有限数据下的带口音文本转语音合成

Xuehao Zhou，*IEEE*学生会员，Mingyang Zhang，*IEEE*会员，Yi Zhou，*IEEE*会员，
Zhizheng Wu，*IEEE*会员，Haizhou Li，*IEEE*会士，

arXiv:2305.04816v1 [eess.AS] 8 May 2023

摘要—**本文提出了一种在训练数据有限的情况下进行带口音文本转语音（TTS）合成的框架。我们研究了口音呈现的两个方面：语音（音素差异）和韵律（音高模式与音素时长）变化。所提出的带口音文本转语音框架由两个模型组成：用于字素到音素（G2P）转换的带口音前端，以及集成了基频和时长预测器、用于音位到梅尔频谱图预测的带口音声学模型。带口音前端直接建模语音变化，而带口音声学模型则显式控制韵律变化。具体而言，两个模型都先在大量数据上进行预训练，然后仅使用目标口音的有限数据微调口音相关层。在实验中，我们使用三种英语口音的语音数据进行预训练，即通用美式英语、爱尔兰英语和英国英语标准发音。随后分别使用苏格兰语和通用澳大利亚英语口音对预训练模型进行微调。客观和主观评估结果均表明，使用小型带口音语音词典（$5k$ 个词）微调的带口音TTS前端能够有效处理口音的语音变化；而使用有限的带口音语音数据（约3分钟）微调的带口音TTS声学模型能够有效改善韵律呈现，包括基频和时长。整体口音建模有助于提升语音质量和口音相似度。**

索引词—**文本到语音（TTS）、口音、语音变化、韵律变化**

I. 引言

STATE-of-the-art 神经端到端（E2E）文本到语音（TTS）系统 [1]–[5] 能够根据文本输入生成如真人般自然的语音。在有声读物、客户服务和电影配音等许多应用场景中，为了提升用户体验，需要带有口音的语音。不同于通过话语级说话人表征进行条件控制、生成目标说话人声音的多说话人TTS [6]–[8]系统，也不同于通过话语级情感嵌入来表征特定情绪的情感TTS [9]–[11]系统，口音的特征体现在话语的不同层面 [12]。因此，同时克隆说话人和口音颇具挑战，尤其是在目标口音可用数据有限的情况下。本研究旨在解决口音文本转语音合成中的此类问题。

当第一语言（L1）母语者由于 L1 与 L2 [13], [14] 之间语言系统的差异而使用第二语言（L2）进行表达时，便会产生外国口音。音段结构和超音段结构的发音模式会影响人们对外国口音的感知 [15], [16]。外国口音的属性主要可分为几类。进入音位和韵律的变化 [16], ，二者都是口音表征的必要组成部分 [17], [18]。

以英语为例，词汇词的语音转写会因口音而异。这类差异可从两个方面来描述 [19]: (a) 音位集合之间的差异，例如，英式和美式英语口音中都不存在元音集合 /ax, ea, ia, ua, oh/；(b) 语音转写之间的差异，例如，单词 ‘day’ 在美国口音中转写为 /dei/，而在澳大利亚口音中转写为 /dæi/。另一方面，基频 [20] 和时长等韵律属性已被证明有助于口音变换 [19]。一项关于土耳其语 [21], 的研究结果表明，基频 ($F0$) 和时长在口音呈现中起着关键作用。词汇重读音节与非重读音节在 $F0$ 峰值方面存在明显差异。在 [22][23]中，通过有选择地移植 $F0$ 和时长，观察到这两种韵律线索会显著影响口音特征。

要构建带口音文本转语音系统，研究问题在于如何有效建模特定口音的语音和韵律模式。在本文中，我们认为，口音可以由三个关键属性来刻画，即语音变化、音高模式和音素时长。我们提出了一种带口音文本转语音框架，由用于字素到音素（G2P）转换的带口音前端，以及用于音位到梅尔频谱预测的带口音的声学模型组成。随后，我们研究了如何利用有限的训练数据构建带口音文本转语音系统。具体而言，我们利用预训练模型，将从大规模通用数据中学到的共享知识迁移到带口音文本转语音系统中。

在本文中，我们提出了一种用于建模特定口音语音变化的带口音前端。该前端 [24] 学习将输入文本转换为目标口音下的音位序列。同时，我们还将基频预测器和时长预测器集成到声学模型中，以调节$F0$ 和音素时长，从而实现口音呈现。我们在两种场景下验证了所提出的技术。在第一种场景中，我们假设只有一个小规模的带口音的语音词典，并探讨如何处理TTS前端中的语音变化。在第二种场景中，我们进一步假设不仅只有一个小规模的带口音的语音词典，而且只有少量带口音语音数据。在语音词典和语音数据均有限的情况下，我们探讨如何在建模语音变化的同时建模韵律变化。本文的主要贡献总结如下:

- 我们通过处理语音的

三个关键属性来构建口音呈现问题，即语音变化、音高模式和音素时长。

- 我们通过使用小型带口音音标词典微调TTS前端，解决语音变异问题。
- 我们通过在TTS声学模型中纳入 $F0$和时长预测器来控制语音的韵律，并可使用有限的带口音语音数据对其进行微调。我们还将$F0$ 和时长作为额外输入提供给基于注意力的解码器。

本文其余部分安排如下。第二节介绍相关工作，为本研究作铺垫。第三节介绍我们提出的方法和框架。第四节和第五节分别展示实验设置和结果分析。第六节总结全文。

II. 相关工作

A. 传统带口音*TTS*合成

针对多种语言的带口音语音合成已有相关研究。在 [25], [26], 中，研究者使用单元选择工具Festvox研究了印度口音语音合成。Kolluru等人[27]提出了一种方法，通过构建口音空间，为单个说话人生成多种口音的发音。他们通过联合序列模型插值，将标准形式的音素序列转换为带口音的序列。插值权重作为口音空间中的点，用于表示不同口音。该方法能有效改变不同口音下音素的发音。然而，口音属性中的韵律变化，例如基频和音素时长，并未得到显式建模。

Anumanchipalli 等人 [28] 提出了一种语调模型，可为统计参数语音合成系统自动预测文本中适当的语调曲线。该语调模型将口音组（一系列语调事件）作为建模单元。这项工作揭示，口音成分的单元（例如 $F0$）与文本中的语言模式具有极高的相关性，这促使我们考虑使用 $F0$ 来控制口音呈现中的韵律变化。

B. 神经端到端带口音文本转语音合成

随着先进的基于神经网络的方法的发展，Abeysinghe 等人 [29]采用端到端TTS架构进行带口音语音生成。他们首先使用美国英语语音语料库对母语TTS模型进行预训练，然后使用非母语语音语料库进行微调。他们对微调阶段的元音空间变化进行了可视化和分析，并声称微调模型生成的语音，其元音空间比起母语语音，更接近非母语语音。Liu 等人 [30] 提出了一种可控口音文本转语音框架。他们提出了一种口音强度建模方法，用于量化每个样本的口音强度，并提出了一种受总文本转语音训练损失约束的一致性约束损失。结果表明，使用强度控制进行口音呈现是有效的。Melechovsky 等人[31] 提出了一种基于

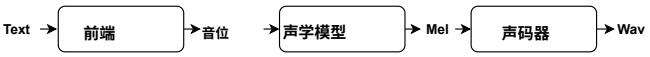


图 1. 典型文本转语音合成流水线的系统示意图。梅尔频谱指声学特征梅尔频谱图，而语音波形指语音波形。

条件变分自编码器的可控语音合成系统。该方法能够生成具有任意目标口音的特定说话人的语音。

然而，以上工作均未探讨英语口音的语音变化，而这正是口音变换中的一个重要方面。此外，先前的研究通常依赖大量训练数据，这限制了其在低资源环境中的应用范围。

本文旨在利用有限的训练数据，对带口音的文本转语音中的语音变化和韵律变化进行建模。为了对韵律信息进行建模，Yasuda 等人 [32] 通过在文本编码器中添加一个额外的自注意力层，捕获与音高重音相关的长程依赖关系，从而开发了一种端到端日语语音合成系统。此外，Yasuda 等人 [33] 研究发现，预训练 PnG BERT 能够为日语文本转语音系统捕获与音高重音相关的信息。他们将 PnG BERT 与文本转语音系统和声调预测任务结合进行微调，以促使 PnG BERT 丰富其音高重音信息。因此，我们考虑从文本编码器中控制口音呈现的韵律信息，并将其扩展到有限数据场景下的带口音语音生成。

III. 方法

如图 1 所示，文本转语音系统通常由三个组件构成一条流水线：前端、声学模型和声码器。前端在文本转语音系统中发挥着至关重要的作用，为其提供所需的音素相关语言知识 [34]–[36]。我们提出了一种由带口音前端和带口音声学模型组成的带口音文本转语音系统。带口音前端模型将文本输入转换为带口音的音素表示，以处理带口音的语音变异。我们采用当前最先进的端到端文本转语音系统之一——Tacotron 2 [2] 架构，作为带口音声学模型的骨干网络，以预测具有丰富口音表示的梅尔频谱图。我们采用 Parallel WaveGAN [37] 作为神经声码器，将预测得到的梅尔频谱图转换为时域语音波形。

我们研究带口音文本转语音框架中的两种情形：a) 目标口音仅有少量带口音的音标词典可用；b) 目标口音既有少量带口音的音标词典，也有数量有限的带口音语音样本可用。

A. 带口音文本转语音框架

1) 带口音前端：基于 Transformer 的模型 [38]以能够将字素转换为音素而闻名，且效果显著。然而，构建标准的基于 Transformer 的字素转音素模型需要大量的音标词典，这对于各种英语口音来说并不现实。据报道 [39], [40], 通过

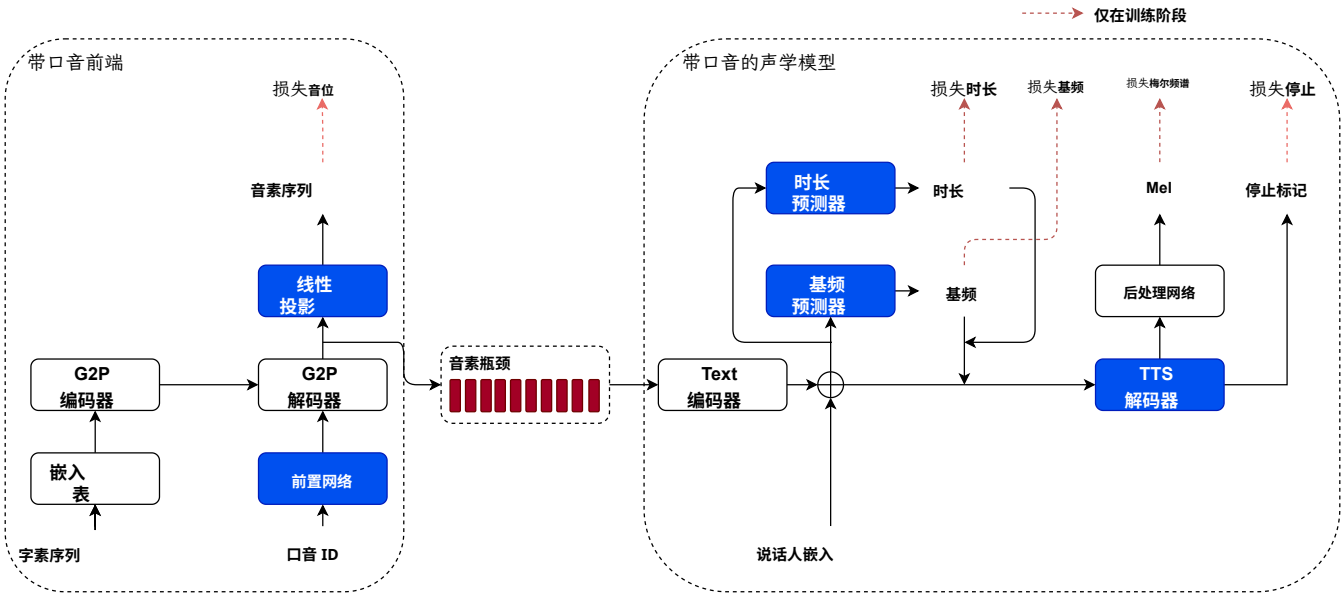


图 2. 所提出的带口音文本转语音框架的框图。左侧面板是用于字素转音素转换的带口音前端，右侧面板是用于音位到梅尔频谱图预测的带口音声学模型。从字素到音素解码器输出中提取的音素瓶颈，作为声学模型编码器的输入。前端和声学模型分别经过预训练，然后单独进行微调。图中所示的所有模块都参与预训练阶段。在微调阶段，仅更新蓝色模块中的参数，白色模块中的其他模块则保持固定。为简洁起见，省略了说话人编码器和神经声码器。

迁移预训练模型中的共享知识，可以提升低资源字素转音素模型的性能。因此，我们采用先预训练、再微调的技术，以规模有限的带口音语音词典构建带口音前端。

受多语言和多说话人语音合成研究 [41]–[43]，的启发，我们提出使用多个英语音标词典预训练多口音G2P模型，并使用目标口音的小型带口音语音词典微调口音相关层。多口音G2P模型与多语言模型类似，因为二者都以不同程度处理语音变化。早期的一项研究 [44]将语言 ID 作为额外输入来训练多语言G2P模型。本文使用口音 ID 来预训练多口音G2P模型。

图 2 的左侧面板是用于字素转音素转换的多口音前端。字素到音素模型将英语字素序列和口音 ID 作为输入，并生成音素序列作为输出。实际上，字素序列通过嵌入表转换为嵌入序列，然后通过一组 Transformer 编码器编码为潜在文本表示序列。潜在表示序列的长度与输入英语字素序列的长度相同。

对于每种口音，我们都有一份带口音的语音词典。一个英语字素序列可能对应多种语音转写，每种口音各对应一种。为了预测带口音的音素序列，我们将口音 ID 作为额外输入传递给解码器。音素序列和口音 ID 分别通过前置网络中的音素嵌入表和口音嵌入表转换为连续空间表示。口音嵌入是一个适用于整个音素序列的向量。具体而言，口音嵌入会在音素级进行复制，并与音素嵌入拼接。音素嵌入和口音嵌入的组合用作基于注意力的解码器的输入。

一组 Transformer 解码器将编码后的潜在文本表示与口音嵌入结合起来，以自回归方式生成输出序列。在当前步骤中，将前一步预测的音位作为解码器的输入。在训练过程中，G2P 解码器以教师强制方式将真实值音素序列作为输入。在运行时，解码器则将预测的音素序列作为上下文输入。最后，带有 softmax 函数的线性投影层将输出序列转换为离散的音素序列。损失函数定义为生成的音素序列与真实值之间的交叉熵。

2）带口音的声学模型：图 2 右侧展示了文本转语音编码器—解码器声学模型的一种变体。在标准的文本转语音系统中，嵌入表会将音位映射为连续空间中的嵌入向量。然而，一种口音中的某些音位可能不存在于另一种口音中。为了将针对一种口音训练的带口音前端泛化到另一种口音，我们提出使用字素到音素解码器输出的嵌入向量，即音素瓶颈，作为 TTS 声学模型的输入，而不是离散的音素 ID。通过这种方式，即使没有目标口音的语音训练数据，我们也希望能够处理未见音素的发音。

所提出的声学模型同样采用预训练和微调策略。它采用 Tacotron 2 架构，这是一种数据驱动模型，在文本转语音领域代表着最先进的性能。然而，Tacotron 2 并未明确

对口音的韵律属性进行建模，例如音高模式和音素时长。因此，我们将基频和时长预测器纳入 TTS 声学模型，使模型能够预测准确的基频和音素时长，使其接近真实值。

如图 2 所示，文本编码器以音位瓶颈特征作为输入，生成隐藏的音位表示。我们使用基于 d 向量的说话人嵌入对多说话人TTS进行预训练，以表示说话人身份。由于基频模式和音素时长也受单个说话人的影响，因此在基频和时长预测器之前，将说话人嵌入添加到文本编码器的输出中。文本编码器处理的音位建模是带口音语音表示的重要组成部分。与 FastSpeech 2 [5]，不同，本研究同时采用时长预测器和注意力机制，以便在语音数据有限的情况下更准确地确定音素边界；此外，将基频和时长预测器与文本编码器结合，在音位级预测 $F0$ 和时长。通过嵌入表和全连接层，分别将 $F0$ 和音素时长融入 TTS声学模型。基于注意力的解码器网络将文本编码器输出、说话人嵌入、 $F0$ 和音素时长组合起来，用于预测梅尔频谱图和停止标记标签。

在训练阶段，以教师强制的方式使用梅尔频谱图、$F0$和音位时长的真实值，而在运行时则使用预测值。对于梅尔频谱图、 $F0$和音位时长的预测，采用均方误差作为损失函数，而停止标记损失函数定义为二元交叉熵。所提出的TTS声学模型的总损失函数为：

$$Loss_{Total} = \alpha Loss_{Mel} + \beta Loss_{Stop} + \gamma Loss_{F0} + \delta Loss_{Dur} \quad (1)$$

其中， $Loss_{Mel}$ 是后处理网络前后梅尔频谱图重构损失之和。

B. 有限数据下的带口音*TTS*训练

1) 仅有少量带口音的音标词典：当有少量适用于目标口音的带口音音标词典可用时，我们只微调预训练多口音G2P模型中的口音相关层。口音相关层定义为与音位生成相关的层。与 [40], 类似，我们使用预训练模型进行初始化，并固定字素嵌入表、G2P编码器和G2P解码器。在 [40]，中，音素嵌入表和线性投影层会在微调阶段重新初始化，因为预训练阶段和微调阶段使用的音位集合不同。相比之下，我们在预训练和微调阶段对所有涉及的口音都使用统一的音位集合。因此，我们不进行重新初始化，而是基于预训练参数微调以下层：前置网络中的口音嵌入表和音素嵌入表，以及线性投影层。微调后的模块以蓝色绘出，如带口音前端

表 I 用于构建带口音的声学模型的数据库概览。

	预训练	微调	
数据库	VCTK	CMU ARCTIC	Google TTS API
口音数量	3	1	1
说话人数	45	1	1
话语数量	16,600	50 - 300	50 - 300
总时长	9.66 小时	2.63 - 14.96 分钟	3.10 - 18.29 分钟

模型的图 2 所示，而其他模块在微调阶段保持固定，并由所有口音共用。

2) 小型带口音语音词典与有限的带口音语音样本：除了微调预训练的字素到音素模型，我们还利用目标口音的有限语音样本，进一步处理生成的带口音语音中的韵律变化。具体而言，我们使用预训练的TTS声学模型进行初始化，并微调口音相关层以调整韵律成分。如图 2 所示，带口音的声学模型中经过微调的模块以蓝色绘制。

此前的一项研究 [45], 报告指出，由于语言信息分布不均，使用有限的语音数据微调文本编码器可能会降低文本转语音的性能。因此，我们在微调阶段冻结文本编码器。为了控制音素时长，我们微调时长预测器和注意力层。为了改变音高模式，我们微调音高预测器和解码器。解码器之后的后处理网络能够在预训练阶段提升模型在大量梅尔频谱图上的重建泛化能力，因此在微调阶段，我们固定预训练的后处理网络，并将其应用于所有口音。公式 1 中的所有损失函数都在微调阶段进行优化。

IV. 实验设置

我们开展实验，以验证所提出的带口音前端模型和带口音的声学模型在口音呈现方面的有效性。两个模型均先在大量数据上进行预训练，然后使用有限的带口音数据进行微调。

A. 带口音前端

我们首先使用通用美式英语、爱尔兰英语和英国英语标准发音的带口音的语音词典，对多口音G2P模型进行预训练。然后，分别使用苏格兰语和通用澳大利亚英语音标词典对预训练的G2P模型进行微调。所有音标词典均取自 Unisyn 词典 [46] ，该词典采用统一的音素符号集。在微调阶段，我们改变 $1k$、 $5k$、 $10k$和$40k$ 个最常用词的词典规模，并与约含 $120k$ 个词的完整词典规模进行比较，以观察其影响。训练G2P模型时，输入是一段话语，而非像 [40]中那样的单个词。文本转录选自 LibriTTS[47]语料库。过短或过长的话语，以及包含任何根据音标词典判定的词汇表外词的话语，

表 II
模型架构与网络概要
所提出的带口音文本转语音框架的配置
由带口音的字素到音素模型和带口音的文本转语音组成
声学模型。FC-X 表示具有 X 个单元的全连接层。
单元。CONV1D-K-C 表示宽度为 K、通道数为 C 的一维卷积
输出通道。LN 表示层归一化 ，以及 BN
表示批归一化。

字素到音素嵌入	字素嵌入表：256 维 位置编码器：256 维 → 丢弃（0.1）
字素转音素编码器	编码器堆栈数量：3 多头注意力：8 个 256 维的头 →丢弃(0.1) → 层归一化 前馈网络：全连接层-512-修正线性单元 → 丢弃(0.1) →FC-256 →丢弃(0.1) → 层归一化
字素转音素注意力	多头注意力：8 个 256 维的头 →丢弃(0.1) → 层归一化
字素转音素预网络	音素嵌入表：256 维 口音嵌入表：32 维 线性投影：FC-256 位置编码：256-D → 丢弃(0.1)
字素到音素解码器	解码器堆栈数量：3 掩码多头注意力：8 个 256 维注意力头 →丢弃(0.1) → 层归一化 前馈网络：全连接层-512-修正线性单元 → 丢弃(0.1) →FC-256 →丢弃(0.1) → 层归一化
字素到音素投影	FC-82
文本转语音编码器	音素线性投影：FC-512-软符号函数 3 层一维卷积-5-512-一维批归一化 → 双向长短期记忆网络-512 说话人线性投影：FC-512-软符号函数
TTS 注意力	位置层：一维卷积-31-32 →线性投影：全连接层-128
TTS 解码器	前置网络：FC-256-ReLU →丢弃(0.5) →FC-256-ReLU → 丢弃(0.5) 循环神经网络：长短期记忆网络-1024 → 丢弃(0.1) →长短期记忆网络-1024 → 丢弃(0.1) 梅尔线性投影：FC-80 停止标记线性投影：全连接层-1
TTS 后网络	4 层一维卷积-5-512-一维批归一化 → 一维卷积-5-80-一维批归一化
基频预测器	一维卷积-3-512-ReLU-层归一化 → 丢弃(0.5) → 一维卷积-3-256-ReLU-层归一化 → 丢弃(0.5) → 线性投影：全连接层-1 音高嵌入表：512-D
时长预测器	一维卷积-3-512-ReLU-层归一化 → 丢弃(0.5) → 一维卷积-3-256-ReLU-层归一化 → 丢弃(0.5) → 线性投影：全连接层-1 时长投影：全连接层-512-ReLU → 丢弃（0.5）
说话人编码器	3 层长短期记忆网络-768 → 线性投影：FC-256

根据音标词典，数据集中删除了包含任何词汇表外词的话语。

实现了两个前端模型以进行比较：

- **SA-G2P：** 这是一个仅使用通用美式英语训练的预训练单口音字素到音素模型。
- **MA-G2P：** 这是一个预训练多口音字素转音素模型，使用多种带口音的音标词典进行训练，包括通用美式英语、爱尔兰英语和英国英语标准发音。口音 ID 用作多口音G2P模型的额外输入。

基于 Transformer 的字素到音素模型的架构和网络配置如表 II 所示。每个字素到音素模型都经过 100 个训练轮次的预训练，并针对每个实验组再微调 50 个训练轮次。该模型采用预训练阶段验证损失最低的模型作为微调阶段的初始化模型，并使用微调阶段验证损失最低的模型进行测试。预训练和微调阶段的学习率均为 $5e$-4，批量大小均为 128。

B. 带口音的声学模型

我们首先使用 CSTR VCTK [48] 语音语料库的一个子集预训练多说话人TTS声学模型。为捕捉不同口音之间共享的知识，我们使用多种口音的语音数据预训练模型。我们选择与预训练字素到音素模型时相同口音的数据集，即通用美式英语、爱尔兰英语和英国英语标准发音。随后，我们分别使用 CMU ARCTIC [49] 语料库中的一位苏格兰口音说话人的数据，以及由 Google TTS API[1]，生成的一位澳大利亚口音说话人的数据，对预训练声学模型进行微调。在微调阶段，我们随机选取 50、100、200 和 300 个话语作为训练数据。相关数据库的详情汇总于表 I。所有语音信号均降采样至 $16k$ Hz，并剪去每个话语开头和结尾处的静音。我们使用对数尺度下的 80 维梅尔频谱图作为声学特征，该特征以 12.5 ms 的帧移和 50 ms 的帧长提取。

为训练音位级基频和时长预测器，我们使用自动语音识别（ASR）系统进行强制对齐，以获取音素边界。由于通用 ASR 模型无法很好地适应不同口音，我们训练了口音相关的 ASR，用于强制对齐[2]。音素时长表示为属于同一音位的帧数，并转换为对数尺度。使用 pyworld[3] 提取 $F0$，帧移为 12.5 毫秒。对 $F0$中的清音帧采用线性插值法。根据音素时长，将帧级 $F0$ 下采样至音位级。具体而言，我们对属于同一音位的帧上的$F0$ 值取平均。随后，对语音数据中的音位级 $F0$ 进行归一化，使其在预训练时具有零均值和单位方差。从预训练的说话人编码器中提取 256 维说话人嵌入。该说话人编码器使用 AISHELL-2 数据集 [50]，按照 [51]进行训练。

以下实现了六种带口音的文本转语音系统，以供比较：

- **Char-AM：** 这是一种多说话人TTS声学模型，以字符序列作为输入。
- **US G2P-AM：** 这是一种带口音文本转语音框架，由一个美式字素转音素模型和一个多说话人TTS声学模型组成。该美式字素转音素模型是预训练多口音字素转音素模型的一部分，并以通用美式口音 ID 为条件。
- **SCOT G2P-AM：** 这是一种带口音文本转语音框架，由苏格兰字素转音素模型和多说话人

[1]https://cloud.google.com/text-to-speech [2] https://montreal-forced-aligner.readthedocs.io/en/latest [3] https://pypi.org/project/pyworld

TTS声学模型组成。苏格兰字素转音素模型是经过预训练的多口音字素转音素模型，使用包含 $5k$ 个单词的苏格兰语音标词典进行了微调。

- **澳大利亚G2P声学模型：** 这与 SCOT G2P-AM 类似，不同之处在于，该字素到音素模型使用包含 $5k$ 个单词的通用澳大利亚英语音标词典进行微调。
- **苏格兰G2P基频时长声学模型：** 这是一种带口音文本转语音框架，由苏格兰字素转音素模型和集成了基频和时长预测器的多说话人TTS声学模型组成。苏格兰字素转音素模型是使用包含 $5k$ 个词的苏格兰语音标词典微调后的预训练多口音字素转音素模型。
- **澳大利亚G2P基频时长声学模型：** 这与 SCOT G2P-F0 时长声学模型相似，只是该字素转音素模型使用包含 $5k$ 个词的通用澳大利亚语音标词典进行了微调。

在以下实验中，TTS 系统名称后缀“-L”表示仅使用小型带口音的语音词典，“-S”表示仅使用有限的带口音语音数据，而“-LS”表示两者兼用。

基于 Tacotron 2 的声学模型的模型架构和网络配置总结于表 II 中。每个文本转语音声学模型均使用 Adam优化器 [52] 预训练 800 个训练轮次，并在每个实验组中再微调 100 个训练轮次。公式 1 中的所有权重 α、β、γ 和 δ 均设为 1。预训练阶段和微调阶段的学习率分别为 $1e$-3 和 $1e$-4，批量大小分别为 32 和 8。

C. 波形生成

在所有实验中，我们选择 Parallel WaveGAN [37] 作为神经声码器，将预测的梅尔频谱图重建为时域语音波形，因为它能够快速生成高质量的语音波形。该声码器使用 CSTR VCTK [48]语料库进行预训练，帧移和帧长与文本转语音声学模型相同。

V. 实验分析

我们进行了客观评估和主观评估。本节报告了两种场景下的实验：a) 仅有小型带口音的语音词典可用；b) 小型带口音的语音词典和有限的带口音语音样本均可用。

A. 仅有一个小型带口音的音标词典

在这种情况下，我们只对预训练的多口音字素到音素模型进行微调，同时冻结预训练的多说话人TTS声学模型。我们通过前端模型输出音素的准确度以及完整系统的输出语音，评估口音呈现中的语音变化。

1) 客观评估：我们通过音位错误率（音位错误率）和词错误率（词错误率）报告字素到音素模型的性能。音位错误率和词错误率分别表示预测序列与参考序列在音位和词级别上的莱文斯坦距离。较低的音位错误率和

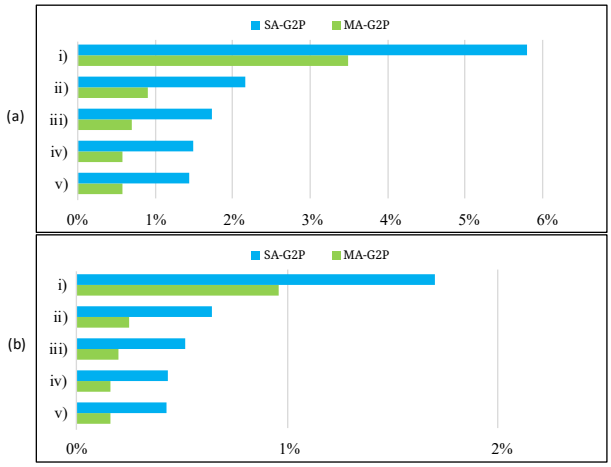


图 3. MA-G2P 和 SA-G2P 使用不同规模的苏格兰语音标词典进行微调后的词错误率和音素错误率表现。(i)–(v)分别表示词典规模为 $1k$、 $5k$、 $10k$、 $40k$ 个词以及完整词典。(a)词错误率(b)音素错误率。

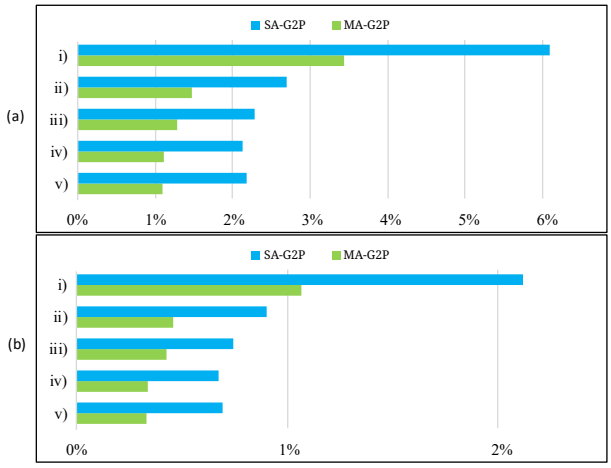


图 4. MA-G2P 和 SA-G2P 分别使用不同规模的通用澳大利亚语音标词典进行微调后，在词错误率和音素错误率方面的表现。(i)–(v)分别表示词典规模为 $1k$、 $5k$、 $10k$、 $40k$ 个词以及完整词典。(a) 词错误率 (b) 音素错误率。

WER 能够更准确地反映预测音素序列的情况。为了评估文本转语音系统输出语音的语音变化，我们计算生成语音与参考语音之间语音后验图的 Kullback-Leibler散度[53] （库尔贝克-莱布勒散度），以衡量语音分布的相似度。该值越低，性能越好。

i) 带口音前端的准确率：我们比较了两种字素转音素模型：通过微调 MA-G2P 得到的带口音的字素转音素模型，以及通过微调 SA-G2P [40] 在不同规模词典上得到的模型。苏格兰语和澳大利亚口音的结果分别见图 3 和图 4，结果显示，微调后的 MA-G2P 在所有测试案例中的表现均优于微调后的 SA-G2P。苏格兰语和澳大利亚口音的结果也呈现出一致的趋势。

这表明，MA-G2P 能够从不同口音之间的共享语音知识中获益。此外还观察到，苏格兰语的字素转音素效果优于澳大利亚的对应系统，这可以解释为苏格兰英语口音更接近

表 III 苏格兰口音／通用澳大利亚口音与通用美式口音之间的语音转写比较

	苏格兰语	澳大利亚的
共享词（%）	26.62	22.22
带口音的《Word》（%）	73.38	77.78
元音变异率（%）	24.07	41.16
辅音变体（%）	22.04	19.66

表 IV 三种对比文本转语音系统在苏格兰口音带口音的词语上的元音分布库尔贝克-莱布勒散度（KLD）

系统	KLD
Char-AM	6.60
US G2P-AM-L	5.36
SCOT G2P-AM-L	**5.15**

爱尔兰英语和英国英语标准发音，而非澳大利亚英语口音。

由于使用包含 $5k$ 个词的词典对 MA-G2P 进行微调后，性能已达到合理水平（苏格兰口音的音素错误率为 0.25%、词错误率为 0.90%；澳大利亚口音的音素错误率为 0.46%、词错误率为 1.46%），与使用完整规模词典时的性能相当（苏格兰口音的音素错误率为 0.16%、词错误率为 0.58%；澳大利亚口音的音素错误率为 0.33%、词错误率为 1.08%），因此，在数据有限的场景下，后续所有实验均选用使用包含 $5k$ 个词的带口音语音词典微调后的字素转音素模型作为带口音前端。

ii) 语音变化：我们对带口音的语音词典进行统计分析，以了解苏格兰口音／通用澳大利亚口音与通用美式口音之间的差异。我们统计两个词典中语音转写相同的共有词所占比例，以及两个词典中语音转写不同的带口音词语所占比例。我们还从元音和辅音变化的角度进一步分析带口音词语。具体而言，我们分别通过计算两个词典中带口音词语之间的莱文斯坦距离，得出元音和辅音差异率。如表 III 所示，元音变化比辅音变化更为显著。以往关于带口音语音语料库的研究也报告了相同发现 [19], [54] 。

因此，我们通过聚焦于带口音词语中的元音来评估口音呈现。从苏格兰口音测试集中选取了 191 个带口音的词语，其中共有 304 种元音变体。我们从预训练的说话人无关自动语音识别声学模型中提取语音后验图。随后，根据强制对齐得到的音素边界，将帧级后验图下采样至音位级。我们计算音位级后验图的库尔贝克-莱布勒散度，以比较元音分布的相似性，如表 IV 所示。我们可以观察到，生成的
SCOT G2P-AM-L 合成的语音，其元音分布比 US G2P-AM-L 合成的语音更接近带有苏格兰口音的参考语音。我们还观察到，G2P-AM-L 的表现均优于 Char-AM。

根据上述观察，我们可以得出以下两点结论：(a) 与字符输入相比，采用音素输入的文本转语音系统更

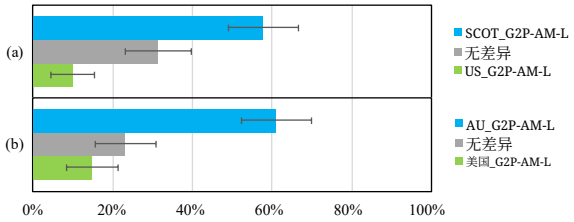


图5. 使用小型带口音的音标词典对字素转音素进行预训练和微调的带口音文本转语音系统在苏格兰口音和澳大利亚口音上的口音发音相似度 XAB测试结果，置信区间为95%。(a) 苏格兰口音。(b) 澳大利亚口音。

能准确地建模语音变化。(b) 使用小型带口音的语音词典对预训练的字素转音素进行微调，可以有效改变文本转语音系统输出语音的语音变化。

2) 主觀評估：主觀評估是透過人類受試者進行聽力測試來完成的。每組聽力測試實驗均有 20 位聽眾參與。所有聽眾皆為精通英語的大學生及教職工。 [4] 我們分別針對蘇格蘭語和澳大利亞口音中的帶口音的詞語進行 XAB偏好測試，以評估其口音發音相似度。在 XAB測試中，我們將每段話語中的帶口音的詞語標記為紅色。聽眾只需專注於標記的帶口音的詞語，並忽略其他詞語。每位聽眾會先聆聽參考語音，然後從兩個不同的樣本中選擇一個發音較接近參考語音的樣本。每位聽眾在單次實驗中會聆聽 18 個樣本，因此總計為 36 個（18×2 （口音數量）= 36）樣本。

如图5所示，将 US G2P-AM-L 与 SCOT G2P-AM-L 进行比较，以展示带口音的字素转音素对苏格兰口音的影响。此外，还将 US G2P-AM-L 与 AU G2P-AM-L 在澳大利亚口音方面进行比较。可以看出，在口音发音相似度方面，采用微调后的字素到音素模型的文本转语音系统显著优于采用预训练的美国英语字素转音素模型的系统，而且这一结果在苏格兰口音和澳大利亚口音中均一致。这些结果有力地支持了带口音的字素转音素建模有助于解决语音变异问题这一观点；该方法仅依赖包含 $5k$ 个词的小型带口音语音词典，无需在训练期间使用带口音语音样本。

B. 小型带口音语音词典与有限的带口音语音样本

在研究了使用小型带口音语音词典微调预训练的字素到音素模型所产生的影响后，我们进一步使用有限的带口音语音数据微调文本转语音声学模型，以研究韵律变化对口音呈现的影响。我们根据音高模式、音素时长、语音质量和口音相似度，对苏格兰口音和澳大利亚口音进行评估。所有实验系统均在相同的 100 条未见话语上进行测试。

[4]所有语音样本均可在以下网址获取：https://xuehao-marker.github.io/taslp G2P-TTS/

表V：在苏格兰语和澳大利亚口音上，使用不同规模的带口音语音数据微调的对比文本转语音系统的基频和时长客观评估结果。

口音	语句数 / 时长	系统	基频			时长	
			RMSE（Hz）	相关性	U/V（%）	扰动（帧）RMSE（ms）	
苏格兰语	-	Char-AM	30.21	0.67	16.35	21.53	45.48
		SCOT G2P-AM-L	35.33	0.67	15.27	19.18	51.49
		SCOT G2P-$F0$ 时长 AM-L	43.57	0.68	14.98	12.34	46.94
	50 / 2.63 分钟	Char-AM-S	20.09	0.75	14.08	13.10	47.40
		SCOT G2P-AM-LS	19.78	0.77	12.78	10.51	44.55
		SCOT G2P-$F0$ 时长 AM-LS	**17.47**	**0.81**	**11.61**	**10.24**	**42.66**
	100 / 5.03 分钟	Char-AM-S	17.74	0.78	13.10	13.04	43.90
		SCOT G2P-AM-LS	17.62	0.80	11.88	10.13	42.21
		SCOT G2P-$F0$ 时长 AM-LS	**16.64**	**0.83**	**11.63**	**9.85**	**40.79**
	200 / 9.93 分钟	Char-AM-S	16.85	0.74	12.93	13.25	45.22
		SCOT G2P-AM-LS	16.40	0.80	12.00	10.15	**40.46**
		SCOT G2P-$F0$ 时长 AM-LS	**12.92**	**0.82**	**11.28**	**9.50**	41.31
	300 / 14.96 分钟	Char-AM-S	15.93	0.78	12.66	12.89	44.02
		SCOT G2P-AM-LS	15.51	0.81	11.83	10.04	40.88
		SCOT G2P-$F0$ 时长 AM-LS	**12.26**	**0.82**	**10.45**	**9.47**	**40.60**
澳大利亚的	-	Char-AM	40.88	0.37	19.04	12.85	40.11
		澳大利亚 G2P-AM-L	42.24	0.47	17.03	13.44	36.36
		澳大利亚 G2P-$F0$ 时长 AM-L	40.48	0.54	17.73	9.97	41.01
	50 / 3.10 分钟	Char-AM-S	36.44	0.67	16.71	12.26	31.96
		AU G2P-AM-LS	31.91	0.73	15.25	11.30	30.05
		澳大利亚 字素转音素-$F0$ 时长 AM-LS	**29.02**	**0.78**	**15.12**	**7.13**	**28.98**
	100 / 5.94 分钟	Char-AM-S	35.07	0.70	16.41	11.59	30.38
		AU G2P-AM-LS	30.91	0.76	15.33	7.32	29.75
		澳大利亚 G2P-$F0$ 时长 AM-LS	**27.43**	**0.81**	**15.09**	**6.28**	**28.74**
	200 / 12.21 分钟	Char-AM-S	33.49	0.73	16.20	12.03	28.02
		澳大利亚G2P声学模型-LS	27.97	0.80	14.94	6.74	27.31
		澳大利亚 G2P-$F0$ 时长 AM-LS	**26.42**	**0.83**	**14.67**	**5.75**	**26.93**
	300 / 18.29 分钟	Char-AM-S	31.76	0.75	16.10	11.15	28.00
		澳大利亚G2P-AM-LS	27.17	0.81	14.73	6.83	27.02
		澳大利亚G2P-$F0$ Dur AM-LS	**25.58**	**0.83**	**14.57**	**5.47**	**26.39**
平均值	-	Char-AM-S	25.92	0.74	14.77	12.42	37.36
		G2P-AM-LS	23.41	0.78	13.59	9.13	35.28
		字素转音素-$F0$ 时长 AM-LS	**20.97**	**0.82**	**13.05**	**7.96**	**34.55**

1) 客观评估：我们计算 $F0$ 均方根误差（RMSE）、对数尺度 $F0$ 相关系数 [55], 以及清音/浊音（U/V）错误率，以评估基频幅度和轨迹趋势相似度。我们计算帧扰动 [56] 和音素时长均方根误差，以评估预测时长的准确性。使用动态时间规整（DTW） [57] 对生成的梅尔频谱与参考频谱进行对齐。为进一步证明在带口音语音数据有限的情况下，基频和时长预测器对微调的有效性，我们分别使用50、100、200和300个话语对TTS声学模型进行微调。平均客观测试结果汇总于表V。

i) 基频： $F0$ RMSE 定义如下：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{2}$$

其中， N 是帧的总数， x_i 和 $\hat{x}_i$ 分别是参考语音和生成语音在第 i^{th} 帧处的$F0$ 值。 $F0$ RMSE 越低，表示基频幅度误差越小。我们不计算对数 $F0$ RMSE [58], ，而是计算对齐后的生成语音和参考语音中所有浊音帧的原始 $F0$ RMSE。

对数尺度下的 $F0$ 相关系数定义为：

$$C = \frac{\sum_{i=1}^{N}(y_i - \overline{y}_i)(\hat{y}_i - \overline{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}}_i)^2}} \tag{3}$$

其中，$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y$ 和 $\overline{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$、 y_i 和 $\hat{y}_i$ 分别是参考语音和生成语音在第 i^{th} 帧的对数尺度 $F0$ 值， N 是参与计算的帧数。我们仅在对齐后的参考语音和生成语音中 $F0$ 的浊音帧上计算相关系数。相关系数的取值范围为 $[-1, 1]$ ，且越接近 1，表示基频轨迹趋势相似度越高。

我们根据清音/浊音错误率比较生成语音与参考语音的清音/浊音标签，该错误率是帧差异数占总帧数的比例。清音/浊音错误率越低，基频重建效果越好。

在表V中，我们有三点观察：(a) 所有使用带口音语音数据微调的TTS系统，性能都优于未使用此类数据微调的系统。这证实了在TTS声学模型中进行韵律建模的有效性。这也表明，基于未见过的d-向量的说话人嵌入能够表示说话人特征，但缺乏口音相关韵律

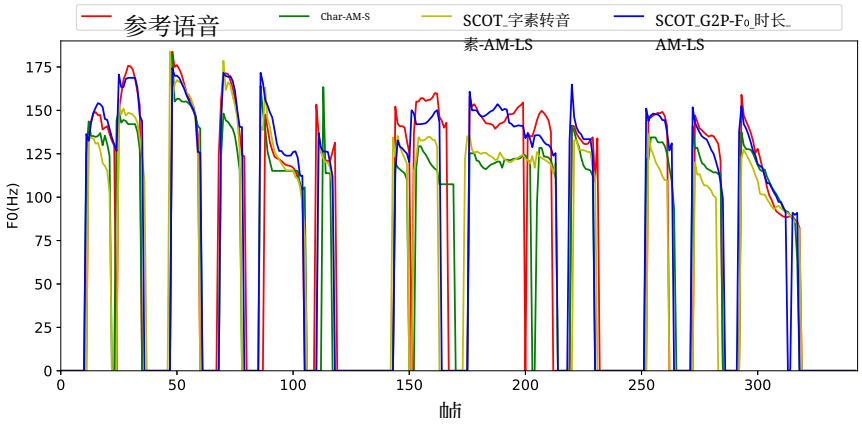


图 6。文本转写“I just do appreciate it without being able to express my feelings”对应的一段苏格兰语话语中的 $F0$ 轮廓。这些轮廓提取自参考语音，以及三个使用 300 条带口音话语进行微调的对比系统所生成的语音。

信息。(b) 在四种使用不同数量带口音语音数据进行微调的情形中，我们注意到，所有G2P-AM-LS的基频表现都优于Char-AM-S。这证实了字素到音素模型对韵律呈现的贡献。(c) 值得欣喜的是，我们提出的G2P-F0 Dur AM-LS在每种微调情形下，始终在所比较的三个系统中取得最低的 $F0$RMSE、最高的相关性以及最低的U/V错误率。这一观察有力地表明，即使目标口音的语音数据有限，音高预测器也有助于在生成语音中获得更准确的基频轨迹。在苏格兰语和澳大利亚语测试案例中，观察结果均一致。

我们进一步对三个对比系统生成的基频（$F0$）轮廓与参考语音进行了可视化。图6展示了一条苏格兰语话语的$F0$ 轮廓。我们发现，在所有系统中，SCOT G2P-F0 Dur AM-LS 的 $F0$ 轮廓始终与参考 $F0$ 最为相似，这也印证了表V中的客观评估结果。

ii）时长：帧扰动定义如下：

$$Disturbance = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_t - \hat{p}_t)^2} \qquad (4)$$

其中，p_t 和 $\hat{p}_t$ 分别表示参考语音与生成语音在 i^{th} 帧处对齐的路径，N 表示对齐帧的数量。扰动值越低，参考语音与生成语音之间的时长失真就越小。

音素时长均方根误差的定义与公式2类似，区别在于此处的 x_i 和 $\hat{x}_i$ 分别表示参考语音与生成语音中第 i^{th} 个音素的时长，N 表示音素的数量。时长均方根误差越低，音素时长重建效果越好。

在表V最后两列中，我们报告了苏格兰语和澳大利亚口音的时长评估结果。我们观察到的情况与音高评估相同：（a）使用有限的带口音语音数据进行微调后，时长重建变得更加准确；（b）整体表现

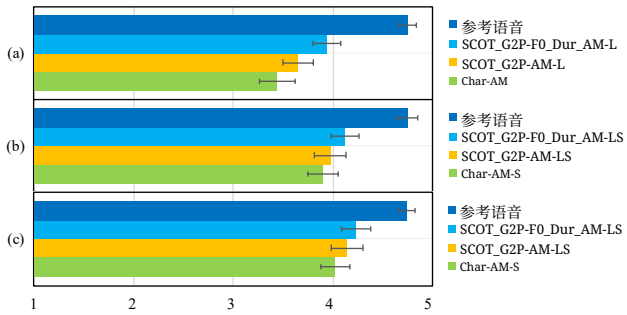


图7. Char-AM、SCOT G2P-AM-L、SCOT G2P-F0 Dur AM-L 和参考语音在苏格兰口音下自然度的MOS测试结果，以及95%的置信区间。(a) 未使用带口音语音数据进行微调。(b) 使用50个话语进行微调。(c) 使用300个话语进行微调。

三个经过相对微调的系统按性能从高到低依次为：G2P-F0 Dur AM-LS、G2P-AM-LS 和 Char-AM-S。这一观察结果在苏格兰语和澳大利亚口音上也一致。结果进一步证实了时长预测器对时长重建的贡献。

值得注意的是，基频和时长预测器高度依赖少量带口音语音训练数据。

2) 主观评估：我们通过平均意见分数（MOS） [59]评估语音的自然度，并进一步进行XAB偏好测试和最佳-最差量表（BWS） [60]测试，以评估口音相似度。

i) 语音质量：在MOS测试中，听者需使用1至5的五点评分量表，对所提供样本的自然度进行评分。每位听者在一次实验中评定24个样本，因此样本总数为144个（24×3 （微调案例数） $\times$ 2 （口音数） = 144）。我们在三种不同情形下评估三个对比TTS系统和参考语音：(a) 未使用带口音语音数据进行微调；(b) 使用50个带口音语音话语进行微调；(c) 使用300个带口音语音话语进行微调。评估涵盖苏格兰口音和澳大利亚口音。结果分别见图7和图8。

图7显示，SCOT G2P-F0 Dur AM-L在三个方面的得分均最高，其次是SCOT G2P-AM-L和Char-AM。这表明，音素

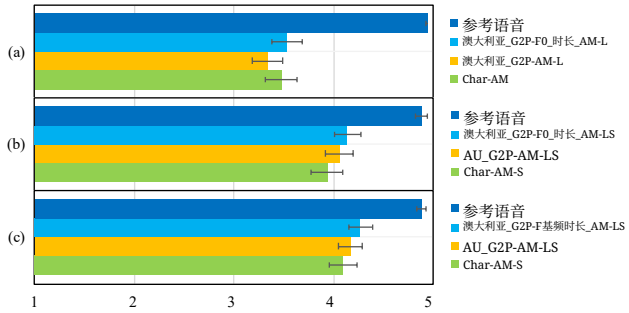


图8. Char-AM、AU G2P-AM-L、AU G2P-F0 Dur AM-L和参考语音在澳大利亚口音下的自然度MOS测试结果，附95%置信区间。(a) 未使用带口音语音数据进行微调。(b) 使用50条话语进行微调。(c) 使用300条话语进行微调。

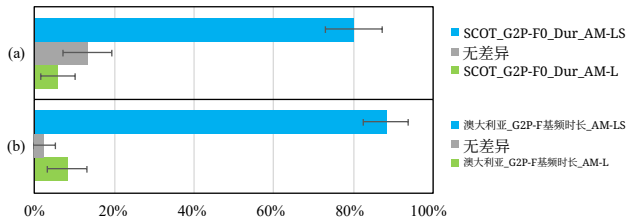


图9. G2P-F0 Dur AM-L和G2P-F0 Dur AM-LS在苏格兰口音和澳大利亚口音下的口音相似度XAB测试结果，附95%置信区间。(a) 苏格兰口音。(b) 澳大利亚口音。

输入、基频和时长预测器有助于生成质量更高、自然度更好的语音。从图8也可以得出澳大利亚口音的类似结论。不过，我们注意到，在图8（a）中，AU G2P-F0 Dur AM-L的表现并不明显优于Char-AM，甚至AU G2P-AM-L的得分还略低于Char-AM。而在图7（a）中，SCOT G2P-AM-L的表现则明显优于Char-AM。我们推测，与通用美式英语、爱尔兰英语和英国英语标准发音的预训练数据相比，澳大利亚口音的音素表示比苏格兰口音更具区别性。澳大利亚口音中的未见音素在一定程度上对语音质量和自然度产生了不利影响。这或许也能解释为什么图7（a）中的两个G2P-AM-L模型都比图8（a）中的对应模型表现更好。

ii) 口音相似度：在XAB偏好测试和BWS测试中，所有名称带有“-LS”的TTS系统均表示其TTS声学模型使用50条带口音语音话语进行了微调。

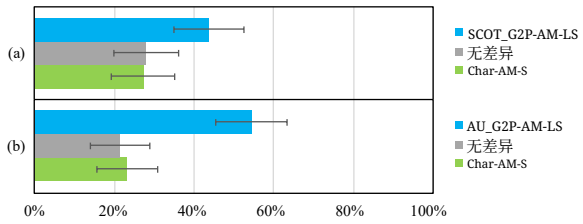


图10. Char-AM-S和G2P-AM-LS在苏格兰口音和澳大利亚口音下的口音相似度XAB测试结果，附95%置信区间。(a) 苏格兰口音。(b) 澳大利亚口音。

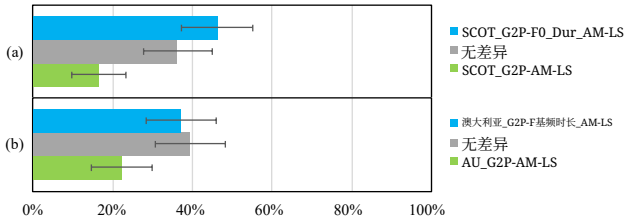


图11. G2P-AM-LS和G2P-F0 Dur AM-LS在苏格兰口音和澳大利亚口音下的口音相似度XAB测试结果，附95%置信区间。(a) 苏格兰口音。(b) 澳大利亚口音。

表VI 苏格兰口音和澳大利亚口音下四种对比系统的口音相似度BWS测试结果。

口音	系统	最佳（%）	最差（%）
苏格兰语	SCOT G2P-F0 Dur AM-L	7.5	78.33
	Char-AM-S	13.33	13.33
	SCOT G2P-AM-LS	25.83	5.00
	SCOT G2P-F0 Dur AM-LS	**53.33**	**3.33**
澳大利亚的	AU G2P-F0 Dur AM-L	0.83	86.67
	Char-AM-S	15.00	6.67
	澳大利亚G2P-AM-LS	36.67	4.17
	澳大利亚G2P-F0 Dur AM-LS	**47.50**	**2.50**

在XAB偏好测试中，听者先听一段参考语音，然后根据口音相似度，从两个不同的样本中选出与参考语音更相似的样本。每位听者在一次实验中听取18个样本，因此总计听取108个（18×3（比较对数）$\times 2$（口音数）$= 108$）样本。结果如图9-11所示。图9中显而易见，经过微调的G2P-F0 Dur AM-LS显著优于G2P-F0 Dur AM-L。这表明，使用有限的带口音语音数据微调TTS声学模型，能够改善生成的带口音语音的韵律呈现。图10显示，与使用字符输入相比，使用正确音素表示输入的TTS声学模型所生成的语音具有更高的口音相似度。这表明，与基于字符的编码器相比，基于音素的编码器能够更好地提取韵律相关文本信息。图11中令人欣喜的是，将基频和时长预测器纳入TTS声学模型后，生成语音的口音相似度感知明显提升。这有力地表明，集成基频和时长预测器对口音呈现十分有效。报告的所有结论在苏格兰语和澳大利亚口音上均一致。

在BWS测试中，我们向听众提供5个内容相同的样本。听众需要先听参考语音，然后根据口音相似度，从四个不同的样本中选出与参考语音最相似和最不相似的样本。每位听众在单次实验中听取30个样本，因此总共听取60个（30×2（口音数量）$= 60$）样本。苏格兰语和澳大利亚口音的结果如表VI所示。可以看出，在四个对比系统中，所提出的G2P-F0 Dur AM-LS系统获得了最高的最佳评分和最低的最差评分。我们还注意到，G2P-F0 Dur AM-

L系统的最佳评分明显最低，最差评分则最高。这表明，使用有限的带口音语音数据进行微调有助于口音呈现。从整体趋势来看，各系统在口音相似度方面的性能由低到高依次为：G2P-F0 Dur AM-L、Char- AM-S、G2P-AM-LS和G2P-F0 Dur AM-LS。所得结论与客观评估的结果一致。

VI. 结论

我们提出并验证了一种带口音文本转语音框架，同时处理口音呈现中的语音和韵律变化。研究表明，语音变化可以通过带口音前端建模，并且只需少量带口音的语音词典。同时，当可用的带口音语音数据有限时，可以通过带口音文本转语音声学模型，并显式控制基频和时长，来建模韵律变化。研究还表明，带口音前端也有助于带口音韵律呈现。本研究的关键发现是，只需有限的带口音词汇和语音数据，就可以有效地建模目标口音。未来，我们将探索对带口音前端和声学模型进行联合训练，以实现带口音文本转语音合成。

参考文献

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio 等，“Tacotron：迈向端到端语音合成”，*arXiv* 预印本 *arXiv:1703.10135*，2017。[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan 等，“通过以梅尔频谱预测为条件的 WaveNet 实现自然文本转语音合成”，载于*2018 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018，第4779–4783页。[3] N. Li, S. Liu, Y. Liu, S. Zhao, 和 M. Liu，“基于 Transformer 网络的神经语音合成”，载于*AAAI*人工智能会议论文集，第33卷，第01期，2019，第6706–6713页。[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, 和 T.-Y. Liu，“FastSpeech：快速、稳健且可控的文本转语音”，神经信息处理系统进展，第32卷，2019。[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, 和 T.-Y. Liu，“FastSpeech 2：快速、高质量的端到端文本转语音”，*arXiv* 预印本 *arXiv:2006.04558*，2020。[6] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, 和 J. Miller，“Deep Voice 3：利用卷积序列学习扩展文本到语音”，*arXiv* 预印本 *arXiv:1710 07654*，2017。。[7] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, 和 T.-Y. Liu，“Multispeech：基于 Transformer 的多说话人文本转语音”，*arXiv* 预印本 *arXiv:2006.04664*，2020。[8] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu 等，“从说话人验证迁移学习至多说话人文本转语音合成”，神经信息处理系统进展，第31卷，2018。[9] M. Whitehill, S. Ma, D. McDuff, 和 Y. Song，“基于对抗循环一致性的多参考神经文本转语音风格化”，*arXiv* 预印本 *arXiv:1910.11958*，2019。[10] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, 和 H. Meng，“借助跨域语音情感识别，使用无情感标签数据集实现情感可控语音合成”，载于*ICASSP 2021—2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021，第5734–5738页。[11] T. Li, S. Yang, L. Xue, 和 L. Xie，“端到端语音合成中的可控情感迁移”，载于*2021*年第*12*届中文口语语言处理国际研讨会（*ISCSLP*）。IEEE，2021，第1–5页。[12] M. J. Munro，“外国口音与语音可懂度”，音系学与第二语言习得，第5卷，第193–218页，2008年。

[13] J. E. Flege，《第二语言语音习得：理论、发现与问题》，《语音感知与语言经验：跨语言研究中的问题》，第92卷，第233–277页，1995年。[14] C. T. Best等，《婴儿母语音系影响的出现：一种知觉同化模型》，《语音感知的发展：从语音到口语词的转变》，第167卷，第224期，第233–277页，1994年。[15] A. Behrman，《口音管理的音段与韵律方法》，《美国言语语言病理学杂志》，第23卷，第4期，第546–561页，2014年。[16] J. J¨ugler、F. Zimmerer、J. Trouvain和B. M¨obius，《母语韵律移植对第二语言语音感知的影响：以带法国口音的德语为例》，载于《Interspeech 2016论文集》。美国旧金山。2016年9月8日至12日。国际语音通信协会，2017年，第67–71页。[17] L. Loots和T. Niesler，《不同英语口音发音之间的自动转换》，《语音通信》，第53卷，第1期，第75–84页，2011年。[18] P. B. De Mare¨uil和B. Vieru-Dimulescu，《韵律对外国口音感知的贡献》，《语音学》，第63卷，第4期，第247–267页，2006年。[19] Q. Yan、S. Vaseghi、D. Rentzos、C.-H. Ho和E. Turajlic，《英式、澳大利亚和美国口音的声学相关特征分析》，载于2003年IEEE自动语音识别与理解研讨会（IEEE目录号03EX721）。IEEE，2003年，第345–350页。[20] D. L. Bolinger，《英语音高重音理论》，《Word》，第14卷，第2-3期，第109–149页，1958年。[21] S. V. Levi，《土耳其语词汇重音的声学相关特征》，《国际语音学会会刊》，第35卷，第1期，第73–97页，2005年。[22] S. Winters和M. G. O’Brien，《感知口音程度与可懂度：基频和时长的相对贡献》，《语音通信》，第55卷，第3期，第486–507页，2013年。[23] L. Rognoni和M. G. Bus`a，《测试音段与超音段语音线索对外国口音评定的影响：一项采用韵律移植的实验》，载于《第二语言语音习得国际研讨会论文集》，《康考迪亚应用语言学工作论文集》，第5卷，2014年，第547–560页。[24] A. Deri和K. Knight，《适用于（几乎）任何语言的字素到音素转换模型》，载于《计算语言学协会第54届年会论文集》（第1卷：长篇论文），2016年，第399–408页。[25] M. Waseem和C. Sujatha，《使用Festvox合成带印度口音语音的系统》，《国际科学工程与技术研究杂志》，ISSN，第2319–8885页，2014年。[26] S. Kayte、M. Mundada和D. C. Kayte，《使用Festvox合成带马拉地语口音语音的系统》，《国际计算机应用杂志》，第130卷，第6期，第38–42页，2015年。[27] B. Kolluru、V. Wan、J. Latorre、K. Yanagisawa和M. J. Gales，《使用联合序列模型插值为文本转语音生成多口音发音》，载于国际语音通信协会第十五届年会，2014年。[28] G. K. Anumanchipalli、L. C. Oliveira和A. W. Black，《通过口音组建模改善统计参数语音合成中的韵律》，载于2013年IEEE声学、语音与信号处理国际会议。IEEE，2013年，第6890–6894页。[29] B. Abeysinghe、J. James、C. I. Watson和F. Marattukalam，《通过元音空间可视化文本转语音系统的模型训练》，arXiv预印本 arXiv:2208.09775，2022年。[30] R. Liu、B. Sisman、G. Gao和H. Li，《可控的带口音文本转语音合成》，arXiv预印本 arXiv:2209.10804，2022年。[31] J. Melechovsky、A. Mehrish、B. Sisman和D. Herremans，《基于条件变分自编码器的带口音文本转语音合成》，arXiv预印本 arXiv:2211.03316，2022年。[32] Y. Yasuda、X. Wang、S. Takaki和J. Yamagishi，《针对音高重音语言，研究采用自注意力的增强型Tacotron文本转语音合成系统》，载于ICASSP 2019-2019 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2019年，第6905–6909页。[33] Y. Yasuda和T. Toda，《研究用于音高重音语言文本转语音合成的日语PNG BERT语言模型》，《IEEE信号处理专题精选期刊》，2022年。[34] A. Parlikar、S. Sitaram、A. Wilkinson和A. W. Black，《用于字素到音素转换的Festvox Indic前端》，载于WILDRE：印度语言数据资源与评测研讨会，2016年。[35] J. Pan、X. Yin、Z. Zhang、S. Liu、Y. Zhang、Z. Ma和Y. Wang，《用于普通话文本转语音合成的统一序列到序列前端模型》，载于ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议

声学、语音与信号处理国际会议（ICASSP）。IEEE，2020年，第6689–6693页。[36] S. Bansal、A. Mukherjee、S. Satpal和R. Mehta，《利用混语字素到音素模型改进代码混合语音合成》，载于INTERSPEECH，2020年，第2957–2961页。[37] R. Yamamoto、E. Song和J.-M. Kim，《Parallel WaveGAN：一种基于生成对抗网络和多分辨率频谱图的快速波形生成模型》，载于ICASSP 2020-2020 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2020年，第6199–6203页。[38] Y. Sevinj、N. G´eza和G.-T. B´alint，《基于Transformer的字素到音素转换》，《Interspeech 2019论文集》，2099，2019年，第2095页。–。[39] L. Dong、Z.-Q. Guo、C.-H. Tan、Y.-J. Hu、Y. Jiang和Z.-H. Ling，《利用预训练字素模型进行神经字素到音素转换》，载于ICASSP 2022-2022 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2022年，第6202–6206页。[40]E. Engelhart、M. Elyasi和G. Bharaj，《用于方言迁移学习的字素到音素Transformer模型》，arXiv预印本 arXiv:2104.04091，2021年。[41] Q. Yu、P. Liu、Z. Wu、S. K. Ang、H. Meng和L. Cai，《利用多语言BLSTM学习跨语言信息，以合成低资源语言语音》，载于2016年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2016年，第5545–5549页。[42] A. Gutkin，《用于低资源语言统计参数语音合成的统一多语言多说话人声学模型》，2017年。[43] J.-H. Kim、S.-H. Lee、J.-H. Lee、H.-G. Jung和S.-W. Lee，《Gc-TTS：采用几何约束的少样本说话人适应》，载于2021年IEEE系统、人机与控制论国际会议（SMC）。IEEE，2021年，第1172–1177页。[44] A. Sokolov、T. Rohlin和A. Rastrow，《用于多语言字素到音素转换的神经机器翻译》，arXiv预印本 arXiv:2006.14194，2020年。[45] Z. Zhang、Q. Tian、H. Lu、L.-H. Chen和S. Liu，《Adadurian：基于Durian的神经文本到语音少样本适应》，arXiv预印本 arXiv:2005.05642，2020年。[46] S. Fitt，《英语发音词典的形态学方法》，国际语音通信协会，2001年。[47] H. Zen、V. Dang、R. Clark、Y. Zhang、R. J. Weiss、Y. Jia、Z. Chen和Y. Wu，《LibriTTS：从LibriSpeech衍生的文本转语音语料库》，arXiv预印本 arXiv:1904.02882，2019年。[48] C. Veaux、J. Yamagishi、K. MacDonald等，《CSTR VCTK语料库：用于CSTR声音克隆工具包的英语多说话人语料库》，爱丁堡大学。语音技术研究中心（CSTR），2017年。[49] J. Kominek和A. W. Black，《CMU Arctic语音数据库》，载于第五届ISCA语音合成研讨会，2004年。[50] J. Du、X. Na、X. Liu和H. Bu，《AISHELL-2：将普通话自动语音识别研究推向工业规模》，arXiv预印本 arXiv:1808.10583，2018年。[51] L. Wan、Q. Wang、A. Papir和I. L. Moreno，《说话人验证的广义端到端损失》，载于2018年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2018年，第4879–4883页。[52] D. P. Kingma和J. Ba，《Adam：一种随机优化方法》，arXiv预印本 arXiv:1412.6980，2014年。[53] J. M. Joyce，《Kullback-Leibler散度》，载于《国际统计科学百科全书》。Springer，2011年，第720–722页。[54] Q. Yan、S. Vaseghi、D. Rentzos和C.-H. Ho，《英式、澳大利亚和美国口音共振峰空间的分析与合成》，《IEEE音频、语音与语言处理汇刊》，第15卷，第2期，第676–689页，2007年。[55] H. Kameoka、K. Tanaka、D. Kwa´sny、T. Kaneko和N. Hojo，《ConvS2S-VC：全卷积序列到序列语音转换》，《IEEE/ACM音频、语音与语言处理汇刊》，第28卷，第1849–1863页，2020年。[56] R. Liu、B. Sisman、G. Gao和H. Li，《利用帧与风格重构损失训练表现力文本到语音系统》，《IEEE/ACM音频、语音与语言处理汇刊》，第29卷，第1806–1818页，2021年。[57] M. M¨uller，《动态时间规整》，《音乐与动作信息检索》，第69–84页，2007年。[58] H. Ze、A. Senior和M. Schuster，《使用深度神经网络进行统计参数语音合成》，载于2013年IEEE声学、语音与信号处理国际会议。IEEE，2013年，第7962–7966页。[59] R. C. Streijl、S. Winkler 和 D. S. Hands，“平均意见分数（MOS）再探：方法与应用、局限性及替代方案”，多媒体系统，第 22 卷，第 2 期，第 213–227 页，2016 年。[60] J. A. Lee、G. Soutar 和 J. Louviere，“最佳–最差尺度法：施瓦茨价值观调查的一种替代方案”，人格评估杂志，第 90 卷，第 4 期，第 335–347 页，2008 年。