 B 都拥有所有部分，我们仍然可以在它们之间交换个别部分，以生成最多 14 种额外的组合。为了确保一致性，我们仅在每个部分的分数与目标 CEFR 分数相差不超过 0.5 时才应用交换采样。这项技术使我们仅使用原始 S & I 语料库就能构建更大且更平衡的数据集。我们使用这种方法构建了三个变体的训练数据：

- 标准：1,684 个完整的样本，包含所有四个部分。
- 4x 互换采样（4x）：通过互换生成大约多四倍的样本来扩展训练集。这些样本是在 0.5 分数范围内随机选择的。
- 通过增加接近边缘的样本数量（CEFR ≤ 2.75 或 ≥ 4.75）到每个分数点 80 个样本，从而对边缘分数进行过采样（OE），同时将中间范围的分数减少到 40 个。在分数从 2.0 到 5.5 按 0.125 递增的情况下，这导致总共 1,720 个样本，包括原始样本和合成样本。

在同时使用 TRAIN 和 DEV 集合进行训练时，我们使用交换采样法构建了一个合成的开发集，以确保这些组合不会与训练中使用的重叠。此外，我们创建了一个 OE $_d$ 开发集，以模仿 OE 训练集的数据分布。由于 OE 和 OE $_d$ 都侧重于合成的边界情况样本，因此它们需要的说话人 ID 显著减少（表 1 ）。

3. 实验

我们的模型架构基于 Whisper-small 编码器。相比文本模型，例如 BERT 需要额外的声学模型来进行 ASR，我们选择了声学模型。通过直接使用声学模型进行评分，我们消除了中间转录的需求，从而减少了模型的复杂性和推理时间。在声学模型中，选择 Whisper 是因为它在零样本环境中具有鲁棒性 [13] ，这对于处理 L2 语者时特别重要。

架构概览如图 1 所示。由于 Whisper 一次只能处理 30 秒的音频 [13] ，而完整的多部分响应可以超过 240 秒，我

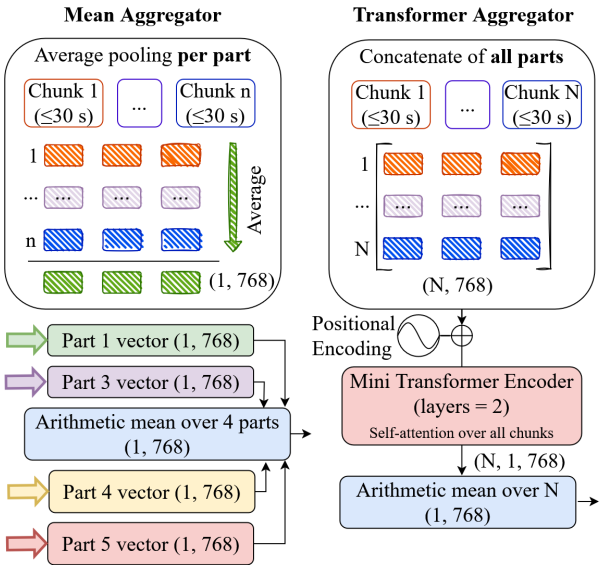

Figure 2: 两个聚合策略的示意图。*AVG* 对每个部分的块嵌入进行平均，而 *TF* 则使用 *Transformer* 编码器组合所有块。

们的挑战在于高效聚合编码器输出。我们将每个样本划分为连续的、非重叠的 30 秒块，保持短块而不进行掩码或填充。然后将这些块级表示传递给聚合器，聚合器将它们合并为用于 CEFR 预测的固定大小的嵌入。聚合器的设计对我们模型的性能至关重要。虽然一个大而复杂的聚合器可以保留更多 Whisper 编码器的细节，但它也会引入大量的计算成本。相反，我们的聚合器是轻量级的，并输出一个单一嵌入，然后通过一个全连接层生成最终预测。

Whisper-small 编码器的输出是一个具有两个维度的张量：编码器时间步（ T ）和每一步大小为 768 的特征向量。为了保持模型的轻量级，我们对时间维度 T 进行平均，对每个 30 秒的片段生成一个单一的 768 维向量。此步骤与后续由聚合器执行的池化不同。它将所有帧级信息压缩成一个嵌入，交换细粒度时间细节以提高计算效率。结果是，高度局部化的现象——例如短暂但严重的发音错误——被平滑处理，可能对最终表示的贡献较小。

ASA 模型预测 CEFR 分数为 2.0 到 5.5 的连续值。虽然我们最初尝试使用均方误差（MSE）来更严厉地惩罚较大的误差，但在 16 位精度下，MSE 值可能会变得过小，导致梯度下溢问题 [16] ，因此我们选择了均方根误差（RMSE）作为损失函数。所有模型以学习率 5×10^{-5} 训练了 15 个周期。通过合成开发集或 OE $_d$ 开发集选择表现最佳的周期。

3.1. 聚合器

我们实验了两种聚合策略：平均聚合器（AVG）和变压器聚合器（TF）（图 2 ）。首先，AVG 通过在两个阶段应用平均池化，实现了最轻量级的模型。在聚合器之前，每个块已经被时间平均成为一个 768 维向量，匹配了贯穿 Whisper 小模型使用的隐藏大小。然后，AVG 在每个部分内对这些块的嵌入进行平均，产生每个部分一个 768 维向量。一旦四个部分（1, 3, 4, 5）都处理完，AVG 计算这四个部分级向量的算术平均，最终得到一个单一的 768 维向量，用于总结整个响应。这个最终嵌入通过全连接层以预测 CEFR 分数。AVG 的设计目的是与语料库的评分方法相一致，该方法通过平均部分级分数来得出最终的整体分数。

AVG 聚合器预计会失去语义细节，因为它在计算答案

Table 2: 模型的性能及 *SLA* 挑战基线。使用 *4x*、*OE* 和 *OE* $_d$ 的模型采用了第 *2* 节中描述的训练和开发配置。我们为最佳提交选择的模型用 * 标记。

	RMSE	RMSE $_e$	% ≤ 0.5	% ≤ 1.0
Baseline	0.440		73.7	96.7
AVG	0.384	0.698	81.3	99.7
AVG-4x*	0.383	0.717	81.3	99.3
AVG-OE	0.387	0.613	81.3	98.7
AVG-OE+OE $_d$	0.386	0.618	80.3	99.7
TF	0.372	0.637	81.7	99.7
TF-4x	0.384	0.675	79.7	99.0
TF-OE*	0.388	0.626	79.0	98.7
TF-OE+OE $_d$	0.383	0.597	79.3	99.3
Submission	0.384	0.716	81.3	99.3

中的所有块嵌入的简单平均值时，没有保留它们的顺序。因此，时间结构被舍弃，个别块的不同贡献可能被稀释。实际上，AVG 主要捕捉的是表达方式的方面（即如何说某事），而不是语音内容（即说了什么）。这种结构信息的丢失可能限制模型评估高级熟练程度的能力 [11] 。值得注意的是，这些较高的熟练程度水平还与在第 2 节讨论的不平衡边界情况重叠，从而进一步降低了 C1 或以上的 CEFR 水平的性能。

为了更好地保存语义和结构信息，我们引入了基于 Transformer 模型 [17] 编码器的 TF 聚合器。为了保持效率，TF 使用了轻量级配置：768 维输入嵌入，4 个注意力头，$4 \times 768 = 3072$ 维度的前馈层，以及 2 个编码器层。TF 聚合器旨在保留关于块顺序和跨部分依赖的粗略信息，同时保持计算上的高效性。显然

4. 结果

表 2 比较了基线解决方案的性能与我们模型的性能，包括我们提交给 SLA 挑战的最佳方案。主要的评估指标是 RMSE，但我们也报告 RMSE $_e$ ，其衡量的是在第 2 节中描述的边缘案例上的性能。我们评估了不同的聚合策略和训练集变体，以测试交换采样是否能提高整体性能，及目标过采样是否有助于处理罕见的熟练程度。

从 4x 模型性能来看，我们观察到增加交换样本的数量并没有改善 RMSE 指标，它保持在大约 0.383。我们假设这是因为大多数新增样本是没有新信息的合成组合，只有少数表示极端情况。此外，合成开发集并没有始终如一地帮助选择表现最佳的检查点，如图 3 所示。

相比之下，过采样边缘（OE）模型在两个聚合器上的边缘案例中表现始终更好。这表明有针对性的过采样可以部分缓解数据不平衡的影响，与先前的研究 [18, 19] 一致。值得注意的是，使用过采样边缘开发集（OE+OE $_d$ ）验证的 OE 模型也在整体性能上与其他配置相当，尽管使用的说话人 ID 显著减少。这个结果突显了它们在数据效率方面的优势，我们将在第 4.1 节中进一步讨论。

虽然我们表现最佳的单个模型是具有 0.372 RMSE 的 TF 模型，但它并未用于我们的正式 SLA 提交。在开发过程中，我们发现 TF 模型在各个训练时期的性能波动更大——有时会达到更低的 RMSE，但也可能根据检查点选择而出现明显退化（参见图 3 ）。相比之下，AVG 模型更加稳定：由于其平均设计，无论选择哪个时期，它们总是稳定地产生约 0.383 的类似 RMSE 分数。这种稳定性在评估设置中尤其有利，特别是在测试集访问受限的情况下，如同在 SLA 挑战中一样。

正如在第 2 节中讨论的那样，TF 模型倾向于在边缘

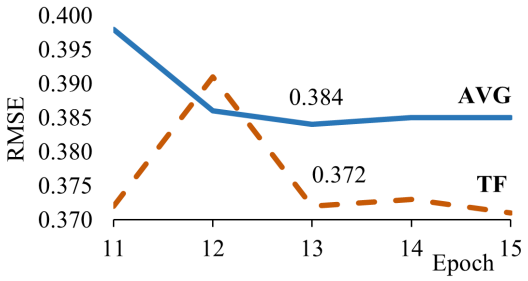

Figure 3: *AVG* 和 *TF* 在训练时期上的 *EVAL* 集的均方根误差（*RMSE*）。基于开发集的表现，第 *13* 轮被选为最佳检查点。

情况下优于 AVG 模型，在各种训练配置中实现更低的 RMSE $_e$ 。为了利用这一点，我们实现了一种受专家混合 [20] 启发的模型组合策略，其中 TF 模型处理边缘案例样本的预测，而 AVG 处理其余部分。我们提交的模型结合了 AVG-4x（以保证整体稳定性）和 TF-OE（以获得更好的边缘案例预测）。然而，边缘案例的改进表现并不一定转化为更好的总体评分，因为官方测试集中仅包含了 21 个边缘样本（7 % ）。此外，尽管我们的 TF-OE 模型在边缘情况下表现更好，但仍然预测出超出边缘分数范围的值（例如，对真实分数 5.5 预测出 4.9）。这些错误导致这些预测被我们专家混合方案中的选择机制排除，最终限制了 TF-OE 在最终提交中的影响。

4.1. 数据和计算效率

我们从两个角度评估效率：数据使用和计算成本。首先，在数据效率方面，OE+OE $_d$ 模型使用的样本明显少于其他配置——仅使用了官方 TRAIN 和 DEV 集中 44.8 % 的说话人 ID（表 1 ）——同时在性能上仍具竞争力，特别是在边缘情况下。作为参考，由于架构限制，即使是我们的标准模型也只能使用语料库的 77.2 %，因为它需要对所有四个部分提供完整的响应。这些结果引发了一个更广泛的问题：未来的系统是否可以从选择性地收集代表性不足的边缘案例中获益更多，而不是扩展那些已经可以通过不同采样技术模拟的常见样本？

第二，就计算效率而言，我们的模型架构是为了低成本推理而设计的。基准系统依赖 Whisper 进行转录，并使用四个独立的 BERT 模型为每个部分评分，然后进行分数平均。相比之下，我们的系统仅使用一个 Whisper 编码器和一个轻量级聚合器直接预测整体分数，完全跳过 Whisper 解码器和 ASR。

在标准的 CPU 设置(Intel Xeon E5-2680 v3 2.50 GHz, 4 GB RAM，无 GPU）下，我们对一个总计 240 秒音频的 4 部分响应的推理时间约为 60 秒。这表明我们的模型可以作为实时或大规模 CALL 系统的实用、低成本解决方案。在使用 GPU 的情况下，可以几乎立即提供反馈；例如，在 RTX 4070 上进行推理耗时不到一秒。

在模型规模方面，我们的 AVG 配置使用 154M 的参数，而 TF 模型使用 168M，均显著小于完整的 Whisper-small 模型（244M）。虽然我们目前为未来的研究加载了 Whisper 解码器（88M）（见第 4.2 节），但在推理过程中并不使用它。

4.2. ASA 中的信度和效度

在语言测试中，口语评估的质量通常通过三个关键原则来评估：可靠性、有效性和实用性 [21, pp.175–186] 。尽管这些标准最初是为人工评分制定的，但它们为评估 ASA 系统提供了一个有用的框架。我们的架构通过模型简单性

Table 3: 插入虚拟或交换语音后预测分数的变化（更高 = 内容敏感性更强）。

	$\Delta\uparrow$	$\%\geq 0.25\uparrow$	$\%\geq 0.5\uparrow$
Dummy speech			
AVG	0.120	16.7	1.0
TF	0.308	63.0	10.3
Part swapping			
AVG	0.000	0.0	0.0
TF	-0.012	0.0	0.0

和低推理成本来解决实用性问题。本文聚焦于 (i) 可靠性，解释为分数的稳定性，以及 (ii) 有效性，这里通过一个单一标准进行评估：模型对缺失或不匹配内容的反应 [22, 23]，这为结构涵盖提供了一个有用但部分的指标。

可靠性。图 3 绘制了训练历元期间在 EVAL 集上的 RMSE。AVG 在第 12 个历元后收敛，变化很小，而 TF 则显示出较大的波动，包括在同一历元出现的一个明显尖峰。这表明 AVG 对训练波动更加稳健，这在测试集无法进行调整的盲评估设置中特别重要。AVG 的双重平均池化有助于平滑训练噪声，可能有助于这种一致性，

有效性。我们设计了两个对照测试来评估模型对语音内容的敏感性，所有测试都在 EVAL 集合上进行。在第一个测试（虚拟语音）中，我们用 60 秒的静音替换了一部分。在第二个测试（部分交换）中，我们将第 1 部分和第 3 部分之间、第 4 部分和第 5 部分之间的答案进行了交换，以模拟与主题无关的答案。对每个测试，我们计算了 $\Delta = \hat{y}_{\text{orig}} - \hat{y}_{\text{edit}}$，即原始输入与编辑后输入的预测分数差异。表格 3 报告了 $\Delta \geq 0.25$ 和 $\Delta \geq 0.5$ 的样本百分比。

虚拟语音测试显示，TF 对内容缺失显著更为敏感。当所有四个部分都静音时，TF 预测的 CEFR 得分为 2.02（A2 - 该数据集可能的最低得分），而 AVG 仍然预测 3.16。然而，这两个模型对部分交换不敏感，这证实了先前的发现，即仅基于声学特征的 ASA 系统主要评估语音传递，而不是语音内容 [11, 24]。

这些研究结果突显了一个关键的局限性：ASA 系统容易被流利但离题或无关的内容误导。尽管对基于声学的 ASA 的兴趣越来越大，但在有效性上的差距仍未得到充分探索。对于高风险的使用案例，我们建议将声学 ASA 与一个专门的内容评估模块配对，例如 [25] 中提出的那样。在低风险的计算机辅助语言学习设置中，我们最初计划使用 Whisper 解码器和 BERT 来使评分系统对问题进行条件化。然而，由于技术问题，在 SLA 挑战期间无法访问任务问题。

5. 结论

在本文中，我们提出了一种简单的架构，该架构使用单一的 Whisper 编码器从多部分回答中预测整体口语评分，消除了对每部分单独模型的需求。我们的系统优于 SLA 挑战基线，该基线使用 Whisper small 和四个基于 BERT 的模型。结合我们的交换过采样技术，模型展现了强大的数据效率：它使用不到 S & I 语料库中可用说话者的一半，同时保持了竞争性的性能，使其成为不平衡数据的一个有前景的解决方案。

然而，与其他仅限声学的 ASA 系统一样，我们的模型对语音内容仍不敏感。未来的工作应该探索整合内容感知组件，使模型能够评估所说的内容以及说话的方式，从而提高在计算机辅助语言学习应用中的有效性和反馈质量。

6. 致谢

部分研究资金来自芬兰研究委员会的资助，编号为 322625、345790、355587 和 365233。计算资源由 Aalto Science-IT 提供。

7. References

[1] L. F. Bachman, Fundamental considerations in language testing . Oxford university press, 1990.

[2] G. Ling, P. Mollaun, and X. Xi, “A study on the impact of fatigue on human raters when scoring speaking responses,” Language Testing , vol. 31, no. 4, pp. 479–499, 2014.

[3] L. Davis, “The influence of training and experience on rater performance in scoring spoken language,” Language Testing , vol. 33, no. 1, pp. 117–135, 2016.

[4] M. S. Park, “Rater effects on L2 oral assessment: Focusing on accent familiarity of L2 teachers,” Language Assessment Quarterly , vol. 17, no. 3, pp. 231–243, 2020.

[5] H. Qiao and A. Zhao, “Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context,” Frontiers in Psychology , vol. Volume 14 - 2023, 2023.

[6] S. Crossley and D. McNamara, “Applications of text analysis tools for spoken response grading.” Language Learning & Technology , vol. 17, no. 2, pp. 171–192, 2013.

[7] Y. Wang, M. Gales, K. Knill, K. Kyriakopoulos, A. Malinin, R. Van Dalen, and M. Rashid, “Towards automatic assessment of spontaneous spoken English,” Speech Communication , vol. 104, pp. 47–56, Nov. 2018.

[8] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, “End-to-End Neural Network Based Automated Speech Scoring,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . Calgary, AB: IEEE, Apr. 2018, pp. 6234–6238.

[9] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo, “Automatic Rating of Spontaneous Speech for Low-Resource Languages,” in 2022 IEEE Spoken Language Technology Workshop (SLT) , 2023, pp. 339–345.

[10] S. Bannò and M. Matassoni, “Proficiency Assessment of L2 Spoken English Using Wav2Vec 2.0,” in 2022 IEEE Spoken Language Technology Workshop (SLT) , 2023, pp. 1088–1095.

[11] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. Gales, “Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning,” in 9th Workshop on Speech and Language Technology in Education (SLaTE) . ISCA, Aug. 2023, pp. 126–130.

[12] M. Qian, K. Knill, S. Banno, S. Tang, P. Karanasou, M. J. F. Gales, and D. Nicholls, “Speak & Improve Challenge 2025: Tasks and Baseline Systems,” 2024. [Online]. Available: https://arxiv.org/abs/2412.11985

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in Proceedings of the 40th International Conference on Machine Learning , vol. 202. PMLR, 2023, pp. 28 492–28 518.

[14] K. Knill, D. Nicholls, M. Gales, M. Qian, and P. Stroinski, “ Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback,” Apollo - University of Cambridge Repository, Tech. Rep., Dec. 2024. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/377542

[15] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment . Cambridge University Press, 2001. [Online]. Available: https://rm.coe.int/1680459f97

[16] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in International Conference on Learning Representations , 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems , vol. 30, 2017.

[18] H. Rathpisey and T. B. Adji, "Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods," in 2019 5th International Conference on Science in Information Technology (ICSITech) . Yogyakarta, Indonesia: IEEE, Oct. 2019, pp. 193–198.

[19] T. M. Lun, E. Voskoboinik, R. Al-Ghezi, T. Grosz, and M. Kurimo, "Oversampling, augmentation and curriculum learning for speaking assessment with limited training data," in Interspeech 2024 , 2024, pp. 4019–4023.

[20] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: a literature survey," Artificial Intelligence Review , vol. 42, pp. 275–293, 2014.

[21] S. Luoma, Assessing speaking . Cambridge University Press, 2004.

[22] J. Cheng and J. Shen, "Off-topic detection in automated speech assessment applications," in Interspeech 2011 . ISCA, Aug. 2011, pp. 1597–1600.

[23] S.-Y. Yoon and C. Lee, "Content modeling for automated oral proficiency scoring system," in Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications , 2019, pp. 394–401.

[24] X. Xi, D. Higgins, K. Zechner, and D. M. Williamson, "Automated scoring of spontaneous speech using SpeechRater v1.0," ETS Research Report Series , vol. 2008, no. 2, pp. i–102, 2008.

[25] N. Phan, A. von Zansen, M. Kautonen, E. Voskoboinik, T. Grosz, R. Hilden, and M. Kurimo, "Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task," in Interspeech 2024 , 2024, pp. 317–321.