arXiv:2602.07211v1 [cs.CL] 6 Feb 2026

赋予大语言模型定向语音理解能力

Ju Lin，Jing Pan，Ruizhi Li，Ming Sun，Yuzong Liu，Alaa Hassan，Jing Zheng，Florian Metze

Meta，美国

摘要

近期研究表明，通过音频编码提示大语言模型（LLM），可以实现有效的语音理解能力。然而，大多数语音大语言模型都基于单通道、单说话人数据进行训练，这使得它们难以直接应用于多说话人和多通道语音理解任务。在本研究中，我们全面探讨了如何为大语言模型实现定向多说话人语音理解能力，具体面向智能眼镜使用场景。我们提出了两种将指向性融入大语言模型的新方法：（1）利用源分离前端模块的级联系统；（2）采用序列化输出训练的端到端系统。所有方法均利用嵌入智能眼镜中的多麦克风阵列，以优化指向性的解读与处理，并实现流式处理。实验结果表明，我们提出的方法能够有效赋予大语言模型定向语音理解能力，并在语音识别和语音翻译任务中取得了良好性能。

索引词— 大语言模型，定向语音理解，智能眼镜，序列化输出训练，源分离

1. 引言

在相距数英尺的距离上自动转录对话伙伴的语音，是自动语音识别（ASR）领域一项至关重要的新兴应用场景，在多说话人环境中尤为如此。智能眼镜等可穿戴设备可以利用这项技术生成实时字幕，从而支持包括语音翻译在内的一系列创新应用，并改善听力障碍人士的沟通体验。近期在 [1, 2, 3, 4, 5]定向自动语音识别（DASR）方面取得的进展表明，利用多麦克风阵列识别并分离说话者（包括佩戴者、对话伙伴和背景旁观者）是有效的；与传统的单通道波束成形方法相比，这种方法还表现出更强的抗噪鲁棒性。

最近，语音语言模型（小型语言模型）的出现已展现出巨大潜力，在广泛的一系列语音相关任务中，包括自动语音识别（ASR）、语音翻译、说话人分离和语音合成[6, 7, 8]。然而，这些任务中的一个关键挑战是处理存在重叠的多说话人语音。序列化输出训练（SOT）方法 [9] 被提出以应对这一挑战。序列化输出训练按照说话人起始时间对参考转录文本进行序列化，并在参考片段之间插入特殊的说话人切换标记。最近的多项研究提出了基于大语言模型的创新性序列化输出训练方法，用于多说话人自动语音识别。例如，Meng等人[10]提出了一种基于大语言模型的序列化输出训练方法，该方法利用预训练语音编码器和大语言模型，并在单通道多说话人数据集上对其进行微调。同样，Xie等人[11]提出了Directional-SpeechLlama，该方法利用序列化定向输出训练来执行语音识别和声源定位任务。此外，Shi等人 [12]提出了序列化输出提示（SOP），这是一种自适应提示方法，通过根据输入指明多说话人内容如何混合以及如何分离，从而提升在重叠情形下的性能。

在本文中，我们提出两种新颖方法，以增强小型语言模型处理和理解定向语音的能力，重点面向智能眼镜应用。具体而言，我们考虑佩戴者（近场声源）与远处的对话伙伴（远场声源）进行对话的场景，由此产生可通过多通道处理加以利用的显著空间声学差异。我们提出一种两阶段方法：首先采用流式多通道定向声源分离模型，将音频流分离为各个说话人片段，然后应用后处理，为每个较小的音频块预测说话人标签。随后，预测得到的说话人标签用于选择特定任务提示，例如在翻译任务中使用“翻译为目标语言”，从而使模型能够针对特定说话人和当前任务生成有针对性的响应。此外，我们提出一种基于 SOT 的方法，即使用 SOT 风格训练数据对大语言模型进行微调，使其能够接受多通道音频作为输入。该方法使模型能够纳入来自波束形成信号的定向信息，从而遵循针对各个说话人的指令，并促进个性化的、依赖说话人的处理。此外，对于这两种方法，我们均为小语言模型启用流式推理，

支持同时进行语音识别和翻译。

2. 提出的方法

2.1. 流式定向源分离模型

如图1所示，我们提出一个两阶段系统，使小型语言模型能够理解定向语音，该系统由源分离（源分离）模块和后处理模块组成。我们的源分离神经网络基于编码器-解码器架构 [13]。该过程首先从 K + 1波束成形通道中提取短时傅里叶变换（短时傅里叶变换）特征。这些时频特征随后被送入编码器模块，该模块由多个卷积块组成，其间采用门控线性单元（门控线性单元）激活 [14] 函数和Dropout层。编码输出随后通过一个3层长短期记忆网络，再经过一组卷积解码层。解码器输出随后被送入门控函数，该函数从参考音频中返回与佩戴者和对话伙伴语音相关联的STFT掩码。在我们提出的架构中，第一个音频通道被直接用作参考音频。最后一步是计算与佩戴者和对话伙伴相对应的掩蔽时频输出，然后使用逆短时傅里叶变换将其转换为各自的语音信号。源分离建模的优化目标是三种损失函数的组合：L1损失、STFT损失和对数SI-SDR损失。这种多目标方法使网络能够有效分离佩戴者与对话伙伴的语音，从而提升语音质量和可懂度。

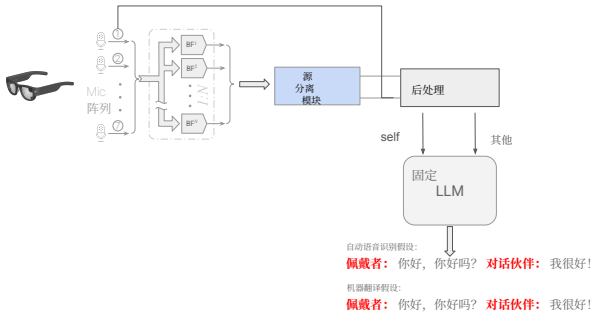


图 1。级联定向系统的示意图，该系统由源分离和小型语言模型组成。

然而，语音分离模型不可避免地会给下游任务带来语音失真。为缓解这一问题，我们不直接将分离后的音频用作小型语言模型的输入，而是利用分离后的音频以逐块方式预测说话人标签（例如，600 毫秒的音频块）。鉴于我们关注的是佩戴者与对话伙伴之间的对话，我们采用了一种简单而有效的方法：比较分离后的双通道信号（分别对应佩戴者和对话伙伴）的均方根（RMS）比值通道，以确定每个音频块中哪位说话人占主导。此外，我们使用语音活动检测模型过滤静音片段。由此得到的说话人标签

随后利用该结果为各种任务筛选合适的提示词。总之，最终的说话人标签通过以下公式确定：

$$\text{rms_ratio} = \frac{\text{RMS(self)}}{\text{RMS(other)}} > \alpha \tag{1}$$

α 是用于判断何时将说话人识别为佩戴者的阈值。分配说话人标签后，会将参考通道（单通道信号）输入小型语言模型，再由模型处理相应的提示。不过，这种方法有一个局限：无法处理重叠语音。端到端方法可以解决这一问题，例如基于序列化输出训练的方案，下一节将对此进行详细讨论。

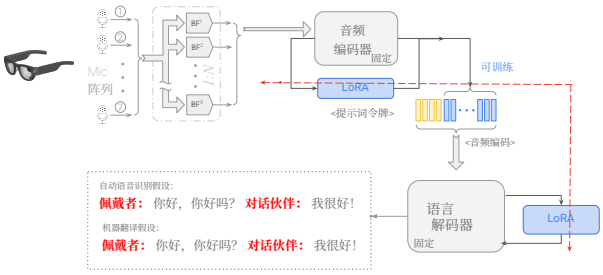


图 2。基于多通道序列化输出训练的小语言模型后训练示意图。

2.2. 序列化输出训练后训练

使小语言模型能够检测说话人变化的另一种方法是引入序列化输出训练（SOT）。这项技术旨在识别佩戴者与目标说话人之间的说话人转换，并识别部分重叠的语音。我们使用适当的提示，以序列化输出训练目标对小型语言模型进行微调，旨在让模型不仅能够检测说话人变化，还能执行语音识别或语音翻译等专门任务。为解决定向语音识别和定向语音翻译问题，我们提出了一种新颖的波束成形技术，详见 [5]。具体而言，我们采用非线性约束最小方差（NLCMV）波束成形来处理多通道音频输入。这些波束成形器使用预设系数并保持多通道结构，其中每个通道都聚焦于不同的方向。

为了保留 SLM 预训练音频编码器的原始状态，我们仅选择一个嘴部波束作为 SLM 的输入，而不使用多个波束。单个嘴部波束形成信号已经增强了佩戴者与对话伙伴信号之间的差异。先前研究已证明，低秩适配（LoRA） [15]有助于高效微调小型语言模型。如图 2 所示，我们在 SOT 后训练阶段将 LoRA 应用于音频编码器和解码器。

3. 实验与结果

3.1. 数据集

由于缺乏充足的真实世界多通道训练数据，本研究中的所有多通道训练数据均为仿真生成。仿真基于一个仿照 Aria 眼镜 [16]建模的五麦克风原型的几何结构。为引入空间多样性，我们采用了来自真实环境的房间脉冲响应（RIR），以 30° 为间隔模拟 12 个不同方向。为更好地贴合典型对话场景，我们为对话伙伴设定了五个关注的正面方向：-60°、-30°、0°、30° 和 60°。仿真训练数据通过对单通道音频片段进行空间定位生成，以表示两个主要角色：佩戴者及其对话伙伴。此外，我们还尝试模拟双语对话。有关仿真过程的更多详情，请参阅 [1]。基线系统和小语言模型微调用到的训练音频来源，是语音处理领域广泛采用的数据集：Common Voice 数据集 [17] 和多语言 LibriSpeech（MLS）数据集[18]。本研究关注四种语言：英语、法语、西班牙语和意大利语。评估使用了 Fleurs 数据集 [19] 。为训练翻译任务，我们使用教师机器翻译模型，将 Common Voice 和多语言 LibriSpeech 的转录文本自动翻译成目标语言。训练和测试语源分离模型时，使用了 LibriSpeech 数据集 [20]。

3.2. 模型设置

为更好地了解小型语言模型的方向能力，我们采用两个方向模型作为基线。第一个基线是基于 RNN-T 的多通道自动语音识别系统 [1], , 其模型配置与原始工作基本一致。针对每个波束形成器方向，提取 80 维 log-Mel 滤波器组特征，并输入卷积前端。该前端由两个含 5 个通道的二维卷积块组成，滤波器尺寸为 2×5，步幅设置为 1×2。随后，将输出与连续的六帧堆叠，形成一个 512 维向量，从而有效地将序列长度缩短 6 倍。接下来是 30 层 Emformer[21], , 每层包含 4 个注意力头和 2048 维前馈层。该模型约有 1.92 亿个参数。第二个基线是 JSTAR [22], , 这是一种端到端语音翻译系统，能够以流式方式同时输出自动语音识别结果和机器翻译（MT）结果。我们采用 [22] 中详述的相同模型架构。JSTAR 模型的参数总数为 1.21 亿，其中包括波束形成器和特征提取模块。这两个方向基线模型均使用上述数据集从头开始训练。

SS 模型架构改编自 [3]。对于每个波束形成器方向或原始麦克风通道，我们提取了 257 维复数短时傅里叶变换特征。多个方向或通道的输入特征会进行拼接。我们使用 Adam 优化器和三阶段学习率调度器。我们对语音分离模型进行了 60 个训练轮次，学习率设为 4e-4，预热为 10k 次迭代，并在 10 个训练轮次后强制退火。SS 模型总大小约为 6M 个参数。推理时，模型以分块解码模式运行，块大小为 600ms。

本研究采用的 SLM 是开源模型 Gemma-3n 4B [1]，它是一种经过指令微调的多模态大语言模型，可接受多模态提示并以文本作答。在使用多通道模拟数据进行 SOT 微调阶段时，我们将音频编码器和语言解码器的 LoRA 秩均设为 64。这样一来，更新的参数约占模型总参数的 1.9%。

3.3. 评估指标

源分离模型采用常用的语音质量感知评估（PESQ）分数 [23, 24, 25], 、短时客观可懂度（STOI）分数 [26] 和尺度不变信噪失真比（SI-SDR） [27]进行评估。此处使用的定向系统还需识别说话者（佩戴者或对话伙伴）。这一能力通过修改后的词错误率（WER）指标体现；除替换、插入和删除错误外，该指标还将归属于错误说话者的词计为错误。这类错误以 SA 表示，即“说话人归属”错误。对于翻译任务，则使用 BLEU [28] 分数评估模型性能。

3.4. 小语言模型流式解码

使用这个预训练小语言模型（SLM）进行同步语音翻译与识别，受到两个主要限制：模型不支持真正的流式输入，因此需要按轮次处理音频；每次推理处理的音频最长不得超过 30 秒。为在不重新训练 SLM 的情况下模拟流式体验，我们将输入音频流切分为 600 毫秒的片段，并可选择使用语源分离模型处理这些片段，以区分不同说话人（例如佩戴者和对话伙伴）。随后将这些片段累积到滑动窗口中，确保提供给 SLM 的音频上下文总时长不超过 30 秒。与此同时，我们持续记录此前生成的文本，并将其截断为最近的 50 个词，为模型提供相关的语言线索。尽管这种方法无法保证音频上下文与文本上下文完全对齐，但实证结果表明，SLM 仍能生成连贯、准确的转写文本和译文。每次推理时，我们会发送两个不同的提示：一个指示模型以源语言转写音频，另一个指示模型将音频翻译为目标语言，从而使系统能够以低延迟和高响应速度同时输出自动语音识别结果和译文。

[1] https://huggingface.co/google/gemma-3n-E4B-it

模型	佩戴者				对话伙伴			
	WER[%]↓	插入/删除/替换	句子准确率[%]↓	BLEU↑	WER[%]↓	插入/删除/替换	句子准确率[%]↓	BLEU↑
佩戴者：英语，对话伙伴：西班牙语								
多通道自动语音识别 [1]	16.5	1.5/5.7/9.2	0.0	-	13.2	1.0/5.9/6.3	0.0	-
JSTAR [22]	16.7	1.3/5.9/9.4	0.0	18.6	13.2	0.1/5.4/6.7	0.0	18.3
源分离+小语言模型	12.5	1.2/6.0/5.3	0.0	22.0	10.6	1.0/5.0/4.6	0.6	25.3
多通道序列化输出训练+小语言模型	17.3	5.2/5.9/5.8	0.0	19.6	16.5	1.7/7.3/5.7	1.8	22.6
佩戴者：英语，对话伙伴：法语								
多通道自动语音识别 [1]	16.7	1.5/5.9/9.3	0.0	-	22.0	1.1/8.1/12.7	0.0	-
JSTAR [22]	16.5	1.4/5.7/9.4	0.0	19.3	21.9	1.1/7.7/13.2	0.0	16.5
源分离+小语言模型	12.9	1.0/6.7/5.2	0.0	36.9	22.8	1.8/8.4/12.2	0.4	28.4
多通道序列化输出训练+小语言模型	15.5	2.8/7.1/5.5	0.2	34.7	25.9	2.5/9.3/11.4	2.5	26.5
佩戴者：英语，对话伙伴：意大利语								
多通道自动语音识别 [1]	16.1	1.4/5.8/8.9	0.0	-	22.3	0.8/14.6/6.8	0.0	-
JSTAR [22]	16.5	1.7/5.6/9.1	0.0	19.3	17.6	1.0/7.9/8.6	0.0	16.5
源分离+小语言模型	13.3	1.6/6.4/5.3	0.0	22.7	10.2	1.0/5.2/3.9	0.0	24.8
多通道 序列化输出训练+小语言模型	16.8	4.3/6.3/5.7	0.0	21.6	15.4	1.3/6.8/5.1	2.3	21.1

表1。simulatedFLEURS测试数据上的说话人归属（SA）误差以及归属后（“佩戴者”、“对话伙伴”）的词错误率（WER）。

指标	佩戴者		对话伙伴	
	混合	分离	混合	分离
PESQ↑	1.60	2.91	1.52	1.74
STOI↑	0.91	0.97	0.70	0.81
尺度不变信号失真比↑	5.85	19.56	-13.28	8.66

表2。语源分离模型在 PESQ、STOI 和尺度不变信号失真比方面的性能评估。

3.5. 结果

我们首先评估了SS模型的性能。如表2所示，根据本研究采用的三项指标，SS模型在佩戴者和对话伙伴的说话人分离方面均表现良好。例如，对话伙伴说话人的尺度不变信号失真比（SI-SDR）显著提升，从 -13.28 dB 提高到 8.66 dB。这表明，SS模型能够有效分离并增强目标说话人的音频，同时抑制背景噪声。

接下来，我们对基线方法和所提出的方法进行了比较分析，重点关注语音识别和语音翻译。表1所示结果表明，在说话人归属错误方面，基于SLM的定向系统展现出更强的方向理解能力。与专为智能眼镜设计的定向基线系统相比，在模拟多通道FLEURS测试集上，SS+SLM和SOT+SLM对佩戴者的句子准确率几乎达到完美水平，与基线系统的表现相当。不过，使用基于SOT的SLM时，我们观察到对话伙伴的归属错误率较高。尽管如此，所有语言对的错误率仍低于2.5%。这主要是因为SLM不太擅长遵循指令，可能是由于我们对模型进行了微调，使其同时输出转录文本和译文。模型似乎混淆了转录与翻译，这可能导致其难以准确归属说话人。在BLEU分数方面，所提出的两种系统均优于多通道自动语音识别和JSTAR系统。具体而言，在“对话伙伴说西班牙语”的场景中，与JSTAR相比，SS+SLM和SOT+SLM将BLEU分数从18.3分别提高到25.3和22.6。

对于语音识别任务，SS+SLM 方案的表现似乎优于所有系统。然而，在某些情况下，与基线系统相比，SOT+SLM 方案的表现欠佳。详细的错误分析显示，性能下降主要是由说话人归属错误造成的。说话人归属错误导致“对话伙伴”的删除错误率更高，同时也导致“佩戴者”的插入错误率更高。为了进一步提升性能，一种可行的方法是探索替代策略，使模型能够准确地遵循指令。

对这两种 proposed 系统的比较表明，SS+SLM 在语音识别和语音翻译任务中的表现均优于 SOT+SLM。不过，需要注意的是，SS+SLM 方法以当前形式存在局限性，因为它无法处理重叠语音。该方法非常适合语音翻译应用场景，因为可以合理地预期对话中的重叠语音较少——参与者通常会等待翻译完成后再回应他人。相比之下，SOT+SLM 的主要优势之一在于能够有效处理多位说话人的重叠语音。

4. 结论

本文提出了两种方法，使小语言模型具备针对智能眼镜场景的方向理解能力。实验结果表明，基于小语言模型的解决方案能够准确区分佩戴者（近场）和对话伙伴（远场）的语音，在说话人归属错误方面表现出色。此外，所提出的基于源分离的小语言模型和基于SOT的小语言模型在语音识别和语音翻译任务中均表现出优异性能。

5. 参考文献

[1] Ju Lin、Niko Moritz、Ruiming Xie、Kaustubh Kalgaonkar、Christian Fuegen 和 Frank Seide，“用于说话人消歧和串音抑制的定向语音识别”，载于*INTERSPEECH 2023* 论文集，2023年，第 3522–3526 页。[2] Aswin Shanmugam Subramanian、Chao Weng、Shinji Watanabe、Meng Yu、Yong Xu、Shi-Xiong Zhang 和 Dong Yu，“定向自动语音识别：一种结合声源定位的端到端多说话人语音识别新范式”，载于*2021*年*ICASSP—2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第 8433–8437 页。[3] Tiantian Feng、Ju Lin、Yiteng Huang、Weipeng He、Kaustubh Kalgaonkar、Niko Moritz、Li Wan、Xin Lei、Ming Sun 和 Frank Seide，“面向智能眼镜鲁棒语音识别的定向声源分离”，*arXiv* 预印本 *arXiv:2309.10993*，2023年。[4] Ju Lin、Yiteng Huang、Ming Sun、Frank Seide 和 Florian Metze，“具备全双工能力的定向语音识别”，载于*INTERSPEECH 2025* 论文集，2025年，第 2570–2574 页。[5] Ju Lin、Niko Moritz、Yiteng Huang、Ruiming Xie、Ming Sun、Christian Fuegen 和 Frank Seide，“Agadir：迈向阵列几何无关的定向语音识别”，载于*ICASSP 2024—2024*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第 11951–11955 页。[6] Siddhant Arora、Kai-Wei Chang、Chung-Ming Chien、Yifan Peng、Haibin Wu、Yossi Adi、Emmanuel Dupoux、Hung-Yi Lee、Karen Livescu 和 Shinji Watanabe，“口语语言模型的发展概况：全面综述”，*arXiv* 预印本 *arXiv:2504 08528*，2025年。.[7] Shujie Hu、Long Zhou、Shujie Liu、Sanyuan Chen、Lingwei Meng、Hongkun Hao、Jing Pan、Xunying Liu、Jinyu Li、Sunit Sivasankaran 等，“WavLLM：迈向鲁棒且自适应的语音大语言模型”，*arXiv* 预印本 *arXiv:2404.00656*，2024年。[8] Yuan Gong、Hongyin Luo、Alexander H Liu、Leonid Karlinsky 和 James Glass，“聆听、思考与理解”，*arXiv* 预印本 *arXiv:2305.10790*，2023年。[9] Naoyuki Kanda、Yashesh Gaur、Xiaofei Wang、Zhong Meng 和 Takuya Yoshioka，“用于端到端重叠语音识别的序列化输出训练”，*arXiv* 预印本 *arXiv:2003.12687*，2020年。[10] Lingwei Meng、Shujie Hu、Jiawen Kang、Zhaoqing Li、Yuejiao Wang、Wenxuan Wu、Xixin Wu、Xunying Liu 和 Helen Meng，“大语言模型能够通过灵活的指令转录多说话人场景中的语音”，载于*2025*年*ICASSP—2025*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第 1–5 页。[11] Jiamin Xie、Ju Lin、Yiteng Huang、Tyler Vuong、Zhaojiang Lin、Zhaojun Yang、Peng Su、Prashant Rawat、Sangeeta Srivastava、Ming Sun 等，“以指向性为思考维度：用于多说话人定向语音识别的语音大语言模型”，*arXiv* 预印本 *arXiv:2506.14973*，2025年。[12] Hao Shi、Yusuke Fujita、Tomoya Mizumoto、Lianbo Liu、Atsushi Kojima 和 Yui Sudo，“用于基于大语言模型的多说话人语音识别的序列化输出提示”，*arXiv* 预印本 *arXiv:2509.04488*，2025年。[13] Ju Lin、Kaustubh Kalgaonkar、Qing He 和 Xin Lei，“低比特率语音编解码器的语音增强”，载于*ICASSP 2022—2022*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第 7777–7781 页。[14] Yann N Dauphin、Angela Fan、Michael Auli 和 David Grangier，“基于门控卷积网络的语言建模”，载于国际机器学习会议。PMLR，2017年，第 933–941 页。[15] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等，“LoRA：大语言模型的低秩适配”，*ICLR*，第 1 卷，第 2 期，第 3 页，2022年。

[16] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith 等，“Project aria：一种用于第一人称视角多模态人工智能研究的新工具”，*arXiv* 预印本 *arXiv:2308.13561*，2023年。[17] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Fran- cis 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*，2019年。[18] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve 和 Ronan Collobert，“MLS：一个用于语音研究的大规模多语言数据集”，*arXiv* 预印本 *arXiv:2012.03411*，2020年。[19] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axel- rod, Siddharth Dalmia, Jason Riesa, Clara Rivera 和 Ankur Bapna，“FLEURS：语音通用表征的小样本学习评估”，载于*2022*年*IEEE*口语技术研讨会*(SLT)*。IEEE，2023年，第798–805页。[20] Vassil Panayotov, Guoguo Chen, Daniel Povey 和 Sanjeev Khudan- pur，“LibriSpeech：基于公有领域有声读物的自动语音识别语料库”，载于*2015*年*IEEE*国际声学、语音与信号处理会议*(ICASSP)*。IEEE，2015年，第5206–5210页。[21] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Ju- lian Chan, Frank Zhang, Duc Le 和 Mike Seltzer，“Emformer：一种基于高效记忆变换器的声学模型，用于低延迟流式语音识别”，载于*2021*年*ICASSP-2021*年*IEEE*国际声学、语音与信号处理会议*(ICASSP)*。IEEE，2021年，第6783–6787页。[22] Niko Moritz, Ruiming Xie, Yashesh Gaur, Ke Li, Simone Merello Zee- shan Ahmed, Frank Seide 和 Christian Fuegen，“快速与慢速的转录和翻译：联合语音翻译与识别”，载于*2025*年*ICASSP-2025*年*IEEE*国际声学、语音与信号处理会议*(ICASSP)*。IEEE，2025年，第1–5页。[23] Anthony W. Rix, John G. Beerends, Michael P. Hollier 和 An- dries P. Hekstra，“语音质量感知评估 (PESQ)——一种用于电话网络和编解码器语音质量评估的新方法”，盐湖城，犹他州，2001年5月，第749–752页。[24] “语音质量感知评估 (PESQ)：一种用于窄带电话网络和语音编解码器端到端语音质量评估的客观方法”，2001年2月。[25] “感知客观听音质量预测”，2018年3月。[26] Cees H. Taal, Richard C. Hendriks, Richard Heusdens 和 Jesper Jensen，“一种对时频加权含噪语音进行可懂度预测的算法”，第19卷，第7期，第2125–2136页，2011年9月。[27] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan 和 John R. Her- shey，“SDR——半生不熟还是恰到好处？”，2019年，第626–630页。[28] Kishore Papineni, Salim Roukos, Todd Ward 和 Wei-Jing Zhu，“BLEU：一种机器翻译自动评估方法”，载于计算语言学协会第*40*届年会论文集，2002年，第311–318页。