通过同音字扩展和统一书写提升低资源粤语语音识别中的罕见词识别

HoLam Chung[1,3,†], *Junan Li*[1,2], *Pengfei Liu*[1,*], *Wai-Kim Leung*[1], *Xixin Wu*[1,2], *Helen Meng*[1,2]

[1]感知与交互智能中心（CPII），香港[2]香港中文大学，香港[3]国立台湾大学电机工程研究所，台湾holam.chung@protonmail.com，{jali,pfliu,williamleung,wuxx,hmmeng}@cpii.hk

arXiv:2302.00836v1 [cs.CL] 2 Feb 2023

摘要

同音字在普通话和粤语等声调音节语言中十分常见。在低资源环境下，数据密集型端到端自动语音识别（ASR）系统更容易误识别同音字和罕见词。针对低资源粤语语音识别问题，本文提出一种新颖的同音字扩展方法，将同音字词库中的人工知识融入结合语言模型重评分的束搜索解码过程。此外，我们还提出一种自动统一书写方法，用于合并粤语汉字的变体并规范语音标注指南。通过为合并后的汉字提供更多样本，该方法能够更高效地利用已标注语句。实验结果表明，同音字扩展和统一书写均显著提升了域内和域外测试集上的识别性能，字符错误率（CER）绝对降低约5%和18%。

索引词：低资源粤语语音识别、词典、语言模型重评分、同音字

1. 引言

端到端深度学习模型 [1–5] 由于架构更简单，且性能优于经典的模块化系统或混合系统，因此在开发语音识别系统时越来越受欢迎。然而，这些模型通常需要大量数据，在低资源设置下的表现较差。随着自监督学习在语音与语言处理领域的近期进展，微调预训练模型已成为低资源语音识别中的常见做法。然而，在带标注数据极其有限的情况下微调模型仍然非常具有挑战性，尤其难以正确识别训练集中的罕见词或未见词（例如实体名称）。当测试数据与训练数据集差异很大，即属于域外数据时，性能会进一步下降。例如，我们观察到，字符错误率从域内测试集中的 13% 显著上升至域外测试集中的 50%；在该域外测试集中，许多字符或词在训练集中较为罕见或未曾出现。

本文聚焦于粤语语音识别，旨在提升训练集中出现频率较低或未出现过的词语的识别性能。与英语相比，粤语是一种声调语言，存在大量同音字和同音词，这给使用有限数据训练的自动语音识别模型带来了更大的挑战。如表 1 所示，不同的粤语汉字具有相同的发音（即同音字），并以粤拼标注。粤拼是香港语言学学会制定的一种粤语罗马化系统。此外，一个汉字可能有多种异体写法（即发音和含义相同，但写法不同）。

表 1：粤语同音字，以及粤拼和语义相同的异体字示例。

粤拼	同音字
zo2	左、阻、俎、租、诅、座
sai3	世、細、勢、婿、貰、些、僿、埶、棲
wong4	王、黃、皇、簧、煌、蝗、惶、磺、凰
粤拼	**书写变体**
zoeng3	帳, 賬
lei5	裏, 裡
曾	淨, 凈

在粤语语音识别的低资源场景中，同音字和书写变体带来两个难点：（1）某个汉字可能被误识别为在有限训练集中出现频率更高的同音字；（2）同一汉字的不同变体可能导致不同标注者或不同数据集中的转写文本不一致，从而减少该汉字的训练样本数量，并给解码过程带来不必要的复杂性。在域外数据集上评估训练好的模型时，这两个问题可能会更加严重，因为域外数据集中的罕见或未见汉字、词语可能比训练集中更多。

一般而言，我们将这两个问题视为罕见词识别问题，并提出两种方法来提升识别性能，即同音字扩展（HE）和统一书写（UW）。对于同音字扩展，我们利用粤语词典找出同音字（这些字在训练集中可能较为罕见），并将其加入束搜索过程，随后通过语言模型（LM）进行重评分。因此，在训练期间罕见或未见的词语也能在推理时以一定概率得到识别。对于统一书写方法，我们合并粤语汉字的各种变体，并统一采用其中最常见的形式，以避免变体带来的不必要复杂性。这一步也增加了特定汉字的训练样本数量。

我们在三个公开可用的粤语数据集上评估了上述两种方法，并确认两者都能提升识别性能，尤其是在罕见词更多的域外测试集上，字符错误率从50.58%降至31.87%。

[†]在 CPII 实习期间完成的工作。

[*]通讯作者。

2. 相关工作

近期，使用预训练自监督学习语言模型的趋势已证明，这类模型能够有效处理不同的下游任务。因此，文献中的许多研究开始探索预训练模型在语音处理中的迁移学习。例如， [6] 提出了Wav2Vec2.0，该模型采用语音预训练技术，使模型仅凭语音音频就能学习强大的表征。沿着这一研究方向， [7]在多语言环境中使用Wav2Vec2.0模型，并提出了在128种不同语言上预训练的XLS-R模型。在XLS-R模型上微调的下游自动语音识别任务，其表现优于其他预训练语音模型，尤其是在低资源语言上。为提升粤语语音数据的数量和质量， [8] 提出了一个名为多领域粤语语料库（MDCC）的新语料库，数据采自香港的有声读物，旨在尽可能覆盖更多领域。他们还通过手动统一书写方式，解决了含义和发音相同的粤语字符异体字问题。

近期一些研究指出，仅通过微调预训练语音模型来识别罕见词面临诸多挑战。例如， [9]提出了一种语言模型多任务训练方法，将意图和槽位用作额外的目标，并在第二遍重评分中使用该语言模型，以提升特定交互式语音助手应用中罕见词（如槽位）的识别性能。此外， [10] 采用单元语法浅层融合方法，在解码过程中遇到罕见词时施加固定奖励。然而，如果罕见词未在解码过程中出现，该方法便无法奏效。 [11, 12] 探索了使用混合自回归转导器结合神经语言模型，对声学模型的输出重新排序，从而提升罕见词识别质量的思路。除了使用语言模型外， [13] 还提出采用语音合成技术合成罕见词或域外词的语音，并利用合成语音微调自动语音识别模型，以提高识别准确率。最后，[14]提出了一种名为 Phoebe 的方法。该方法通过将发音注入上下文模块，扩展了上下文化的 Listen, Attend, and Spell (CLAS) 模型，并实现了相较于 CLAS 16% 的相对词错误率降低。Phoebe 同时利用偏置短语的文本形式及其对应发音，从而提升罕见词或未见词的识别性能。

3. 方法

微调预训练语音模型是开发自动语音识别系统的一种常见做法。然而，由于缺少带标注数据，罕见词或未见词问题仍无法仅通过使用有限数据微调 XLS-R 模型得到妥善解决。为缓解这一问题，我们提出同音字扩展（HE）方法：将基于粤语词典的罕见词相关人类知识注入束搜索过程，随后进行语言模型重评分。这样，在训练期间罕见或未见的词语也能在推理时以一定概率得到识别。此外，我们还将粤语汉字的异体字统一为使用频率最高的字形，以避免异体字带来的不必要复杂性，并增加特定汉字的训练样本数量。我们将这一过程称为自动统一书写（UW）；实验表明，该方法还能降低字符错误率。

3.1. 基线模型

我们选择对名为 XLS-R [7] 的预训练模型进行微调，并将其作为基线模型。该模型在预训练模型之上采用一个线性层，并使用连接时序分类（CTC）损失进行训练。解码过程采用束搜索算法，在每个解码步骤中保留排名前 B 的候选项，并根据输出序列树剪除不太可能的路径。在束搜索解码过程中还融合了 n 元语言模型，以浅层融合的方式生成语义关联度更高的结果。然而，在识别罕见词和未见词方面，性能提升并不显著。由于缺少涵盖这些词语的训练数据，排名前 B 的候选项中可能不包含目标词。

3.2. 同音字扩展

我们提出一种同音字扩展（HE）方法，以提升低资源训练集中罕见词或未见词的识别效果。其核心思想是在束搜索过程中注入同音字词典，并为其指定初始分数，使同音罕见词能够进入搜索空间并重新估算分数。为了确定给定汉字的同音字，我们采用了一种公开可用的粤拼 [15] 输入法词库，该词库将粤拼编码（例如 wong2）与对应汉字关联起来，这些汉字称为同音字。

如图1所示，将一段语音语句输入一个经过微调的 XLS-R 模型，它会为每一帧跨度输出概率分布，并在束搜索算法中剪枝后保留排名前 B（例如，图中的 B=3 ）个候选字符。对于每个候选字符，我们通过添加其同音字来扩展搜索空间。随后，采用语言模型对所有扩展后的候选项进行重评分，并输出最终的新候选项。值得注意的是，所提出的同音字扩展方法与任何用于语音识别的语言模型重评分方法都不相上下。考虑到这些同音字的概率可能很低，无法出现在排名靠前的结果中，我们进一步将每个同音字的输出概率调整为对应源字符的输出概率；源字符的输出概率远高于同音字。

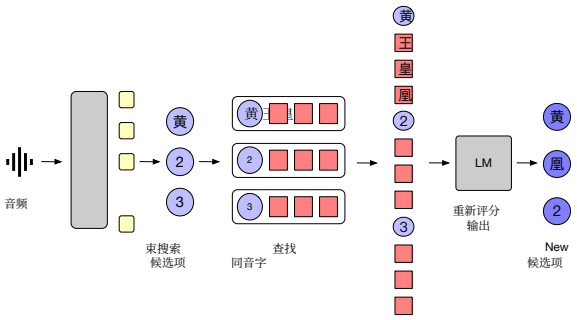

图 1：在束搜索过程中注入同音字扩展方法，并进行语言模型重评分。

具体而言，对于给定的字符 c，令 $a_p(c)$ 表示 CTC模型给出的声学概率， $H_c = \{h_i^c\}$ 为 c的同音字集合，其中 $i \in \{1, \ldots, M\}$。一个字符可能有多种不同的发音，称为多音字；我们用 $N(h_i^c)$ 表示字符 h_i^c的发音数量，并从粤拼中获取与其共享相同带声调编码的 h_i^c。此外， $q(h_i^c)$ 表示该多音字在 CTC模型中的声学概率。我们还引入一个折扣

因子（$1 - \log_{10} N(h_i^c)$）来调整 $q(h_i^c)$，以降低发音数量较多的多音字的概率；同时引入权重因子 γ，以平衡 $a_p(c)$ 和 $q(h_i^c)$ 的概率。此外，我们采用最大值运算，确保更新后的概率 $p(h_i^c)$ 不低于原始概率 $a_p(c)$，如公式（1）所示：

$$p(h_i^c) = \max\{a_p(c), (1-\gamma) \cdot a_p(c) + \gamma \cdot q(h_i^c) \cdot (1 - \log_{10} N(h_i^c))\} \quad (1)$$

3.3. 统一书写

粤语转写标准缺乏严谨性，而许多字在相近的书写或输入代码中读音和含义相同。此外，不同数据集之间的语言风格也各不相同，合并不同数据集时，标注差异会进一步增加。例如，裏面（里面）和裡面（里面）是两个含义相同的词，在 MDCC [16] 数据集中只会出现“裏面”，而在 Common Voice 数据集中则会同时出现“裏面”和“裡面”。这一问题有两个关键之处：首先，尽可能找出某个字的各种异体；其次，检查将该字替换为异体后是否会影响原意。

我们提出一种名为统一书写（UW）的自动化方法，用于处理不同数据集之间的标注差异，而非依赖需要语言学专家知识且核查成本高昂的人工替换。该方法包括两个步骤：第一步是发现统一配对，第二步是通过重写语义检查器进行替换。我们假设标注差异源于标注者使用的输入系统。粤语有许多输入法，例如粤拼、仓颉、四角号码等。这些输入方式可分为发音编码或字形编码。发音或字形相近的字，其对应的发音或字形输入代码也会相近。

设 $C_{0..L}$ 为词库中的所有汉字，其中 L 为词库中的条目总数。统一字对发现过程将遍历总计 $L \cdot (L-1)$ 种汉字组合。在遍历过程中，我们使用归一化编辑距离计算粤拼的发音编码相似度，然后筛除粤拼编辑距离为 $d > 0$ 的汉字组合，这意味着它们的发音不同。对于字形编码相似度，我们分别使用仓颉五代、速成五代、BSM、CKC、QCode、G6Code、五笔画、嘸蝦米、大易四代和四角号码五代等输入法生成输入码，然后筛除归一化编辑距离为 $d > 0.25$ 的所有结果。

经过两轮筛选后，我们可以得到一组发音相同且字形相似的汉字对。为确保汉字的用法相似，我们进一步使用 fastText [17, 18]，根据余弦相似度值 > 0.5筛除语义上不相似的汉字对。相似度值接近 1，意味着这两个汉字更有可能出现在相同的语境中。筛选后剩余的汉字对即为统一字对发现步骤的输出；由于它们符合所有汉字相似度指标，因此可能可以相互替换。

下一步是通过改写语义检查器后，根据频率进行实际替换。字符频率高，意味着它在目标应用场景中更常见。我们将低频字符替换为高频字符。替换后，我们还需要验证统一书写后的含义。我们采用 BERT 分数 [19]进一步评估原始转写文本与替换后文本之间的语义相似度，其中 BERT 分数越高，表示两份转写文本的相似度越高。我们会筛除相似度分数低于 0.9 的结果。

4. 实验

4.1. 实验数据集

我们在三个公开可用的数据集上评估所提出的方法，分别是 MDCC [16], Common-Voice [20], 和 Yueqie[1]。表 2 显示了各数据集中的话语数量，其中 MDCC 使用默认划分，Common-Voice 则使用 Mozilla[2]收集的粤语子集。我们使用 Yueqie 的全部数据（共 171 条话语），在域外设置下评估所提出的方法。粤切数据集共有 171 个四字词，其中 25 个仅在训练集中出现一次，另有 4 个出现两次。

表 2：各实验数据集中的话语数量。

	训练	Dev	Test
MDCC [16]	65120	5663	12492
Common-Voice [20]	8392	5578	5578
Yueqie	-	-	171

4.2. 字符分词器

我们使用大型字符分词器构建自动语音识别系统，以尽可能涵盖更多粤语汉字，并纳入数据集中可能不会出现的罕见汉字。否则，推理时这些未见汉字会被识别为未知字符，也无法支持模型适配到包含新词的新领域。此外，我们还需要更大的分词器来支持同音字扩展方法，因为该方法可能会在解码过程中引入一些罕见词或未见词。我们从一个为输入法提供 cin 表的开源项目中收集所有汉字；cin 表是一种将输入代码与汉字关联起来的查询表[3]。我们使用中文（香港）的 cin 表，并获取其中所有不重复的汉字。该表涵盖了标注人员可能使用的所有汉字，最终的分词器包含 32,693 个汉字。

4.3. 实现细节

遵循 XLS-R 的微调设置 [7], ，我们使用 xls-r-300m 作为骨干模型进行实验，并使用 CTC 损失对其进行微调，以构建我们的自动语音识别系统。模型使用 6 个图形处理器进行训练，批量大小为 17，梯度累积步数为 20。学习率为 2e-3，并在整个训练过程中线性衰减。我们训练模型 150 个周期，并根据开发集选择最佳检查点进行后续测试。对于语言模型集成，我们使用粤语维基百科数据集训练了一个 5-gram KenLM 模型[4]。基于束搜索的解码超参数设置为 $\alpha = 0.45, \beta = 1.55$，束宽设为 20，其中 α 是语言模型浅层融合的权重，β 是基于解码序列长度进行分数调整的权重

[1] http://www.livac.org/yueqie/ [2] https://huggingface.co/datasets/mozilla-foundation/common_voice_9_0 [3] https://github.com/chinese-opendesktop/cin-tables [4] https://dumps.wikimedia.org/zh_yuewiki

的解码序列长度 [21]。此外，公式（1）中用于同音字扩展的 γ 设为 0.5。

4.4. 实验结果

表3展示了不同测试集之间的字符错误率（%）性能对比，其中 Yueqie 是一个不同的域外数据集，模型在该数据集上的字符错误率非常高。

表3： 不同测试集之间的性能对比。

测试集	字符错误率（%）
MDCC	9.32
Common-Voice	24.93
MDCC + Common-Voice	13.88
Yueqie	50.58

表4展示了不同方法在域内数据集（即 MDCC + Common-Voice）和域外数据集（即 Yueqie）上的性能比较。可以看出，同音字扩展和统一书写均能有效降低字符错误率，尤其是在粤切数据集上。请注意，尽管在某些设置下我们对训练集采用了统一书写，但为确保结果可比，所有实验均未对测试集应用统一书写处理。此外，我们还在经过统一书写处理的测试数据上评估了这些方法。该处理会自动更改 MDCC 和 Common-Voice 中的部分转写文本，但由于没有匹配字符，粤切数据集保持不变。我们观察到，所有方法组合均能降低字符错误率，最高达 7.73，这得益于通过统一书写合并训练样本和测试转写文本。

表4： 不同方法在 *MDCC + Common-Voice*（域内）和 *Yueqie*（域外）上的性能比较。

数据集与方法	字符错误率 (%)
MDCC + Common-Voice 测试集上未使用统一书写	
基线	13.88
基线 + 束搜索 + 语言模型	10.66
基线 + 束搜索 + 语言模型 + 同音字扩展	10.28
基线 + 束搜索 + 语言模型 + 统一书写	8.61
基线 + 束搜索 + 语言模型 + 同音字扩展 + 统一书写	**8.07**
MDCC + Common-Voice 在测试集上使用统一书写	
基线 + 束搜索 + 语言模型	11.17
基线 + 束搜索 + 语言模型 + 同音字扩展	10.82
基线 + 束搜索 + 语言模型 + 统一书写	8.05
基线 + 束搜索 + 语言模型 + 同音字扩展 + 统一书写	**7.73**
测试集上使用/不使用统一书写的 *Yueqie*	
基线	50.58
基线 + 束搜索 + 语言模型	39.33
基线 + 束搜索 + 语言模型 + 同音字扩展	35.82
基线 + 束搜索 + 语言模型 + 统一书写	33.78
基线 + 束搜索 + 语言模型 + 同音字扩展 + 统一书写	**31.87**

4.5. 案例研究

为了更好地了解同音字扩展如何提升罕见词的识别效果，我们分析了粤切数据集中的一些示例，并比较了使用和不使用同音字扩展时的结果。如表5所示，前四个话语在使用同音字扩展后均被正确识别，尽管最后两个话语从第一个字开始就识别错误。事实上，同音字扩展提升了171个粤语四字成语中20个的识别效果，但未能改善其余6个成语（例如表5中的第5个和第6个），因为首字识别错误很可能导致后续的语言模型重评分失败。请注意，空白的红色方块表示相应的字未被识别。

表5: 采用和不采用同音字扩展的预测示例。

	Ground truth	Without HE	With HE
1	阻頭阻勢	左頭左細	阻頭阻勢
2	無憑無據	無憑無居	無憑無據
3	飛擒大咬	飛擒大□	飛擒大咬
4	縮手縮腳	□□□腳	縮手縮腳
5	搭頭搭尾	搭求你	答求答你
6	擔天望地	擔天望地	眈天望地

表6展示了采用和不采用统一书写（UW）的一些预测示例。蓝色标出的汉字有对应的异体字。如表6第二列所示，如果不采用统一书写，模型会错误识别这些汉字（以红色标出）。然而，如果将这些词统一为其异体字——这些异体字在语料库中出现得更频繁——模型便能够将其识别为对应的异体字（以绿色标出）。由于真实转录文本仍使用汉字异体字进行标注，表6中的示例字符错误率并未降低。不过，采用统一书写后的预测实际上是正确的，而且这些异体字在粤语书写中更为常见。

表6: 采用和不采用统一书写的预测示例。

	Ground truth	Without UW	With UW
1	如果要在户外活動或工作	如果要在牖活動或工作	如果要在戶外活動或工作
2	以便市民查閱	以便市民查月	以便市民查閱
3	惟有人才及科技兩者善用	為有人才及科技兩者善用	唯有人才及科技兩者善用

5. 结论

本文提出了一种新颖的同音字扩展方法和一种自动统一书写方法，以提升低资源场景下粤语语音识别中训练数据集里罕见或未见汉字的识别能力。借助粤语同音字词典，如果解码过程中遇到某个罕见汉字的同音字，该方法便能使这个罕见汉字进入束搜索空间。通过采用统一书写方法，我们能够更高效地利用带标注语句，并提升粤语汉字异体的识别能力。实验结果表明，所提出的两种方法显著提升了识别性能，尤其是在域外数据集上，字符错误率从 50% 大幅降至 31%。本文聚焦于低资源场景下的粤语语音识别问题，主要贡献总结如下：

(1) 提出一种同音字扩展方法，将同音字词典中的人工知识融入束搜索过程，并通过罕见词的同音字提升其识别效果；(2) 引入一种自动统一书写方法，合并粤语汉字的不同变体，并通过增加该汉字的训练样本以及缩小解码空间来提升性能；(3) 通过实验证明，所提出的两种方法均能显著提升域内和域外数据集上的识别性能。

6. 致谢

本研究部分获得感知与交互智能中心（CPII）有限公司的资助。该中心由香港中文大学牵头，隶属创新科技署 InnoHK 计划；此外，研究还部分获得香港特别行政区研究资助局优配研究金资助（参考编号：14207619）。我们感谢国立台湾大学的 Jeff Chen 为本文提出的宝贵建议和意见。

7. 参考文献

[1] A. Graves and N. Jaitly, 《通过循环神经网络实现端到端语音识别》，载于国际机器学习会议。PMLR，2014年，第1764–1772页。[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Bat- tenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen 等，《Deep Speech 2：英语和普通话的端到端语音识别》，载于国际机器学习会议。PMLR，2016年，第173–182页。[3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, 《用于端到端语音识别的混合 CTC/注意力架构》，*IEEE*《信号处理精选主题期刊》，第11卷，第8期，第1240–1253页，2017年。[4] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, 《探索用于端到端语音识别的神经转导器》，载于*2017 IEEE* 自动语音识别与理解研讨会（*ASRU*）。IEEE，2017年，第206–213页。[5] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, and Y. Wu, 《实现快速准确的流式端到端自动语音识别》，载于*ICASSP 2020-2020 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，第6069–6073页。[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, 《wav2vec 2.0：用于语音表征自监督学习的框架》。 [在线]。网址：http://arxiv.org/abs/2006. 11477[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino 等，《XLS-R：大规模自监督跨语言语音表征学习》，*arXiv* 预印本 *arXiv:2111.09296*，2021年。[8] T. Yu, R. Frieske, P. Xu, S. Cahyawijaya, C. T. S. Yiu, H. Lovenia, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. E. Shi, and P. Fung, 《粤语自动语音识别数据集：综述与新数据集》。 [在线]。网址：http://arxiv.org/abs/2201.02419[9] C.-H. H. Yang, L. Liu, A. Gandhe, Y. Gu, A. Raju, D. Fil- imonov, and I. Bulyko, 《通过多任务语言建模改善罕见词的语音识别》，载于*2021 IEEE* 自动语音识别与理解研讨会（*ASRU*）。IEEE，2021年，第1087–1093页。[10] V. Ravi, Y. Gu, A. Gandhe, A. Rastrow, L. Liu, D. Fil- imonov, S. Novotney, and I. Bulyko, 《通过单元语法浅层融合提高循环神经网络转导器对罕见词的识别准确率》，*arXiv* 预印本 *arXiv:2012.00133*，2020年。[11] T. N. Sainath, Y. R. He, A. Narayanan, R. Botros, R. Pang, D. J. Rybach, C. Allauzen, E. Variani, J. Qin, A. Gruenstein 等，《高效的流式非循环端侧端到端模型及其罕见词建模改进》，*Interspeech*，2021年。[12] W. Wang, T. Chen, T. N. Sainath, E. Variani, R. Prabhavalkar, R. Huang, B. Ramabhadran, N. Gaur, S. Mavandadi, C. Peyser 等，《通过语言模型感知的最小词错误率训练改进罕见词识别》，*arXiv* 预印本 *arXiv:2204.07553*，2022年。[13] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, 《利用合成音频改进端到端自动语音识别系统对未登录词的识别》，载于*IEEE* 国际声学、语音与信号处理会议，*ICASSP 2021*，加拿大多伦多，*2021*年*6*月*6*日至*11*日。IEEE，2021年，第5674–5678页。 [在线]。网址：https: //doi.org/10.1109/ICASSP39728.2021.9414778

[14] A. Bruguier, R. Prabhavalkar, G. Pundak, and T. N. Sainath, 《Phoebe：面向端到端语音识别的发音感知上下文建模》，载于*ICASSP 2019-2019 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第6171–6175页。[15] L. S. of Hong Kong, 《香港粤拼字表》，1997年。[16] T. Yu, R. Frieske, P. Xu, S. Cahyawijaya, C. T. S. Yiu, H. Lovenia, W. Dai, E. J. Barezi, Q. Chen, X. Ma 等，《粤语自动语音识别数据集：综述与新数据集》，*arXiv* 预印本 *arXiv:2201.02419*，2022年。[17] A. Joulin, ´E. Grave, P. Bojanowski, and T. Mikolov, 《高效文本分类的实用技巧集》，载于欧洲计算语言学协会第*15*届会议论文集：第*2*卷，短篇论文，2017年，第427–431页。[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, 《利用子词信息丰富词向量》，计算语言学协会汇刊，第5卷，第135–146页，2017年。[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, 《BERTScore：使用 BERT 评估文本生成》，*arXiv* 预印本 *arXiv:1904.09675*，2019年。[20] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, 《Common Voice：大型多语言语音语料库》，载于第*12*届语言资源与评测会议（*LREC 2020*）论文集，2020年，第4211–4215页。[21] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, 《将外部语言模型融入序列到序列模型的分析》，载于*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第1–5828页。