基于帧级适配器和多重 wait-k 知识蒸馏的端到端同步构音障碍语音重建

吴明辉*、唐海涛†、范家欢†、廖瑞智†、张彦勇*
* 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

arXiv:2603.01382v1 [cs.SD] 2 Mar 2026

摘要—构音障碍语音重建（DSR）通常采用级联系统，将自动语音识别（ASR）与句子级语音合成（TTS）相结合，把构音障碍语音转换为正常韵律语音。然而，构音障碍患者往往语速较慢，导致此类系统的响应时间过长，在长语音场景中不切实际。基于流式自动语音识别和增量式文本转语音的级联构音障碍语音重建系统有助于降低延迟。然而，不同构音障碍严重程度的患者对同一文本表现出显著的发音变异性，导致 ASR 鲁棒性较差，并限制了重建语音的可懂度。此外，增量式文本转语音因感受野有限而存在韵律特征预测不佳的问题。在本研究中，我们提出一种端到端同步构音障碍语音重建系统，包含两项关键创新：1）引入帧级适配器模块以衔接 ASR 与 TTS。通过显式-隐式语义信息融合和模块联合训练，提升 TTS 对 ASR 输出的错误容忍度。2）设计多重 wait-k 自回归语音合成模块，通过多视角知识蒸馏缓解韵律退化。我们的系统在 Tesla A100 上的平均响应时间为 1.03 秒，平均实时因子（RTF）为 0.71。在 UASpeech 数据集上，其平均意见分（MOS）达到 4.67，与当前最优水平相比，词错误率（WER）相对降低 54.25%。我们的演示见：https://wflrz123.github.io/

一、引言

构音障碍语音重建（DSR）旨在将异常韵律转换为正常韵律，同时保留原始语义和说话人音色 [1], [2], [3]。一种常见的解决方案是采用级联系统 [4], [5] ，该系统整合了自动语音识别（ASR）[6] 和句子级语音合成（TTS）[7], ，如图 1 所示。在这种方法中，TTS 每仅在自动语音识别完整识别整句之后才进行快速合成。然而，构音障碍患者在构音方面面临挑战，通常说话更慢 [2], [8]。因此，句子级级联系统的重建时延会显著增加。这一限制阻碍了其在长语音通信场景中的实际应用，例如会议或电话交谈。

为降低句子级时延，研究者提出了基于流式自动语音识别和增量式文本转语音的同步级联系统 [9], [10] （如图 1 所示）。这些

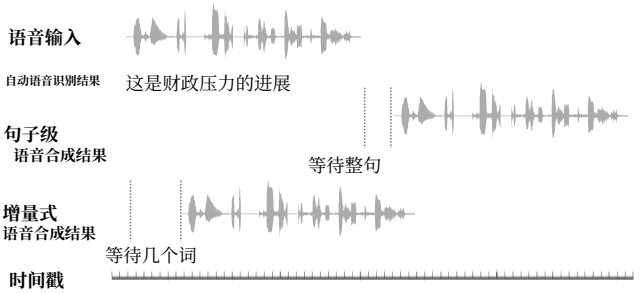

图 1. 基于自动语音识别和文本转语音的级联构音障碍语音重建系统。

系统无需等待自动语音识别输出完整结果，即可生成重建正常语音。流式自动语音识别通常采用 RNN 转换器（RNNT）模型 [11], [12], ，而增量式文本转语音则利用上下文约束文本输入进行实时合成 [13], [14]。然而，构音障碍语音带来了诸多挑战。由于难以采集，这类数据被视为低资源数据 [15] 。不同严重程度的患者对相同文本的发音差异显著 [15], ，导致语音-文本强制对齐（FA）的映射空间更大。构音障碍语音自动语音识别任务面临重大挑战，严重限制了重建语音的可懂度 [16]。此外，尽管增量式文本转语音能够有效估计倒谱和非周期特征，但由于感受野有限，在韵律特征预测方面仍表现不佳 [13], [14]。这会对重建语音的质量、自然度和表现力造成不利影响。

为应对上述挑战，我们提出了一种端到端同步构音障碍语音重建（E2E-SDSR）系统，该系统融合了流式自动语音识别与增量式文本转语音的级联架构。为减轻自动语音识别错误对可懂度的影响，我们引入适配器模块，以衔接流式自动语音识别与增量式文本转语音。鉴于构音障碍语音与文本之间的映射空间很大，适配器采用门控激励算子，将来自 RNNT 编码器的高层隐式语义表示与来自 RNNT 解码器的帧级显式语义信息融合。通过扩展语义搜索空间，这种融合机制增强了文本转语音模块对 ASR 识别错误的容错能力。为提升增量式文本转语音的上下文建模

能力，我们设计了一个多等待步长自回归文本转语音模块。该模块有效利用适配器的帧级输出，处理长度各异的语义帧序列 k。此外，我们将基于知识蒸馏（KD）的多任务学习（MTL）融入多等待步长策略，使感受野更大的模型能够指导较小的模型，并确保后者的合成质量接近前者。由此，所提出的 E2E-SDSR 系统在确保重建语音可懂度与自然韵律的同时，也能在长语音场景中保持实时性能。

II. 所提出的方法

本节将详细介绍所提出的系统，该系统由三个模块组成。首先，流式自动语音识别模块负责提取构音障碍语音的语义内容特征。其次，帧级适配器模块连接自动语音识别与文本转语音，将融合语义信息从自动语音识别传递至文本转语音。第三，后端包含一个基于多等待步长策略的增量式自回归文本转语音模块，该模块将自动语音识别提取的语义内容与音色及正常韵律相结合，以实现重建。为增强系统在低资源构音障碍语音数据上的鲁棒性，我们采用两阶段训练策略：首先使用正常语音对重建系统进行预训练，随后使用构音障碍语音对自动语音识别进行微调，以驱动后端重建正常韵律。

A. 基于 *Conformer-RNNT* 的自动语音识别模块

为确保构音障碍语音重建系统具备同步处理能力，我们采用具有固有流式能力的 RNNT 模型作为自动语音识别模块。RNNT 模型由编码器、解码器和联合网络 [17], [18] 组成。编码器将声学特征与语义特征解耦，其中较低层捕捉说话人特征，较高层则表示语义和声学信息 [12]。编码器中集成了 Conformer层，以发挥其全局和局部序列建模能力 [19], [20]。解码器使用多层长短期记忆网络 [21], 处理上下文，联合网络则通过全连接（FC）层 [22] 整合编码器和解码器的输出，并通过 RNNT损失 [23] 学习显式语义信息。图 2 展示了 Conformer-RNNT 的详细步骤，可表示如下：

$$\mathbf{h}_t^{enc} = encoder(\mathbf{x}_t) \tag{1}$$

$$\mathbf{h}_u^{dec} = decoder(y_u) \tag{2}$$

$$\mathbf{h}_{t,u}^{joint} = joint(\mathbf{h}_t^{enc}, \mathbf{h}_u^{dec}) \tag{3}$$

$$Q(\hat{y}_{u+1}|x_t, y_u) = softmax(\mathbf{h}_{t,u}^{joint}) \tag{4}$$

其中，声学特征$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$表示前一个词元$\mathbf{Y} = [y_1, y_2, \cdots, y_U]$，$\mathbf{h}_t^{enc}$是声学编码器在第 t帧提取的声学表示，而$\mathbf{h}_u^{dec}$是解码器网络中第 u个子词单元的文本表示。$\mathbf{h}_{t\,u}^{joint}$，是联合网络的输出。为了降低 DSR 系统的延迟，Conformer网络采用

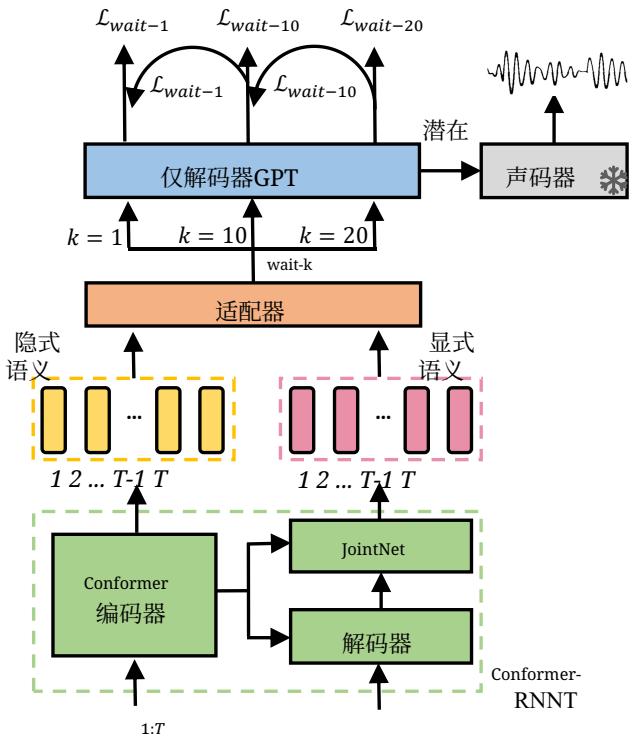

图 2. 端到端同步构音障碍语音重建系统。

截断上下文，左侧为 64 帧，右侧为 0 帧，通过在自注意力中使用简单掩码实现 [24]。

B. 帧级适配器模块

鉴于构音障碍语音属于低资源数据，且与正常语音相比具有更广阔的语音-文本映射空间，RNNT 模型的识别性能受到限制，进而降低了 DSR 系统中重建输出的可懂度。为缓解自动语音识别错误导致的可懂度下降，我们设计了一个适配器模块，用于连接 RNNT 和增量式文本转语音，如图 2 所示。该模块遵循信息性和适应性原则。我们引入语义软表示以扩大语义搜索空间，从而增强后端 TTS 模块的错误容忍度。具体而言，适配器模块融合显式语义信息$\mathbf{h}_{t\,u}^{joint}$，与隐式语义表示 $\mathbf{h}_t^{enc}$。考虑到这两种特征之间的维度差异，我们使用前向算子将 $\mathbf{h}_{t\,u}^{joint}$，与帧级 $\mathbf{h}_t^{exp}$ [17],[18] 对齐。$\mathbf{h}_t^{enc}$主要包含由语音帧和静音帧构成的声学特征 [25]。在融合阶段，这些无关的静音帧会被视为语义信息，从而增加语义空间的搜索成本。为消除这种冗余，我们引入门控激励算子，以激活相关帧并抑制无关帧 [26]。适配器模块的具体形式如下：

$$\mathbf{h}_t^{exp} = \begin{cases} \mathbf{W}^{exp} \cdot \mathbf{h}_{t+1,u}^{joint} & \text{if } \hat{y}_{u+1} \text{ is } \phi, \\ \mathbf{W}^{exp} \cdot \mathbf{h}_{t+1,u+1}^{joint} & \text{else} \end{cases} \tag{5}$$

$$\mathbf{h}_t^{imp} = \mathbf{W}^{imp} \cdot \mathbf{h}_t^{enc} + \mathbf{b}^{imp} \tag{6}$$

$$\mathbf{h}_t^{apt} = \lambda \cdot \mathbf{h}_t^{exp} + (1-\lambda) \cdot SwitchGLU(\mathbf{h}_t^{imp}) \tag{7}$$

其中，$\mathbf{h}_t^{exp}$ 和$\mathbf{h}_t^{imp}$ 分别表示第 t 帧的显式和隐式语义软表示。$\mathbf{h}_t^{apt}$ 是适配器模块的输出。$\mathbf{W}^{exp}$ 是可训练映射矩阵。 ϕ 表示 RNNT 对齐过程中使用的空白符号，用于处理声学帧与输出词元之间的长度不匹配。$\mathbf{h}_t^{enc}$通过单层神经网络投影到潜在空间。$\mathbf{W}^{imp}$和$\mathbf{b}^{imp}$ 是可训练参数。 $SwitchGLU(\cdot)$表示开关门控线性单元 [27]。 λ是$\mathbf{h}_t^{exp}$ 的可训练权重，初始化时设为 0.5。适配器模块在 DSR 系统中具有两项主要优势：1）与仅使用显式 ASR 输出相比，它提供了更广阔的语义搜索空间，从而提高 TTS 模块对 ASR 识别错误的容忍度。2）它支持梯度反向传播，并可实现 ASR 和 TTS 模块的联合微调，从而有效缩小两者之间的性能差距。

C. 基于多重 *Wait-k* 的自回归 *TTS* 模块

传统的增量式文本转语音系统通常依赖词级语义输入 [13], [14]。然而，ASR 模块需要足够的声学特征才能解码这些语义，这往往导致 DSR 系统的首帧响应时间过长 [12]。为缓解这一问题，我们引入了一种基于仅解码器 GPT架构的自回归 TTS 模型 [28],，如图 2 所示。它巧妙地利用适配器模块的帧级语义输出，来预测重建语音的量化码本条目。该模块的设计包含三个关键方面：$\mathbf{h}_t^{apt}$ 与量化码本$\mathbf{c}_t$ 具有相同的帧移，从而确保序列长度一致并实现自然对齐；通过量化向量量化 [29] （VQ）模型生成的量化码本，相比梅尔频谱图具有更强的鲁棒性，并已被广泛用作现代 TTS 系统中的预测目标 [29]；输入和输出序列均包含语义和部分声学特征，从而使学习过程更加容易。

增量式 TTS 系统常因感受野有限，导致重建语音的韵律和质量不佳。为解决这一问题，我们将基于多任务学习的多等待步长策略引入自回归 TTS 模型，使其能够通过不同的 k 值适应不同感受野。虽然较大的 k 值通常能提升合成质量，但也会增加延迟。为了兼顾较短的响应时间和较高的合成质量，我们将知识蒸馏融入多重 wait-k 框架，使感受野较大的模型能够指导感受野较小的模型进行训练，从而确保后者的合成语音质量与前者十分接近。训练目标定义如下：

$$\mathcal{L}_{wait-k}^{CE} = CE(P(\hat{\mathbf{c}}_{t+1}|\mathbf{h}_{<t+k}^{apt}, \mathbf{c}_{<t}; \boldsymbol{\theta}), \mathbf{c}_{t+1}) \tag{8}$$

$$\mathcal{L}_{wait-k}^{KD} = D_{KL}(P(\hat{\mathbf{c}}_{t+1}|\mathbf{h}_{<t+k}^{apt}, \mathbf{c}_{<t}; \boldsymbol{\theta}) \;||\; P(\hat{\mathbf{c}}_{t+1}|\mathbf{h}_{<t+k+10}^{apt}, \mathbf{c}_{<t}; \boldsymbol{\theta})) \tag{9}$$

此处， $\boldsymbol{\theta}$ 表示基于仅解码器GPT的自回归模型，该模型在多任务学习中共享于不同感受野。为适应不同的感受野，$\mathcal{L}_{wait-k}^{CE}$ 采用三个不同的 wait-k 参数：$k \in \{1, 10, 20\}$。$\mathbf{h}_{<t+k}^{apt}$来自适配器模块的输出与历史码本$\mathbf{c}_{<t}$ 拼接并融合，作为 $\boldsymbol{\theta}$的输入。当 $t < k$时，$\mathbf{c}_t$ 设为固定常数。 $P(\hat{}\mathbf{c}_{t+1}|\mathbf{h}_{<t+k}^{apt}, \mathbf{c}_{<t}; \boldsymbol{\theta})$表示 $\boldsymbol{\theta}$的软预测输出。 $D_{KL}(\cdot)$是 Kullback-Leibler（KL）散度损失 [30]。较大感受野产生的软预测作为教师，指导感受野较小的学生模型。这种方法无需增加额外延迟，即可提升韵律和语音质量。最后，增量式 TTS 模型的损失函数 $\mathcal{L}^{TTS}$由 $\mathcal{L}_{wait-k}^{CE}$ 和 $\mathcal{L}_{wait-k}^{KD}$ 组成。

$$\mathcal{L}^{CE} = \mathcal{L}_{wait-1}^{CE} + \mathcal{L}_{wait-10}^{CE} + \mathcal{L}_{wait-20}^{CE} \tag{10}$$

$$\mathcal{L}^{KD} = \mathcal{L}_{wait-1}^{KD} + \mathcal{L}_{wait-10}^{KD} \tag{11}$$

$$\mathcal{L}^{TTS} = (1-\alpha) \cdot \mathcal{L}^{CE} + \alpha \cdot \mathcal{L}^{KD} \tag{12}$$

这里， α 并默认设为 0.2。为确保整个 DSR 系统具备同时处理能力，潜在预测会被切分为多个块，并传入基于 Hifi-GAN 的声码器 [31]进行波形重建。

D. 两阶段训练策略

由于构音障碍语音属于低资源数据，我们采用两阶段训练策略来增强系统的鲁棒性。第一阶段，我们联合训练 E2E-SDSR系统（包括 RNNT、适配器和文本转语音模块），通过优化 RNNT损失和 TTS损失，为正常语音奠定重建基础。随后，固定训练所得的参数，并用其生成声码器训练所需的输入，以确保声码器 [28]的稳定性。第二阶段，考虑到预训练的 E2E-SDSR 系统依赖 RNNT 来纠正构音障碍语音的语义，我们对 RNNT 模型进行微调，以进一步提升其对构音障碍语音的处理性能。~ ~我们使用一批 n 样本 $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i), (\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{n}$, ~ 其中$\mathbf{X}$ 表示构音障碍语音，$\mathbf{X}$ 表示正常语音。这种平衡组合可确保 RNNT 适应构音障碍语音，同时使正常语音的输出分布与后端模块保持一致。

III. 实验

A. 实验设置

对于该数据集，训练和测试均使用商业构音障碍语音数据和 UASpeech 数据集。该商业数据集包含 500 小时来自中国自由交谈场景的真实录音，涉及中国 35 座城市的 3,000 名说话人。语音可懂度分为

五个等级，其中等级 1 最为严重，等级 5 则接近正常。本研究聚焦于等级 3 和 4。UASpeech 数据集 [32] 是规模最大的开源英语构音障碍语音数据集，包含 29 名说话人。我们选取了四名语音可懂度水平各异的说话人进行测试：M05（中等）、F04（中等）、M07（较低）和 F02（较低）。

对于 RNNT 模块，我们采用对数梅尔滤波器组作为输入声学特征。编码器采用 [12] 中所述的 12 层 Conformer 架构。解码器由嵌入层和一个 LSTM 层组成。对于 TTS 模块，仅解码器 GPT 模型包含 15 个 Transformer 层 [28]。VQ 模型遵循 [29] 中的设计，并针对 16 kHz 语音采用与 RNNT 编码器相同的 640 ms 帧移。

在评估指标方面，我们使用 Whisper-Base 模型 [33] 获得识别结果，并通过词错误率（ WER ）计算可懂度。采用 5 级平均意见分（MOS） [28] 评估自然度。

B. 结果与分析

适配器消融。表 I 比较了适配器模块在商业数据集上的性能。为评估对自动语音识别错误的容忍度提升情况，在配置 E2 中，将自动语音识别的词级输出直接输入句子级文本转语音模块。自动语音识别模块采用流式 RNNT 模型，而句子级文本转语音模块采用仅解码器 GPT 架构的 XTTS 模型 [28]。这种组合确保前端和后端模块与我们提出的系统具有可比性。在 E3 中，适配器模块仅将公式（5）中的显式表示用作增量文本转语音输入。在 E4 中，它仅使用公式（6）中的隐式表示。在 E5 中，通过可学习的 λ 融合显式表示和隐式表示。E6 在 E5 的基础上，对隐式表示应用门控激励算子。在 E3~E6 中，多个 wait-k 固定为 10，且不采用多任务学习，以消除不同感受野的影响。我们观察到，与 E2 相比，E3 和 E4 均提高了文本转语音对自动语音识别错误的容忍度。由于隐式表示中包含声学特征，E4 在 3 级和 4 级的 MOS 分数更高。E5 和 E6 表明，结合两种表示比单独使用任一种效果更好。在 E6 中引入 SwitchGLU 后，重建性能令人满意，可与 E1 相媲美。我们还分析了 Whisper-Base 对 E6 和 E2 重建语音的识别结果。E6 主要改善了相似语音单位的识别，例如 /d/ 和 /g/、/in/ 和 /ing/，从而增强了词语区分能力。

多重 Wait-k 消融。表 II 展示了多任务 wait-k 策略对增量式文本转语音的影响。我们主要从三个方面进行比较：不同的 wait-k 超参数、多任务联合训练以及知识蒸馏。E6~E8 表示不同感受野下的结果。可以观察到，较大的感受野在第 3 级和第 4 级取得了更低的词错误率和更高的 MOS分数。E9 采用多任务学习来共享增量式文本转语音，使其能够适应不同的感受野。E10 在 E9 的基础上进行了扩展，整合了

表 I 适配器模块在商业数据集上的结果。

ID	模型	第三级 词错误率	第三级 平均意见分	第四级 词错误率	第四级 平均意见分
E1	原始	49.42	2.08	33.71	2.48
E2	自动语音识别+句子级文本转语音	25.93	3.32	12.91	4.71
E3	采用公式（5）的适配器	25.63	3.27	12.89	4.24
E4	采用公式（6）的适配器	25.07	3.41	12.23	4.53
E5	适配器，采用式 (5)、式 (6)	23.67	3.64	11.55	4.76
E6	E5 采用 SwitchGLU	22.58	3.66	10.84	4.79

表 II 多任务Wait-k在增量式文本转语音上的结果。

ID	模型	第三级 词错误率	第三级 平均意见分	第四级 词错误率	第四级 平均意见分
E1	原始	49.42	2.08	33.71	2.48
E7	wait-k = 1	24.83	3.28	14.17	4.02
E6	wait-k = 10	22.58	3.66	10.84	4.79
E8	wait-k = 20	21.07	4.44	9.73	4.84
E9	wait-k = 1 与多任务学习	24.61	3.30	14.08	4.05
	wait-k = 10 与多任务学习	22.58	3.73	10.84	4.81
	wait-k = 20 与多任务学习	20.75	4.53	9.84	4.85
E10	wait-k = 1 与多任务学习，KD	22.25	3.82	10.37	4.81
	wait-k = 10 结合多任务学习，知识蒸馏	21.52	3.93	10.09	4.82
	wait-k = 20 结合多任务学习，知识蒸馏	20.95	4.41	10.05	4.84

表 III 不同 DSR 系统的性能。

标识符	参数量（M）	浮点运算次数（M）	系统响应时间（秒）	RTF
E2	248.92	400.48	2.67	1.39
E7			0.61	0.70
E6	232.83	350.15	1.03	0.71
E8			1.38	0.71

知识蒸馏策略，即由大感受野模型指导小感受野模型的训练。与 E6~E8 相比，小感受野模型受益于大感受野模型的先验知识，从而提升了识别性能以及重建语音的 MOS分数。例如，E10（wait-k=0）取得了更接近 E9（wait-k=10）的性能。

我们在图3中分析了增量式文本转语音在不同感受野下的韵律。(a)为感受野较小的模型，(b)为感受野较大的模型。为便于理解，这里使用了对应梅尔频谱图的汉字和拼音。可以看出，(a)和(b)中的语音时长和停顿保持不变。在(b)中，韵律和自然度显著提升。红框中的错误声调得到明显纠正，绿框中的病理性延长片段也被重构为正常片段。

效率比较。 表 III 比较了级联系统和 E2E-SDSR 系统在 Tesla A100

参考文献

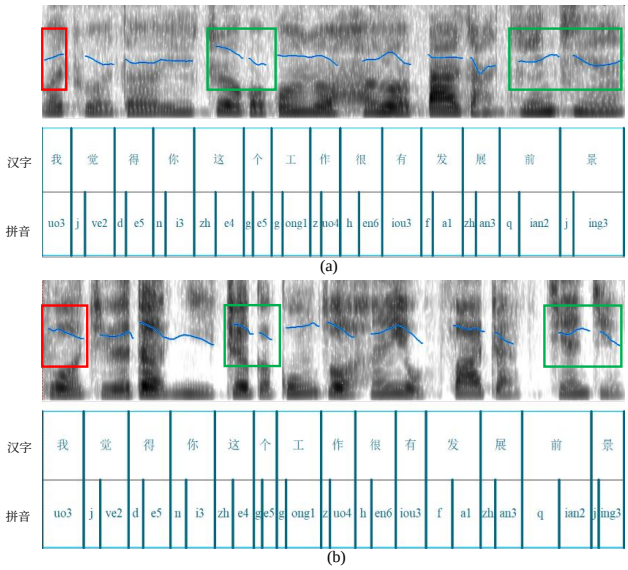

图3. 增量式文本转语音在不同感受野下的韵律比较。(a) 为 E7；(b) 为 E8。

表 IV E2E-SDSR 与最先进水平在 UASPEECH 上的 5 级平均意见得分测试分数和词错误率（%）比较。

ID	模型	指标	M05（中等）	F04（中等）	M07（低）	F02（低）
E11	原始	词错误率	81.70	81.70	95.60	95.90
		MOS	2.86	2.42	1.75	1.95
E12 端到端语音识别		词错误率	69.80	69.30	73.10	72.00
		MOS	3.61	3.47	3.72	3.87
E13 ASA-DSR		词错误率	62.50	76.60	62.70	65.80
		MOS	4.19	3.98	3.71	4.26
E14 Unit-DSR		词错误率	64.40	65.50	62.10	68.30
		MOS	4.52	4.65	4.62	4.55
E15 E2E-SDSR		词错误率	28.25	31.26	26.16	33.42
		MOS	4.65	4.72	4.68	4.65

平台上的性能。在参数量和浮点运算次数相近的情况下，所提出的系统比级联系统具有更快的响应时间和更低的实时因子。这是因为我们提出的系统无需等待完整句子结束，即可同时校正和重建语音。比较 E6~E8 可知，感受野大小的选择主要影响响应时间。

最先进水平对比。 表 IV 展示了 UASPEECH 上平均意见分和词错误率的对比结果。E11 是原始语音。E12 来自 [2]。E13 来自 [1]。E14 来自 [3]。E15 是我们在 E10 上采用 E2E-SDSR（wait-k=10）得到的结果。与最先进水平相比，E15 的平均意见分达到 4.67，平均词错误率相对降低了 54.25%。

IV. 结论

在本研究中，我们提出一种端到端同步 DSR 系统，专注于在有限感受条件下优化重建语音的可懂度与自然度野。该系统引入两项关键创新：（1）帧级适配器模块，通过显式-隐式信息融合与模块联合训练，缓解由自动语音识别错误导致的可懂度损失。（2）基于 Wait-k 策略的增量式文本转语音模块，在低延迟与改善韵律之间取得平衡，从而应对受限感受野所带来的局限。最后，所提出的方法将平均响应时间降低了 50% 以上。在 UASpeech 数据集上，相较于最先进水平，词错误率相对降低了 54.25%。在商业数据集与开源数据集上，相较于原始构音障碍语音，其可懂度与自然度均有显著提升。

参考文献

[1] D. Wang 等，“通过跨模态知识蒸馏实现构音障碍语音重建的端到端语音转换”，载于 *ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2020年，第 7744–7748 页。[2] D. Wang 等，“通过对抗式说话人适配保持构音障碍语音重建中的说话人身份”，载于 *ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2022年，第 6677–6681 页。[3] Y. Wang、X. Wu、D. Wang、L. Meng 和 H. Meng，“Unit-dsr：使用语音单元归一化的构音障碍语音重建系统”，载于 *ICASSP 2024- 2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2024年，第 12 306–12310 页。[4] X. Chen、D. Yang、D. Wang、X. Wu、Z. Wu 和 H. Meng，“Colm-dsr：利用神经编解码器语言建模进行多模态构音障碍语音重建”，*arXiv* 预印本 *arXiv:2406.08336*，2024年。[5] K. Fatemeh、M. T. Rahil 和 Z. Hassan，“通过对比学习增强构音障碍语音重建”，*arXiv* 预印本 *arXiv:2410.04092*，2024年。[6] R. Turrisi，“关于运用深度学习策略解决构音障碍语音自动语音识别（asr）问题”，2021年。[7] F. Rudzicz，“调整构音障碍语音信号以提高可懂度”，*Computer Speech &Language*，第 27 卷，第 6 期，第 1163–1177 页，2013年。[8] D. Liu、Y. Lin、H. Bu 和 M. Li，“用于零样本构音障碍语音重建的两阶段自监督语音转换”，载于*2024*年亚洲语言处理国际会议（*IALP*），IEEE，2024年，第 423–427 页。[9] K. Sudoh、T. Kano、S. Novitasari、T. Yanagita、S. Sakti 和 S. Nakamura，“采用神经增量式自动语音识别、机器翻译和文本转语音的同步语音到语音翻译系统”，*arXiv* 预印本 *arXiv:2011.04845*，2020年。[10] L. Barrault 等，“Seamless：多语言、富有表现力的流式语音翻译”，*arXiv* 预印本 *arXiv:2312.05187*，2023年。

[11] M. Ghodsi、X. Liu、J. Apfel、R. Cabrera 和 E. Weinstein，“采用无状态预测网络的 RNN 转导器”，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2020年，第 7049–7053 页。[12] H. Tang 等，“通过自适应两阶段知识蒸馏缩小流式与非流式基于转导器的自动语音识别之间的差距”，载于*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2023年，第 1–5 页。[13] B. Stephenson、T. Hueber、L. Girin 和 L. Besacier，“交替结尾：利用预测的未来文本输入改善增量式神经文本转语音的韵律”，*arXiv* 预印本 *arXiv:2102.09914*，2021年。[14] T. Saeki、S. Takamichi 和 H. Saruwatari，“采用蒸馏上下文预测网络的低延迟增量式文本转语音合成”，载于*2021*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*），IEEE，2021年，第 749–756 页。[15] Y. Lin、J. Dang、L. Wang、S. Li 和 C. Ding，“兼顾低资源和异常发音的语音识别”，*SpeechCommunication*，第 155 卷，第 103002 页，2023年。[16] J. Shor 等，“利用有限数据为构音障碍语音和带口音语音实现个性化自动语音识别”，*arXiv* 预印本 *arXiv:1907.13511*，2019年。[17] A. Graves，“使用循环神经网络进行序列转导”，*CoRR*，第 abs/1211.3711 卷，2012年。[18] T. N. Sainath、Y. He、B. Li 等，“一种流式端侧端到端模型，在模型质量和延迟方面超越服务器端传统模型”，载于*2020*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP 2020*），巴塞罗那，西班牙，*2020*年*5*月*4*日至*8*日，IEEE，2020年，第 6059–6063 页。[19] A. Gulati、J. Qin、C. Chiu 等，“Conformer：用于语音识别的卷积增强型 Transformer”，H. Meng、B. Xu 和 T. F. Zheng 编，第 5036–5040 页，2020年。[20] Q. Zhang、H. Lu、H. Sak 等，“Transformer 转导器：一种采用 Transformer 编码器和 RNN-T 损失的可流式语音识别模型”，载于*2020*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP 2020*），巴塞罗那，西班牙，*2020*年*5*月*4*日至*8*日，IEEE，2020年，第 7829–7833 页。[21] K. Rao、H. Sak 和 R. Prabhavalkar，“探索用于 RNN 转导器流式端到端语音识别的架构、数据和单元”，第 abs/1801.00841 卷，2018年。[22] A. Graves、A. Mohamed 和 G. E. Hinton，“使用深度循环神经网络进行语音识别”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP 2013*），温哥华，不列颠哥伦比亚省，加拿大，*2013*年*5*月*26*日至*31*日，IEEE，2013年，第 6645–6649 页。[23] G. Huang、D. Zhou、S. Dan 和 F. Chi，“在自动语音识别中使用 CTC 损失训练 RNN-T”，期刊《应用数学与计算》，第6卷，第2期，第256–262页，2022年。[24] X. Chen、Y. Wu、Z. Wang等，“在大规模数据集上开发用于语音识别的实时流式Transformer 转导器”，载于*IEEE*国际声学、语音与信号处理会议，*ICASSP 2021*，加拿大安大略省多伦多，*2021*年*6*月*6*日至*11*日，IEEE，2021年，第5904–5908页。[25] W. R. Huang等，“双通道级联编码器自动语音识别模型中的端到端分割”，载于*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2023年，第1–5页。[26] M. Wu等，“通过高效联合优化在说话人自适应自动语音识别中隐式增强目标说话人”，载于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），IEEE，2024年，第10 051–10055页。[27] G.-B. Zhou、J. Wu、C.-L. Zhang和Z.-H. Zhou，“用于循环神经网络的最简门控单元”，《国际自动化与计算期刊》，第13卷，第3期，第226–234页，2016年。[28] E. Casanova等，“XTTS：大规模多语言零样本文本转语音模型”，*arXiv*预印本 *arXiv:2406.04904*，2024年。[29] J. Betker，“通过扩展规模改进语音合成”，*arXiv*预印本 *arXiv:2305.07243*，2023年。[30] T. Van Erven和P. Harremos，“Rényi散度与Kullback–Leibler散度”，*IEEE*信息论汇刊，第60卷，第7期，第3797–3820页，2014年。[31] J. Kong、J. Kim和J. Bae，“HiFi-GAN：用于高效高保真语音合成的生成对抗网络”，《神经信息处理系统进展》，第33卷，第17 022–17033页，2020年。[32] J. Yu等，“面向UA语音语料库的香港中文大学构音障碍语音识别系统开发”，载于*Interspeech*，2018年，第2938–2942页。[33] C. Graham和N. Roll，“评估OpenAI的Whisper自动语音识别：跨不同口音和说话人特征的性能分析”，《*JASA*快报》，第4卷，第2期，2024年。