arXiv:2301.04606v1 [eess.AS] 11 Jan 2023

无须口音专用 TTS 前端的低资源口音建模

Georgi Tinchev[*1], *Marta Czarnowska*[*1], *Kamil Deja*[2], *Kayoko Yanagisawa*[1], *Marius Cotescu*[1]

[1]Amazon Research，英国 [2]华沙理工大学，波兰

摘要

本研究聚焦于建模一种没有专用文本转语音（TTS）前端（包括字素转音素（G2P）模块）的说话人口音。以往的口音建模研究假定目标口音有可用的音标转写，而低资源地区口音可能并非如此。我们提出一种方法：首先通过语音转换扩增目标口音数据，使其听起来像供体语音；然后基于录音与合成数据的组合训练多说话人多口音文本转语音模型，从而生成以目标口音说话的供体语音。在整个流程中，我们使用为同一种语言的另一种口音开发的文本转语音前端。我们通过定性和定量分析表明，与其他生成模型相比，所提出的策略取得了最先进的结果。本研究证明，只需相对少量的数据，且无需开发口音专用文本转语音前端，即可对低资源口音进行建模。我们模型转换为多种口音的音频样本可在我们的网页上获取[3]。

索引词— 文本转语音，口音转换

1. 引言

近年来，文本转语音（TTS）领域取得了进展，研究者开始探索声学模型的多语言训练 [1, 2]。这类方法通常依赖大量多语言多说话人数据，并可通过针对说话人和口音信息进行条件控制，实现不同口音之间的转换。尽管已有研究尝试直接以字符作为模型输入 [3]，，许多方法仍然依赖音素输入，因为它（a）可以避免模型因未能正确习得某些词的发音而引入发音错误，并降低神经网络任务的难度 [4]；（b）可以对合成语音进行语音学控制 [5]。音素序列可以通过将文本输入文本转语音前端来生成。后者由多个模块组成，包括文本规范化和字素转音素（G2P）转换，通常依赖特定语言或口音。由于需要专业的语言知识，开发专用前端成本高昂 [6]。

本文提出一种文本转语音系统，能够使用选定的供体语音对目标口音进行建模，无需专用前端生成与目标口音相符的语音学转写文本。我们的模型在自然度和口音相似度方面取得了最先进的结果。即使训练所用的音素序列由并非专门针对该口音开发的前端生成，模型仍能泛化并学习不同口音之间音系对立的差异。总而言之，我们的贡献如下：1）据我们所知，我们提出了首个口音建模方法，无需明确地以针对目标口音的特定口音字素转音素模型所生成的音素作为模型的条件信息。与采用归一化流、变分自编码器或扩散模型构建的现有多语言多口音基线模型相比，我们的方法取得了最先进的结果；2）无论选择哪种文本转语音前端，我们都可以通过增强数据可靠地对口音进行建模；3）我们证明，仅使用目标口音的2000年条话语，即可通过合成数据实现口音建模。

2. 相关工作

本节总结了与口音建模任务相关的文献综述。综述分为两个部分——字素到音素（G2P）研究和口音转换研究。

文本转语音模型通常使用音素序列作为输入来建模语音，因为直接从字符输入进行建模会增加一项额外挑战：神经网络需要学习字符与发音之间的关系。字素转音素（G2P）研究传统上采用基于规则的方法 [7, 8, 9]，，随后发展为基于机器学习的方法，使用 n 元语法 [10] 或神经网络 [11, 12]。近期，研究者尝试利用距离度量和语言学专业知识，将这些方法应用于低资源语言 [13]。我们的工作重点是为一种没有专用前端的口音开发文本转语音系统，因为开发新的前端需要语言学专业知识，而且难以扩展。

历来，口音转换（AC）主要通过两种方式实现：利用语音变形组合频谱特征 [14, 15, 16]，以及通过频谱和音素帧匹配进行语音转换（VC）适配 [17, 18, 19]。此外，还有一类方法依赖外部特征，并在 VC 步骤后进行帧匹配 [20, 21, 22, 23, 24, 25]。语音变形方法的主要缺点是需要

[*]表示贡献相同。 [2]在亚马逊工作期间完成。[3] https://bit.ly/3V52ZrF

平行数据。对于帧匹配，很难转换目标语音的时长和语速。与现有的基于语音转换的方法进行口音转换不同，我们不需要自动语音识别模型来分离口音信息。相反，我们利用语音转换作为第一步，生成供体语音使用目标口音的数据，因为这样可以保留口音 [17]。

与我们类似，在带外国口音的语音识别领域，其他研究发现，通过语音和口音转换进行数据增强能够提升模型性能 [26, 27]。与我们最为相似的是， [28] 也利用数据增强进行口音迁移。关键的是，我们的模型无需目标口音的文本到音素转写文本，而且只需更少的数据即可对该口音建模——仅需 2.7 万条话语，而 [28] 中使用了 42.5 万条话语。

3. 方法论

我们的方法旨在为一种新的低资源口音（下文称为目标口音）训练文本转语音模型，无需特定口音前端来生成语音转写文本。为此，我们使用一个多说话人、多口音数据集。我们从非目标口音说话人中选取一位供体说话人。我们希望合成语音听起来像供体说话人，但带有目标口音。我们需要针对数据集中一种口音的字素转音素模型，并用它提取所有话语的音素序列。关键的是，对于其他任何口音，包括目标口音，我们都不需要特定口音字素转音素模型。本节将介绍该模型，以及如何为带有目标口音的供体语音生成增强数据。

3.1. 架构

我们将方法分为两个主要步骤——语音转换（VC）和文本转语音（TTS）。我们使用语音转换生成合成数据，让供体 语音以目标口音说话。随后，我们将这些扩增数据与原始录音一起用于训练多说话人多口音文本转语音系统。下面简要介绍这两个模型。

3.1.1. 语音转换

为了扩增文本转语音模型的训练数据，我们使用基于流的语音转换模型 [29, 30]。训练期间，模型将输入的梅尔频谱图编码为潜在向量 z，并以音素序列、说话人嵌入、基频和二元有声/清音标志作为条件信息。基频按话语进行归一化，以将说话人信息与话语韵律解耦。我们使用基于 [的预训练说话人嵌入

31]。模型使用负对数似然损失进行优化。在推理过程中，我们使用相应的条件信息，将目标口音说话人的梅尔频谱图编码为潜变量 z，然后再将其解码回梅尔频谱图，同时将说话人嵌入更改为供体说话人的说话人嵌入。我们首先使用整个数据集训练语音转换模型。利用该模型，我们将目标口音说话人的语音转换成供体说话人的声音。这样，我们就能够改变说话人身份，同时保留其余部分

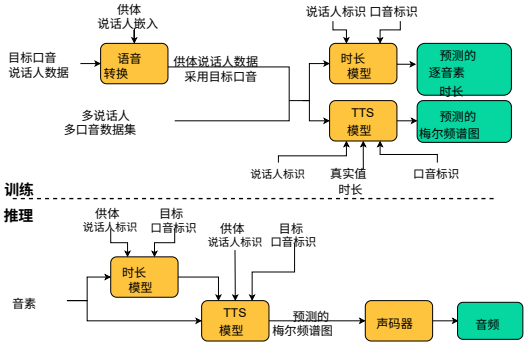


图1：文本转语音模型的训练与推理流程。说话特征，即口音、风格和时长，与 [17] 保持一致。随后，我们使用原始数据集和由语音转换模型生成的合成数据训练文本转语音声学模型和时长模型，具体内容见下一节。

3.1.2. 文本转语音

我们的文本转语音（TTS）架构是一个带有显式时长 [3, 32]的序列到序列模型。它包含两个独立的模型：声学模型和时长模型。

声学模型由三个部分组成：编码器、上采样层和解码器。编码器架构基于 Tacotron2 [3]，解码器则由残差门控卷积层和一个 LSTM 层组成。我们使用 Flow-VAE参考编码器 [2]。我们使用语音转换生成的目标口音下供体说话人的 2000年条话语训练该架构。然而，我们通过实验发现，在多说话人多口音数据集上训练能提升自然度（第4.4节）。因此，我们修改架构，加入说话人和口音条件信息。我们使用与模型共同训练的说话人嵌入和口音嵌入，为编码器和解码器提供条件信息。训练时，使用真实音素时长对编码后的音素序列进行上采样；推理时，则使用时长模型预测的时长。声学模型通过真实值与预测梅尔频谱图之间的 L2损失以及变分自编码器的 KL散度进行优化。

时长模型采用与声学模型相同的编码器架构，后接一个投影层。该模型还以说话人和口音信息为条件。模型通过真实值与预测音素时长之间的 L2 损失进行优化。在推理过程中，我们以目标口音和供体 说话人为条件，预测音素的时长。随后，我们运行声学模型进行推理，提供相同的说话人和口音条件、预测的时长，以及根据语音转换模型生成的合成数据计算得到的 VAE质心。我们使用通用声码器对预测的梅尔频谱图进行声码 [33]。

3.2. 数据集

在实验中，我们选择使用英国英语（en-GB）供体说话人来建模爱尔兰英语（en-IE）口音。我们使用 2.5 万条话语作为供体说话人 的数据，并为其余说话人使用 0.5 千至

表 1：与最先进水平的 MUSHRA 比较。在口音相似度评估中，我们选取了 4 位爱尔兰英语使用者的 25 个测试案例作为参考音频。我们使用了两个上限锚点：i）参考说话人的另一段录音；ii）另一位爱尔兰英语使用者的录音。

系统	自然度	口音相似度 所有说话人	口音相似度 说话人1	口音相似度 说话人2	口音相似度 说话人3	口音相似度 说话人4
上方锚点	74.42	82.24	72.56	89.33	84.79	79.24
不同说话人的上限锚点	N/A	52.55	43.40	49.42	57.10	54.89
所提方法	**64.61**	**50.73**	**61.40**	**40.81**	**45.16**	**50.44**
Grad-TTS	63.23	**50.68**	59.29	**40.77**	**46.32**	**50.71**
Flow-TTS	56.46	**50.32**	59.09	**40.28**	**45.19**	**51.21**

表 2：与最先进水平的 MUSHRA 比较。

系统	自然度	口音相似度
不同说话人的上限锚点	90.46	65.85
所提方法	**72.33**	**74.14**
Polyglot	47.24	49.29
下方锚点（英式英语）	N/A	12.17

表3：不同数据配置的消融研究。

系统	自然度	口音相似度
语音转换多说话人多口音	78.58	60.85
文本转语音多说话人多口音	**65.07**	**58.12**
文本转语音单说话人爱尔兰英语微调	59.07	**57.32**
下方锚点（英式英语）	N/A	11.50

4.5 千条话语，涵盖 6 种支持口音：英国英语（en-GB）、美国英语（en-US）、澳大利亚英语（en-AU）、印度英语（en-IN）、威尔士英语（en-GB-WLS）和加拿大英语（en-CA）[34]。对于 en-IE，也就是我们的目标口音，我们有由 12 位说话人录制的相同 2000年条话语。无论是语音转换模型还是文本转语音模型，都不要求具备这种类型的平行数据。在第4.4节中，我们将证明，仅使用单个说话人的 2000年条话语进行训练，就足以让我们的方法对该口音进行建模。我们使用英式英语前端提取音标转写文本，该口音与供体说话人的口音相同。我们对音素采用统一表示方式——将不同口音中的相同音素映射为相同的输入词元 [2]。

4. 实验

在本研究中，我们选择使用英式英语供体说话人对爱尔兰英语口音进行建模。本节将展示，与其他文本转语音模型相比，我们的方法如何实现最先进的结果。我们还使用两个不同的字素转音素模型生成音标转写文本（英式英语和美式英语），对所提出的模型进行定性和定量评估。最后，通过消融实验，我们将证明可以利用低资源数据对口音进行建模。

4.1. 评估方法

我们采用多刺激隐去参考音频与锚点（MUSHRA）测试，将所提方法与最先进的基线系统进行比较 [35]。每次评估均包含100个独特的测试用例，这些用例未在模型训练期间出现。24位爱尔兰语母语者对这些测试用例进行了评估，并以0至100分为各系统打分。我们评估了样本的自然度和口音相似度。评估自然度时，测试者需要判断样本听起来是否像真人说话；评估口音相似度时，则需判断样本与给定参考样本有多相似。参考音频和隐去的上方锚点均为爱尔兰英语说话者的录音。为确保结果具有统计显著性，我们进行了配对t检验，并采用Holm-Bonferroni校正（$p \leq 0.05$）。粗体结果表示各列中表现最佳的系统，差异达到统计显著性为止。

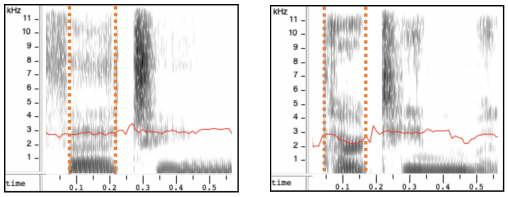
4.2. 与最先进水平的比较

我们将所提方法与 Polyglot——一种多说话人、多口音的基于序列到序列注意力机制的模型进行了比较 [34]。该模型使用特定口音的字素转音素生成音标转写，取得了最先进的结果。我们使用英式英语字素转音素模型提取的语音特征，分别训练了我们的方法和基线模型。表格2 显示，我们的方法在自然度和口音相似度方面的结果显著更好，这表明若没有特定口音的音标转写，仅靠口音和说话人条件信息不足以实现口音迁移。在口音相似度评估中，我们使用了另一位爱尔兰英语说话者作为参考音频和上方锚点，两者具有相同种类的爱尔兰英语口音。评估者善于辨别这种细微差异，因此给我们的模型打分高于上方锚点。因此，在后续实验中，我们引入了第二个上方锚点，使用参考说话人的其他录音。

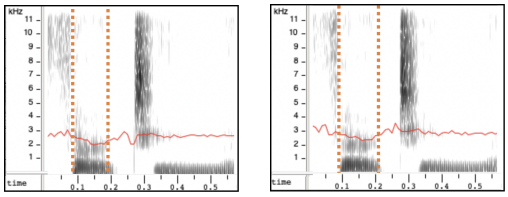
我们还将本方法与另外两个基线模型进行了比较。这两个模型使用相同的数据集进行训练，并使用英式英语字素转音素工具为每种口音提取音素：1）Grad-TTS [36] ，采用显式的说话人和口音条件；2）基于 Flow-TTS 的口音转换模型 [37, 29]。结果见表1。在自然度方面，本方法的表现显著优于这两个基线系统。由于 Grad-TTS 和 Flow-TTS 基线系统建模的是通用的爱尔兰英语口音，而非特定说话人的爱尔兰英语口音，我们决定使用 4 位不同的爱尔兰英语使用者作为参考说话人。当参考说话人与目标相同（我们使用说话人1的口音进行数据增强）时，本方法显著更优；但使用不同参考说话人（说话人2、3、4）时，模型之间的差异并不显著。这证实了使用语音转换样本能够使我们提出的方法更适用于特定说话人的口音。

4.3. 卷舌性分析

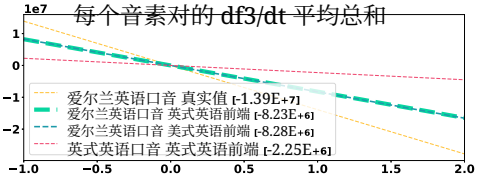
接下来，我们展示在使用缺少某项特征的口音所对应的前端时，如何复现目标口音 的特征。在上述所有实验中，我们展示了



(a) en-GB 合成 (b) en-IE 真实值



(c) 使用 en-GB G2P 的 en-IE (d)使用 en-US G2P 的 en-IE



(e) F3 导数平均和

图 2：(a-d) 第三共振峰轮廓（红色）显示了 /ɝ/ 或 /ɜː/ 元音在 “thirteen” 中的共振峰降低情况，橙色虚线之间标出了该位置。(e) 元音后 /r/ 语境平均斜率的示意图。

使用 en-GB G2P 训练的模型（“使用 en-GB G2P 的 en-IE”）的结果，该模型采用供体语音的口音。该模型能够生成口音更接近 en-IE 而非 en-GB（下方锚点）的样本，如表 2所示。en-IE 与 en-GB 的一个主要区别在于卷舌性：当 /r/ 后面没有元音时，它会在元音后语境中发音（例如，/kɑːr.pɑːrk/ 在“停车场”中）。另一方面，en-GB 是非卷舌音（/kɑː.pɑːk/）。尽管存在这一差异，使用 en-GB G2P 训练的模型仍能在合成的 en-IE 样本中复现卷舌性。

为了量化这一现象，我们分析了存在卷舌性对立的片段中的第三共振峰（F3），因为 F3 降低在声学上与音素 /r/ [38] 相关。作为比较，我们还训练了一个额外的模型，使用 en-US G2P（同样属于卷舌口音）提取音素（“使用 en-US G2P 的 en-IE”）。为了确定对立出现的区域，我们使用 en-GB 和 en-US 前端为测试集提取音素，并使用基于音素相似度的代价函数，通过动态时间规整（DTW） [39] 对齐音素序列。我们利用对齐结果，找出卷舌口音中存在 /r/、而非卷舌口音中不存在 /r/ 的语境。例如，在短语“car park”中，en-GB /ɑː/ 与 en-US /ɑːr/对齐。我们使用 Kaldi 外部对齐器 [40] 确定音频文件中各音素的位置，并通过 LPC 分析提取这些语境中的 F3，再计算其斜率。

图。 2 显示，单语英式英语（非卷舌音）模型生成的样本没有出现第三共振峰下降（2a），而



图 3：使用英式英语和美式英语前端训练的模型样本偏好测试结果。

真实值爱尔兰英语（卷舌音）录音则出现了这种现象（2b）。使用我们的方法训练的模型生成的爱尔兰英语样本显示出第三共振峰下降，无论模型是使用英式英语 G2P（2c）还是美式英语 G2P（2d）进行训练。这一点在图2e 中得到了证实，该图展示了在 134 个具有卷舌音对比的语境中第三共振峰斜率的平均梯度。“使用 en-GB G2P 的 en-IE”和“使用 en-US G2P 的 en-IE” 之间的梯度在统计上没有显著差异。

随后，我们对“使用 en-GB G2P 的 en-IE”和“使用 en-US G2P 的 en-IE”进行了偏好测试，24 名听众各自聆听了 100 对测试案例。图3 显示，听众显著偏好使用 en-US G2P 训练的模型，不过多数人没有偏好。这些结果表明，即使我们使用的前端所属口音不具备某项特征，我们仍能复现目标口音中的特征；但由于口音在许多方面都可能有所不同，我们可以通过谨慎选择用于数据标注的前端，进一步提升模型性能。

4.4. 消融实验

作为最后一项实验，我们针对两种采用不同数据配置的模型进行了消融实验。我们仅使用供体语音训练文本转语音模型，再使用目标口音合成数据进行微调。表3显示，多说话人、多口音模型与使用供体说话人训练并使用目标口音合成数据微调的模型，在口音相似度方面没有显著差异。与 [34], 一致，在多说话人、多口音场景下进行训练能提升自然度。这项实验的结果表明，对于口音建模任务，使用来自单个目标口音说话人的数据就已足够。

5. 结论

本文提出了一种口音建模方法，通过语音转换进行数据增强，并使用录音与合成数据相结合的数据训练文本转语音模型。重要的是，我们的方法不需要使用特定口音字素转音素模型生成的转录文本。我们表明，该模型能够可靠地建模目标口音的特征，且与现有口音建模策略相比，达到了最先进的性能。我们还表明，在目标口音数据较少的低资源场景中，我们的方法依然有效。我们还发现，为字素转音素选择更合适的口音，可以进一步提升性能。如何选择最适合字素转音素的口音，将留待未来研究。我们还计划将该方法扩展到英语以外语言的口音。

6. 致谢

感谢 Andre Canelas、Tom Merritt、Piotr Bilinski、Maria Corkery 和 Dan McCarthy 提供的反馈与见解。

7. 参考文献

[1] Y. Zhang、R. J. Weiss、H. Zen、Y. Wu、Z. Chen、R. Skerry-Ryan、Y. Jia、A. Rosenberg 和 B. Ramabhadran，“学习流利地使用外语交流：多语言语音合成与跨语言语音克隆”，*arXiv* 预印本 *arXiv:1907.04448*，2019年。[2] A. Sanchez、A. Falai、Z. Zhang、O. Angelini 和 K. Yanagisawa，“统一并征服：语音特征表示如何影响多语种文本转语音（TTS）”，发表于*Interspeech*，2022年。[3] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，“通过以梅尔频谱图预测为条件信息的 WaveNet 实现自然 TTS 合成”，发表于*ICASSP*。IEEE，2018年。[4] R. Skerry-Ryan、E. Battenberg、Y. Xiao、Y. Wang、D. Stanton、J. Shor、R. Weiss、R. Clark 和 R. A. Saurous，“利用 Tacotron 探索表现力语音合成的端到端韵律迁移”，发表于第*35*届国际机器学习会议论文集，J. Dy 和 A. Krause 编。2018年7月10日至15日，第80卷，机器学习研究论文集，第4693–4702页，PMLR。[5] A. Perquin、E. Cooper 和 J. Yamagishi，“探究字素嵌入与 Tacotron 系统发音之间的关系”，2021年。[6] M. Gakuru，“肯尼亚英语文本转语音系统的开发：为此前尚未定义的英语方言开发 TTS 的一种方法”，发表于*Speech Comm.*，2009年。[7] R. Weide 等，“卡内基梅隆发音词典”，*0.6*版，*www. cs. cmu. edu*，1998年。[8] C. R. Rosenberg 和 T. Sejnowski，“NETtalk：一种学习朗读的并行网络”，技术报告，约翰斯·霍普金斯大学出版社，1986年。[9] P. Kingsbury、S. Strassel、C. McLemore 和 R. MacIntyre，“Call-home 美式英语词典（Pronlex）”，语言数据联盟，1997年。[10] J. R. Novak、N. Minematsu 和 K. Hirose，“Phonetisaurus：在加权有限状态转换器框架中探索采用联合 n 元语法模型的字素到音素转换”，*Natural Language Engineering*，第22卷，第6期，第907–938页，2016年。

[11] G. Zhao、S. Ding 和 R. Gutierrez-Osuna，《用于字素到音素转换的 S2s 神经网络模型》，载于 *Interspeech*，2015 年。

[12] K. Rao、F. Peng、H. Sak 和 F. Beaufays，《使用长短期记忆循环神经网络进行字素到音素转换》，载于*2015 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，2015 年，第 4225–4229 页。[13] A. Deri 和 K. Knight，《适用于几乎任何语言的字素到音素模型》，载于 *ACL)*，2016 年。[14] M. Huckvale 和 K. Yanagisawa，《通过口音变换实现口语转换》，2007 年。[15] D. Felps、H. Bortfeld 和 R. Gutierrez-Osuna，《计算机辅助发音训练中的外国口音转换》，*Speech Comm.*，2009 年。[16] S. Aryal、D. Felps 和 R. Gutierrez-Osuna，《通过语音变形实现外国口音转换》，载于 *Interspeech*，2013 年。

[17] S. Aryal 和 R. Gutierrez-Osuna，“语音转换能否用于减轻非母语口音？”，载于*ICASSP*。IEEE，2014年。

[18] S. Liu、D. Wang、Y. Cao 等人，“无需使用母语者语句的端到端口音转换”，发表于 *ICASSP*。IEEE，2020年。

[19] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, 和 X. Li, 《多对多语音转换中的口音与说话人解耦》，载于 *ISCSLP*。IEEE，2021年，第 1–5 页。

[20] G. Zhao、S. Sonsaat、J. Levis、E. Chukharev-Hudilainen 和 R. Gutierrez-Osuna，《使用音素后验图进行口音转换》，载于 ICASSP。IEEE，2018年。[21] G. Zhao、S. Ding 和 R. Gutierrez-Osuna，《通过从音素后验图合成语音进行外国口音转换》，载于 Interspeech，2019年。[22] G. Zhao 和 R. Gutierrez-Osuna，《使用基于音素后验图的帧配对进行分段口音转换》，IEEE/ACM《音频、语音与语言处理汇刊》，2019年。[23] S. Ding、G. Zhao 和 R. Gutierrez-Osuna，《Accentron：利用零样本学习将外国口音转换为任意非母语说话人的口音》，Computer Speech & Language，第 72 卷，第 101302 页，2022年。[24] S. Ding、C. Liberatore、S. Sonsaat、I. Lučić、A. Silpachai、G. Zhao、E. Chukharev-Hudilainen、J. Levis 和 R. Gutierrez-Osuna，《黄金说话人构建器——一种用于发音训练的交互式工具》，Speech Comm.，第 115 卷，第 51–66 页，2019年。[25] C. Liberatore 和 R. Gutierrez-Osuna，《一种用于母语者与非母语者语音转换的范例选择算法》，载于 Interspeech，2021年，第 841–845 页。[26] T. Fukuda、R. Fernandez、A. Rosenberg、S. Thomas、B. Ramabhadran、A. Sorin 和 G. Kurata，《数据增强提高了对带外国口音语音的识别效果》，载于 Interspeech，2018年，9月，第 2409–2413 页。[27] H. Huang、X. Xiang、Y. Yang、R. Ma 和 Y. Qian，《面向带口音英语语音识别挑战赛的 Aispeech-SJTU口音识别系统》，载于 ICASSP。IEEE，2021年。[28] L. Finkelstein、H. Zen、N. Casagrande、C.-a. Chan、Y. Jia、T. Kenter、A. Petelin、J. Shen、V. Wan、Y. Zhang 等，《从合成数据训练文本转语音系统：口音迁移任务的实用方法》，载于 Interspeech，2022年，9月。[29] C. Miao、S. Liang、M. Chen、J. Ma、S. Wang 和 J. Xiao，《Flow-TTS：一种基于流的非自回归文本转语音网络》，载于 ICASSP。IEEE，2020年。[30] P. Bilinski、T. Merritt、A. Ezzerg、K. Pokora、S. Cygert、K. Yanagisawa、R. Barra-Chicote 和 D. Korzekwa，《使用归一化流创建新声音》，载于 Interspeech，2022年，第 2958–2962 页。[31] L. Wan、Q. Wang、A. Papir 和 I. L. Moreno，《用于说话人验证的广义端到端损失》，载于 ICASSP。IEEE，2018年。[32] R. Shah、K. Pokora、A. Ezzerg、V. Klimkov、G. Huybrechts、B. Putrycz、D. Korzekwa 和 T. Merritt，《采用显式时长建模的非自回归文本转语音：面向低资源、高表现力语音》，载于 Interspeech研讨会（SSW11），2021年。[33] Y. Jiao、A. Gabryś、G. Tinchev、B. Putrycz、D. Korzekwa 和 V. Klimkov，《使用并行WaveNet进行通用神经声码生成》，载于 ICASSP。IEEE，2021年。[34] Z. Zhang、A. Falai、A. Sanchez、O. Angelini 和 K. Yanagisawa，《Mix and Match：关于多语种文本转语音（TTS）训练语料库构成的实证研究》，载于 Interspeech，2022年。[35] I. Recommendation，《中间音质主观评估方法（MUSHRA）》，ITU，BS，2001年。[36] V. Popov、I. Vovk、V. Gogoryan、T. Sadekova 和 M. Kudinov，《Grad-TTS：一种用于文本转语音的扩散概率模型》，载于 ICML，2021年。[37] A. Ezzerg、T. Merritt、K. Yanagisawa、P. Bilinski、M. Proszewska、K. Pokora、R. Korzeniowski、R. Barra-Chicote 和 D. Korzekwa，《Remap, warp and attend：采用归一化流的非平行多对多口音转换》，载于 SLTW。IEEE，2023年。[38] C. Espy-Wilson、S. Boyce、M. Jackson、S. Narayanan 和 A. Alwan，《美式英语 /r/ 的声学建模》，《美国声学学会杂志》，第 108 卷，第 343–56 页，2000年8月。[39] M. M¨uller，《动态时间规整》，《音乐与动作信息检索》，第 69–84 页，2007年。[40] D. Povey、A. Ghoshal、G. Boulianne、L. Burget、O. Glembek、N. Goel、M. Hannemann、P. Motlicek、Y. Qian、P. Schwarz 等，《Kaldi语音识别工具包》，载于 ASRU研讨会。IEEE，2011年。