面向自动语音识别的合成交叉口音数据增强

Philipp Klumpp[1,2], *Pooja Chitkara*[1], *Leda Sarı*[1],*Prashant Serai*[1],*Jilong Wu*[1],*Irina-Elena Veliche*[1], *Rongqing Huang*[1],*Qing He*[1]

[1]Meta AI[2]模式识别实验室，FAU

Erlangen-N¨urnbergphilipp.klumpp@fau.de,
ledasari@meta.com

arXiv:2303.00802v1 [cs.CL] 1 Mar 2023

摘要

近年来，人们对自动语音识别数据集或模型存在偏见的关注显著增加。即使是英语，尽管可用训练数据十分庞大，系统对非母语者语音的识别效果仍然较差。

在这项工作中，我们改进了一种口音转换模型（ACM），该模型可将美国本土英语语音转换为带口音的发音。我们在 ACM 训练中纳入语音学知识，以便准确评估合成波形对特定发音模式的还原效果。此外，我们还研究了使用学习得到的口音表示而非静态嵌入的可行性。随后，我们使用生成的数据训练了两个最先进的自动语音识别系统。

我们在英语母语者和非英语母语者数据集上评估了该方法，发现合成的带口音数据有助于自动语音识别系统更好地理解来自已见口音的语音。这一观察结果并未推广到未见口音，而且对于仅使用母语语音进行预训练的模型，也没有观察到这一现象。

索引词：语音识别、口音转换、包容性自动语音识别

1. 引言

现代自动语音识别（ASR）模型使用数千小时的录制语音样本进行训练。尽管自动语音识别数据集规模不断扩大，为模型提供了更多探索机会，但底层数据分布很容易引入缺乏代表性的先验。例如，Common Voice [1] 的开放源代码英语子集在其最新版本中，自我报告为男性的说话人数量是女性的三倍[1]。

数据本身存在偏见，并不一定会导致模型产生偏见。一项研究调查了自动语音识别中的偏见，结果发现，使用 1185 名荷兰语女性说话人和 1678 名荷兰语男性说话人训练的模型，对代表性不足的女性贡献者群体持续取得较低的测试词错误率（WER） [2]。同一项研究还发现，不同年龄组之间的词错误率存在差异，而这种差异与训练数据的年龄分布并不一致。尽管年龄和性别偏见在最终模型性能中的表现并不明显，但非母语者的错误率显著上升

荷兰语使用者。一项关于英式英语自动语音识别中算法偏见的研究也得出了相同的观察结果 [3], , 指出这些模式并不局限于第二语言学习者口音，在带有不同地区口音的母语者中也同样可见。甚至肤色也可能隐含地反映

这项工作是在 Philipp 于 Meta AI 实习期间完成的。联系 Leda Sarı：ledasari@meta.com[1]
https://github.com/common-voice/cv-dataset

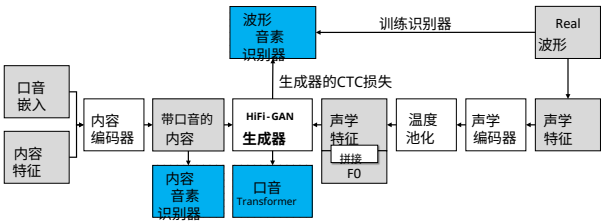

图 1：口音转换模型结合了类似于口音特定发音的内容特征和一组声学特征，以生成口音转换语音。蓝色方块表示新添加的模块。

社会文化背景，并可能因此引发偏见 [4]。Artie Bias Corpus [5] 提供了一个测试数据集，用于评估自动语音识别系统中偏见（尤其是与带口音语音相关的偏见）会在多大程度上出现。

与某种语言的母语者相比，对非母语者的表征不足，会妨碍语音识别系统探索发音或语言结构上的差异，而这些差异源自其原有的外语母语。举例来说，第二语言学习者与婴儿不同，他们已经习得了一套音系库存。其母语音系库存与第二语言音系库存的交集可能产生各种发音变体，从而形成有所变化但仍可理解的话语 [6]。自动语音识别系统无法对非母语者保持较高的识别质量，这表明主要使用母语者样本训练的模型，缺乏应对代表性严重不足的发音模式的鲁棒性。

为自动语音识别系统提供描述源口音的辅助输入，有助于降低带口音语音的词错误率（WER） [7]。辅助输入既可以从学习得到的嵌入中提取，也可以来自 wav2vec 的中间无监督表示 [8]。除了口音嵌入之外，语音识别与口音分类的多任务学习方案也已证明优于基线系统 [9, 10]。通过领域对抗训练 [11],AIPNet [12]能够学习口音不变声学特征向量，随后自动语音识别系统可利用这些向量提升对非母语语音样本的鲁棒性。

多种口音转换方法旨在解决反向问题，即将非母语口音的语音转换为母语口音，例如结合语言和说话者表示，并使用以母语语音为条件的语音合成模型 [13]。另一种方法是计算各个说话人的语音后验图，再将其转换到母语口音域，作为神经声码器的输入 [14]。通过一组瓶颈特征 [13] 或声学和内容特征的口音不变表征 [14]，实现了从母语口音到外国口音的转换；这两项研究都强调了在保留源说话人嗓音的同时实现良好口音转换所面临的挑战。

一组瓶颈特征 [15] 或声学和内容特征的口音不变表征 [16], , 两项工作都强调了在保留源说话人嗓音的同时实现良好口音转换的挑战。

本研究旨在将美国本土英语语音样本转换为带口音语音，以合成外国口音自动语音识别训练数据。我们使用合成数据训练了两个最先进的语音识别器，并将它们的表现与一个在优化过程中未接触过带口音语音的基线系统进行了比较。我们关注的主要问题是："能否利用某种特定口音的合成样本来提升带口音语音的识别效果？"此外，我们通过使用一组口音 A_{train} 训练自动语音识别模型，并在另一组口音 A_{test} 上测试其表现，评估模型对非母语发音变体的泛化能力，使得 $A_{\text{train}} \cap A_{\text{test}} \approx \varnothing$。最后，我们评估了可学习、非静态口音表征的可行性，因为这类表征或许能帮助 ACM 推导出此前未见口音的发音模式。

2. 方法

我们训练了一个 ACM，将美国本土英语转换为带有各种外国口音的语音。随后，我们使用口音转换语音数据训练更稳健的非母语英语自动语音识别模型。此前一项研究 [16] 已介绍该 ACM，图 1 展示了其概况。其生成组件基于 Hifi-GAN [17]，并以声学特征和内容表征的组合作为输入。前者应保留源说话人的原始嗓音、节奏和语调；后者则将 wav2vec 2.0 模型 [18] 预测的已发字符独热序列与口音嵌入相结合，以编码口音特定发音。与原研究 [16], 一样，ACM 采用口音对称的方式进行训练，这意味着源（输入）信号的口音始终与生成波形的目标口音一致。这一要求由重建损失引入

$$L_{mel} = ||M(\mathbf{x}) - M(\hat{\mathbf{x}})||_1 \tag{1}$$

其中，M 表示真实（$\mathbf{x}$）和生成（$\hat{\mathbf{x}}$）波形的 80 维梅尔频谱图。训练期间，源口音与目标口音保持一致，因此，口音相关信息必须从声学特征向量中解耦出来，因为这些信息应仅源自内容特征。

2.1. AccentTransformer

通过引入一个序列分类Transformer模块来提升合成信号的口音程度，该模块用于预测真实音频和生成音频的口音。它以 13 个梅尔频率倒谱系数特征及其一阶和二阶导数为输入，将每个时间步的特征投影到 96 个通道，然后与一个分类标记拼接。经过 10 层 Transformer编码器层（前馈维度为 384，丢弃率为 $p = 0.1$）后，分类标记处的输出被线性投影到 N 种口音。该口音Transformer（AT）仅使用真实音频样本进行训练。预热 20000 步后，也向其输入生成波形，并将交叉熵分类损失反向传播至生成模型，从而反馈特定目标口音在生成的语音信号中保留得如何。由于 AT 执行的是多口音分类，而非母语与非母语分类，我们也相应调整了原始 ACM 的口音判别器。其目的是从声学特征中学习口音分类，然后对这些特征应用对抗损失，以消除其中的任何口音信息。

2.2. 语音学知识

本文提出的 ACM 改进旨在增强对口音特定发音的转换。公式 1 中的重建误差 L_{mel} 对生成器总损失的贡献很大。这有助于提升低频频带的合成质量，而人类对这些频带具有更高的感知分辨率。因此，对于许多辅音音素，尤其是清音变体，该误差的描述能力较弱，因为它们更多地体现在高频区域。重建损失的另一个缺点是它会使音素时长的重要性增加。较长的音素（如元音）不可避免地比塞音等较短音素受到更强的优化。为了更准确地反馈口音发音质量，我们在 ACM 训练中加入了语音学信息。真实标注信息取自对训练样本进行的机器生成序列标注，其中涵盖了各种语音学特征。这些特征对每个语音音素的类型（元音、擦音等）、位置、第一和第二共振峰、时长、发声态、发音、协同发音及圆唇度进行了类别描述。这使我们能够定义由 51 个音素类别构成的封闭集合，每个类别代表一种独特的语音学特征组合。该音素序列有两种不同的用途。我们对预训练的 wav2vec 2.0 基础模型进行微调，使其使用 CTC损失 [19]预测真实音频样本的音素类别序列。与 AT模块一样，该模型随后也用于生成波形，并将梯度反向传播至生成器。我们的想法是，先利用真实波形教会音素识别器每个音素应有的听感，再强制生成的话语采用相同的发音。

我们还利用语音学知识来提升带口音内容特征的质量。因此，我们添加了第二个音素识别器，它需要识别与 wav2vec 2.0 模块相同的真实标注序列，但这一次输入的是内容编码。经过卷积位置编码后，结果被送入十个 Transformer 层，每层有四个注意力头，前馈维度为 128。输出再次被投影到 51 个音素类别及一个空白标记，并使用 CTC损失进行优化。

2.3. 可学习的口音表征

使用静态映射（例如通过独热编码）会带来一些限制。语音信号中的口音程度无法控制，因此无法考虑口音表现的不同强度。另一个缺点是采用监督式设置，这使 ACM 只能处理训练期间见过的口音。除了静态嵌入，也可以尝试从输入信号中学习口音表示。如果能将其与说话人特征或声学环境等所有其他信息成功解耦，那么就可以利用学习到的口音表示来改变次级语音样本的口音。由于所提出的 ACM 侧重于不同口音的发音特征，我们使用了基于波形的音素识别器随时间平均池化的 Transformer 输出，将其投影到较低维度的表示空间，随后再通过另一次投影映射至 N 种口音。该表示模块基于池化后的 Transformer 输出进行口音分类训练，并将中间投影用作口音表示。

2.4. 训练设置

AT模块采用与原始 ACM [16] 中所有其他判别器相同的超参数进行训练。基于波形和基于内容的音素识别器使用 Adam 优化器（$\eta = 3 \cdot 10^{-5}$）进行训练，并在每个优化步骤后采用指数学习率衰减（$r = 0.999995$）。用于监督式表征学习的口音分类模块采用与 AT 相同的配置进行优化。整个 ACM 使用批量大小为 8 的设置进行训练，每个样本长度为 1 秒。

3. 实验设置

3.1. 自动语音识别模型

我们采用了两种不同的模型架构，以评估合成口音语音如何帮助提升自动语音识别在词错误率（WER）方面的性能。输入为 80 个频带的梅尔频谱图，帧移为 10 ms。按照 [20] 中提出的方法，通过屏蔽连续时间步和频带对频谱图进行增强。两个模型均采用由 5000 个句子片段构成的输出词汇表 [21]，该词汇表根据 LibriSpeech [22]（LS）的 960 小时训练集（clean-100 + clean-360 + other-500）估算得出。基础版模型（Base）的编码器先将 80 个梅尔频带线性投影到 128 个通道，随后进行时间降采样，将每四个连续时间步沿特征维度拼接合并。模型通过堆叠 20 层高效记忆Transformer（Emformer [23]）获取时间上下文；每层采用 8 个注意力头、2048 的前馈维度和 GELU 激活函数 [24]。Emformer 的输出被投影到 1024 个层归一化通道 [25]，并作为循环神经转导器（RNNT）的输入。RNNT 预测器使用三个长短期记忆网络 [26]，将此前发出的所有符号嵌入到 512 维空间中。每个循环层的输出均通过丢弃（$p = 0.3$）进行正则化，最后投影到 1024 个层归一化通道。编码器与预测器输出相加后，再投影为覆盖输出词汇表及一个空白标记的概率分布。

HuBERT模型的编码器结构与Base模型相似，只对某些层的通道配置做了细微调整。梅尔频带初始线性投影的维度增至192个通道。随后使用包含31个卷积核元素的残差一维卷积作为相对位置编码。由于初始卷积的通道数增加，时间降采样层的拼接结果为变换器模块提供了768维输入。最后一个变换器模块的输出未进行线性投影（与Base模型不同）。HuBERT模型采用了与Base模型相同的转导器架构，但在每个循环层之后提高了丢弃概率（$p = 0.5$），并增大了最终投影的维度（768个通道），使其与预测器的维度一致。

两个模型均通过最小化转导器损失 [27]进行优化，使用 Adam 优化器 [28]和权重衰减；对于预训练的 HuBERT 模型，权重衰减显著较小（基础版：0.1；HuBERT：10^{-6}）。我们将最大梯度值裁剪为 1.0，将最大梯度范数裁剪为 10.0。每批训练样本解码后，在真实标注中的句子片段令牌数最多为 40,000（基础版）或 5,000（HuBERT）。由于基础版模型使用全部 LS-960h 数据从头开始训练，因此我们将训练轮数增加到 180，而 HuBERT 微调为 120 轮。学习率的基础版（10^{-3}）和 HuBERT（10^{-4}）模型的学习率在前 10,000 个（HuBERT：5,000 个）预热步数内线性增加，随后保持恒定，并在 60 轮训练后按 0.96 的因子指数衰减。

3.2. 数据集

采用了原始ACM的数据分布和权重：L2-Arctic 语料库 [29] 包含阿拉伯语、英式英语、印地语、韩语、普通话、西班牙语和越南语口音的英语语句，每种口音均由两名女性说话者和两名男性说话者录制。通过纳入包含两名女性说话者和两名男性说话者的 Indic TTS [30],，进一步增加了印度口音数据。对于英式英语，我们在VCTK 数据集 [32]之外，还纳入了Speech Accent Archive [31] 中对应的子集。对于美式英语，则使用了针对语音合成应用优化后的LS版本（LibriTTS [33]），具体包括 train-clean-100和train-clean-360两个划分。

基础版自动语音识别模型使用 LS 960h 从初始化状态直接进行优化。另一方面，HuBERT 模型先使用 LS 960h 预训练，以获得自监督 HuBERT 特征 [34]，随后仅在 LS 的 train-clean 100 h 子集上进行微调。
为了在所述自动语音识别训练数据中加入带口音语音样本，我们将 LibriTTS train-clean-100（美国英语）中的语音样本转换成七种外国口音。请注意，在 ACM 的训练过程中从未执行这一流程；当时输入口音始终与输出口音一致。每种口音都可以合成 53 小时的音频。我们选择从上述划分中生成带口音样本，因为它是原始 LS train-clean-100 划分的一个子集，而该划分已用于自动语音识别训练，因此不会引入新的说话人或话语。
随后，我们在 LS 的干净及含噪开发集和测试集划分上测试了所得的自动语音识别模型。此外，我们还在 L2-Arctic 语料库中的三种口音以及 Accented Vox Populi（AVP）数据集 [35] 上评估了自动语音识别性能。AVP 由欧洲议会活动中录制的语音样本组成，因此包含许多欧洲口音。

3.3. 实验

在第一个实验中，我们训练了未使用任何带口音语音数据的基础版自动语音识别系统，并在 LS 的不同划分、L2-Arctic 中的阿拉伯语、印地语和普通话子集，以及 AVP 中的不同口音上评估其性能。随后，我们训练了三个自动语音识别模型，将真实训练数据与 53 小时的合成口音数据结合使用；这些合成口音数据来自用于自动语音识别测试的 L2-Arctic 口音，并采用 80% 真实数据和 20% 合成数据的比例。
两个自动语音识别模型还使用全部七种合成口音进行了优化。真实数据与生成数据的比例与第一个实验保持一致。

在最后一个实验中，我们通过 t-SNE图将 L2-Arctic 样本中学习到的口音嵌入进行可视化，以评估其在 ACM 中编码口音信息的适用性。

4. 结果

表 1 总结了各种实验设置的词错误率结果。与其他数据集相比，LibriSpeech 上的词错误率要低得多，因为子词词典仅根据这部分数据进行估计。因此，自动语音识别词汇表无法完整还原来自

实验	模型	LS 开发集		LS 测试集		L2-Arctic			AVP
		干净	其他	干净	其他	阿拉伯语	印地语	普通话	mean
仅 LibriSpeech	Base	3.8	10.3	4.1	10.2	24.7	23.5	33.9	37.9
	HuBERT	9.5	16.2	9.8	16.9	26.9	26.9	34.2	43.4
LibriSpeech 和阿拉伯语	Base	3.9	10.4	4.1	10.4	23.4	23.3	33.5	38.4
	HuBERT	9.3	15.8	9.6	16.0	31.2	29.5	35.4	42.3
LibriSpeech 和印地语	Base	3.8	10.3	4.1	10.3	24.7	22.0	33.8	38.3
	HuBERT	9.3	15.6	9.7	16.1	26.2	26.4	33.4	42.5
LibriSpeech与普通话	Base	3.9	10.4	4.1	10.5	24.9	23.5	32.7	38.9
	HuBERT	9.3	15.7	9.5	15.9	27.8	30.0	37.5	42.7
LibriSpeech与7种口音	Base	3.6	9.8	3.8	9.7	23.4	22.4	32.2	37.6
	HuBERT	9.5	15.8	9.8	16.1	29.4	28.0	37.3	43.1

表1： 美国英语（*LibriSpeech*）、阿拉伯语、印地语和普通话英语，以及一组欧洲口音（*AVP*）的词错误率。仅使用***LibriSpeech*** 是基线设置，识别器仅使用真实的美国英语数据进行训练。***LibriSpeech***与口音 将原始数据与一种合成口音相结合。***LibriSpeech***与***7***种口音将原始数据与所有七种非美国口音的合成数据相结合。

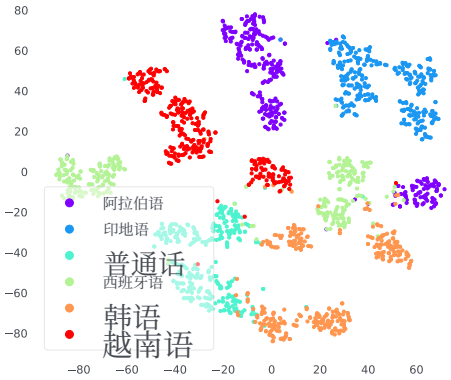

图2： 对来自*L2-Arctic*样本的学习得到的口音表示进行*t-SNE*嵌入。可以观察到一些口音簇，但它们在很大程度上受到说话人信息的干扰。

L2-Arctic 或 AVP。

在自动语音识别训练中加入阿拉伯语、印地语或普通话的合成带口音数据后，我们发现，Base模型在相应口音上的词错误率略有改善。然而，对于HuBERT模型，我们观察到了相反的情况：阿拉伯语和普通话的词错误率大幅上升，只有印地语改善了 0。5 %。与此同时，额外的带口音数据改善了LibriSpeech开发集和测试集部分上的所有词错误率。在欧洲口音（AVP）上，使用合成数据时，Base模型的表现始终更差，而HuBERT的表现始终更好。

在使用真实数据以及来自全部 7 种口音的合成数据训练两个自动语音识别模型后，我们发现 LibriSpeech 的噪声数据划分上的词错误率有所改善。只有 Base模型在干净开发集和测试集划分上也取得了小幅改善。它在 L2-Arctic 的所有口音上也有所改善，而欧洲口音（AVP）的结果则几乎保持不变。尽管 HuBERT模型在微调期间接触过带口音语音，但它在 L2-Arctic 数据上的结果仍然较差。与 Base模型一样，AVP 上的结果几乎没有变化。

图 2 可视化了从 wav2vec 2.0 Transformer输出中学习到的 t-SNE 嵌入口音表征。一些口音簇清晰可见，但另一些似乎毫无连贯性，反而与 L2-Arctic数据集中的每种口音对应的四位说话人各自相似。

5. 讨论

在自动语音识别训练中加入一种合成口音，对该特定口音的识别结果产生了积极影响，这清楚地表明 ACM 能够合成足够程度的口音。另一方面，HuBERT模型在测试带口音语音时受到合成数据的负面影响。一种可能的解释是，预训练期间缺少合成数据或带口音的数据。同时，使用合成数据对 HuBERT 进行微调，似乎略微提升了其在嘈杂的 LibriSpeech 上的鲁棒性。AVP 上的结果差异不大，但鉴于词错误率一开始就很高，已知的 LS 词汇与 AVP 词汇之间显然存在很大差距。AVP 上的结果均不足以让我们得出结论，认为任何模型都能够学会普遍的跨口音鲁棒性。

训练中纳入全部七种合成口音后，Base模型在所有实验中的表现均有所提升。这显然可能是发音变体鲁棒性增强所致。同样，这似乎并不适用于未见口音（AVP）。尽管 HuBERT 变体的架构与 Base模型非常相似，但它在带口音语音上的表现更差，因此，让这类模型在预训练阶段就接触带口音语音或许会有所帮助。由于 Base模型是从头开始优化的，它似乎能从带口音的训练样本中获益更多。

在本研究中，我们从未使用学习到的口音嵌入来生成样本。由于独立说话人的数量非常少，而每位贡献者提供的音频量相对较大，这些嵌入一开始学到的并非口音特征，而是对底层说话人进行分类，因此可能缺乏表现口音特定发音的能力。

6. 结论

合成带口音语音样本有助于提高自动语音识别系统应对系统性发音差异的鲁棒性。这种能力并不局限于单一口音，在训练期间纳入多种口音时也能观察到。然而，这些改进并未体现为普遍的口音鲁棒性，因为对于欧洲口音，并未报告出明显进步。学习得到的口音表示似乎能够编码口音，但其特征主要由说话人聚类决定，这表明所使用的带口音语音数据需要涵盖更多样的说话人。

7. 参考文献

[1] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Hen-retty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in第*12*届语言资源与评测会议论文集, 2020年, pp. 4218–4222.[2] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying bias in automatic speech recognition,” *arXiv*预印本 *arXiv:2103.15122*, 2021年.[3] N. Markl, “Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition,” in *2022*年*ACM*公平性、问责制与透明度会议, 2022年, pp. 521–534.[4] C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, A. Alvarado, C. Hazirbas, and Y. Saraf, “Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions,” in *ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*. IEEE, 2022年, pp. 6162–6166.[5] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, “Artie bias corpus: An open dataset for detecting demographic bias in speech applications,” in 第*12*届语言资源与评测会议论文集, 2020年, pp. 6462–6468.[6] J. E. Flege, “Phonetic approximation in second language acquisition 1,” *LanguageLearning*, vol. 30, no. 1, pp. 117–134, 1980年.[7] J. Li, V. Manohar, P. Chitkara, A. Tjandra, M. Picheny, F. Zhang, X. Zhang, and Y. Saraf, “Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings,” *arXiv*预印本 *arXiv:2110.03520*, 2021年.[8] A. Tjandra, D. G. Choudhury, F. Zhang, K. Singh, A. Conneau, A. Baevski, A. Sela, Y. Saraf, and M. Auli, “Improved language identification through cross-lingual self-supervised learning,” in *2022*年国际声学、语音与信号处理会议 *(ICASSP)*. IEEE, 2022年, pp. 6877–6881.[9] T. Viglino, P. Motlicek, and M. Cernak, “End-to-end accented speech recognition.” in*INTERSPEECH*, 2019年, pp. 2140–2144.[10] A. Jain, M. Upreti, and P. Jyothi, “Improved accented speech recognition using accent embeddings and multi-task learning.” in *INTERSPEECH*, 2018年, pp. 2454–2458.[11] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, “Domain adversarial training for accented speech recognition,” in*2018*年*IEEE*声学、语音与信号处理国际会议 *(ICASSP)*. IEEE, 2018年, pp. 4854–4858.[12] Y.-C. Chen, Z. Yang, C.-F. Yeh, M. Jain, and M. L. Seltzer, “Aip-net: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition,” in *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*. IEEE, 2020年, pp. 6979–6983.[13] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu 等, “End-to-end accent conversion without using native utterances,” in *ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*. IEEE, 2020年, pp. 6289–6293.[14] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, “Accent conversion using phonetic posteriorgrams,” in *2018*年国际声学、语音与信号处理会议 *(ICASSP)*. IEEE, 2018年, pp. 5314–5318.[15] Y. Zhang, Z. Wang, P. Yang, H. Sun, Z. Wang, and L. Xie, “Accentspeech: Learning accent from crowd-sourced data for target speaker tts with accents,” *arXiv*预印本 *arXiv:2210.17305*, 2022年.[16] M. Jin, P. Serai, J. Wu, A. Tjandra, V. Manohar, and Q. He, “Voice-preserving zero-shot multiple accent conversion,” 2022年。[在线]。可访问：https://arxiv.org/abs/2211.13282[17] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” 神经信息处理系统进展, vol. 33, pp. 17 022– 17 033

，2020年。

[18] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，神经信息处理系统进展，第 33 卷，第 12 449–12 460 页，2020年。[19] A. Graves，“连接时序分类”，载于使用循环神经网络进行监督式序列标注。施普林格，2012年，第 61–93 页。[20] D. S. Park、W. Chan、Y. Zhang、C.-C. Chiu、B. Zoph、E. D. Cubuk 和 Q. V. Le，“Specaugment：一种用于自动语音识别的简易数据增强方法”，*arXiv* 预印本 *arXiv:1904.08779*，2019年。[21] 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去分词器”，*arXiv* 预印本 *arXiv:1808.06226*，2018年。[22] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“Librispeech：一个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*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2015年，第 5206–5210 页。[23] Y. Shi、Y. Wang、C. Wu、C.-F. Yeh、J. Chan、F. Zhang、D. Le 和 M. Seltzer，“Emformer：一种用于低延迟流式语音识别的高效记忆变换器声学模型”，载于*2021*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第 6783–6787 页。[24] D. Hendrycks 和 K. Gimpel，“高斯误差线性单元（GELU）”，*arXiv* 预印本 *arXiv:1606.08415*，2016年。[25] J. L. Ba、J. R. Kiros 和 G. E. Hinton，“层归一化”，*arXiv* 预印本 *arXiv:1607.06450*，2016年。[26] S. Hochreiter 和 J. Schmidhuber，“长短期记忆”，神经计算，第 9 卷，第 8 期，第 1735–1780 页，1997年。[27] A. Graves，“使用循环神经网络进行序列转换”，*arXiv* 预印本 *arXiv:1211.3711*，2012年。[28] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于第 *3* 届国际学习表征会议（*ICLR 2015*），*2015*年 *9* 月 *7* 日，美国加利福尼亚州圣迭戈，会议论文集，Y. Bengio 和 Y. LeCun 编，2015年。 [在线]。获取地址：http://arxiv.org/abs/1412.6980 [29] G. Zhao、S. Sonsaat、A. Silpachai、I. Lucic、E. Chukharev-Hudilainen、J. Levis 和 R. Gutierrez-Osuna，“L2-Arctic：一个非母语英语语音语料库”，载于*INTERSPEECH*，2018年，第 2783–2787 页。[30] S. H. A. Baby 和 S. A. Kunath，“印度语言资源”，载于语言使用、语言学习与语言记录中的基于语料库的研究。博睿，2011年，第 265–281 页。[31] N. N. L. Arun Baby、Anju Leela Thomas 和 H. A. Murthy，“Speech Accent Archive：迈向英语口音类型学”，载于社区参与式语言资源建设（文本、语音与对话国际会议），2016年，第 37–43 页。[32] S. Aryal 和 R. Gutierrez-Osuna，“语音转换能否用于减轻非母语口音？”，载于*2014*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2014年，第 7879–7883 页。[33] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen 和 Y. Wu，“LibriTTS：一个从LibriSpeech衍生的文本转语音语料库”，*2019*年 *INTERSPEECH* 论文集，第 1526–1530 页，2019年。[34] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021年。[35] C. Wang、M. Rivi`ere、A. Lee、A. Wu、C. Talnikar、D. Haziza、M. Williamson、J. Pino 和 E. Dupoux，“VoxPopuli：一个用于表征学习、半监督学习和解释的大规模多语种语音语料库”，载于*ACL 2021*——计算语言学协会第 *59* 届年会，2021年。