DARS：面向自动语音识别增强的构音障碍感知的韵律风格合成

吴明辉*，刘雪玲†，范家欢†，唐海涛†，张彦勇*，张越‡
* 中国科学技术大学，中国 电子邮箱：
wmhsky@mail.ustc.edu.cn，yanyongz@ustc.edu.cn
† 科大讯飞股份有限公司，中国
‡ 华为技术，中国

arXiv:2603.01369v1 [cs.SD] 2 Mar 2026

摘要—构音障碍语音具有异常韵律和显著的说话人差异性，给自动语音识别（ASR）带来了持续的挑战。尽管基于文本转语音（TTS）的数据增强展现出应用潜力，现有方法往往无法准确建模构音障碍语音的病理性节奏和声学风格。为此，我们提出了DARS，这是一种基于 Matcha-TTS 架构的构音障碍感知节奏风格合成框架。DARS 引入了一个多阶段节奏预测器，该预测器通过正常语音与构音障碍语音之间的对比偏好进行优化；同时采用构音障碍风格条件流匹配机制，共同增强时间韵律重建和病理性声学风格模拟。在 TORGO 数据集上的实验表明，DARS 的平均倒谱失真（MCD）为 4.29，生成语音与真实构音障碍语音十分接近。使用 DARS 合成的构音障碍语音对基于 Whisper 的 ASR 系统进行微调后，与最先进的方法相比，词错误率（WER）相对降低了 54.22%，证明了该框架在提升识别性能方面的有效性。

I. 引言

构音障碍是一种由神经系统疾病引起的运动性言语障碍，通常表现为构音含糊、语速减慢和韵律异常。这些表现会严重损害言语交流能力，并限制构音障碍患者的社会参与 [1], [2]。作为一种辅助工具，自动语音识别（ASR）有望帮助构音障碍患者 [3], [4]。然而，当前自动语音识别系统对构音障碍语音的识别准确率仍然有限。主要挑战来自两个关键因素：(1) 构音障碍模式的异质性导致说话人差异性显著 [5], [6], ；(2) 数据采集成本高、复杂度大，造成严重的数据匮乏。这些问题阻碍了针对构音障碍语音开发鲁棒自动语音识别系统 [7], [8]。

为应对数据匮乏，近期研究探索了基于文本到语音（TTS）的数据增强，以减少合成构音障碍语音与真实构音障碍语音之间显著的韵律失配 [9], [10], [11], [12], ，尤其是在节奏和风格分布方面。Wagner 等人 [10] 将大语言模型与可控文本转语音及基于x-vector的说话人自适应相结合，提升了重度构音障碍语音的处理效果。然而，该方法高度依赖提示工程，对语音风格的可控性有限。Leung等人 [11]采用基于扩散的方法来合成具有高语音质量和语义一致性的病理性语音，但在构音障碍节奏模式的建模方面不够精确。尽管 Soleymanpour 等人 [9] 引入了严重程度条件式风格控制器和停顿插入机制，但其方法仍难以捕捉轻度构音障碍病例中细微的语音特征。

为解决节奏建模精度和病理风格可控性方面的局限，我们提出了**D**ysarthria-**A**ware**R**hythm-**S**tyle（DARS）合成框架，该框架基于Matcha-TTS架构。DARS包含两种关键机制：

1）一种遵循“先停顿、后时长”策略的**多阶段节奏预测器**，通过正常语音与构音障碍语音之间的对比偏好优化（CPO）进行引导，从而更好地捕捉片段化的韵律模式；

2）一种**构音障碍感知条件流匹配**机制，该机制融入病理风格向量以约束合成过程，从而增强对构音障碍声学变化的建模。

这两种机制协同作用，显著提升了合成构音障碍语音的韵律一致性，为构音障碍自动语音识别系统生成了高度典型的训练数据。

II. 相关工作

A. MATCHA-TTS

鉴于基于扩散概率模型（DPM）的文本转语音系统 [13], [14], 采样过程计算开销较大，我们采用 Matcha-TTS [15]作为主干模型。该非自回归编码器-解码器（ED）框架能够以显著减少的推理步数实现高保真语音合成，同时无需人工标注即可实现输入文本与构音障碍语音之间的自动对齐。

编码器采用基于 Transformer 的架构 [16]，从文本中提取高层语义表示。单调对齐搜索（MAS）[17] 提供强制对齐信息，作为时长预测器的目标。推理期间，时长预测器的输出用于对编码器输出进行上采样，从而得到条件均值 μ。解码器采用基于 U-Net 的流预测网络 [18],，由下采样、中间和上采样模块组成。每个模块均由 Transformer 层构成，并且

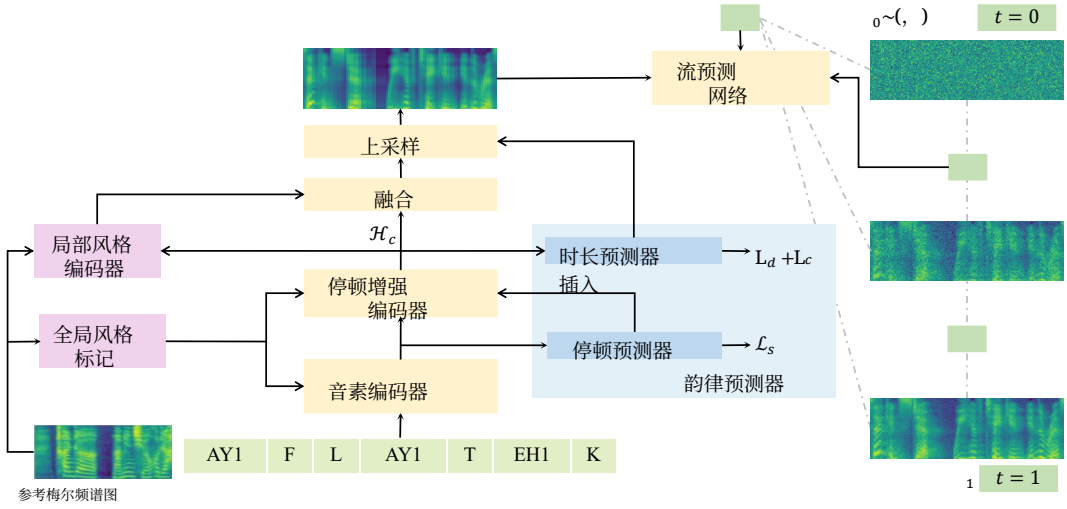


图 1. 基于 Matcha-TTS 模型的 DARS 框架

采用 BigVGAN [19] 中的 Snake-beta 激活函数来加速收敛。

在训练过程中，Matcha-TTS 模型采用最优传输条件流匹配（OT-CFM）策略 [14], [15], [20]，消除了扩散概率模型中冗余的多步去噪过程。该方法通过对条件流匹配向量场进行建模，简化了从随机噪声到目标语音的映射。通过使用一阶欧拉前向法求解相应的常微分方程（ODE），生成声学特征 [21],[22], ，推理步数减少了 50% 以上。

B. 基于 *Whisper* 的自动语音识别

Whisper 是应用最广泛的端到端语音识别框架之一 [23], ，采用基于 Transformer 的编码器-解码器架构。编码器接收梅尔频谱图作为输入，并生成高层语义表示；解码器则以自回归方式生成相应的文本序列。该模型使用约 68 万小时的多语言弱监督语音数据进行训练，因此能够在各种语音相关任务中实现出色的泛化。Whisper 发布了多个参数规模不同的版本。其中，Whisper-large [24]，拥有约 15.43 亿个参数，在一系列自动语音识别基准测试中表现领先，并且在低资源语言、嘈杂环境和多说话人场景中展现出很强的鲁棒性。

然而，由于 Whisper 几乎完全基于正常语音进行预训练，因此将其应用于构音障碍语音时，性能会显著下降。为了解决这一局限性，我们采用所提出的 DARS-TTS 系统生成大规模构音障碍韵律合成数据，用于 Whisper 微调。我们探索了两种适配策略：全参数微调和参数高效的低秩适配（LoRA） [25], ，以促进对构音障碍语音的有效适配。

III. 所提方法

基于 Matcha-TTS，我们引入了专为构音障碍语音设计的节奏和风格适配机制。以下各节将详细阐述这些机制。

A. 通过构音障碍引导的对比偏好优化的多阶段节奏预测器

Matcha-TTS 中的时长预测器用于根据输入文本估计时长。然而，由于文本通常按照正常语音进行预处理，其停顿模式往往与构音障碍语音中的模式存在显著差异。为了解决这一问题，我们引入了一种构音障碍语音节奏预测器（图 1 中的蓝色方框），它结合了多阶段预测和 CPO，以精确控制其独特的韵律模式。

1) 多阶段韵律预测：多阶段节奏预测器通过序列处理流程对病理性语音中的异常韵律模式进行建模。首先，音素编码器在不包含停顿信息的情况下对原始音素序列进行编码。随后，停顿预测器对音素之间可能出现的停顿类型进行分类，并将相应的停顿嵌入插入序列中。然后，停顿增强编码器对增强后的序列重新编码，以获得融入停顿线索的上下文表征。最后，这些表征被输入时长预测器，以估计音素级时长。这种级联设计将停顿显式融入时长建模，从而能够更准确地重构构音障碍语音韵律。

令音素编码器输出的高层音位表征序列为$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N]$。停顿预测器 f_θ^s处理$\mathbf{E}$并输出停顿类别分布$\mathbf{S} = [s_1, s_2, \cdots, s_N]$。真实标签$\mathbf{S}$根据使用蒙特利尔强制对齐器（MFA） [26]获得的对齐结果生成。具体而言，停顿按持续时间分为K 个离散类别：类别 0 表示没有停顿，类别 1 至 $K-1$ 表示

时长逐渐增加的停顿。我们使用交叉熵损失来指导停顿预测：

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\mathbb{I}(s_i = k)\log \hat{s}_{i,k} \tag{1}$$

其中，$\hat{s}_{i,k}$ 表示第 i 个位置对应停顿类别 k 的预测概率。

完成停顿预测后，将停顿嵌入（对应于 $\mathbf{S}$）插入 $\mathbf{E}$，生成增强序列$\mathbf{E}^{\text{aug}}$。停顿增强编码器处理$\mathbf{E}^{\text{aug}}$，以获得融入停顿信息的上下文表征。随后将这些表征输入时长预测器 f_θ^t 来估计时长。我们使用对数尺度时长的均方误差（MSE）训练时长预测器：

$$\mathcal{L}_d = \frac{1}{N'}\sum_{i=1}^{N'}(\log d_i - \log \hat{d}_i)^2 \tag{2}$$

其中，d_i 是第 i个元素的预测时长，d_i是通过 MAS 对齐方法 [13], [15], 获得的参考时长，$N' \geq N$ 表示插入停顿后序列的长度。

2）构音障碍引导的对比偏好优化：考虑到构音障碍语音模式的复杂性和多变性，我们提出一种基于多阶段韵律预测模块的构音障碍引导对比偏好学习机制。该机制促使合成韵律风格更贴近真实构音障碍语音的分布，同时偏离正常语音的分布。我们将对比偏好损失函数 $\mathcal{L}_{cp}$ 定义如下：

$$\mathcal{L}_{cp} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \max\left(0,\, \mathrm{D}(\hat{d}_i, d_i) - \mathrm{D}(\hat{d}_i, d_i^n) + m\right) \tag{3}$$

$$w_i = \begin{cases} \alpha \cdot \hat{s}_{i,s_i} & \text{if } s_i > 0 \\ \beta \cdot \hat{s}_{i,0} & \text{if } s_i = 0 \end{cases} \tag{4}$$

$\mathrm{D}(d_i, \hat{d}_i)$ 表示预测时长 d_i 与构音障碍语音真实时长 $\hat{d}_i$之间的距离，而 $\mathrm{D}(d_i, d_i^n)$ 表示 d_i 与正常语音时长 d_i^n 之间的距离。在我们的实验中，该距离通过绝对差来衡量。间隔 m 是预先设定的阈值，通常设为 0.75。对比损失 $\mathcal{L}_{cp}$ 确保预测时长与构音障碍语音真实时长的距离至少比与正常语音对应时长的距离小 m。动态权重 w_i 根据模型预测的类别概率定义。对于标记为停顿的位置（$s_i > 0$），权重与 $\hat{s}_{i,s_i}$成正比，即真实停顿类别的预测概率。对于非停顿位置（$s_i = 0$），权重取决于 $\hat{s}_{i,0}$，即分配给非停顿类别的概率。参数 α 和 β 控制损失函数中分配给停顿和非停顿位置的相对权重。我们通过实验将 $\alpha \geq \beta$ 设为较大的值，以强调准确的停顿建模；这对于捕捉构音障碍语音特有的非典型韵律模式至关重要。

B. 构音障碍感知声学条件流匹配

Matcha-TTS 仅根据条件均值 $\boldsymbol{\mu}$预测流匹配向量场，这些条件均值由音素序列$\mathbf{E}$ 和说话人身份$\mathbf{v}$得出。然而，在构音障碍语音合成中，这种设计可能会忽略声学特征中蕴含的丰富病理性韵律模式。为解决这一局限，我们提出了一种构音障碍感知声学条件流匹配机制，引入额外的全局和局部声学风格向量 $\mathcal{A}_g$ 和 $\mathcal{A}_l$，并将它们与内容相关条件融合，共同调节流匹配过程。

具体而言，我们从参考梅尔频谱图中提取全局风格标记（GSTs） [27], 作为全局韵律表征 $\mathcal{A}_g$。与此同时，我们通过一个改编自 [28],的局部风格编码器对帧级风格变化进行建模，得到 $\mathcal{A}_l$。该编码器采用向量量化瓶颈，从参考梅尔频谱图中提取风格表征，并通过注意力机制将其与停顿增强编码器的输出隐藏状态 $\mathcal{H}_c$ 对齐。随后，将对齐的局部风格特征$\mathcal{A}_l$ 与 $\mathcal{H}_c$ 融合，以此为条件生成条件均值 $\boldsymbol{\mu}$。完整的向量预测网络定义如下：

$$\boldsymbol{\mu} = \text{Encoder}(\mathbf{E}, \mathcal{A}_g, \mathcal{A}_l) \tag{5}$$

$$\mathbf{v}_t(\mathbf{x} \mid \boldsymbol{\mu}; \boldsymbol{\theta}) = \text{Decoder}(\mathbf{x}, \boldsymbol{\mu}, t) \tag{6}$$

其中，Decoder(·)是一个基于U-Net的向量场预测网络，$\boldsymbol{\theta}$ 表示模型参数，$\mathbf{x}$ 是潜变量，$t \in [0, 1]$ 表示流匹配步。

最终，我们的模型采用了Matcha-TTS [15],中的条件流匹配目标 $\mathcal{L}_{\text{CFM}}$，旨在学习一个向量场，将简单先验分布中的样本映射为构音障碍声学特征。除了与内容相关的信息之外，我们的方法还将构音障碍声学风格表征纳入向量场预测过程，从而增强模型重建病理性语音中典型声学异常的能力。

IV. 实验

A. 实验设置

TORGO 数据库被用作构音障碍语音数据集 [29]。其中包含 8 名构音障碍说话人和 7 名健康对照说话者的录音。8 名构音障碍说话人均被诊断患有肌萎缩侧索硬化症（ALS）或脑性瘫痪（CP）。如表 I 所示，根据他们的言语障碍程度，这些说话人被划分为四个严重程度等级：重度、中重度、中度和轻度。在说话人描述中，“F”和“M”分别表示女性和男性，后面的数字表示其在数据集中的参与者编号。

表 I TORGO 数据库的构音障碍严重程度。

	重度		中重度	中度		Mild
参与者	F01	M01 M02 M04		M05	F03	F04 M03
数量	228	739	772	659	610	1097	675 806
语句						

鉴于TORGO数据集的资源有限，我们研究了Matcha-TTS模型的训练策略，以尽可能减少数据使用量：

- 全说话人（ASp）：使用所有构音障碍说话人的合并训练数据训练单个文本到语音模型。
- 单说话人（SSp）：针对每位构音障碍说话人（F01～M03），仅使用其对应的语音数据分别训练八个独立模型。
- 构音障碍严重程度分组说话人（DSpG）：根据构音障碍的严重程度等级对数据进行分组，并分别独立训练四个文本转语音模型。

对于 Matcha-TTS 模型架构和训练策略，我们采用 [15] 中所述的 MAT-10 配置。不过，我们扩大了文本编码器的参数规模，其包含 8 层，每层具有 4 个注意力头和 1024 个滤波器通道。在我们提出的 DARS 框架中，音素编码器和停顿增强编码器采用与原始 Matcha-TTS 编码器相似的架构。具体而言，音素编码器是一个较小的变体，仅包含 2 层，每层具有 4 个注意力头和 512 个滤波器通道；而停顿增强编码器规模更大，与前文所述的文本编码器相当。停顿预测器和时长预测器采用相同的网络结构。有关 GST 编码器的更多细节，请参阅 [27]，；帧级局部风格编码器的实现则遵循 [28]。在训练 DARS 模型之前，我们使用一个在完整 TORGO 数据集上训练的 MFA 模型 [26] 提取停顿标签。在 CPO 中，正常语音的时长预测器在 LibriSpeech 英语数据集 [30] 上进行预训练，以获得与输入文本对应的参考时长。

对于构音障碍语音识别，我们使用合成语音数据对 Whisper-Large 模型进行微调，并在原始数据上评估其性能。由于 TORGO 数据集未提供预先划分的训练集、验证集和评估集，我们采用多重采样策略，分别选取 10%、20% 和 30% 的数据作为评估集。这些子集经过筛选，以确保使用公开可用的 Whisper-Large 模型 [23]进行评估时，三个划分之间的性能差异最小。随后，将剩余数据按 9:1 的比例划分为训练集和验证集，这些数据也用于训练 Matcha-TTS 模型。接着，使用训练好的 Matcha-TTS 合成训练数据，并将其用于 Whisper-Large 的适配。

我们使用平均倒谱失真（MCD） [11], [31] 评估合成梅尔频谱图与真实梅尔频谱图之间的相似度。在语音合成评估中，MCD 已被证明与主观评估结果相关。对于自动语音识别性能评估，我们使用词错误率（WER） [23]作为主要指标。此外，通过对各个说话人的 WER 分数取平均，计算总体词错误率。

B. 实验结果

不同采样比例下评估集的稳定性。表 II 展示了开源 Whisper-Large 模型在按不同比例抽样的 TORGO 测试集上的表现。这些评估集（E1–E3）经过多次抽样和识别，结果表明其在不同采样比例下均具有稳健的稳定性。基于这一一致性，我们选择 20% 的评估集用于所有后续实验。其余 80% 的数据用作语音合成和识别任务的训练与验证数据。

表 II 按句子数量比较 TORGO 评估集上的词错误率。

标识符	Whisper-Large	词错误率 (%)				
		重度	中重度	中度	轻度	总体
E1	10%	125.67	182.53	32.78	13.91	83.21
E2	20%	**123.36**	**184.14**	**34.69**	**12.71**	**83.19**
E3	30%	121.91	186.25	35.47	11.98	83.24

表 III TORGO 数据集上训练的文本转语音模型的梅尔倒谱失真。

ID	模型	α	β	ASp	DSpG	SSp
E4	Grad-TTS [11]	–	–	6.61	6.71	6.81
E5	Matcha-TTS（基线）	–	–	6.25	6.43	6.64
E6	DARS（E5 含节奏）	–	–	6.09	6.32	6.57
E7	DARS 含 CPO	0.5	0.5	5.83	6.07	6.21
E8		0.7	0.3	5.72	5.93	6.08
E9	DARS 含 CPO + 风格	0.7	0.3	**4.29**	**4.46**	**4.61**

建模机制对合成质量的影响。表 III 展示了三种训练策略（ASp、DSpG 和 SSp）下，文本转语音模型生成的合成语音在验证集上的梅尔倒谱失真（MCD）结果。E4 对应 Grad-TTS 的结果。E5 是使用原始 Matcha-TTS 的基线系统。E6 在 E5 中加入了多阶段韵律预测模块。E7 和 E8 则在 E6 的基础上进一步整合了 CPO 策略。E9 在经过 CPO 增强的模型中引入了声学风格向量。

我们观察到，ASp训练策略取得了最佳性能，这可归因于其能够利用更多训练数据。结合停顿和时长建模以及CPO策略，可显著减少合成语音与真实构音障碍语音之间的声学差异。此外，融入声学风格向量可提供说话人特定的声学引导，进一步提升合成质量。在接下来的实验中，我们将分析基于 E9 衍生的三种训练策略生成的合成数据。

合成语音在提升自动语音识别方面的有效性。表 IV 对比了使用不同训练策略，以文本转语音模型合成的构音障碍语音训练的自动语音识别系统，与使用真实语音训练的系统的识别性能。对于 E9 合成数据，我们主要采用两种适配策略：全参数微调和基于LoRA的微调。可以看出，全参数微调的表现优于 LoRA。这种性能差距可能源于 Whisper模型最初是在非构音障碍语音上进行预训练的。因此，LoRA 的局部参数适配

表 IV 使用 WHISPER-LARGE 模型在不同数据类型和训练策略下的词错误率性能比较。

ID	数据类型	训练策略		词错误率 (%)				
		微调	LoRA	重度	中重度	中度	轻度	总体
E2	-			123.36	184.14	34.69	12.71	83.19
E10	真实数据	✓		18.63	11.88	2.93	2.43	8.85
E11			✓	22.35	13.66	3.24	2.54	10.09
E12	ASp	✓		**18.64**	**11.88**	**2.94**	**2.44**	**8.87**
E13			✓	22.37	13.67	3.24	2.56	10.12
E14	DSpG	✓		21.06	13.07	3.15	2.50	9.96
E15			✓	24.23	15.45	3.21	2.54	11.32
E16	SSp	✓		22.18	13.55	3.21	2.54	10.31
E17			✓	26.10	16.64	3.30	2.61	12.19

表V 与最先进系统的词错误率比较。

ID	数据集	词错误率 (%)				
		重度	中重度	中度	轻度	总体
E18 [9]	TORGO	55.88	49.6	36.8	12.6	39.2
E19 [11]	TORGO	23.3	13.98	3.27	2.57	16.93
E12	TORGO	18.64	11.88	2.94	2.44	8.87
E20	TORGO+ LibriSpeech 文本	**13.98**	**9.79**	**2.62**	**2.31**	**7.75**

在应用于构音障碍语音时表现出有限的建模能力，因为这类语音与典型语音模式存在显著的分布差异。

可以观察到，ASp训练策略（E12）生成的构音障碍语音在声学特征方面与真实语音高度相似，并展现出很强的泛化能力。在所有构音障碍严重程度等级上，使用 E12 训练的自动语音识别模型，其识别性能均与使用真实语音（E10）训练的模型相当。此外，与基线模型（E2）相比，E12 的总体词错误率相对降低了 89.33%。值得注意的是，障碍类别越严重，性能提升越明显，这凸显了 DARS 在应对严重言语障碍方面更高的有效性和适应性。

与最先进系统和 DARS模型的比较。表V展示了 DARS 与最先进系统在 TORGO数据集上不同严重程度等级的识别性能比较。在 E18 [9], 中，文本到语音模型为 FastSpeech 2 [32]，自动语音识别模型为深度神经网络—隐马尔可夫模型 [33] 系统。在 E19 [11], 中，文本到语音模型为 Grad-TTS [13]，自动语音识别模型为 Whisper。

为评估 DARS 的泛化能力，我们使用 LibriSpeech数据集中的文本进行构音障碍语音合成，并从 TORGO 中随机选取参考语音。使用增强的构音障碍语音（E20）调整自动语音识别模型后，总体词错误率达到 7.75%，相比 E19，词错误率相对降低了 54.22%。

V. 结论

本研究表明，使用 DARS 合成的语音对 Whisper模型进行微调，可显著降低各构音障碍严重程度等级的词错误率，其中中重度病例的相对降幅为 93.54%。与结合 Grad-TTS 和 Whisper 的先进系统相比，我们的方法使词错误率相对降低了 54.22%，展现出更优的韵律适应能力和泛化能力。使用 LibriSpeech 文本进行的跨语料合成实验进一步证实了该方法在域外场景中的稳健性。这些结果表明，对病理性节奏和声学风格进行联合建模，不仅能提升合成语音的真实感和可控性，还能显著提高面向构音障碍语音识别的数据增强效果及泛化能力。

参考文献

[1] J. R. Duffy等人，《运动性言语障碍：病理基础、鉴别诊断与管理》。Elsevier Health Sciences，2012年。[2] P. Enderby，“交流障碍：构音障碍”，《临床神经病学手册》，第110卷，第273–281页，2013年。[3] B. Vachhani、C. Bhat和S. K. Kopparapu，“利用健康语音增强构音障碍语音识别的数据”，载于Interspeech，2018年，第471–475页。[4] Z. Qian和K. Xiao，“构音障碍语音自动语音识别综述”，《Electronics》，第12卷，第20期，第4278页，2023年。[5] M. B. Mustafa、F. Rosdi、S. S. Salim和M. U. Mughal，“探究一般因素和特定因素对构音障碍说话人自动语音识别系统识别准确率的影响”，《Expert Systems with Applications》，第42卷，第8期，第3924–3932页，2015年。[6] H. P. Rowe、S. E. Gutz、M. F. Maffei、K. Tomanek和J. R. Green，“构音障碍多样性的表征与自动语音识别：临床视角教程”，《Frontiers in Computer Science》，第4卷，第770210页，2022年。[7] V. Young和A. Mihailidis，“构音障碍说话人自动语音识别面临的困难及其对老年人语音应用的影响：文献综述”，《Assistive Technology》，第22卷，第2期，第99–112页，2010年。[8] C. Bhat和H. Strik，“用于构音障碍语音自动识别与评估的语音技术：概述”，《Journal of Speech, Language, and Hearing Research》，第68卷，第2期，第547–577页，2025年。[9] M. Soleymanpour、M. T. Johnson、R. Soleymanpour和J. Berry，“利用多说话人文本转语音合成构音障碍语音，以用于构音障碍语音识别”，载于《ICASSP 2022-2022 IEEE国际声学、语音与信号处理会议（ICASSP）》，IEEE，2022年，第7382–7386页。

[10] D. Wagner等人，“利用大语言模型生成的转录文本生成可控合成语音，进行个性化微调，以用于构音障碍语音识别”，《arXiv预印本 arXiv:2505.12991》，2025年。[11] W.-Z. Leung、M. Cross、A. Ragni和S. Goetze，“通过文本到构音障碍语音合成增强构音障碍自动语音识别训练数据”，《arXiv预印本 arXiv:2406.08568》，2024年。[12] K. El Hajal、E. Hermann、A. Kulkarni和M. M. Doss，“将构音障碍语音无监督转换为健康语音的节奏与音色，以用于自动语音识别”，载于《2025年IEEE声学、语音与信号处理国际会议研讨会（ICASSPW）》，IEEE，2025年，第1–5页。[13] V. Popov、I. Vovk、V. Gogoryan、T. Sadekova和M. Kudinov，“Grad-TTS：用于文本转语音的扩散概率模型”，载于《国际机器学习会议》，PMLR，2021年，第8599–8608页。[14] R. Huang、Z. Zhao、H. Liu、J. Liu、C. Cui和Y. Ren，“ProDiff：用于高质量文本转语音的渐进式快速扩散模型”，载于《第30届ACM国际多媒体会议论文集》，2022年，第2595–2605页。[15] S. Mehta、R. Tu、J. Beskow、 ´E. Sz´ekely和G. E. Henter，“Matcha-TTS：采用条件流匹配的快速文本转语音架构”，载于《ICASSP 2024-2024 IEEE国际声学、语音与信号处理会议（ICASSP）》，IEEE，2024年，第11 341–11 345页。[16] A. Vaswani等人，“Attention is All You Need”，《神经信息处理系统进展》，第30卷，2017年。[17] J. Kim、S. Kim、J. Kong和S. Yoon，“Glow-TTS：通过单调对齐搜索实现文本转语音的生成式流模型”，《神经信息处理系统进展》，第33卷，第8067–8077页，2020年。[18] R. Rombach、A. Blattmann、D. Lorenz、P. Esser和B. Ommer，“基于潜在扩散模型的高分辨率图像合成”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第10 684–10695页。[19] S.-g. Lee、W. Ping、B. Ginsburg、B. Catanzaro和S. Yoon，“BigVGAN：采用大规模训练的通用神经声码器”，《arXiv预印本 arXiv:2206.04658》，2022年。[20] Z. Ye、W. Xue、X. Tan、J. Chen、Q. Liu和Y. Guo，“ComoSpeech：通过一致性模型实现一步式语音与歌声合成”，载于《第31届ACM国际多媒体会议论文集》，2023年，第1831–1839页。[21] Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon和B. Poole，“通过随机微分方程进行基于得分的生成建模”，《arXiv预印本 arXiv:2011.13456》，2020年。[22] M. S. Albergo和E. Vanden-Eijnden，“利用随机插值器构建归一化流”，《arXiv预印本 arXiv:2209.15571》，2022年。[23] N. Cao、Y.-R. Lin、X. Sun、D. Lazer、S. Liu 和 H. Qu，“Whisper：实时追踪信息扩散的时空过程”，*IEEE* 可视化与计算机图形学汇刊，第 18 卷，第 12 期，第 2649–2658 页，2012 年。[24] H. Ma、Z. Peng、M. Shao、J. Li 和 J. Liu，“通过提示调优扩展 Whisper，以实现目标说话人自动语音识别”，载于*ICASSP 2024-2024*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），IEEE，2024 年，第 12 516–12520 页。[25] Y. Liu、X. Yang 和 D. Qu，“面向低资源语音识别的 Whisper 微调策略探索”，*EURASIP* 音频、语音与音乐处理期刊，第 2024 卷，第 1 期，第 29 页，2024 年。[26] M. McAuliffe、M. Socolof、S. Mihuc、M. Wagner 和 M. Sonderegger，“蒙特利尔强制对齐器：使用 Kaldi 训练文本-语音对齐”，载于*Interspeech*，第 2017 卷，2017 年，第 498–502 页。[27] Y. Wang 等，“风格令牌：端到端语音合成中的无监督风格建模、控制与迁移”，载于第 *35* 届国际机器学习会议论文集，J. Dy 和 A. Krause 编，机器学习研究论文集，第 80 卷，PMLR，2018 年 10 月，第 5180–5189 页。[28] R. Huang、Y. Ren、J. Liu、C. Cui 和 Z. Zhao，“Generspeech：面向可泛化域外文本转语音的风格迁移”，载于神经信息处理系统进展，S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh 编，第 35 卷，Curran Associates, Inc.，2022 年，第 10 970–10 983 页。[29] F. Rudzicz、A. K. Namasivayam 和 T. Wolff，“TORGO 数据库：构音障碍说话者的声学与发音语音数据库”，语言资源与评估，第 46 卷，第 523–541 页，2012 年。[30] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“LibriSpeech：基于公有领域有声书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。[31] J. Kominek、T. Schultz 和 A. W. Black，“以平均梅尔倒谱失真校准新语言的合成语音质量”，载于*SLTU*，2008 年，第 63–68 页。[32] Y. Ren 等，“FastSpeech 2：快速、高质量的端到端文本转语音”，*arXiv* 预印本 *arXiv:2006.04558*，2020 年。[33] L. Li 等，“基于深度神经网络—隐马尔可夫模型（DNN-HMM）的混合语音情感识别”，载于*2013*年人类情感计算与智能交互协会会议，IEEE，2013 年，第 312–317 页。