使用 LLMs 进行语音到语音翻译的计划式交叉语音文本训练

Hayato Futami[1], *Emiru Tsunoo*[1], *Yosuke Kashiwagi*[1], *Yuki Ito*[1], *Hassan Shahmohammadi*[2], *Siddhant Arora*[3], *Shinji Watanabe*[3]

[1]Sony Group Corporation, Japan
[2]Sony Europe, Germany
[3]Carnegie Mellon University, USA

Hayato.Futami@sony.com

Abstract

语音到语音翻译（S2ST）在大型语言模型（LLM）的推动下得到了进步，这些模型在离散语音单元上进行了微调。在这种方法中，从文本到语音的模态适应一直是一个问题。LLM 是在仅有文本的数据上训练的，这在将它们适应到具有有限语音到语音数据的语音模态时提出了挑战。为了解决训练困难，我们在本研究中提出了计划的交错语音–文本训练。我们在训练过程中使用交错的语音–文本单元而不是语音单元，其中对齐的文本标记在词级别进行交错。我们随着训练的进展逐步减少文本的比例，以促进从文本到语音的渐进式模态适应。我们通过在 CVSS 数据集上对 LLaMA3.2-1B 进行微调，进行实验评估用于 S2ST。我们表明，所提出的方法持续改善了翻译性能，特别是对于训练数据有限的语言。

Index Terms: speech-to-speech translation, large language model, modality adaptation

1. 引言

语音到语音翻译（S2ST）是一种有前景的技术，可以将一种语言的语音转换为另一种语言，这对于与不共享相同语言的人进行交流很有用。传统上，S2ST 是通过级联的方法解决的，包括自动语音识别（ASR）、机器翻译（MT）和文本到语音（TTS）合成 [1] 。相比之下，端到端 S2ST 系统已经出现 [2–9] ，它们通过联合优化将源语言的语音翻译为目标语言的语音。它们的优势在于简化的架构和通过联合优化避免错误传播。

最近，端到端系统采用了离散语音单元，而不是连续的梅尔频谱图特征 [4–9] 。离散单元主要有两种类型：语义单元和声学单元。语义单元捕捉语音中的语义信息，从自监督（SSL）语音编码器 [10,11] 获取。声学单元捕捉声学细节，从神经编解码器 [12,13] 获取。对于 S2ST，语义语音语言模型用于语义翻译，声学语言模型连接用于高质量语音生成，通常保留讲话者身份 [7–9] 。

在自然语言处理领域取得成功后，许多工作专注于微调预训练的 LLMs 以进行语音处理。它们主要应用于 ASR 和语音转文本翻译任务。使用离散语音单元为基于预训练 LLMs 的 S2ST 铺平了道路。AudioPaLM 是这方面的一个代表性研究，其通过微调文本 LLM（PaLM2）在语义单元上实现。由于 LLMs 强大的语言理解能力，AudioPaLM 在广泛的语言中优于现有的语音转文本和语音转语音翻译系统。在这项研究中，我们构建了一个遵循 AudioPaLM 的 S2ST 系统。

在对语音任务进行文本大语言模型（LLM）的微调时，模态适应一直是一个问题。由于 LLM 是在大规模的仅文本数据上训练的，因此用有限的监督数据将 LLM 适应到语音领域是具有挑战性的。语音和文本模态之间存在长度和表示差异 [14,15] ，这可以通过长度压缩 [15,16] 、模态适配器模块 [17] 和训练目标 [15] 来解决。在 AudioPaLM [7] 中，一个 SSL 语音编码器在 LLM 微调之前通过一个 ASR 任务进行微调。这有助于通过获取类似于文本表示的离散语音单元来减轻模态差距，从而显著提高性能。

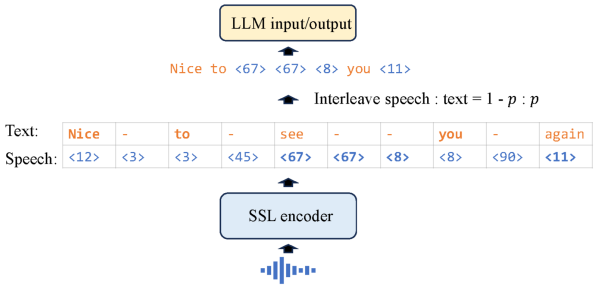


Figure 1: 交错的语音-文本单元被用作 *LLM* 的输入和输出。随着训练的进行，文本比率 p 逐渐减少，这我们称之为计划交错训练。

最近，SpiRitLM [18] 被提出为一种语音-文本多模态语言模型，其在大规模的仅语音、仅文本以及对齐的语音-文本数据上进行预训练。SpiRitLM 使用交错的语音-文本表示在对齐数据上进行训练，其中语音和对齐的文本单元在词级上随机混合。据报道，这种训练实现了更好的模态转换，进而在语音和文本理解任务上带来了性能提升。

在这项研究中，我们提出了一种如图 1 所示的计划交错语音-文本训练。我们专注于使用有限的监督数据进行 S2ST 的微调预训练 LLMs，而不是像 [18] 中的预训练基础语音-文本 LMs。由于预训练的 LLMs 没有解释和生成语音单元的能力，我们建议逐步将 LLMs 适应到语音单元中。在 S2ST 中，我们在输入和输出端都使用词级的交错语音和文本单元。为了进行渐进式适应，我们建议逐渐减少文本单元的比例至零，最终在推理时结果完全为语音单元。

在实验评估中，我们通过对 LLaMA3.2-1B LLM [19] 进行微调，并基于 CVSS 语料库 [20] 构建了一个 S2ST 系统。我们使用了 w2v-BERT 编码器的离散语音单元，该编码器通过 CTC 目标进行 ASR 微调。波形生成方面，我们使用了基于单元的 HiFi-GAN 声码器 [21, 22] 。整个系统如图 2 所示。我们展示了提出的定期交错语音–文本训练提高了 S2ST 的性能，这可以归因于更好的模态适应。所提方法在不同语言中始终有效，尤其对训练数据有限的语言效果显著。

2. 相关工作

2.1. 语音到语音翻译（S2ST）

端到端语音到语音翻译（S2ST）系统一直在积极研究，它作为一个语音到语音任务联合优化 [2–9] 。早期研究通过序列到序列模型直接预测目标语言中的声谱图来解决该任务 [2, 3] 。最近，语音合成已经出现并通过离散语音单元

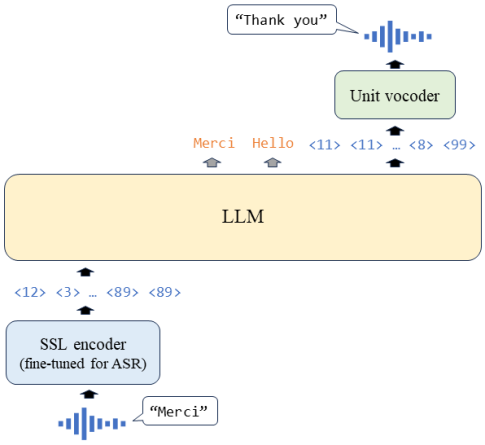


Figure 2: 我们的基于 *LLM* 的语音到语音翻译系统。

上的自回归语言模型进行 [21, 23–26] 。考虑了两种类型的离散语音单元。第一种类型是语义单元，它代表从 SSL 语音编码器获得的语义信息。第二种类型是声学单元，它代表来自神经编解码器的声学细节。生成语音语言建模（GSLM）在语义单元上训练以继续语音 [23] 。AudioLM 在语义单元和声学单元上均进行了训练 [24] 。VALL-E 扩展了在声学单元上训练的语言模型以实现零样本文本到语音（TTS）[25,26] 。根据这一趋势，S2ST 也利用离散语音单元解决。UnitY [5] 使用语义单元，采用语音到文本和文本到单元转换的两阶段架构，这些模型也在大量多语言数据上进行训练，称为 SeamlessM4T [6] 。AudioPaLM 微调预训练的 LLM 在语义单元上，并通过 AudioLM 生成声学单元语音 [24] 。PolyVoice [8] 和 MSLM-S2ST [9] 采用了类似的架构，包括语义和声学语言模型。在这项研究中，我们专注于如何微调预训练的 LLM 以及翻译中的语义准确性。就像在 AudioPaLM 中，我们专注于在语义单元上微调预训练的 LLM，并通过语音单元 HiFi-GAN 声码器生成语音，如 [6,21] [1] 。

2.2. 用于语音处理的 LLM

自然语言处理领域随着大型语言模型的发展取得了显著进展 [19,27,28] 。在多模态大型语言模型的背景下，各种努力已被付诸实践，以将预训练的大型语言模型扩展到语音模态 [16,17,29–33] 。它们添加了一个语音编码器作为大型语言模型的前端以支持语音输入，其中采用了基础的自监督学习或语音识别模型。此外，常常引入长度和模态适配器以填补语音编码器和大型语言模型之间的模态差距 [16,17] 。这些模型使用连续特征作为输入给大型语言模型，主要专注于语音到文本任务。与连续表示不同，几种方法利用离散语音单元并整合语音生成任务 [31,33] ，包括语音到语音翻译（S2ST）[7] 。最近，提出了 SpiRitLM，该模型在交错的语音-文本上预训练，并展示了强大的语音-文本理解和上下文学习能力 [18] 。这也通过合成交错数据进行了扩展 [34] 。我们并没有预训练能够自由混合语音和文本的基础语音-文本语言模型，而是专注于使用有限的监督数据微调预训练的大型语言模型，并提出一种逐渐减少文本比例的调度策略。

[1] 由于简洁性和训练稳定性，用单元声码器替代了 AudioLM

3. 方法

3.1. 语音到语音翻译系统

如图 2 所示，我们采用了一种从 LLM 微调的语音到语音翻译 (S2ST) 系统，类似于 AudioPaLM [7] 。令 S^{src} 表示源语言中的语义单元，S^{tgt} 表示目标语言中的语义单元。端到端 S2ST 通过最小化以下目标函数从 S^{src} 预测 S^{tgt} 进行训练：

$$\mathcal{L} = -\log p(S^{\text{tgt}}|S^{\text{src}};\theta), \tag{1}$$

其中 θ 表示模型参数。如在 [7] 中所做，我们将预测转录文本 T^{src} 和翻译文本 T^{tgt} 作为中间步骤，这被联合建模为：

$$\mathcal{L} = -\log p(S^{\text{tgt}}|T^{\text{tgt}}, T^{\text{src}}, S^{\text{src}};\theta) \\ p(T^{\text{tgt}}|T^{\text{src}}, S^{\text{src}};\theta)p(T^{\text{src}}|S^{\text{src}};\theta), \tag{2}$$

我们称之为链式思维 (CoT) 提示。注意，模型在预测 S^{tgt} 时可以访问 S^{src} ，从而实现端到端优化 [7] 。模型参数初始化为文本 LLM 的参数。其词汇 $\mathcal{V}$ 扩展为语音单元的词汇 $\mathcal{V}_s$ 与来自 LLM 的文本 BPEs 的词汇 $\mathcal{V}_t$ 的组合，即 $\mathcal{V} = \mathcal{V}_s \cup \mathcal{V}_t$ 。

大型语言模型仅在文本标记上进行了预训练，没有见过语音数据。因此，在微调阶段，大型语言模型难以适应语音单元，因为有监督的训练数据通常很有限。语音和文本在长度和表示上存在模态差距 [14] ，这使得跨模态微调变得困难。研究表明，微调自监督学习编码器用于自动语音识别是有效的 [7] ，这会减少模态差距并减轻训练难度。

3.2. 计划交错语音–文本训练

除了对 SSL 进行 ASR 微调外，我们还提出了借鉴 SpiRitLM [18] 的计划交错语音–文本训练，以更好地适应模态。首先，我们计算语音的词级对齐。由于我们采用了针对基于 CTC 的 ASR 微调的语音编码器，我们可以利用 CTC [35] 提供的对齐，而不依赖于任何外部对齐器。然后，我们在词级随机混合语音和对齐的文本单元，以构建交错的语音–文本单元，如图 1 所示。在 LLM 微调期间，我们使用交错的语音–文本单元 I_p^{src} 和 I_p^{tgt} 而不是语音语义单元，其中 p 表示文本比例。方程（2 ）被改写为 I_p^{src} 和 I_p^{tgt} 的形式：

$$\mathcal{L} = -\log p(I_p^{\text{tgt}}|T^{\text{tgt}}, T^{\text{src}}, I_p^{\text{src}};\theta) \\ p(T^{\text{tgt}}|T^{\text{src}}, I_p^{\text{src}};\theta)p(T^{\text{src}}|I_p^{\text{src}};\theta). \tag{3}$$

算法 1 详细解释了交错算法。输入是语音语义单元 S (S^{src} 或 S^{tgt}) 以及相应的单词对齐 A ，表示为 S 中的起始和结束帧索引 a_i, b_i ，还有单词表示 w_i 。输出是交错的语音-文本单元 I_p (I_p^{src} 或 I_p^{tgt})。我们通过从剩余索引 J 中随机选择起始索引 j （第 3 行）和从泊松分布中选择跨度长度 l （第 4 行）来随机选择单词跨度。选定跨度 $s_{a_j}, ..., s_{b_{j+l}}$ 中的语音单元被相应的字节对编码（BPE）标记 BPE($w_j, ..., w_{j+l}$) 替换（第 5 行）。在所选单词的百分比不超过文本比率 p 的情况下应用文本替换（第 2 行）。

在这项研究中，我们考虑根据训练步骤调度文本交织比例 p 。我们建议随着训练的推进（例如每隔 300 步），减少文本比例 p ，以实现从文本 LLM 到语音 LLM 的无缝模态转换。这意味着比文本单元更长的语音单元的比例逐渐增加。其长度和学习表示逐渐适应，如后面图 3 所示，这将缓解模态适应的困难。p 的调度是与 SpiRitLM [18] 的主要不同。此外，我们专注于使用有限监督数据对 S2ST 任务进行微调，而不是像 SpiRitLM 那样进行大规模的预

Algorithm 1 交错的语音-文本单元

Input: speech semantic units $S = \{s_1, ..., s_M\}$, word alignments $A = \{(a_1, b_1, w_1), ..., (a_j, b_j, w_j), ..., (a_N, b_N, w_N)\}$
Output: interleaved speech–text units I_p

Initialize $I_p \leftarrow S, J \leftarrow \{0, ..., N\}$.
while $N - |J| \leq p \cdot N$ **do**
  Randomly select $j \in J$.
  Sample length $l \sim$ Poission(λ) .
  Replace $s_{a_j}, ..., s_{b_{j+l}}$ in I_p with BPE($w_j, ..., w_{j+l}$).
  $J \leftarrow J - \{j, ..., j + l\}$.
end while

Table 1: *w2v-BERT (W2VB) 的 ASR 微调 (FT) 和连锁思维 (CoT) 提示的效果。*

	Fr-en
	ASR-BLEU(↑) / UTMOS(↑)
W2VB	9.6/4.35
ASR FT W2VB (baseline)	**28.8**/4.21
w/o CoT	13.7/4.21

训练。与 SpiRitLM 不同，我们还研究了多语言设置、有 ASR 微调的语义单元，以及与链式思维提示的结合。

我们基于 CVSS [20] 语料库进行语音到语音翻译（S2ST）实验。CVSS 语料库是一种广泛使用的多语言 S2ST 语料库，通过从 CoVoST2 [36] 语音到文本翻译语料库进行语音合成构建。CVSS 包含两个版本：CVSS-C，具有单一高质量规范声音，以及 CVSS-T，从源语音进行语音转移。在实验中，我们使用了 CVSS-C。

如第 3 节所述，我们使用了经过微调的 SSL 编码器中的语义单元。我们针对所有 CoVoST2 和 CVSS 训练数据（21 种语言加上英语）对 w2v-BERT 编码器进行了 ASR 微调。w2v-BERT 的微调使用了 Adam 优化器，学习率为 2 e −5 。我们为 BPE 添加了一个大小为 5 K 的线性层，并使用 CTC 目标训练整个模型。然后，我们使用 CoVoST+CVSS 数据在第 20 层 w2v-BERT 的特征上训练 k -means，聚类大小为 2048 。我们使用来自 k -means 的聚类索引作为语义单元，而不进行去重。我们针对 CVSS 的每种语言子集微调了 LLaMA3.2-1B [2] LM。该 LM 的词汇量总共有约 12 K 文本 BPE（与原始 LLaMA 相同）和 2048 语音单元。LM 微调使用的 Adam 优化器的学习率为 5 e −5 ，丢弃率为 0.2 。在训练和推理期间，我们在公式 (3) 中应用了 CoT 提示，其中语音识别和文本翻译作为中间步骤进行。我们使用 ESPnet 工具箱实现了 w2v-BERT 和 LM 的微调 [37] 。对于波形生成，我们在 CVSS 数据上训练了一个单元 HiFi-GAN 声码器，该数据由合成的英语单人语音组成。在单元 HiFi-GAN 中，我们使用来自 w2v-BERT 的语义单元作为输入，而不是 mel 频谱图。我们按照其 LJSpeech 配方 [3] 使用 ParallelWaveGAN 工具包实现了这一点。

首先，我们在表 1 中测试了 w2v-BERT 的 ASR 微调和 CoT 提示的有效性。我们使用了法语到英语（Fr-en）的 CVSS。我们在 Fr-en 的训练集上微调了语言模型，并在测试集 [4] 上进行了评估。翻译准确性通过 ASR-BLEU 分数进行评估，其中 BLEU 分数是在由 Whisper [38] large-v3 模型生成的语音的 ASR 转录上计算的。音频质量通过 UTMOS [39] 进行评估。我们确认 ASR 微调（ASR FT）对于准确性至关重要，如 [7, 40] 中所示。同时也确认了 CoT 提示的有效性，如 [7] 中所示。在此之后，我们将同时应用它们的语言模型视为基线。

表 2 展示了我们在 CVSS 中对七种语言子集进行的拟议计划交替训练（ILT）的主要实验。我们比较了三种训练方法：没有任何交替的基线（Baseline）、提议的计划交替语音–文本训练（Scheduled ILT）以及不带计划的 ILT（ILT w/o scheduling）。对于 ILT，我们通过 CTC 分段 [35] ，使用微调的 w2v-BERT 编码器获得词对齐。在计划的 ILT 中，算法 1 中的文本比例 p 以 0.9 初始化。文本比例 p 每隔 300 个训练步骤逐渐减少 0.1 。在不带计划的 ILT 中，我们使用了常数 p 为 0.3 。表 2 显示，计划的 ILT 在所有七种语言中在翻译准确度（ASR-BLEU）方面优于基线，证明了我们提出方法的有效性。我们还发现，在训练数据较少的语言（It-en、Ru-en 和 Pt-en）中，改进特别显著。对于语音模态的适应在语音到语音数据有限的情况下是困难的，但提出的训练缓解了这一困难。这对于现实世界的应用非常有利，因为监督训练数据通常是有限的。此外，我们确认生成语音的质量（UTMOS）在高水平上，与训练方法无关。最后，我们将计划的 ILT 与不带计划的 ILT 进行了比较，发现计划对于翻译性能是有效的。

表格 3 展示了计划的 ILT 组件在 CVSS 葡萄牙语到英语（Pt-en）子集上的消融研究。正如在章节 3 中描述的，我们在输入（源）和输出（目标）两侧使用了交错的语音-文本单元（I_p^{src} 和 I_p^{tgt} ）。为了验证在两侧进行交错是否必要，我们将其与仅在输入或输出侧进行交错进行比较（第 2 行或第 3 行）。例如，仅在输入侧交错对应于在等式（3 ）中使用 S^{tgt} 而不是 I_p^{tgt} 。我们表明，相比仅在输入或输出侧进行交错，在两侧进行交错对于翻译性能是必要的。此外，我们查看我们的方法是否不仅仅是一个带有噪声输入和输出的数据增强方法，将其与使用掩码而不是文本标记交错进行比较（第 4 行）。具体来说，我们在算法 1 的第 5 行，用一个特殊的掩码标记（`[MASK]` ）替代选定的语音单元 $s_{a_j}, ..., s_{b_{j+l}}$ ，而不是使用 BPE。我们观察到相比语音-文本交错，掩码的效果更差。最后，为了查看词对齐是否必要，我们使用不对齐的交错（第 5 行）。我们假设词在语音中以等间隔出现。这在算法 1 中表示为 $a_i = i \cdot \lfloor M/N \rfloor$ 和 $b_i = (i+1) \cdot \lfloor M/N \rfloor - 1$ 。我们观察到不进行对齐的交错导致性能下降，这说明了对齐的重要性。

如图 3 所示，所提出的计划 ILT 是如何影响训练的，在长度和在语音与文本之间学习到的表示方面。我们使用了 Pt-En 的基线和计划的 ILT 模型。左图显示的是语音单元长度与文本单元长度的比率，即 $|S|/|T|$ 。不是 S ，而是计划的 ILT 采用了交错的单元 I_p 。我们在源端和目标端都进行了测量。我们发现，语音和文本之间的长度显著不同，其中语音单元比文本长，在源序列中长 21.7 倍，在目标序列中长 13.1 倍。计划的 ILT 逐渐降低文本比率 p ，因此长度逐渐增加，如图 3 [5] 所示，这将减少训练的难度。右图显示了隐藏表示的余弦相似度。我们分析了最后一层 LLaMA 的隐藏状态，平均在每种输入类型对应的序列上（例如 S^{src} ）。我们测量了余弦相似度，包括源语音和文本（src S-T）、源和目标文本（src-tgt T）以及目标文本和语音（tgt T-S）之间的相似度。我们发现，基线模型在早期训练阶段难以学习语音和文本之间的关系。另一方面，在计划的 ILT 中，由于交错中较高比例的文本，语音和文本之间的相似性在早期阶段显得更高。这有助于学习语音和文本之间的互动，并且在最终阶段（无交错，$p = 0$ ）学习到的表示比基线更为相似。相比之下，计划的 ILT 对文本和文本之间的相似性没有产生影响，因为我们的方法提高

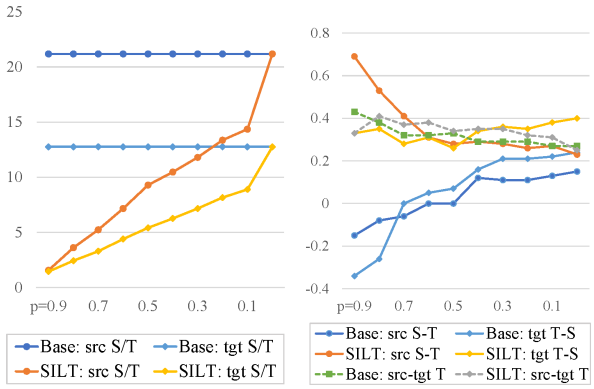
[2] https://huggingface.co/meta-llama/Llama-3.2-1B
[3] https://github.com/kan-bayashi/ParallelWaveGAN
[4] 我们从测试集中随机抽取了 1 K 个样本。在某些语言中，样本数量过多（超过 10 K）。
[5] 在 $p = 0.1$ 和 0 之间观察到了长度的跳跃。实际文本比例大于 p ，因为当文本比例超过 p 时，算法 1 中的循环被停止。

Table 2: 在 *CVSS* 上的语音到语音翻译（*S2ST*）性能，报告中使用 *ASR-BLEU* 和 *UTMOS* 得分。提出的定时交错语音-文本训练（*ILT*）在 *ASR-BLEU* 中优于没有它的基线。

	Fr-en (180 h)	De-en (119 h)	Es-en (97 h)	Ca-en (81 h)	It-en (28 h)	Ru-en (16 h)	Pt-en (7 h)
Baseline	28.8/4.21	27.3/4.22	33.5/4.28	23.8/4.17	12.8/4.23	6.0/4.25	10.3/4.17
+Scheduled ILT	**29.5**/4.23	**29.2**/4.21	**33.9**/4.28	**25.9**/4.21	**19.5**/4.23	**14.1**/4.23	**19.5**/4.17
+ILT w/o scheduling	25.8/4.11	22.1/4.11	21.9/4.29	17.8/4.10	14.4/4.22	6.0/4.12	12.0/4.11

Table 3: 在 *CVSS Pt-en* 上的 *S2ST* 消融研究，报告使用了 *ASR-BLEU* 和 *UTMOS* 评分。

	Pt-en
Scheduled ILT	**19.5**/4.17
- Interleave input only	11.1/4.18
- Interleave output only	12.6/4.18
- Mask instead of interleave	6.2/4.17
- Interleave without alignment	11.5/4.16



(a) 长度 (b) 表征相似性

Figure 3: 在基线（*Base*）和计划的渐进翻译（*SILT*）训练过程中观察到的语音（*S*）与文本（*T*）之间的长度和学习表示差异。文本比例 p 每 300 步逐步以 0.1 递减。

了语音和文本的适应性。

4. 结论

在本研究中，我们提出了一种新的训练方法，称为基于 LLM 的语音到语音翻译（S2ST）的计划交错语音-文本训练。在微调 LLM 期间，我们使用交错的语音-文本单元作为 LLM 的输入和输出，而不是语音单元。我们逐渐减少交错中文本的比例，以便更好地从文本到语音的模态适应。我们在 CVSS 语料库上进行了 S2ST 实验，结果显示我们的方法在七种语言中始终提高了翻译性能。在未来的工作中，我们将把这种方法扩展到其他语音对语音系统，例如口语对话系统。

5. References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *TASLP*, 2006.

[2] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Interspeech*, 2019.

[3] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *International Conference on Machine Learning*, 2021.

[4] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, "Direct speech-to-speech translation with discrete units," in *ACL*, 2022.

[5] H. Inaguma, S. Popuri, I. Kulikov, P.-J. Chen, C. Wang, Y.-A. Chung, Y. Tang, A. Lee, S. Watanabe, and J. Pino, "UnitY: Two-pass direct speech-to-speech translation with discrete units," in *ACL*, 2023.

[6] SeamlessCommunication *et al.*, "SeamlessM4T: Massively multilingual&multimodal machine translation," 2023.

[7] P. K. Rubenstein *et al.*, "AudioPaLM: A large language model that can speak and listen," *ArXiv*, 2023.

[8] Q. qian Dong, Z. Huang, Q. Tian, C. Xu, T. Ko, yunlong zhao, S. Feng, T. Li, K. Wang, X. Cheng, F. Yue, Y. Bai, X. Chen, L. Lu, Z. MA, Y. Wang, M. Wang, and Y. Wang, "PolyVoice: Language models for speech to speech translation," in *ICLR*, 2024.

[9] Y. Peng, I. Kulikov, Y. Yang, S. Popuri, H. Lu, C. Wang, and H. Gong, "MSLM-S2ST: A multitask speech language model for textless speech-to-speech translation with speaker style preservation," *ArXiv*, 2024.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. rahman Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[11] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *ASRU*, 2021.

[12] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *TASLP*, 2021.

[13] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv*, 2022.

[14] Y. Liu, J. Zhu, J. Zhang, and C. Zong, "Bridging the modality gap for speech-to-text translation," *ArXiv*, 2020.

[15] I. Tsiamas, G. Gállego, J. Fonollosa, and M. Costa-jussà, "Pushing the limits of zero-shot end-to-end speech translation," in *ACL*, 2024.

[16] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu, and Y. Wu, "On decoder-only architecture for speech-to-text and large language model integration," in *ASRU*, 2023.

[17] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. MA, and C. Zhang, "SALMONN: Towards generic hearing abilities for large language models," in *ICLR*, 2024.

[18] T. Nguyen, B. Muller, B. Yu, M. R. Costa-jussà, M. Elbayad, S. Popuri, P.-A. Duquenne, R. Algayres, R. Mavlyutov, I. Gat, G. Synnaeve, J. Pino, B. Sagot, and E. Dupoux, "SpiRit-LM: Interleaved spoken and written language model," *TACL*, 2024.

[19] A. Dubey *et al.*, "The llama 3 herd of models," *ArXiv*, 2024.

[20] Y. Jia, M. Tadmor Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in *LREC*, 2022.

[21] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. rahman Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech*, 2021.

[22] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[23] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On generative spoken language modeling from raw audio," *TACL*, 2021.

[24] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *TASLP*, 2023.

[25] C. Wang, S. Chen, Y. Wu, Z.-H. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *TASLP*, 2023.

[26] Z.-H. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *ArXiv*, 2023.

[27] T. B. Brown *et al.*, "Language models are few-shot learners," *ArXiv*, 2020.

[28] R. Anil *et al.*, "PaLM 2 technical report," *ArXiv*, 2023.

[29] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. L. Seltzer, "Prompting large language models with speech recognition abilities," *ICASSP*, 2023.

[30] M. Wang, W. Han, I. Shafran, Z. Wu, C.-C. Chiu, Y. Cao, Y. Wang, N. Chen, Y. Zhang, H. Soltau, P. K. Rubenstein, L. Zilka, D. Yu, Z. Meng, G. Pundak, N. Siddhartha, J. Schalkwyk, and Y. Wu, "SLM: Bridge the thin gap between speech and text foundation models," *ASRU*, 2023.

[31] S. Maiti, Y. Peng, S. Choi, J. weon Jung, X. Chang, and S. Watanabe, "VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks," *ICASSP*, 2023.

[32] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models," *ArXiv*, 2023.

[33] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *EMNLP*, 2023.

[34] A. Zeng, Z. Du, M. Liu, L. Zhang, S. Jiang, Y. Dong, and J. Tang, "Scaling speech-text pre-training with synthetic interleaved data," *ArXiv*, 2024.

[35] L. Kurzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for german end-to-end speech recognition," in *SPECOM*, 2020.

[36] C. Wang, A. Wu, J. Gu, and J. Pino, "CoVoST 2 and massively multilingual speech translation," in *Interspeech*, 2021.

[37] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018.

[38] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2022.

[39] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," *Interspeech*, 2022.

[40] C.-W. Huang, H. Lu, H. Gong, H. Inaguma, I. Kulikov, R. Mavlyutov, and S. Popuri, "Investigating decoder-only large language models for speech-to-text translation," *Interspeech*, 2024.