说、读与提示：以最少监督实现高保真文本转语音

Eugene Kharitonov Damien Vincent Zalán Borsos Raphaël Marinier Sertan Girgin OlivierPietquin Matt Sharifi Marco Tagliasacchi Neil Zeghidour

Google Research
{kharitonov,damienv,neilz}@google.com

arXiv:2302.03540v1 [cs.SD] 7 Feb 2023

摘要

我们推出了 SPEAR-TTS，这是一种多说话人文本转语音（TTS）系统，可在最少监督下进行训练。通过结合两种离散语音表征，我们将 TTS 设定为由两个序列到序列任务组成：从文本到高层语义标记（类似于“阅读”），以及从语义标记到低层声学标记（“说话”）。将这两项任务解耦后，便可利用丰富的纯音频数据训练“说话”模块；同时，通过高效结合预训练与回译，大幅降低训练“阅读”模块对平行数据的需求。为控制说话人身份，我们采用示例提示，使 SPEAR-TTS 仅凭一段 3 秒的短样本便能泛化到未见过的说话人，无需任何显式的说话人表征或说话人 ID 标签。实验表明，仅使用 15分钟的平行数据，SPEAR-TTS 的字符错误率便可与最先进的方法相媲美；主观测试还显示，其自然度和声学质量均与真实语音相当。

1 引言

训练文本转语音（TTS）系统通常需要数百小时的平行数据，即转录语句。结果是，TTS 仅适用于“高资源”语言。此外，这类系统生成的音频多样性取决于其训练所用的平行数据；这些数据应涵盖许多来自不同人口群体、具有各种口音且录音条件各异的说话人。同时，对于包括低资源语言在内的大多数语言，纯音频语音在网上相对丰富，常见形式包括有声读物、播客、广播和电视节目。

在本文中，我们探讨如何利用纯音频数据，减少对训练文本转语音系统所需的监督。我们提出了 SPEAR-TTS，[1] 这是一种多说话人文本转语音系统，仅需单说话人的15分钟平行数据即可训练。此外，SPEAR-TTS 只需3秒语音，无需任何说话人标签或显式的说话人表征，就能合成新的音色。SPEAR-TTS 的核心是利用近期在口语“无文本”建模方面取得的进展（Lakhotia 等人，2021年；Dunbar 等人，2021年；Polyak 等人，2021年；Kreuk 等人，2021年；Kharitonov 等人，2022年；Nguyen 等人，2022年；Borsos 等人，2022年）。这些方法将连续的音频波形表示为有限词汇表中的词元序列，从而将语音生成转化为语言建模任务。具体而言，AudioLM（Borsos 等人，2022年）结合了两类离散令牌：高层语义标记和可映射为音频的低层声学标记。借助这些表征，我们将文本转语音问题视为从文本转录到声学标记的“翻译”，并以语义标记表征作为中间的“语言”（Utiyama 和 Isahara，2007）。这样一来，文本转语音就被简化为两个序列到序列（seq2seq）任务的组合：将文本翻译为语义标记，再将语义标记翻译为声学标记。

将文本转语音任务拆分为这两个子任务的关键优势在于：学习如何将文本映射为中间语义词元表示（“阅读”），以及如何从中生成语音（“说话”）所需的监督得以解耦。虽然“阅读”阶段依赖文本-音频并行数据，但用于训练“说话”组件的音频标记由自监督音频模型生成，因此可以从海量无标签语音数据中提取。ϑედეგად，生成语音的质量和多样性不再取决于可用的平行数据。

[1] SPEAR 是“**说**、**读**和**提示**”的缩写。

将 SPEAR-TTS 的每个阶段都视为序列到序列问题，使我们能够使用标准的 Transformer模型（Vaswani 等人，2017），并且可以轻松借鉴机器翻译社区提出的大量思想，以减少对监督的需求。具体而言，我们将 BART/T5式预训练（Lewis 等人，2020年；Raffel 等人，2020年）与回译（Sennrich 等人，2016）相结合，从而显著减少训练 SPEAR-TTS 所需的平行监督。

为了控制 SPEAR-TTS 生成话语时所使用的音色，我们采用了一种与文本语言模型中的提示机制密切相关的示例提示机制（Brown 等人，2020年）。在这里，我们用代表目标声音的音频片段对“说话”模型进行条件控制，引导模型在生成话语时使用该声音。对于只有单说话人平行数据可用的语言，这一特性可以简化可控多说话人文本转语音系统的构建。

使用序列到序列模型对语音合成进行建模，可以在推理时采用随机采样，从而针对相同输入生成质量各异的输出。我们利用这一点，通过提出一种基于客观质量指标的采样方案来提升合成音频质量。

我们对英语语音开展的实验研究表明，通过在大型数据集（来自 LibriTTS 的 551小时（Zen 等人，2019年））上结合预训练和反向翻译，并仅使用来自单说话人的 15分钟平行数据（LJSpeech（Ito 和 Johnson，2017）），SPEAR-TTS：（a）生成与输入转录文本高度一致的语音——在 LibriSpeech test-clean 上的字符错误率为 1.92%（Panayotov 等人，2015）；（b）合成具有多样音色的语音；（c）在使用目标说话人的 3秒示例时，能够可靠地复现未见说话人的音色；（d）实现较高的声学质量，与真实语句的质量相当（平均意见分为 4.96，而真实语句为 4.92）。[2]

总体而言，我们通过大规模自监督预训练和离散语音表示的回译来构建文本转语音的方法，与现有文本转语音系统的实现方式有很大不同(Shen 等人，2018年；Kong 等人，2020年；Ren 等人，2020年；Kim 等人，2021年；Ao 等人，2022年；Wang 等人，2023年），显著降低了数据收集成本，并有望为数据覆盖不足的语言提供高质量的多说话人

[2]SPEAR-TTS 生成的样例可在演示网站上查看: https://google-research.github.io/ seanet/speartts/examples/。

文本转语音 day.

2 离散语音表示

下面简要介绍 SPEAR-TTS 所必需的自监督音频表征。AudioLM（Borsos 等人，2022年）提出了这些表征的组合用法，详细讨论请参阅该文。

语义标记 语义标记的作用是提供粗粒度的高层条件，以便随后生成声学标记。因此，它们应能表征语音中的语言内容（从语音学到语义学），同时去除说话人身份和声学细节等副语言信息。为获得这种表征，我们基于 w2v-BERT（Chung 等人，2021年）训练了一个自监督语音表征模型。该模型结合了掩码语言建模（Devlin 等人，2019年）和对比学习（van den Oord 等人，2018年）来获得语音表征。训练完成后，我们对特定层输出进行均值方差归一化，并运行 k均值聚类。我们将质心索引用作离散令牌。

声学标记 声学标记是离散的音频表示，可高保真地重建声学细节。我们训练了 SoundStream（Zeghidour 等人，2021年）神经编解码器，在压缩语音的同时对其进行重建，使其仅由少量离散单元表示。SoundStream 通过在卷积自编码器的瓶颈层加入残差量化器来实现这一目标。为了在序列中表示残差量化器的层级结构，我们将对应不同层级的词元交错排列并展平（Borsos 等人，2022年）。

3 SPEAR-TTS 概述

SPEAR-TTS 扩展了 AudioLM（Borsos 等人，2022年），使文本能够作为一种条件输入。SPEAR-TTS 分为两个主要阶段，如图1所示。在第一阶段（$\mathcal{S}_1$），文本输入会被转换为一系列离散的语义标记。第二阶段（$\mathcal{S}_2$）将语义标记映射为声学标记，再由 SoundStream 解码器（Zeghidour 等人，2021年）将其解码为语音。这样，$\mathcal{S}_1$学习将文本映射到语义标记所提供的内部表征（“阅读”），而$\mathcal{S}_2$ 则负责根据这一中间表征生成语音（“说话”）。

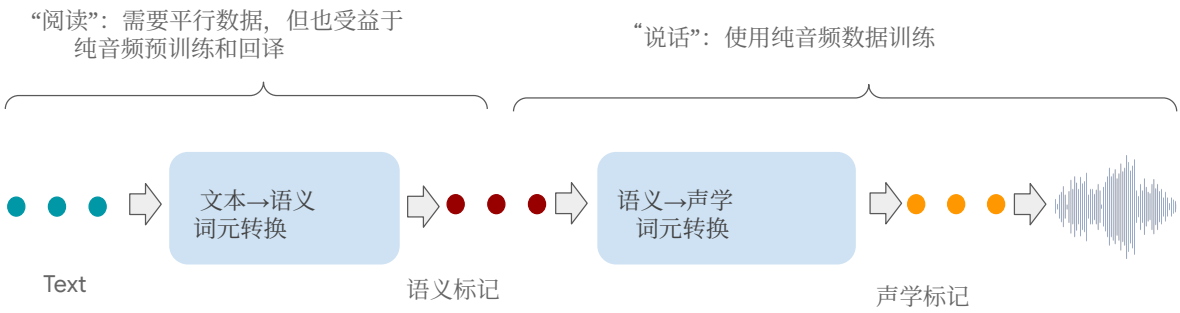

图1：**SPEAR-TTS**。第一阶段 $\mathcal{S}_1$（“阅读”）将词元化文本映射为语义标记。第二阶段$\mathcal{S}_2$（“说话”）将语义标记映射为声学标记。声学标记被解码为音频波形。

通过使用语义标记作为中间表征，我们实现了两个目标。首先，语义标记提供了一种语音表征，其中编码的主要是语音内容，而韵律和说话人信息有限，从而弥合文本与声学标记之间的差距。因此，与声学标记相比，我们的中间表征更接近文本。由此，相较于直接学习文本与声学标记之间的映射，学习从文本转录到语义标记的映射更为容易。其次，由于语义标记和声学标记都源自自监督模型，因此第二阶段 $\mathcal{S}_2$ 可以使用纯音频数据进行训练。事实证明，这对训练 $\mathcal{S}_2$极为有益，因为通常可获取的纯音频数据规模远大于平行数据。[3] 进而，将$\mathcal{S}_1$ 与 $\mathcal{S}_2$ 分离，使我们能够使用作用于语义标记的去噪前置任务对前者进行预训练，进一步利用纯音频数据。

与 AudioLM（Borsos 等人，2022年）类似，可以添加可选的第三阶段，旨在通过预测对应于细粒度残差向量量化级别的声学标记来提高合成语音的质量（附录 A）。

4 $\mathcal{S}_1$：提高监督效率

第一阶段 $\mathcal{S}_1$ 将词元化文本映射为语义标记。我们使用平行的文本与语义标记数据来学习这一映射。我们从一个文本-音频文本转语音数据集入手，并从音频中提取语义标记。因此，$\mathcal{S}_1$ 被简化为一个序列到序列任务，可通过编码器-解码器或仅解码器Transformer架构来实现（Vaswani 等人，2017；Raffel 等人，2020年）。

训练 Transformer 序列到序列模型可能需要大量平行数据，而低资源语言的这类数据可能极其稀缺。

[3]以英语为例，像 LibriTTS 这样的大型数据集包含 580 小时的平行数据（Zen 等人，2019年），而 LibriLight 则包含 60,000 小时的未转录语音（Kahn 等人，2020年）。

下面将讨论两种用于缓解这一局限的方法：目标领域预训练（第 4.1 节）和回译（第 4.2 节）。

4.1 预训练

我们从 BART 和 T5 汲取灵感，并在去噪前置任务上对编码器-解码器Transformer进行预训练（Lewis 等人，2020年；Raffel 等人，2020年）。在此任务中，模型接收一个被破坏的原始语义标记序列，并以生成相应的未破坏词元序列为目标。

典型的损坏方法包括随机替换、删除和掩码，可作用于单个词元或整个词元序列（Raffel 等人，2020年）。在初步研究中，我们发现，以固定概率独立删除单个词元，效果优于其他方法。

在模型 $\mathcal{P}$ 完成去噪任务预训练后，我们对其进行微调以完成 $\mathcal{S}_1$ 任务。为此，我们冻结编码器的上层和解码器的所有参数（解码器与编码器交叉注意力层中使用的参数除外），并更新编码器的下层。要调整的层数是一个超参数。

4.2 回译

同一段文本可以通过多种方式转换为音频，音色、口音、韵律、情感内容和录音条件各不相同。这种一对多关系使文本转语音问题具有高度不对称性——这与文本翻译不同，例如，英语到法语的翻译和反向翻译的难度大致相当。因此，采用回译（Sennrich 等人，2016；Edunov等人，2018年）极具吸引力，也就是说，利用现有的平行数据训练语音转文本模型，并用它从纯音频语料库生成合成平行数据。

SPEAR-TTS 的两阶段架构

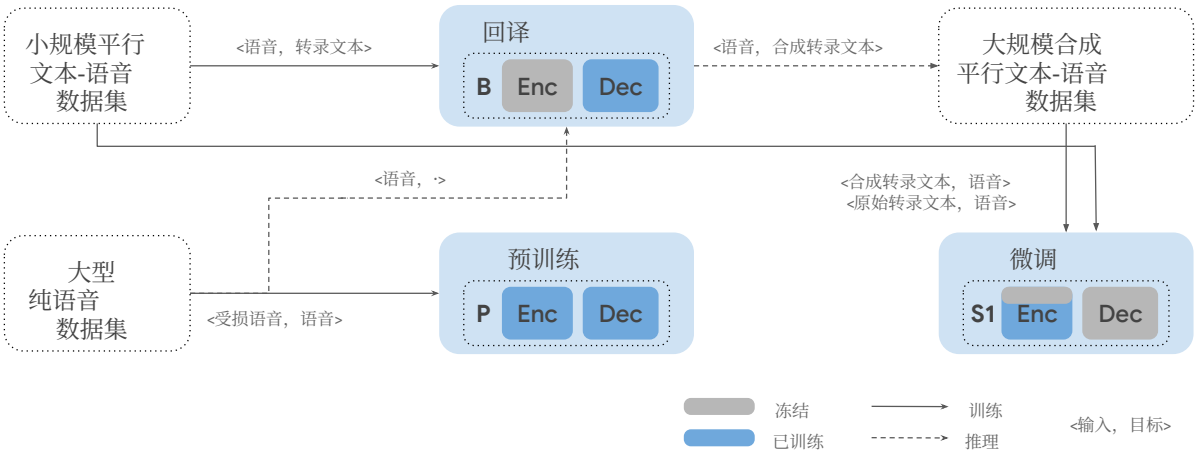

图2：**训练 $\mathcal{S}_1$，结合预训练和回译**。我们首先使用纯语音数据的语义词元表示，在去噪任务上对编码器-解码器模型 $\mathcal{P}$ 进行预训练。接下来，我们使用一个小型平行数据集微调其解码器，使其能够进行回译（将语义标记转换为转录文本）。然后，我们使用该模型转录纯语音数据集，得到一个合成平行数据集。随后，我们使用这个合成数据集以及原始的小型平行数据集，对 $\mathcal{P}$ 的编码器进行微调，使其能够沿正向进行“翻译”（将文本转录转换为语义标记）。

尤其适用于回译，因为它可以实现为语义标记与文本之间的翻译。其优势有两方面：(a) 无需处理原始音频或冗长的声学标记序列，从而降低计算复杂度，[4] 以及 (b) 在训练从语义标记到文本转录的“反向”模型时，能够利用相同的语义词元级预训练（第 4.1 节）。

为了获得回译模型，我们从与上述相同的预训练模型 $\mathcal{P}$ 开始。不过这次我们冻结编码器，只微调解码器。之后，我们使用该模型转录纯音频数据。接下来，我们使用合成平行数据训练文本转语音系统的第一阶段，而该阶段本身也是通过微调另一份 $\mathcal{P}$ 获得的（见第4.1节）。在合成数据上微调后，我们继续在原始平行数据上微调。[5]图2 展示了这一结合预训练与回译的过程。

[4]在我们的设置中，一个话语的声学 token 表示至少 6× 比对应的语义 token 表示更长。[5]另一种选择是像（Edunov等人，2018年）那样，使用合成数据与原始数据的混合进行训练，这会引入混合权重这一额外的超参数。

5 $\mathcal{S}_2$：控制生成过程

第二阶段模型 $\mathcal{S}_2$ 将语义标记映射为声学标记。为了训练这一阶段，我们从纯音频数据集中的每个话语提取语义标记和声学标记序列对。接下来，我们训练一个 Transformer模型，在这两个标记序列之间执行序列到序列翻译。第二阶段生成的话语在音色、语速和录音条件上随机变化，从而再现训练数据中观察到的特征分布。由于 $\mathcal{S}_1$ 和$\mathcal{S}_2$ 的训练相互独立，因此即使在单说话人数据集上训练 $\mathcal{S}_1$ ， $\mathcal{S}_2$ 生成的语音仍保留了这种多样性。

为了控制说话人音色的特征，我们结合了 AudioLM 的两项发现（ Borsos 等人，2022年）：(a) 只要语音前缀仅由语义标记表示，AudioLM 就会在每次生成后续内容时采样一种不同的随机音色；(b) 然而，当条件中也包含声学标记时，AudioLM 在生成后续内容时会保留声学标记所捕捉的语音特征。与 AudioLM 不同，我们在训练

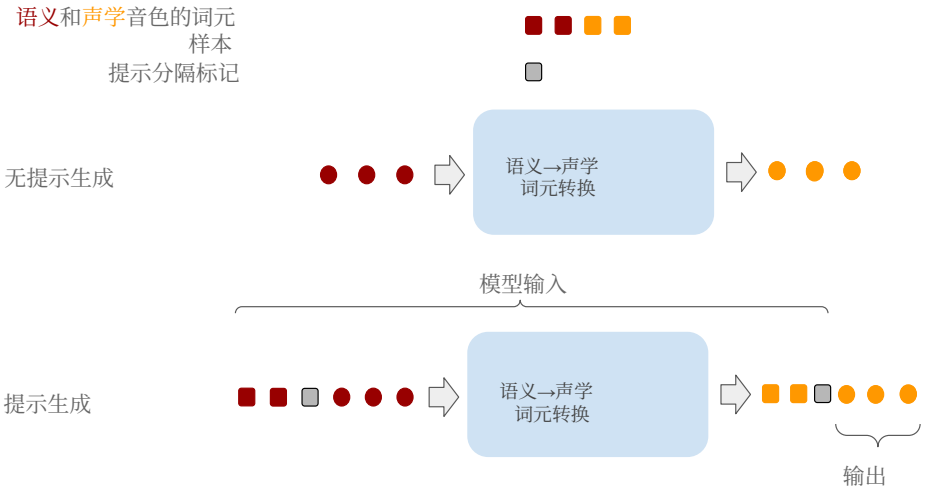

图3：**通过示例提示控制** $\mathcal{S}_2$。对于提示生成，我们按以下顺序连接标记序列：提示中的语义标记、目标中的语义标记，以及提示中的声学标记。随后，模型生成与目标语义标记相对应的声学标记，同时保留提示声学标记中的音色和说话条件。

过程中明确融入了这一能力，如图3所示。在训练过程中，我们从每个训练样本中随机选取两个互不重叠的语音窗口，并据此计算语义标记和声学标记序列。我们将其中一个窗口作为提示，另一个作为目标输出。接下来，按照以下顺序拼接这些序列：(a) 提示中的语义标记，(b) 目标中的语义标记，(c) 提示中的声学标记，以及 (d) 目标中的声学标记。在训练 $\mathcal{S}_2$时，(a)-(c) 用作前缀，模型则学习生成目标声学标记 (d)，同时保留提示中的声学标记所捕捉的说话人身份。在推理时，将 (a)-(c) 作为输入，并自回归地生成 (d)。

需要特别指出的是，我们在每个片段边界处添加了一个特殊的分隔符令牌，以告知模型预期存在的不连续性。这可以防止在不使用分隔符时有时会产生的边界伪影。请注意，不需要提供提示语的文本转录。

$\mathcal{S}_2$ 生成的语音样本可能包含一些背景噪声，因为训练数据中通常存在这类噪声。我们考虑了两种在推理时控制合成语音噪声水平的方法。首先，在提示生成的情况下，可以选择包含更清晰语音的提示。其次，我们可以使用随机采样（例如温度采样），针对同一输入生成多个序列，然后使用无参考音频质量指标，选择包含最少噪声的样本。为此，我们使用了一个类似于 DNSMOS 的平均意见分估计模型（Reddy 等人，2021年）。

6 实验设置

本节介绍实验研究中使用的数据集、指标和基线系统。

6.1 训练和验证数据

声学和语义词元： 我们使用 Libri-Kahn 等人，2020 年）来训练自监督表征模型（SoundStream 和 w2v-BERT），以及用于将 w2v-BERT 嵌入离散化为语义词元的 k。我们使用 LibriLight 最大的 unlab-60k 子集，其中包含约 60,000 小时的英语有声书音频，由 7,000 多位说话人朗读。

第一阶段 $\mathcal{S}_1$**：**为了在低资源条件下进行实验，我们在 $\mathcal{S}_1$ LJSpeech（Ito 和 Johnson，2017）上进行训练。LJSpeech 是一个包含 24 小时平行数据的单说话人数据集。我们仅使用 LJSpeech 作为平行数据来源，同时也表明，即使平行训练数据本身仅包含一位说话人，我们的方法仍能泛化到多位说话人。由于 LJSpeech 未指定标准的训练集/开发集/测试集划分，我们遵循 Liu 等人（2022、2020）的方法，随机选取 300 条话语作为开发集，再选取另外 300 条话语作为测试集（各 30 分钟），其余部分用作训练数据。为模拟可用数据极其有限的情形，我们从训练集中均匀抽取 12、3、2、1 小时以及 30 和 15分钟的子集。作为参考，15分钟子集包含约 21k 个语义标记和 2k 个词语。

预训练：为了在序列破坏任务（第 4.1 节）上预训练模型，我们从 LibriLight 中提取语义标记（Kahn 等人，2020年），因为预训练只需要音频数据。

回译： 在我们的回译实验中，我们使用 LibriTTS（Zen 等人，2019年）作为未标注语音的来源（忽略转录文本）。我们汇总 LibriTTS 的所有训练子集，获得一个包含 551小时语音的纯音频数据集。使用 LibriTTS 作为回译的纯音频数据来源，使我们能够将 SPEAR-TTS 与 $\mathcal{S}_1$ 在原始 LibriTTS 转录文本和回译后的 LibriTTS 转录文本上训练的模型进行比较。

第二阶段 $\mathcal{S}_2$： 为了训练 $\mathcal{S}_2$，我们从 LibriLight（Kahnetal.，2020年）中提取语义标记序列和声学标记序列对。

6.2 评估数据

我们使用 LibriSpeech test-clean（Panayotov 等人，2015年）来测量字符错误率（CER）（见第 6.4节）。由于 LJSpeech 仅包含短于 10 秒的序列，我们从 LibriSpeech test-clean 中筛除时长超过这一限制的序列。最终，我们获得了 2,007 个话语，总时长约为 3 小时。值得注意的是，LibriSpeech test-clean 与我们使用的任何训练数据或验证数据均不存在交集。

6.3 预处理

为准备训练数据，我们展开 LJSpeech 中使用的标准缩写。接着，我们应用 G2p_en 音素化器（Park 和 Kim，2019年）。从其输出中移除词汇重音信息后，我们得到一个由 47 个词元组成的词汇表中的字符串表示（39 个来自 CMU词典的音素、空白字符和标点符号）。

由于在低监督场景中无法假设音素化器普遍可用，因此在附录G 中，我们尝试使用字素输入。

6.4 指标

我们关注 SPEAR-TTS 的以下理想特性：

- 生成语音应符合输入；
- 即使$\mathcal{S}_1$ 使用单说话人数据进行训练，也应具备语音多样性；
- 当输入一段来自未见过的目标说话人的话语时，SPEAR-TTS 应合成与其音色相符的语音；
- 生成语音应具有高质量。

下面我们将讨论用于评估这些属性是否得到满足的指标。

字符错误率（CER）我们使用内部的自动语音识别系统转录 SPEAR-TTS 合成的话语，并通过测量字符错误率（CER）来评估其对输入转录文本的忠实度。我们使用 LibriSpeech test-clean 数据集（Panayotov 等人，2015年）计算 CER，因为与所采用的自动语音识别系统的输出相比，该数据集所需的后处理最少。作为参考，在原始真实音频上，CER 为 0.98%。

语音多样性为了衡量一组合成语音话语中的语音多样性，我们使用一个说话人分类器，为每段话语分配一位说话人，并测量所有话语中检测到的说话人的经验分布熵。我们使用与 Borsos 等人（2022年）相同的说话人分类器。该分类器在 LibriSpeech train-clean-100 和 test-clean 的并集上训练，分别包含 251 位和 40 位说话人，并对 291 个说话人类别进行预测。更多详情请参阅附录D。

语音保留 使用简短话语提示模型时，我们评估提示与生成语音之间说话人音色的一致性。为此，我们使用与上文相同的说话人分类器，并测量从生成语音预测出的说话人标签与从提示预测出的标签相匹配的频率。

质量 我们通过收集平均意见分数（MOS），依靠人工评估来评估 SPEAR-TTS 的感知质量。在此情境下，人工评审员听取单个音频片段，并从差（1）到优秀（5）的量表对其音频质量和语音自然度进行评分。

6.5 基线系统

作为主要基线，我们采用一种专门针对低监督场景训练的系统。具体而言，我们使用 FastSpeech2 的一种改进版本（Ren 等人，2020年），这是一种非自回归模型，使用辅助时长、音高和能量预测器。具体来说，在我们的实验中，我们考虑了Pine 等人(2022 年提出的向低资源场景的适配方案。该模型以文本的音素表示为输入，并预测频谱图，随后使用 HiFi-GAN（Kong 等人，2020年）进行声码器合成。我们

将这一改进版本称为 FastSpeech2-LR。在Pine 等人(2022年)报告的一项主观评估中，使用 1（3）小时平行数据训练的 FastSpeech2-LR，表现与使用 10（24）小时平行数据训练的 Tacotron2（Shen 等人，2018年）开源实现相当。我们使用作者分享的检查点，这些检查点分别采用 LJSpeech 的 15分钟、30分钟、1 小时、3 小时和 24 小时子集进行训练。[6]

我们还将 SPEAR-TTS 与 VALL- E（Wang 等人，2023年）进行了比较。VALL-E 是一款近期推出的文本转语音系统，在零样本语音适应方面展现了最先进的成果。与 SPEAR-TTS 类似，它能够利用 3 秒的语音提示进行语音迁移。VALL-E 将输入文本映射为粗粒度声学标记，并使用非自回归的精细化阶段来预测细粒度声学标记。VALL-E 使用 LibriLight 的经自动语音识别转录版本进行训练（Kahn 等人，2020年），其中包含约 60,000 小时的平行数据。由于该模型尚未公开，我们根据其演示页面提供的样本进行比较。

7 超参数与训练详情

7.1 离散语音表示

我们遵循 AudioLM（Borsos 等人，2022年）的设置来计算语义标记和声学标记，但做了一些调整。语义标记通过量化 w2v-BERT 第 7 层输出的嵌入得到，所用码本大小为 512。因此，1 秒音频由 25 个语义标记表示，词汇表大小为 512，对应的比特率为 $25 \times \log_2 512 = 225$ bit/s。与 Lakhotia 等人 (2021年）Borsos 等人 (2022年）一样，我们会依次移除重复的语义标记。

我们从 SoundStream神经编解码器（Zeghidour 等人，2021年）中提取声学标记。该编解码器具有 3 个量化层级，每层的码本大小为 1024。我们使用包含 3×1024 个唯一词元的词汇表，并将每一帧表示为一个平铺的词元序列，依次交错排列第一、第二和第三量化层的词元。因此，1 秒音频表示为 50 Hz 的 $\times 3 = 150$ 声学标记，对应的比特率为 1500 bit/s。

7.2 第一阶段（$\mathcal{S}_1$）

在所有实验中，我们使用 Adafactor 优化器（Shazeer 和 Stern，2018年），并采用逆平方根学习率衰减。作为正则化方法，我们使用标签平滑，并将平滑参数设为 0.1；但在预训练时，由于可用数据量很大，因此不采用这一方法。

预训练 预训练任务设置为以 0.6 的概率删除单个词元。我们通过网格搜索选择这一参数，并检查 $\mathcal{S}_1$ 微调后的验证准确率。我们采用丢弃概率 0.5，并将批量大小设为 256。预训练进行了 1M 次更新步数，并将所得检查点 $\mathcal{P}$ 用于所有实验。模型建筑采用 T5-Large（Raffel 等人，2020年），这是一个包含 24 层的编码器-解码器序列到序列模型（见附录 F）。

微调 同一个预训练检查点 $\mathcal{P}$会针对不同目的进行微调（图 2）。在所有情况下，我们都会对丢弃率（{0.1, 0.3, 0.5}）和待微调的层数进行网格搜索，并选择验证准确率最高的组合（采用教师强制）。具体而言，在真实值平行数据上进行微调时（作为消融实验），我们冻结编码器的上层和整个解码器，同时更新编码器嵌入和下层的权重。待调整的下层数在{4, 6, 8}中搜索。在合成平行数据上微调时，我们在{4, 6, 8, 10, 12, 24}中搜索编码器待微调的下层数。接下来，我们在原始平行数据上微调编码器的下4层（以避免数据极少时出现过拟合）。最后，为回译而微调解码器时，我们微调 N 顶部和N 底部层，采用 $N \in \{2, 3, 4, 12\}$. 微调期间，我们选择验证准确率最高的检查点。

从头开始训练 作为一项消融实验，我们训练 $\mathcal{S}_1$ 从头开始，并根据可用数据量，尝试不同的 T5 架构变体（Raffel 等人，2020年）。我们采用不对输入序列使用因果掩码的仅解码器模型（Raffel 等人，2020年），在我们的初步实验中取得了更好的结果。我们对以下超参数进行网格搜索：丢弃概率 {0.1, 0.3, 0.5}；架构规模（T5-small 或 T5-base）；层数（T5-small：2、4、6、8；T5-base：4、6、8、12）。更多细节见附录 F。

[6] https://github.com/roedoejet/FastSpeech2_ACL2022_可复现性

7.3 第二阶段（$\mathcal{S}_2$）

对于 $\mathcal{S}_2$，我们使用一个 12 层的仅解码器 Transformer 模型，每层包含 12 个维度为 64 的注意力头，嵌入维度为 768，前馈网络大小为 2048。优化器和学习率调度与 $\mathcal{S}_1$ 相同。

7.4 推理

我们使用束搜索从 $\mathcal{S}_1$ 中采样，并使用温度采样从 $\mathcal{S}_2$中采样。这种组合既能确保忠实于转录文本，又能生成更多样、更自然的语音。我们将束宽设为 10，因为更大的值不会改善字符错误率，却会增加计算成本。生成回译数据时，我们沿用 $\mathcal{S}_1$的设置，不再进行额外的超参数搜索。对于 $\mathcal{S}_2$，我们尝试了采样温度 $T \in \{0.50, 0.55, ..., 0.95, 1.0\}$，并选择使 LJSpeech 验证数据集上的字符错误率最低的 T =0。75。在这种情况下，$\mathcal{S}_1$ 模型使用通过回译获得的合成平行数据进行训练，而回译模型则使用 LJSpeech 的 15 分钟数据划分进行训练。

为了控制合成语音中的噪声水平，我们采用采样技术（第 5 节），对 n_s 与输入对应的音频语句进行采样，并根据类似于 DNSMOS 的无参考音频质量模型（Reddy 等人，2021年）选取质量最高的一条。我们将 n_s 设为 3，以平衡音频质量和计算成本（附录B）。

8 实验

我们从多个维度评估 SPEAR-TTS。首先，我们衡量在不同训练场景以及不同数量的可用平行数据下，生成语音与输入转录文本的忠实度（第 8.1 节）。接着，我们发现，SPEAR-TTS 能够生成比训练期间所用平行数据中的语音更多样的语音（第 8.2 节）。最后，我们展示 SPEAR-TTS 能够成功控制说话人音色，同时不会降低转录文本的忠实度（第8.3 节）。

8.1 可懂度与监督效率

在评估 SPEAR-TTS 时，我们考虑 $\mathcal{S}_1$ 的以下训练设置：（*a*）使用平行数据从头开始训练；（*b*）使用平行数据微调预训练检查点 $\mathcal{P}$；（*c*）微调预训练检查点 $\mathcal{P}$ 以获得回译模型，然后在合成生成的数据上从头开始训练前向模型；（*d*）与（c）相同，但反向模型和前向模型均通过微调 $\mathcal{P}$ 获得，此外还在原始平行数据上对前向模型进行微调。

表1 报告了字符错误率（CER）的主要结果，并以此作为生成语音可懂度的衡量指标。我们观察到，随着平行数据量减少，从头开始训练（a）会导致错误率非常高。相反，得益于预训练（b），SPEAR-TTS 在仅使用 2 小时平行数据时，仍能将 CER 维持在相对较低的水平（$\leq 4\%$）。这一结果与使用 24 小时数据但不进行预训练时达到的 CER 相近。回译（c）总体上有积极影响，尤其是在平行数据量减少时；仅使用 15分钟数据，CER 就达到 2.88%。将回译与预训练结合（d），使用相同数量的平行数据时，CER 进一步降至 2.21 %。这表明，固定解码器有助于应对通过回译获得的合成生成训练数据中存在的噪声。因此，使用 3 小时数据（结合预训练和回译）训练的 SPEAR-TTS，其 CER 与使用 LibriTTS-train 的原始转录文本从头开始训练时的结果相同；后者使用了 551小时平行数据（见附录 C）。

我们还将 SPEAR-TTS 与 FastSpeech2-LR 进行了比较，发现使用 24 小时的平行数据时，两种系统的表现大致相当（FastSpeech2-LR：1.99%，SPEAR-TTS：2.06%）。然而，随着平行数据量减少，FastSpeech2-LR 的字符错误率会迅速上升。因此，当可用数据仅有 15分钟时，两者的差距显著：FastSpeech2-LR 为 4.90%，而 SPEAR-TTS 为 2.21%。

总之，预训练与回译相结合，使 SPEAR-TTS 即使在只有 15分钟平行数据的情况下，也能合成符合输入转录文本的语音。

8.2 语音多样性

SPEAR-TTS 能够以多种音色生成话语，其中包括平行数据中未出现过的说话人。例如，使用 LJSpeech 数据集（Ito 和 Johnson，2017）作为平行数据来源时，尽管该数据集

表 1：根据训练场景以及 LJSpeech 中可用的平行数据量，SPEAR-TTS 与我们的基线系统的<style id='1'>可懂度。我们在 LibriSpeech test-clean 上测量 CER（%，越低越好）。 ±表示通过自助法获得的 95%置信区间。“×”表示会生成不可懂语音的模型。

		SPEAR-TTS			
		训练		回译	
并行训练数据	FastSpeech2-LR	从头开始训练 (a)	预训练 (b)	从头开始（c）	预训练（d）
24 h	$1.99_{\pm 0.20}$	$3.67_{\pm 0.21}$	$2.38_{\pm 0.13}$	$2.26_{\pm 0.14}$	$2.06_{\pm 0.12}$
12 h	-	$4.31_{\pm 0.28}$	$2.54_{\pm 0.14}$	$2.27_{\pm 0.14}$	$2.03_{\pm 0.12}$
3 h	$2.52_{\pm 0.25}$	$20.1_{\pm 0.74}$	$3.07_{\pm 0.15}$	$2.21_{\pm 0.12}$	$2.01_{\pm 0.12}$
2 h	-	$24.7_{\pm 0.71}$	$3.73_{\pm 0.17}$	$2.22_{\pm 0.13}$	$2.09_{\pm 0.12}$
1 h	$2.74_{\pm 0.27}$	×	$5.51_{\pm 0.21}$	$2.23_{\pm 0.13}$	$2.16_{\pm 0.13}$
30 分钟	$3.18_{\pm 0.28}$	×	$21.3_{\pm 0.43}$	$2.52_{\pm 0.15}$	$2.20_{\pm 0.12}$
15 分钟	$4.90_{\pm 0.34}$	×	×	$2.88_{\pm 0.19}$	$2.21_{\pm 0.12}$

表2：**语音多样性（比特）**。我们通过说话人分类器检测到的语音经验分布的熵来衡量语音多样性。

	LJSpeech	LibriTTS		
# 说话人	1	61	123	247
真实值	2.55	5.82	6.71	7.68
SPEAR-TTS	6.11	6.22	6.16	6.28
FastSpeech2-LR	0.66	-	-	-

只有一位说话人，但模型仍能生成多种不同的音色。在以下实验中，我们将对生成语音的音色多样性进行定量测量。

为此，我们在说话人数不同的平行数据集上训练 $\mathcal{S}_1$，并验证合成声音的多样性是否保持稳定。我们考虑 1 位说话人（LJSpeech），以及 61、123 和 247 位说话人（来自 LibriTTS）。具体而言，我们使用完整的 LibriTTS train-clean-100，其中包含 247 位说话人，以及其中两个说话人数约为其 1/2 和 1/4 的子集。我们使用 LibriSpeech test-clean 中的转录文本。

表2 说明，随着说话人数从 247 人减少到 1 人，真实语音的声音变异性自然降低（从 7.68 降至 2.55）。请注意，LJSpeech 的音色对于所使用的说话人分类器而言属于域外，因此测得的声音变异性不为零。相比之下，SPEAR-TTS 的声音变异性不受说话人数显著影响（最小值：6.16，最大值：6.28），且显著高于 FastSpeech2-LR（6.11 对 0.66）。

该实验表明，SPEAR-TTS 合成的语音种类与用于训练的平行数据中包含的不同说话人声音数量无关 $\mathcal{S}_1$。

8.3 提示生成

如第 8.3 节所述，SPEAR-TTS 能够通过示例提示来控制说话人音色。我们在 零样本 场景下评估 SPEAR-TTS。在这种场景中，用于提示的音色在训练期间从未被 $\mathcal{S}_1$ 或 $\mathcal{S}_2$ 见过，而 $\mathcal{S}_2$ 必须从单个提示示例中复现其特征。具体而言，我们固定 $\mathcal{S}_1$，使用在 15 分钟 LJSpeech 数据上训练的模型，并将 LibriSpeech test-clean 中的全部 40 位说话人作为目标说话人。对于每位说话人，我们从同一数据集中随机选取 5 条时长各为 3 秒的语音提示及其转录文本。对于每条语音提示和文本转录，我们重复合成 5 次，并对生成样本的指标取平均值。

表 3报告了说话人准确率，即提示和生成语音中检测到相同音色的频率。我们观察到，Top-1 准确率为 92。4%，这表明提示使 SPEAR-TTS 能够保留说话人的音色。此外，重复进行推理时，合成语音也很稳定，这体现在声音变异性较低（0.41 比特）。此外，我们观察到，使用提示生成时，SPEAR-TTS 的字符错误率为 1.92%，低于不使用提示时的 2.21%。我们认为，这一改进源于提示采用了更清晰的录音，从而引导 $\mathcal{S}_2$ 模型生成更清晰的语音，进而减少自动语音识别错误。

我们还比较了 SPEAR-TTS 与 VALL-E（Wang 等人，2023年）在音色保留能力方面的表现。按照 Wang 等人（2023年）的方法，我们使用一个基于 WavLM（Chen 等人，

表 3：**提示生成中的语音保留（分类器准确率）**。该 $\mathcal{S}_1$ 模型使用 15 分钟的平行数据进行训练。

字符错误率（%）	说话人准确率（%） Top-1	Top-3	语音多样性（比特）
1.92	92.4	98.1	0.41

表 4：**与基线比较语音保留情况（余弦相似度）**。YourTTS 和 VALL-E 的结果摘自（Wang 等人，2023年，表 2）。

模型	并行训练数据	余弦相似度
YourTTS	~ 600 小时	0.34
VALL-E	60,000 小时	0.58
SPEAR-TTS	15分钟	0.56

2022年）。[7] 这是 Wang 等人 （2023年）所使用的同一模型，因此我们的测量结果可以直接与其论文中报告的分数进行比较。从表 4所报告的结果来看，我们观察到 SPEAR-TTS 显著优于 YourTTS（Casanova 等人，2022年）（0.56 对 0.34），并且说话人相似度几乎达到 VALL-E（0.58）的水平，尽管其训练所使用的平行数据少了 240，000× 。

9 主观评估

最终，我们借助人工评审员进行主观测试，以比较 SPEAR-TTS 与基线系统以及真实自然语音的质量。我们聚焦于监督最少的场景，并使用 $\mathcal{S}_1$ 采用 15分钟 LJSpeech（Ito 和 Johnson，2017）子集训练的模型。作为基线系统，我们使用在 LJSpeech 的 15分钟、1小时和 24小时子集上训练的 FastSpeech2-LR 模型（Ren 等人，2020年；Pine 等人，2022年）。

为确保评估句子不属于 SPEAR-TTS 或 FastSpeech2-LR 模型的训练集，我们从 2022年发布的一段有声读物章节中提取句子，该章节由与 LJSpeech 中相同的音色朗读。[8] 该章节的出版时间晚于我们使用的所有数据集。我们从中提取了 20 个句子，每个句子的时长在 3 到 11 秒之间，总计 133 秒。我们从对应书籍的文本中获取这些句子的转录文本。[9] 我们提供转录文本见表 12 附录。

基线系统是经过训练、能够生成单一音色的文本转语音系统。为确保公平比较，我们使用从 LJSpeech 数据集中提取的话语作为提示输入 $\mathcal{S}_2$，使 SPEAR-TTS 生成具有相同音色的语音。为此，我们从 LJSpeech 中随机选取 3 秒的语音样本，并滤除静音超过 1 秒的样本，将剩余样本用作提示。

我们逐一向评审者播放样本，并请他们按从差（1）到优秀（5）的量表评估音频质量和语音自然度。开始前，评审者会收到各评分等级对应的话语示例。每个音频样本由 20 位评审者评估。对于每种处理条件，我们对所有评分取平均值，计算平均意见分（MOS）。

表5 报告了主观测试的结果。我们发现，即使基线系统在训练期间使用了更多平行数据，SPEAR-TTS 的音频质量仍显著高于基线系统。SPEAR-TTS 的平均意见分（4.96）与真实语音的平均意见分（4.92）相当，这证实了生成语音具有很高的质量，尽管该模型仅使用 15 分钟的平行数据进行训练。

我们还使用演示页面上提供的示例，通过一项小规模主观测试，对 SPEAR-TTS 和 VALL-E（Wang 等人，2023年）进行了比较。[10]这些示例由 8 条转录文本分别搭配 3 条提示生成，共得到 24 个语音话语。我们使用上文所述的同一 SPEAR-TTS 实例（采用 15 分钟的单说话人 LJSpeech 进行训练），用相同的转录文本和提示合成 24 个话语，并按照上文所述的同一方案开展主观测试。表6 显示，在这些示例中，尽管 SPEAR-TTS 使用的监督数据少得多（15 分钟的平行数据，由 1 位说话人录制，而 VALL-E 使用了约 60,000 小时的平行数据，由 7,000 多位说话人录制），但其自然度显著更高，语音质量也更好（平均意见分为 4.75，而 VALL-E 为 3.35）。

10 相关工作

10.1 离散化语音处理

Lakhotia 等人（2021年）在生成式口语建模（生成式口语语言模型）方面的工作，开创了在离散化语音表征上使用语言模型的方法。主要任务包括 Lakhotia 等人（2021年）

[7] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification# 预训练模型，“WavLM large”模型。[8] https://librivox.org/ predecessors-of-cleopatra-by-leigh-north/，§10。[9] https://www.gutenberg.org/cache/epub/58236/pg58236.txt

[10] https://valle-demo.github.io/，“更多示例”。

表 5：**平均意见分（MOS）评估**。所有对比系统均在 LJSpeech（Ito 和 Johnson，2017）的子集上训练。 ±表示通过自助法获得的 95% 置信区间。

并行训练数据	FastSpeech2-LR 15 分钟	FastSpeech2-LR 1 h	FastSpeech2-LR 24 h	SPEAR-TTS 15 分钟	真实值 -
MOS	$1.72_{\pm 0.04}$	$2.08_{\pm 0.04}$	$2.11_{\pm 0.04}$	$\mathbf{4.96}_{\pm 0.02}$	$4.92_{\pm 0.04}$

表 6：**提示生成的平均意见分（MOS）评估**。两个系统的提示以及 VALL-E 的样本均取自 VALL-E 的演示页面。 ±表示通过自助法获得的 95%置信区间。

系统	VALL-E	SPEAR-TTS（15 分钟）
MOS	$3.35_{\pm 0.12}$	$\mathbf{4.75}_{\pm 0.06}$

关注不受约束的语音生成和语音续接。他们的工作为一系列应用和扩展奠定了基础，包括情感迁移（Kreuk 等人，2021年）、韵律（Kharitonov 等人，2022年）以及对话（Nguyen 等人，2022年）建模。SPEAR-TTS 与 AudioLM（Borsos 等人，2022年）相关。AudioLM 是该研究方向近期的一项进展，在口语建模方面实现了更优的质量，同时也具备很高的音频质量。

10.2 低资源与半监督文本转语音

能够利用纯音频数据是 SPEAR-TTS 的一项鲜明特点。Guided-TTS 是另一种具备这一能力的文本转语音系统，由 Kim 等人 (2021年) 提出。Guided-TTS 的核心是结合了 (a) 一种去噪扩散概率模型（DDPM），用于学习对纯音频数据进行建模；以及 (b) 一个音素分类器，用于引导生成过程，使其生成具有指定转录文本的话语。Guided-TTS 2（Kim 等人，2022年）在 Guided-TTS 的基础上进行了扩展，支持通过微调或以零样本方式实现说话人适应性，方法是使用经过专用说话人嵌入模块处理的 10 秒语音样本。另一个基于 DDPM、支持适应性的文本转语音系统由Levkovitch 等人 (2022年) 提出，该系统使用分类器引导机制，以零样本方式引导生成过程，使其生成特定音色的语音。

与 SPEAR-TTS 相比，上述方法依赖更强的监督：（a）Guided-TTS 使用在 LibriSpeech 960 上训练的音素分类器；（b）Guided-TTS 2 依赖预训练的说话人验证系统。相比之下，SPEAR-TTS 使用一种直观且无需参数的提示机制，无需任何说话人标签。

Liu 等人 (2020年) 将序列自编码器与向量量化和时间分割机制结合起来，以学习类似音素的离散语音表示，并使用序列到序列模型将这些表示映射到音素。与 SPEAR-TTS 类似，该系统几乎无需监督即可训练；然而，它只能生成单说话人的语音，质量也远低于真实音频（2.33 对比 4.81 ，见其实验）。这与 SPEAR-TTS 不同：尽管只接受极少的单说话人监督，它仍能生成任意音色的语音，同时达到真实语音的质量。

接下来，还有一类研究利用未配对文本的可用性。像 SPEAR-TTS 所做的那样，对纯音频数据进行反向翻译，可以理解为使用自动语音识别系统为文本转语音生成训练数据。 Tjandra 等人(2017) 提出同时训练自动语音识别和文本转语音模型：文本转语音模型根据自动语音识别的输出重构波形，而自动语音识别模型则识别由文本转语音模型合成的音频。 Chung 等人 (2019年) 讨论了一系列 Tacotron 文本转语音系统的预训练方法，包括逐帧自回归解码器预训练，以及为编码器预训练词嵌入。 Ao 等人 (2022年) 提出了 SpeechT5，这是一种可结合文本和纯音频数据进行预训练的系统。

10.3 提示式音频生成

当一句话前面加上用通俗英语表达的情感提示时，例如 [我真的很难过，] Tortoise TTS（Betker，2022年）会以悲伤的音色合成文本。

AudioLM（Borsos 等人，2022年）展示了语音提示能力：声学令牌前缀会促使模型在生成语音续接时，保持提示中的说话人特征和录音条件。我们通过提出 $\mathcal{S}_2$的提示感知训练，扩展了 AudioLM 的提示能力。

Wang 等人（2023年）提出了 VALL-E，这是一种文本转语音系统，支持对合成语音的音色和情感进行基于提示的条件控制。与 SPEAR-TTS 的两阶段架构不同，VALL-E 直接根据文本的音素表示预测相当于声学标记的内容。因此，它需要提示的转录文本；例如，提示音频嘈杂时，这可能会带来困难。这与 SPEAR-TTS 不同，后者仅使用自监督音频词元提示模型，因此不需要相应的转录文本。另一个区别在于所用平行训练数据的规模：VALL-E 使用经自动语音识别转录的 LibriLight 的 60,000 小时数据进行训练（Kahn 等人，2020年）。第8.3 节和9 节表明，即使仅使用 15分钟的平行数据进行训练，SPEAR-TTS 也能提供相近的零样本提示能力，并实现更高的音频质量。

11 结论与未来工作

在这项工作中，我们介绍了 SPEAR-TTS，这是一种具有两项突出特点的多说话人文本转语音系统。首先，该系统训练所需的平行数据极少：只需使用来自单说话人的 15分钟平行数据进行训练，就能合成高保真且具有丰富语音多样性的语音。其次，SPEAR-TTS 能够利用一段 3 秒长的音色示例，合成保留此前未见说话人语音特征的语音。

这些能力源于对丰富纯音频数据的利用。能够解锁此类数据用途的关键组件，是语音的分层离散表示，它将高层语义标记与低层声学标记相结合。借助这些表示，SPEAR-TTS 将文本转语音问题拆解为两个序列到序列任务：“阅读”（从词元化文本到语义标记）和“说话”（从语义标记到声学标记）。

SPEAR-TTS 通过三种方式使用纯音频数据：(a) 训练“说话”模型，使困难的语音生成任务受益于大规模数据；(b) 将其作为预训练领域，训练出的模型随后用作文本到语义标记模型和语义词元到文本模型的基础；(c) 生成用于回译的合成平行数据。

我们对英语数据进行的实验研究（第8节）表明，将 LibriTTS 的纯音频数据（Zen 等人，2019年）与从 LJSpeech（Ito 和 Johnson，2017）中采样的 15 分钟平行数据结合后，SPEAR-TTS 的可懂度可媲美在 LJSpeech 全部 24 小时数据上训练的适配版 FastSpeech2-LR（在 LibriSpeech test-clean 上的字符错误率分别为 1.92 % 和 1.99%（Panayotov 等人，2015））。与此同时，即使仅使用单说话人的平行数据进行训练，SPEAR-TTS 也能合成音色多样的语音（第 8.2节）。

接下来，第 8.3 节中的实验表明，SPEAR-TTS 能以零样本方式准确保留此前未见说话人的语音特征。事实上，我们的测量结果显示，从 LibriSpeech test-clean 中取一段 3 秒长的说话人语音示例后，SPEAR-TTS 在合成留出的文本转录时，根据说话人分类器的评估，保留该音色的准确率可达 92.4%。此外，在测量提示与生成语音之间的说话人相似度时，SPEAR-TTS 的余弦相似度为 0.56，接近 VALL-E（Wang 等人，2023年）报告的分数，且显著高于 YourTTS（Casanova 等人，2022年）（分别为 0.58 和 0.34）。

语音自然度的主观评估表明，即使使用 96× 更少的平行数据进行训练，SPEAR-TTS 的质量仍显著高于强大的单语音基线（平均意见分为 4.96 对 2.11）。此外，SPEAR-TTS 的平均意见分与自然语音相当（4.92）。在比较零样本语音转换任务中合成语音的质量时，SPEAR-TTS 的平均意见分显著高于 VALL-E（4.75 对 3.35），且使用的数据少了 240,000×。

我们相信，我们在有限监督（无论数量还是质量方面）下实现高质量文本转语音的工作，为那些因使用“低资源”语言和方言而被排除在外的社群提供文本转语音技术铺平了道路。SPEAR-TTS 带来的另一项令人振奋的潜在应用，是让有言语障碍的人使用自己过去的音色录音进行口头交流。与此同时，我们也承认，初步研究存在某些局限，也可能遭到滥用（第 12 和13节）。

我们认为，将研究发现应用于为真正的低资源语言构建文本转语音系统，并进一步减少对监督的需求，是后续工作的主要方向。

12 局限性

虽然我们的目标是让低资源语言能够实现高质量、多样化且可控的文本转语音，但我们从英语入手开展研究，这使我们得以利用一系列经过充分研究的数据集来解决这一问题。

接下来，我们使用 LibriLight（Kahn 等人，2020年）作为纯音频数据集，它提供了足够多样的音频。不过，LibriLight 仅包含采样率为 16 kHz 的音频样本，因此 SPEAR-TTS 需要额外步骤才能合成采样率更高的语音（附录 A）。此外，平均而言，LibriLight 中的音频质量低于经过筛选的数据集。不过，这些并非 SPEAR-TTS 的局限性，而是我们所使用数据的局限性。此外，只需更换用于生成声学标记的神经编解码器，便可提升 SPEAR-TTS 的质量，而无需更改 $\mathcal{S}_1$ 和 $\mathcal{S}_2$。

最后，SPEAR-TTS 的灵活性源于采用相对大型的 Transformer 模型，而这些模型在训练和推理时需要大量计算资源。我们认为，可以通过模型蒸馏和量化分别解决这一问题（Polino等人，2018年；Fan等人，2020年）。

13 更广泛的影响

我们认为，我们在高质量文本转语音方面的工作所需的监督极少（无论数量还是质量），可以成为一块重要的垫脚石，让目前因使用“低资源”语言（即缺乏训练深度学习模型所需的大型平行语料库的语言）而未能从文本转语音解决方案中受益的社群也能使用这项核心语音技术。即使对于英语等高资源语言，利用未转录数据生成语音，也能生成现有文本转语音系统尚未覆盖的口音和方言语音。SPEAR-TTS 另一个令人振奋的潜在应用，是让有言语障碍的人使用自己音色的录音来提示 SPEAR-TTS。

同时，我们也承认，模仿音色可能被用于多种恶意用途，包括绕过生物特征识别以及冒充他人（Delgado等人，2021年；Casanova 等人，2022年）。因此，采取防范措施以避免滥用至关重要。作为第一步，我们验证了分类器能够可靠地检测 SPEAR-TTS 生成的语音，在平衡数据集上的准确率为 82.5%（见附录 E）。未来可以探索其他检测合成语音的方法，例如音频水印。

致谢

作者感谢 Ron Weiss 和 Matthieu Geist 对本文初稿提出的反馈。我们也感谢 Aidan Pine 帮助我们获取并运行来自Pine 等人(2022年）。

参考文献

Junyi Ao、Rui Wang、Long Zhou、Chengyi Wang、Shuo Ren、Yu Wu、Shujie Liu、Tom Ko、Qing Li、Yu Zhang 等。2022年。SpeechT5：用于口语语言处理的统一模态编码器-解码器预训练。发表于*TACL*。

James Betker。2022年。TorToiSe 文本转语音。

Zalán Borsos、Raphaël Marinier、Damien Vincent、Eu- gene Kharitonov、Olivier Pietquin、Matt Sharifi、OlivierTeboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour。2022年。AudioLM：一种用于音频生成的语言建模方法。*arXiv* 预印本 *arXiv:2209.03143*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared DKaplan、PrafullaDhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。*NeurIPS*。

Edresson Casanova、Julian Weber、Christopher D Shulby、Arnaldo Candido Junior、Eren Gölge 和 Moacir A Ponti。2022年。YourTTS：迈向人人可用的零样本多说话人文本转语音和零样本语音转换。发表于*ICASSP*。

Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yanmin Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei。2022年。Wavlm：面向全栈语音处理的大规模自监督预训练。*IEEE J.Sel.Top.Signal Process.*，16(6):1505–1518。

Yu-An Chung、Yuxuan Wang、Wei-Ning Hsu、Yu Zhang 和 RJ Skerry-Ryan。2019年。通过半监督训练提升端到端语音合成的数据效率。发表于*ICASSP*。

Yu-An Chung、Yu Zhang、Wei Han、Chung-Cheng Chiu、James Qin、Ruoming Pang 和 Yonghui Wu。2021年。W2V-Bert：结合对比学习与掩码语言建模的自监督语音预训练。发表于*ASRU*。

Héctor Delgado、Nicholas Evans、Tomi Kinnunen、Kong Aik Lee、Xuechen Liu、Andreas Nautsch、Jose Patino、Md Sahidullah、Massimiliano Todisco、Xin Wang 等。2021年。Asvspoof 2021：自动说话人验证欺骗攻击与对策挑战赛评估方案。*arXiv* 预印本 *arXiv:2109.00535*。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。发表于*NAACL-HLT*。

Ewan Dunbar、Mathieu Bernard、Nicolas Hamilakis、TuNguyen、MaureendeSeyssel、PatriciaRozé、Morgane Rivière、Eugene Kharitonov 和 Emmanuel Dupoux。2021年。2021年零资源语音挑战赛：口语语言建模。*IEEEPAMI*。

Sergey Edunov、Myle Ott、Michael Auli 和 David Grangier。2018年。深入理解大规模回译。发表于*EMNLP*。

AngelaFan、PierreStock、BenjaminGraham、Edouard Grave、Rémi Gribonval、Herve Jegou 和 Armand Joulin。2020年。通过量化噪声训练实现极致模型压缩。*arXiv*预印本 *arXiv:2004.07320*。

Sergey Ioffe 和 Christian Szegedy。2015年。批归一化：通过减少内部协变量偏移加速深度网络训练。载于第*32*届国际机器学习会议论文集（*ICML 2015*），法国里尔，*2015*年*7*月*6*日至*11*日，第37卷机器学习研究期刊研讨会与会议论文集，第448–456页。JMLR.org。

KeithIto 和 Linda Johnson。2017年。LJ语音数据集。https://keithito.com/LJ-Speech-Dataset/。

Jacob Kahn、Morgane Rivière、Weiyi Zheng、Evgeny Kharitonov、Qiantong Xu、Pierre-Emmanuel Mazaré、Julien Karadayi、Vitaliy Liptchinsky、Ronan Collobert、Christian Fuegen 等。2020年。Libri-Light：一项针对有限或无监督条件下自动语音识别的基准。在*IEEE ICASSP*。

Eugene Kharitonov、Ann Lee、AdamPolyak、Yossi Adi、Jade Copet、Kushal Lakhotia、Tu Anh Nguyen、Morgane Riviere、Abdelrahman Mohamed、Emmanuel Dupoux 和 Wei-Ning Hsu。2022年。无文本、韵律感知的生成式口语模型。在*ACL*。

Heeseung Kim、Sungwon Kim 和 Sungroh Yoon。2021年。Guided-TTS：利用未转录语音进行文本转语音。*arXiv*预印本 *arXiv:2111.11755*。

Sungwon Kim、Heeseung Kim 和 Sungroh Yoon。2022年。Guided-TTS 2：一种利用未转录数据实现高质量自适应文本转语音的扩散模型。*arXiv* 预印本 *arXiv:2205.15370*。

Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。2020年。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。神经信息处理系统进展（*NeurIPS*，33：17022–17033。Felix Kreuk、Adam Polyak、Jade Copet、Eugene Kharitonov、Tu-Anh Nguyen、Morgane Rivière、Wei-Ning Hsu、Abdelrahman Mohamed、Emmanuel Dupoux 和 Yossi Adi。2021 年。使用分解式离散表征进行无文本语音情感转换。*arXiv* 预印本 *arXiv:2111.07402*。Kushal Lakhotia、Eugene Kharitonov、Wei-Ning Hsu、Yossi Adi、Adam Polyak、Benjamin Bolte、Tu-Anh Nguyen、Jade Copet、Alexei Baevski、Abdelrahman Mohamed 等。2021年。关于基于原始音频的生成式口语语言建模。*TACL*。Alon Levkovitch、Eliya Nachmani 和 Lior Wolf。2022年。用于去噪扩散文本转语音模型的零样本语音条件控制。*arXiv* 预印本 *arXiv:2206.02246*。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：面向自然语言生成、翻译和理解的序列到序列去噪预训练。在*ACL*。

Alexander H Liu、Cheng-I Jeff Lai、Wei-Ning Hsu、Michael Auli、Alexei Baevski 和 James Glass。2022年。简单而有效的无监督语音合成。*arXiv*预印本 *arXiv:2204.02524*。

Alexander H Liu、Tao Tu、Hung-yi Lee 和 Lin-shan Lee。2020年。通过量化语音表示学习实现无监督语音识别和语音合成。发表于*ICASSP*。

Tu Anh Nguyen、Eugene Kharitonov、Jade Copet、Yossi Adi、Wei-Ning Hsu、Ali Elkahky、Paden Tomasello、Robin Algayres、Benoit Sagot、Abdelrahman Mohamed 等。2022年。生成式口语对话语言建模。*arXiv* 预印本 *arXiv:2203.16502*。

Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015年。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。发表于*ICASSP*。

Kyubyong Park 和 Jongseok Kim。2019年。g2p。https://github.com/Kyubyong/g2p。

Aidan Pine、Dan Wells、Nathan Brinklow、Patrick Lit- tell 和 Korin Richmond。2022年。语言复兴中低资源语音合成的需求与动机。载于*ACL*。

Antonio Polino、Razvan Pascanu 和 Dan Alistarh。2018年。通过蒸馏和量化实现模型压缩。*arXiv* 预印本 *arXiv:1802.05668*。

Adam Polyak、Yossi Adi、Jade Copet、Eugene Kharitonov、Kushal Lakhotia、Wei-Ning Hsu、Ab- delrahman Mohamed 和 Emmanuel Dupoux。2021年。基于离散解耦自监督表征的语音重合成。*arXiv* 预印本 *arXiv:2104.00355*。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊，21。

Chandan KA Reddy、Vishak Gopal 和 Ross Cutler。2021年。DNSMOS：一种用于评估噪声抑制器的非侵入式感知客观语音质量指标。载于*ICASSP*。

Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。2020年。FastSpeech 2：快速且高质量的端到端文本转语音。载于*ICLR*。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年。利用单语数据改进神经机器翻译模型。载于*ACL*。Noam Shazeer 和 Mitchell Stern。2018年。Adafactor：具有次线性内存开销的自适应学习率。载于 *ICML*。Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan 等。2018年。通过在梅尔频谱图预测结果上进行条件建模的 WaveNet 实现自然语音合成。载于*ICASSP*。Andros Tjandra、Sakriani Sakti 和 Satoshi Nakamura。2017。边听边说：基于深度学习的语音链。载于*ASRU*。Masao Utiyama 和 Hitoshi Isahara。2007年。基于短语的统计机器翻译中枢轴方法的比较。载于*NAACL*。Aäron van den Oord、Yazhe Li 和 Oriol Vinyals。2018年。利用对比预测编码进行表征学习。*arXiv:1807.03748*。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。你所需要的只是注意力机制。*NIPS*。Chengyi Wang、Sanyuan Chen、Yu Wu、Ziqiang Zhang、Long Zhou、Shujie Liu、Zhuo Chen、Yanqing Liu、Huaming Wang、Jinyu Li 等。2023年。神经编解码器语言模型是零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2301.02111*。Neil Zeghidour、Alejandro Luebs、Ahmed Omran、Jan Skoglund 和 Marco Tagliasacchi。2021年。SoundStream：端到端神经音频编解码器。*IEEE/ACM* 音频、语音与语言处理汇刊，30。

Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J Weiss、YeJia、Zhifeng Chen 和 Yonghui Wu。2019年。LibriTTS：一个源自 LibriSpeech 的文本转语音语料库。*arXiv* 预印本 *arXiv:1904.02882*。

# 样本数, n_s	1	2	3	5	10
CER	$2.10_{\pm 0.07}$	$1.99_{\pm 0.07}$	$1.93_{\pm 0.07}$	$1.90_{\pm 0.06}$	$1.90_{\pm 0.07}$
音频质量	$3.68_{\pm 0.44}$	$3.86_{\pm 0.31}$	$3.94_{\pm 0.26}$	$4.02_{\pm 0.21}$	$4.11_{\pm 0.18}$

表 7：**通过重采样控制音频质量及其对字符错误率的影响**。我们测量 SPEAR-TTS 生成的话语的保真度和质量，并考察其与采样步数的关系。字符错误率是在 LibriSpeech dev-clean 上计算的，音频质量则通过对 DNSMOS 模型（Reddy 等人，2021年）的改进版本在平均意见分量表上测量。 ± 表示一个均值标准误（SEM）。

带宽扩展：从 16 kHz 到 24 kHz

虽然将 LibriLight 作为我们的非配对数据集，使我们能够在 $\mathcal{S}_2$ 中对多样化的说话人和条件进行建模，但该数据集仅包含 16 kHz 音频，而在许多文本转语音应用中，24 kHz 音频更为理想。我们通过带宽扩展提供了一种简单的方法，使 SPEAR-TTS 能够生成 24 kHz 的语音，同时仍能受益于 LibriLight 的多样性。

我们将带宽扩展视为序列到序列任务，即将 SoundStream 编解码器在 16 kHz 下生成的词元（第 7.1 节）映射为 SoundStream 编解码器在 24 kHz 下生成的词元。我们使用 4 个残差向量量化层、每层 1024 个码本条目以及 50 Hz 的嵌入速率，在 LibriTTS（Zen 等人，2019年）上训练后者，从而得到一个 2000 bit/s 的编解码器。为该任务创建训练数据时，我们从 LibriTTS 中提取 SoundStream 词元序列对：目标词元由 24 kHz 编解码器对目标音频样本进行编码生成，输入词元则由 16 kHz 编解码器对音频样本进行编码生成；编码前，先对音频样本应用截止频率在 5 至 8 kHz 之间随机变化的低通滤波器。

由于带宽扩展的序列到序列建模方式很容易融入我们的框架，我们使用 T5-small 编码器-解码器来完成这项任务。需要指出的是，这一阶段的训练数据比 $\mathcal{S}_2$ 小两个数量级（LibriTTS 与 LibriLight），因此采用这种方法，我们可以同时受益于大型低分辨率数据集的声学多样性，以及小型高分辨率数据集的质量。

B 通过采样控制音频质量

如第5节所述，在没有提示的情况下，SPEAR-TTS 生成的音频质量与用于训练 $\mathcal{S}_2$ 的训练数据相当。由于 LibriLight 数据集（Kahn 等人，2020年）包含由志愿者使用个人设备录制的有声读物，因此录音质量差异很大。

在此，我们验证第 5 节提出的采样技术能否让我们控制生成语音的质量，并研究其如何影响所用自动语音识别系统的识别错误。在此实验中，对于 LibriSpeech dev-clean 中的每个音素输入，我们从 SPEAR-TTS 中采样 n_s 次（$n_s \in \{1, 2, 5, 10\}$），并选取平均意见分估计值最高的样本；该估计值由经过修改的 DNSMOS 模型给出（Reddy 等人，2021年）。我们使用选定的样本计算字符错误率。

我们在表7中报告了该实验的结果。我们观察到，提高 n_s 会带来更高的估计质量。此外，更高的音频质量使 SPEAR-TTS 能够实现更低的字符错误率。根据表7中的结果，我们在所有实验中使用 $n_s = 3$，以在计算复杂度和估计质量之间取得平衡。

在 LibriTTS 上进行 C 训练 $\mathcal{S}_1$

在本节中，我们研究使用 $\mathcal{S}_1$ 在 LibriTTS 上训练的 SPEAR-TTS 的可懂度。我们通常采用与 LJSpeech（Ito 和 Johnson，2017）实验中相同的超参数网格，这些实验已在第7节中报告。然而，由于 LibriTTS 比 LJSpeech 更大，我们也对编码器-解码器模型进行了实验。对于 LibriTTS 最大的训练子集（551 小时），我们还使用了具有 T5-Large 规模的编码器-解码器架构和仅解码器架构（24 层）进行实验。对于编码器-解码器模型，我们始终将编码器和解码器中的层数设为相同。

表 8 报告了 SPEAR-TTS 两种变体的字符错误率：一种是 $\mathcal{S}_1$ 从头开始训练，另一种是从预训练检查点 $\mathcal{P}$开始训练，与主要实验中使用的方式相同。我们考察了 Li- briTTS（Zen 等人，2019年）的三个不同规模子集（54、241 和 551小时）。首先，我们注意到，在最大的子集（551 小时）上，SPEAR-TTS 的错误率低至 2.04%；在这种情况下，预训练几乎没有带来提升。然而，配对数据越少，预训练就越重要：当有 241 小时配对数据时，预训练开始发挥作用；而在仅使用 54 小时配对数据进行训练时，预训练则带来显著收益（字符错误率为 2.61%，而非 2.13%）。

数据集大小	从头开始训练	预训练
551 小时	2.04	2.01
241 小时	2.08	1.92
54 h	2.61	2.13

表 8：**在 LibriTTS 上训练时，SPEAR-TTS 在 LibriSpeech test-clean 上的字符错误率**。我们根据训练方案评估 SPEAR-TTS 的保真度。

D 说话人分类器

我们使用的说话人分类器与 Borsos 等人（2022年）相同。该分类器是一个以对数梅尔频谱图为输入的卷积网络。频谱图的窗口大小为 25 毫秒，帧移为 10 毫秒，包含 64 个梅尔频带。该网络包含 6 个模块，每个模块依次进行卷积，使用的卷积核大小为 3x1 和 1x3。每个模块之后接 ReLU 非线性激活和批归一化（Ioffe 和 Szegedy，2015）。各模块的通道数为 [64, 128, 256, 256,512, 512]。分类器的输入跨度为 1 秒；为对更长的话语进行分类，我们以 250 毫秒的帧移运行滑动窗口，并对各窗口的预测结果取平均。

E 检测合成语音

在本节中，我们将展示 SPEAR-TTS 生成的语音能够与真实人类语音成功区分。为此，我们使用了一个经过训练、用于检测 AudioLM 生成语音的分类器（Borsos 等人，2022年）。该分类器采用与说话人分类器（附录 D）相同的架构，并经过训练，可将经过 SoundStream（Panayotov 等人，2015）压缩的 LibriSpeech train-clean-100 示例与 AudioLM 生成的语音区分开来（Zeghidour 等人，2021年）。

为了评估该分类器在 SPEAR-TTS 合成的语音上的效果，我们遍历 LibriSpeech dev-clean 中的样本，并为每个样本生成两段话语：(a) 使用 SPEAR-TTS 将文本合成为语音；(b) 通过声学标记重新合成真实音频。[11] 我们发现，在这组样本上，该分类器区分生成语音与自然语音的准确率达到 82.5%。我们认为，通过直接使用 SPEAR-TTS 的输出训练分类器，可以进一步提升这一结果

[11]我们不会将其与未压缩的真实音频进行比较，因为这项任务对分类器来说过于简单：分类器可以借此专注于表面的编码伪影，从而更容易被绕过。

	嵌入维度	前馈网络维度	注意力头维度	注意力头数
T5-small	256	512	64	6
T5-base	768	2048	64	12
T5-Large	1024	2816	64	16

表 9：**架构详情**。我们报告了 T5-small、T5-base 和 T5-Large 层的详细信息。所使用的层数通过网格搜索确定（参见第 7节）。

	并行训练数据						
	24 h	12 h	3 h	2 h	1 h	30 分钟	15 分钟
音素	2.06	2.03	2.01	2.09	2.16	2.20	2.21
字素	1.79	1.79	2.13	2.27	2.46	2.71	3.45

表 10：**使用基于字素的文本表示时 SPEAR-TTS 的字符错误率（%）**。SPEAR-TTS 使用 LJSpeech 的 15分钟子集作为平行数据，并通过预训练和回译进行训练。

通过直接使用 SPEAR-TTS 的输出训练分类器。

F 架构详情

我们报告了表9中所用 Transformer 层的参数。

G 评估基于字素的 SPEAR-TTS

正文中，我们报告了一个 SPEAR-TTS 变体的评估结果，该变体使用文本的音素表示进行训练。在某些情况下，尤其是对于低资源语言，可能没有可用的音素化器。因此，我们通过评估使用转录文本的基于字素的表示进行训练的 SPEAR-TTS，来补充实验研究。结果见表 10。将这些结果与表1进行比较后，我们发现，当平行数据极少时，基于音素的表示能带来显著优势（例如，使用 15 分钟数据时，字符错误率为 3.45%，而非 2.21%）。相比之下，当平行数据超过 2 小时时，使用音素化器的优势会缩小；当数据达到 12 小时或更多时，基于字素的训练效果优于基于音素的模型。

H 数据规模对 $\mathcal{S}_2$的影响

在本实验中，我们测量 $\mathcal{S}_2$ 对用于训练它的数据量的敏感程度。为此，在训练 $\mathcal{S}_2$ 模型之前，我们将 LibriLight（Kahn 等人，2020年）分别按 1 倍、2 倍、5 倍和 10 倍进行下采样。所有模型都采用相同的建筑，并使用相同的更新步数进行训练，我们

下采样因子	1	2	5	10
字符错误率 (%)	1.99	1.99	2.36	2.92

表 11：**SPEAR-TTS 在 LibriSpeech 开发集清洁语音上的字符错误率与 $\mathcal{S}_2$ 训练数据规模**。我们测量训练前对 LibriLight（Kahn 等人，2020年）进行下采样 $\mathcal{S}_2$ 对字符错误率（%）的影响。

选择验证准确率最高的检查点。接下来，我们将选定的检查点与 $\mathcal{S}_1$ 在 LibriTTS（Zen 等人，2019年）（经过预训练）上训练的模型结合起来，并测量 SPEAR-文本转语音在 LibriSpeech dev-clean 上的可懂度。我们在表11中报告结果。我们注意到，将数据量减少 5 倍开始影响性能。

他的统治年代可能是从他首次与姐姐或继母共同掌握王权时开始计算的。他和前任一样，对建筑颇感兴趣，修建并扩建了神庙，而且在扩建部分展现出独到的品味。其筹备工作耗费了许多年。显然，她没有继承母亲的特征，而且可能并未活很长时间。还有一尊他晚年时的跪姿雕像，手中拿着一个球形花瓶。我们对她的了解比对几乎任何一位王后都少，只知道她延续了王室血脉，而她的名字似乎只留下了简短的记载。它位于两只伸展的爪子之间，其中一只爪子上发现刻有国王的名字。梦似乎在这个家族的历史中发挥了特殊作用。因此，这位年轻的国王被视为某种意义上的神之子。事实上，他能娶到妻子正是得益于这位神，因为他是在一次狩猎之行中遇见她并爱上她的。通常，男人被绘成棕色，女人被绘成黄色，但偶尔也有例外。或许折中的说法最接近事实。埃及雕塑家的名字鲜少留存，不过此例是个例外。石材呈黄褐色，很难加工。一位访客说，雕像的表面原本打磨得十分光滑。这些以石材雕刻而成的非凡作品出自一位艺术家之手。据一些权威人士称，人们曾在二百二十年间听到这些声音。她比丈夫多活了许多年；她丈夫的统治时期很短，不超过八九年。据说他性情和善、慷慨大方，对母亲和妻子都十分尊重、深情厚爱。他似乎是最令人喜爱的埃及国王之一。

表 12：用于主观评估的转录文本。