arXiv:2602.01717v1 [cs.CL] 2 Feb 2026

BBPE16：用于改进多语言语音识别的基于 UTF-16 的字节级字节对编码

Hyunsik Kim、*Haeri Kim*、*Munhak Lee*、*Kyungmin Lee*

三星研究院{hyunsik777.kim, haeri.kim, mun-hak.lee, k.m.lee}@samsung.com

摘要

多语言自动语音识别需要能够高效覆盖多种书写系统的分词。采用 UTF-8 的字节级 BPE（BBPE）因其与语言无关的设计以及对 Unicode 的完整覆盖而被广泛采用，但其变长编码会使中文、日文和韩语（中日韩）等非拉丁文字的词元序列膨胀。更长的序列会增加计算负载和内存使用。我们提出 BBPE16，这是一种基于 UTF-16 的 BBPE 分词器，它以统一的 2 字节代码单元表示大多数现代文字。BBPE16 保留了 BBPE 与语言无关的特性，同时显著改善了跨语言词元共享。在单语、双语和三语自动语音识别中，以及在多语言持续学习设置下，BBPE16 取得了相当或更优的准确率；对于中文，它将词元数量最多减少 10.4%，并将解码迭代次数最多降低 10.3%。这些减少加快了微调与推理，并降低了内存使用量，使 BBPE16 成为多语言自动语音识别的一种实用分词选择。

索引词— 字节级 BPE、UTF-16、分词、语音识别、多语言自动语音识别

1. 引言

随着自动语音识别（ASR）的发展，关注重点已从单语言系统转向多语言自动语音识别。目标不仅是提高准确率 [1] ，还要扩大语言覆盖范围——从数十种 [2] 到某些情况下的数百种语言 [3, 4]。为实现如此广泛的覆盖，模型必须具备能够表示多种文字系统的输出词汇表；因此，分词成为一项首要的设计考量。

对于单语自动语音识别，字符级或子词级字节对编码（BPE） [5]通常已足够；而多语言场景通常采用字节级 BPE（BBPE），以确保完整覆盖 Unicode 并避免针对特定语言的预处理。当代 BBPE 实现主要基于 UTF-8 [6], ，这带来了诸如兼容 ASCII 和鲁棒性等实际优势。然而，UTF-8 在跨文字体系场景中也会带来一些问题。其变长编码会导致不同语言的词元边界不均匀；对于中文、日文和韩语等非拉丁文字（中日韩），还会增加序列长度。序列变长会增加计算量和内存用量，进而提高模型和运行时的复杂度。这些影响使高效、可扩展的多语言自动语音识别更具挑战性，因此有必要探索更适合跨语言场景的替代 BBPE 设计。

*贡献均等。

在实际部署中，多语言自动语音识别系统最初会在大规模多语言语料上进行预训练，随后进行持续学习——即在保留既有知识的同时，逐步适应新领域、方言或代表性不足的语言——以维持或提升其在特定目标场景中的性能。在这类场景中，模型必须纳入新数据，同时避免灾难性遗忘此前学到的语言。在此背景下，多语言自动语音识别的持续学习已成为一个活跃的研究领域 [7, 8]。这些持续学习场景对分词器设计提出了严格要求：分词器应促进高效的跨语言共享，并在逐步加入新语言或领域时保持稳定。

在本研究中，我们提出 BBPE16，这是一种基于 UTF-16 的 BBPE 分词器，旨在解决基于 UTF-8 的字节级字节对编码的局限，同时保持与现有基础设施的兼容性。对于基本多文种平面（BMP）中的绝大多数字符而言，UTF-16 提供统一的 2 字节表示；该平面涵盖了大多数现代文字，包括中日韩字符。值得注意的是，在我们的三语设置中，BBPE16 展现出显著更强的跨语言词元共享能力，能够生成英语、韩语和中文共用的词元，而 BBPE 则无法生成任何此类词元。我们的主要贡献如下：

- 一种新型的基于 UTF-16 的 BBPE 分词器，在实现与现有方案相当的自动语音识别性能的同时，显著提升了词元效率和跨语言词元共享能力
- 在多语言持续学习场景中，中文词元数量最多可减少 10.4%，解码迭代次数也随之减少，从而切实提升训练和推理速度

2. 背景

2.1. 字节级 BPE（BBPE）

现代基于 Transformer [9] 的模型（例如 BERT [10], 、GPT-3 [11]）使用数千亿个词元进行训练，参数量达数百亿。在如此规模下，词级词汇表会面临严重的数据稀疏问题和词汇表规模膨胀问题。BPE [5] 和 SentencePiece [12] 等子词方法能够减少稀疏性，同时涵盖广泛的形态变化。然而，经典 BPE 仍需要针对特定语言进行预处理，例如使用词分词器和处理空白字符。在处理多语言语料库，或包含大量低频字符的语料库时，维护特定语言的预处理规则十分繁琐。

BBPE 首先将输入文本转换为字节序列，然后对该序列应用字节对编码。这使分词器不受语言限制：只要文本采用 UTF-8 编码，任何 Unicode 字符都可以通过同一流程处理。BBPE 有若干

版权所有 2026 IEEE。发表于 ICASSP 2026——2026 年 IEEE国际声学、语音与信号处理会议（ICASSP），会议计划于 2026 年 5 月 4 日至 8 日在西班牙巴塞罗那举行。允许个人使用本材料。但如需为广告或宣传目的转载/再版本材料，或将其用于创作新合集作品以供转售或分发至服务器或列表，或在其他作品中重复使用本作品中任何受版权保护的部分，必须事先获得 IEEE 的许可。联系地址：版权与许可经理 / IEEE 服务中心 / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA。电话： +国际区号 908-562-3966。

优于传统字节对编码的优势：在多语言场景中，它能在整个语料库中更均匀地分配词元用量，并产生一组更多语言共享的词元 [13]。此外，由于 BBPE 的词汇表涵盖所有可能的字节值，因此它从根本上消除了未登录词。

2.2. 基于UTF-8的字节级字节对编码的局限性

尽管UTF-8凭借与ASCII的向后兼容性和容错性，已成为事实上的文本编码标准，但它也给BBPE带来了若干挑战。

变长编码：字符根据其Unicode码点的不同，需要占用1至4个字节，从而导致分词模式不一致。为缓解变长编码特性所导致的、在构建英汉双语数据的BBPE时出现的差异， [14] 引入了长度惩罚和字母表惩罚。

易出现解码错误：尽管UTF-8通过拒绝无效字节序列实现了严格验证，但这也使其更容易出现解码错误。为解决这些解码错误，一些研究采用了动态规划算法 [6, 13],，另一些研究则探索了基于向量量化的解决方案 [15]。

尽管已有研究探索了不同方法来缓解基于UTF-8的字节级字节对编码的局限性，但本文通过简单的修改，引入了一种能够解决这些缺陷的新型分词器。

3. BBPE16：基于 UTF-16 的分词 N

3.1. 基于 UTF-16 的 BBPE 的动机

UTF-16 使用 16 位（2 字节）码元作为基本单位。BMP 中的字符（范围为 U+0000-U+FFFF）由单个码元（2 字节）表示。BMP 包含大多数现代文字系统，例如：（1）拉丁字符、（2）中日韩统一表意文字、（3）谚文音节，以及（4）阿拉伯文、希伯来文和天城文等主要文字系统。因此，每个字符都能以与语言无关且统一的 2 字节形式表示，而 UTF-8 则为每个字符采用长度可变的字节编码。

表1。韩文谚文音节'한'的 UTF-8 与 UTF-16（小端序）编码比较

示例：한(U+D55C)			
编码	**字节 1**	**字节 2**	**字节 3**
UTF-8	**1110**1101 (ED)	**10**010101 (95)	**10**011100 (9C)
UTF-16	01011100 (5C)	11010101 (D5)	-

表1比较了非拉丁文字（重点是中日韩字符）的 UTF-8 和 UTF-16 编码，并以韩文谚文音节"한"作为代表性示例。本研究假设采用小端格式的 UTF-16 编码，因此忽略字节顺序标记（BOM）。粗体和带下划线的位（例如，**1110**）表示 UTF-8 编码规范规定的前缀位，普通位则表示数据位。在 UTF-8 中，使用多个位作为表示长度的前缀，而在 UTF-16 中，所有位都用于编码字符数据。由于 UTF-16 无需长度前缀位的开销，基于 UTF-16 的分词方案也能比基于 UTF-8 的 BBPE 更紧凑、高效地构建词汇表。

3.2. 基于UTF-16的BBPE

我们提出了 BBPE16，这是一种基于 UTF-16 而非传统 UTF-8 编码运行的 BBPE分词器。BBPE16 遵循标准的 BPE合并算法，但所有合并操作都在由 UTF-16代码单元生成的字节序列上进行。BBPE16 的处理流程概述如下：

1. **编码：**将输入文本从 UTF-8 转换为 UTF-16（小端序）
2. **字节提取：**获取原始 UTF-16 字节并丢弃字节顺序标记
3. **BPE训练：**针对 UTF-16 字节序列学习合并规则 4. **分词：**将学习到的合并规则应用于新的 UTF-16 编码文本 5. **解码：**从词元重建 UTF-16 文本，并转换回 UTF-8 文本

如第 3.1 节所述，我们采用 UTF-16 小端序编码，因此在预处理步骤中会丢弃字节顺序标记。输入和输出仍采用 UTF-8 编码，以确保与现有系统兼容。只有内部的分词过程使用 UTF-16，因此 BBPE16 是基于 UTF-8 的 BBPE 的直接替换方案。

4. 实验设置

4.1. 数据集

我们在多种语言设置下评估了 BBPE16：英语使用 LibriSpeech [16]，韩语使用 KsponSpeech [17]，中文使用 AISHELL-1 [18]。在持续学习实验中，英语使用华尔街日报（WSJ） [19]，韩语使用 Zeroth-Korean [20]，中文使用 Common Voice [21] 中文数据集。

对于 LibriSpeech 语料库，我们采用速度扰动（系数为 0.9×、1.0× 和 1.1×），将原有的 281,241 条训练语句扩增至共 843,723 条。速度扰动仅用于英语单语训练，不用于双语或三语设置。文本使用 ESPnet [22] 流程进行规范化，仅保留必要的词元和撇号。对于 KsponSpeech，使用发音标注，并将英语字母转换为韩语音标。对于 Zeroth-Korean，我们将训练数据随机划分为 90% 的训练集和 10% 的开发集。对于 Common Voice 中文，嵌入的英语字母均转换为大写。所有数据集均经过筛选，剔除时长超过 30 秒或规范化文本为空的话语。这一筛选从 KsponSpeech 中移除了 496 条语句，从 WSJ 中移除了 9 条。

4.2. 模型架构与推理设置

我们使用 ESPnet [22] 训练了采用 E-Branchformer [23] 编码器的基于注意力的编码器-解码器（AED）模型。编码器由 17 个模块组成，输出维度为 512，包含 1,024 个线性单元和 8 个采用相对位置编码的注意力头。解码器使用 6 个 Transformer [9] 模块，包含 2,048 个线性单元和 8 个注意力头。所有模型采用相同的架构，仅分词策略不同。除第 5.3 节所述的持续学习模型外，所有模型均训练了 80 个周期。持续学习模型以三语模型为起点，训练了 30 个周期。解码时使用束搜索，束宽为 4。

4.3. 分词器配置

我们比较了三种词元化方法：

- **字节对编码：** 标准字符级BPE
- **BBPE：** 基于 UTF-8 编码的字节级 BPE
- **BBPE16：** 我们提出的基于UTF-16的BBPE

5. 结果与分析

5.1. 单语和双语场景

5.1.1. 自动语音识别性能

为验证 BBPE16 不会降低自动语音识别性能，我们使用多种分词器训练并评估了单语英语、单语韩语和双语（英语与韩语）自动语音识别模型，并比较了由此得到的识别准确率。

表 2。单语与双语场景下不同分词器的词错误率（%）对比

语言	词表 Size	测试集	BPE	BBPE	BBPE16
英语	1000	干净测试集	2.1	2.2	2.1
		其他测试集	4.8	4.7	4.6
韩语	3000	干净评估集	18.5	18.7	18.6
		其他评估集	21.5	21.8	22.0
双语（英语和韩语）	5000	干净测试集	2.5	2.7	2.6
		其他测试集	5.8	6.1	6.0
		干净评估集	19.0	18.9	19.1
		其他评估集	22.1	22.6	22.2

表 2 展示了不同设置下词错误率（WER）的对比。在单语和双语设置中，BBPE16 在任何测试集上的表现与字节对编码和 BBPE 相比均无显著差异。

5.2. 三语场景

为了了解 BBPE16 的跨语言能力和分词器效率，我们比较了在合并数据集（LibriSpeech、KsponSpeech 和 AISHELL-1）上训练的三语分词器。三语分词器的词表大小设为 7,000。

5.2.1. SharedToken 分析

我们分析了使用三语分词器时英语、韩语和中文之间的词元共享模式。表 3 展示了 BBPE 和 BBPE16 在各语言对之间共享的共同词元数量。

表 3. 三语分词器中各语言对之间的共享词元数量

语言对	BBPE	BBPE16
英-韩	0	42
韩中	95	573
中英	0	55
三种语言	0	42

结果表明，BBPE16 在实现跨语言词元兼容性方面具有显著优势。值得注意的是，对于英-韩语言对、中英语言对以及三种语言合并的情况，BBPE 均未产生共享词元，而 BBPE16 则产生了在上述三种情况下，每种情况至少都有 42 个共享词元[1]。BBPE16 显著提升了韩中词元共享水平（95→573），并在其他语言对中也取得了显著进展，展现了其在多语言分词方面的有效性。

5.2.2. 词元数量分析

使用相同的三语分词器，我们评估了各语言之间的压缩效率。表4展示了详细的词元统计数据。

表4。使用三语分词器时各语言每条话语的平均词元数量，最后一列显示了BBPE16相较于BBPE的相对减少量

语言	BPE	BBPE	BBPE16	BBPE16 与 BBPE
英语	76.5	45.4	45.2	-0.4%
韩语	23.5	16.5	16.3	-1.2%
中文	22.3	19.5	18.6	-4.6%

这项分析揭示了 BBPE16 的优势：与 BBPE 相比，英语略有改善（-0.4%），韩语有所改善（-1.2%），中文则显著提升（-4.6%）。值得注意的是，字节对编码将大部分有限的合并词元预算用于覆盖完整的韩语和中文字符范围，因此留给英语的词元相对较少。因此，字节对编码处理英语文本时会产生大量词元，但在韩语和中文方面仍具有竞争力。BBPE16 的高效性直接带来了更快的训练速度和更低的内存使用量。

5.2.3. 词汇覆盖率分析

除了词元效率之外，词汇覆盖率（某种语言使用的词汇表占总词汇表的比例）也是衡量分词器效能的一项关键指标。较高的词汇覆盖率可减少词元槽位的浪费，从而更高效地利用词元空间，并在多语言环境中更充分地利用模型的嵌入参数。表5展示了各三语分词器在不同语言中使用的词汇表百分比。

表5. 三语分词器在各语言中的词汇覆盖率（%）

语言	BPE	BBPE	BBPE16
英语	4.5	44.1	48.1
韩语	35.0	39.9	43.0
中文	60.5	14.7	17.8

BBPE16 在所有语言中都始终实现了更高的词汇覆盖率，与 BBPE 相比，英语、韩语和中文分别提高了 4.0%、3.1% 和 3.1%。BBPE16 更高的词汇覆盖率表明，其词元库存能够在三种语言中得到更高效的利用。

在三语实验中，BPE 下英语词汇的使用比例明显低于韩语和中文。这是因为英语包含的不同字符远少于韩语和中文，而词元预算却必须容纳韩语和中文规模大得多、种类也丰富得多的字符集。

如第 5.2.1 节所述，BBPE16 的跨语言共享使各语言聚合覆盖率超过了 100%。相比之

[1]字节序列”¯G ¯AW”和”¯G ¯AL”是跨语言共享词元的示例，它们出现在英语、韩语和中文中。

表 6. 三语和持续学习场景下 BBPE 与 BBPE16 的性能比较，英语和韩语的词错误率（%），以及中文的字错误率（%）

数据集	三语		持续学习	
	BBPE	BBPE16	BBPE	BBPE16
基础数据集				
LibriSpeech				
干净测试集	2.7	2.6	2.6	2.5
其他测试集	6.1	6.1	5.8	5.8
KsponSpeech				
干净评估集	18.7	19.0	18.7	18.7
其他评估集	22.2	22.4	22.0	21.9
AISHELL-1	5.9	5.7	5.6	5.6
其他数据集				
WSJ	10.7	10.8	4.8	4.2
Zeroth	76.0	47.7	7.6	7.5
CVC	245.7	273.9	15.6	15.6

下，BBPE 中许多预分配字节词元在训练期间从未被选用，因此其聚合覆盖率仍低于 100%。

5.3. 持续学习场景

我们不仅在三语设置中评估了 BBPE16 的有效性，还使用三语分词器在持续学习场景中进行了评估。这些实验表明，在纳入新数据时，BBPE16 具有明显优势。

我们进一步在持续学习设置中评估了 BBPE16：在基础语料库（LibriSpeech、KsponSpeech 和 AISHELL-1）以及另外三个数据集——WSJ（英语）、Zeroth-Korean（韩语）和 Common Voice 中文（中文）上，对三语模型进行微调。

对于这一场景，我们在实验中排除了标准 BPE分词器，因为它处理未登录词的方式使其不适用于这一特定设置。

注意：从此处起，Zeroth-Korean 简称为 Zeroth，Common Voice 中文简称为 CVC。

5.3.1. 自动语音识别性能

表 6 展示了持续学习场景下自动语音识别性能的比较。表 6 中，英语和韩语数据集的指标为词错误率，中文数据集的指标为字错误率（CER）。与三语基线相比，持续学习显著提升了所有新增数据集上的性能。总体而言，BBPE16 的表现与 BBPE 相当：在 CVC 上与 BBPE 持平，同时在 WSJ 和 Zeroth 上取得了小幅提升。

5.3.2. 词元数量分析

表7展示了持续学习场景中每个新增数据集的词元数量对比。BBPE16在所有语言中均表现出一致的优势：英语没有变化，韩语数据集有所改进，而中文数据集则优势显著；与BBPE相比，在CVC上实现了10.4%的词元数量减少。中文数据集的这一显著减少，在对中文进行持续学习时带来了计算成本节省。

表7。持续学习场景中每增加一个数据集时每条话语的平均词元数量，最后一列显示 BBPE16 相较于 BBPE 的相对减少量

数据集	BBPE	BBPE16	BBPE16 与 BBPE 对比
WSJ	28.7	28.7	0.0%
Zeroth	37.0	36.7	-0.8%
CVC	28.9	25.9	-10.4%

表8。持续学习场景下每个测试集的平均话语解码迭代次数，最后一列显示 BBPE16 相较于 BBPE 的相对减少量

测试集	BBPE	BBPE16	BBPE16 与 BBPE 的对比
WSJ	27.3	27.2	-0.4%
Zeroth	36.8	36.5	-0.8%
CVC	27.3	24.5	-10.3%

5.3.3. 推理效率

表8报告了每个话语的平均解码迭代次数，该指标定义为在束搜索下输出序列结束标记所需的解码步骤数量。与训练词元数量的结果一致，BBPE16在所有语言中都减少了解码迭代次数：英语（-0.4%）和韩语（-0.8%）的改进幅度较小，而中文则实现了显著的效率提升（-10.3%）。这表明，BBPE16在所有语言中都能带来一致的效率改进，尤其是在中文上，无论训练还是推理阶段，收益都十分显著。

6. 讨论

在许多实际场景中，训练语料库以大规模英语数据为主，而中日韩等语言的数据量相对较少。因此，通常会用额外数据扩充训练集，以提升非英语语言的性能。在此类情况下，BBPE16 相较于传统BBPE具有优势，它通过跨语言词元共享和2字节编码实现更高效的分词。此外，BBPE16与现有BBPE保持兼容，因为它仅将文本编码中的UTF-8替换为UTF-16。除语音识别之外，这一特性也使BBPE16适用于更广泛的语言建模任务，包括大语言模型。

7. 结论

我们提出了 BBPE16，这是一种基于 UTF-16 的 BBPE 分词器，克服了 BBPE 在多语言场景中的关键局限。通过利用 UTF-16 统一的代码单元表示，BBPE16 在多语言持续学习场景中最多可减少 10.4% 的词元数量，从而缓解多语言自动语音识别的效率瓶颈。

BBPE16 并未牺牲英语词元效率，而是提升了所有语言的分词效果，实现了更好的跨语言词元共享，并降低了计算成本。

随着多语言自动语音识别日益重要，BBPE16 提供了一种实用且高效的解决方案，尤其有利于中日韩等非拉丁文字，同时在各种多语言场景中保持了具有竞争力或更优的性能。

8. 参考文献

[1] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, “采用单一端到端模型的多语言语音识别”，载于*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2018 年，第 4904–4908 页。[2] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, “大规模多语言自动语音识别：50 种语言、1 个模型、10 亿个参数”，载于*Interspeech 2020*，2020 年，第 4751–4755 页。[3] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Stroh man, B. Ramabhadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu, “Google USM：将自动语音识别扩展至 100*01037* 种语言之外”，2023*2303arXiv* 预印本 *arXiv*：。。[4] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, “将语音技术扩展至 1，000+ 种语言”，机器学习研究期刊，第 25 卷，第 1 期，2024 年 1 月。[5] R. Sennrich, B. Haddow, and A. Birch, “利用子词单元对罕见词进行神经机器翻译”，载于计算语言学协会第 *54* 届年会论文集（第 *1* 卷：长篇论文）。德国柏林：计算语言学协会，2016 年 8 月，第 1715–1725 页。

[6] B. Li, Y. Zhang, T. Sainath, Y. Wu, 和 W. Chan，“字节是你所需要的一切：使用字节进行端到端多语言语音识别与合成”，载于*ICASSP 2019 - 2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2019年，第5621–5625页。

[7] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, 以及 M. Ra- vanelli，“CL-MASR：面向多语言自动语音识别的持续学习基准”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 32 卷，第 4931–4944 页，2024 年 10 月。

[8] C. Y. Kwok、J. Q. Yip 和 E. S. Chng，《多语言自动语音识别系统自回归解码器的持续学习优化》，载于 *Interspeech 2024*，2024年，第 1225–1229 页。[9] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin，《注意力机制就是你所需要的一切》，载于 神经信息处理系统进展，第 30 卷。Curran Associates 公司，2017年。[10] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于 *2019*年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文）。明尼阿波利斯，明尼苏达州：计算语言学协会，2019年 6 月，第 4171–4186 页。[11] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、S. Agarwal、A. Herbert-Voss、G. Krueger、T. Henighan、R. Child、A. Ramesh、D. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever 和 D. Amodei，《语言模型是少样本学习器》，载于 神经信息处理系统进展，第 33 卷。Curran Associates 公司，2020 年，第 1877–1901 页。[12] T. Kudo 和 J. Richardson，《SentencePiece：一种用于神经文本处理的简单、语言无关的子词分词器和反分词器》，载于 *2018*年自然语言处理实证方法会议论文集：系统演示。布鲁塞尔，比利时：计算语言学协会，2018年 11 月，第 66–71 页。[13] C. Wang、K. Cho 和 J. Gu，《使用字节级子词的神经机器翻译》，*AAAI*人工智能会议论文集，第 34 卷，第 05 期，第 9154–9160 页，2020年 4 月。[14] L. Deng、R. Hsiao 和 A. Ghoshal，《使用字节级子词的双语端到端自动语音识别》，载于*ICASSP 2022 - 2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年，第 6417–6421 页。[15] R. Hsiao、L. Deng、E. McDermott、R. Travadi 和 X. Zhuang，《优化端到端自动语音识别的字节级表示》，载于 *2024 IEEE*口语语言技术研讨会（*SLT*），2024年，第 462–467 页。[16] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《Librispeech：基于公版有声读物的自动语音识别语料库》，载于 *2015 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2015年，第 5206–5210 页。

[17] J.-U. Bang, S. Yun, S.-H. Kim, M.-Y. Choi, M.-K. Lee, Y.-J. Kim, D.-H. Kim, J. Park, Y.-J. Lee, 和 S.-H. Kim，“Kspon-Speech：用于自动语音识别的韩语自发语音语料库”，应用科学，第 10 卷，第 19 期，2020 年。

[18] H. Bu, J. Du, X. Na, B. Wu, 和 H. Zheng，“AISHELL- 1：一个开源的普通话语音语料库和语音识别基线”，载于*2017* 年东方分会国际语音数据库与语音 *I/O* 系统及评估协调委员会会议（*O-COCOSDA*），2017 年，第 1–5 页。

[19] D. B. Paul 和 J. M. Baker，“基于《华尔街日报》的连续语音识别语料库的设计”，载于《语音与自然语言：*1992* 年*2*月*23-26*日在纽约州哈里曼举行的研讨会论文集》，1992 年。

[20] L. Jo 和 W. Lee，“Zeroth-Korean”，2019 年。[在线]。可访问：https://www.openslr.org/40/[21] R. Ardila、M. Branson、K. Davis、M. Kohler、J. Meyer、M. Henretty、R. Morais、L. Saunders、F. Tyers 和 G. Weber，“Common Voice：一个大规模多语言语音语料库”，载于 第十二届语言资源与评估会议论文集。法国马赛：欧洲语言资源协会，2020年5月，第 4218–4222 页。

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala 和 T. Ochiai，“ESPnet：端到端语音处理工具包”，载于 *Interspeech 2018*，2018 年，第 2207–2211 页。

[23] K. Kim、F. Wu、Y. Peng、J. Pan、P. Sridhar、K. J. Han 和 S. Watanabe，《E-Branchformer：采用增强型融合的 Branchformer 语音识别模型》，载于 *2022* 年 *IEEE* 口语技术研讨会（*SLT*），2023 年，第 84–91 页。