用于有声读物语音合成的基于层级 Transformer 的上下文感知连贯说话风格预测

Shun Lei[1‡]、*Yixuan Zhou*[1†]、*Liyang Chen*[1] *Zhiyong Wu*[1, 2, 4*]、*Shiyin Kang*[3*]、*Helen Meng*[4]

[1] 清华大学深圳国际研究生院，深圳 [2] 鹏城实验室，深圳 n

[3] XVerse Inc.，深圳 [4] 香港中文大学，香港特别行政区{leis21, zhouyx20, cly

21}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, kangshiyin@xverse.cn, hmmeng@se.cuhk.edu.h k

arXiv:2304.06359v1 [cs.SD] 13 Apr 2023

摘要

近年来，文本转语音技术取得了长足进展，显著提升了合成语音的表现力。然而，要为有声读物中的多句文本生成风格恰当且连贯的语音，仍然面临挑战。本文提出了一种用于有声读物语音合成的上下文感知连贯说话风格预测方法。为预测当前话语的风格嵌入，我们设计了一种带有混合注意力掩码的基于分层Transformer的上下文感知风格预测器，同时考虑文本侧的上下文信息和此前语音中的语音侧风格信息。基于此，我们能够逐句生成风格和韵律连贯的长篇语音。在一个普通话有声读物数据集上进行的客观和主观评估表明，与基线方法相比，我们提出的模型在单句和多句测试中均能生成表现力更强、说话风格更连贯的语音[l]。

索引词— 有声读物语音合成、说话风格建模、上下文感知、层次化 Transformer、多句子

1. 引言

文本转语音（TTS）旨在根据文本生成清晰自然的语音。随着深度学习的发展，如今的 TTS 模型能够以中性的说话风格生成高质量、自然的语音 [1, 2]。然而，表现力有限的说话风格仍使合成语音与人类录音之间存在明显差距，因此已无法满足客户日益增长的需求，尤其是在有声读物等场景中。实现有声读物语音合成仍然具有挑战性，因为语料通常包含丰富且复杂的风格变化，受到多种因素的影响（包括上下文信息、说话者意图等），这增加了风格建模的难度。

为了建模这种富有表现力的说话风格，文本预测全局风格词元（TP-GST） [3] 首次提出了从输入文本预测风格嵌入的思路，从而生成比基线 Tacotron [4]具有更丰富音高和能量变化的语音。考虑到话语的风格信息与语义信息密切相关，研究者将从预训练语言模型（如双向 Transformer 编码器表示（BERT） [5], ）中提取的文本嵌入引入 TTS 模型，以更好地预测语音的风格表示 [6–8]。此外，话语中丰富的风格变化也会受到相邻句子的影响，尤其是在连续的有声读物多句文本中 [9, 10]。对此，一些研究通过将跨句文本输入 BERT，获取上下文感知文本嵌入，以改善每个话语的韵律生成 [11, 12]。为了利用广泛上下文中固有的层级结构， [13] 设计了一种层级上下文编码器（HCE），除当前句子外，还从前后若干相邻句子中收集上下文信息，从而显著提升了富有表现力的说话风格预测效果。

然而，将 HCE 合成的富有表现力的话语整合成长篇有声读物语音时，仍会出现一定程度的不自然。例如，相邻句子的韵律变化很大，导致衔接不流畅。这可以解释为 HCE 独立预测每个话语的说话风格，而忽略了它们之间的韵律连贯性。此外，多项研究证实，篇章片段中存在超句子层面的韵律模式，例如音高重置和音高递降；不同的语音风格中也存在此类模式 [14, 15]。因此，为有声读物中的多句文本生成语音时，同时实现符合上下文且连贯的说话风格至关重要。一种直观的方法是构建长篇 TTS 模型，将多个句子甚至一整个段落的音素序列作为输入，直接生成相应的梅尔频谱图 [16, 17]。然而，这种方法需要的内存会随输入长度大幅增加，而且在合成训练语料中未曾出现的超长文本时，性能仍会下降。

本文提出了一种用于有声读物语音合成的上下文感知连贯说话风格预测方法。为了预测当前句子的风格嵌入，我们设计了一种上下文感知风格预测器，同时考虑文本侧的上下文信息和先前语音中的语音侧风格信息。具体而言，该预测器采用带有混合注意力掩码的层次化 Transformer 架构，能够更好地学习上下文信息与风格信息之间的关系。同时，我们使用风格提取器从每个话语的梅尔频谱图中提取风格嵌入，明确引导上下文感知风格预测器的训练。通过这种方式，我们可以逐句生成风格和韵律连贯的长篇语音。在一个普通话有声读物数据集上进行的客观和主观评估表明，无论是在单句还是多句测试中，所提模型在合成语音的表现力和连贯性方面均优于所有基线方法。

2. 方法

我们所提模型的架构如图 1 所示。它由三个主要部分组成：风格提取器、上下文感知风格预测器，以及一个基于 Fast-

‡第一作者在 Xverse Inc. 实习期间完成的工作。

† 共同贡献。 * 通讯作者。[l]语音样本：

https://thuhcsi.github.io/icassp2023-coherent-tts

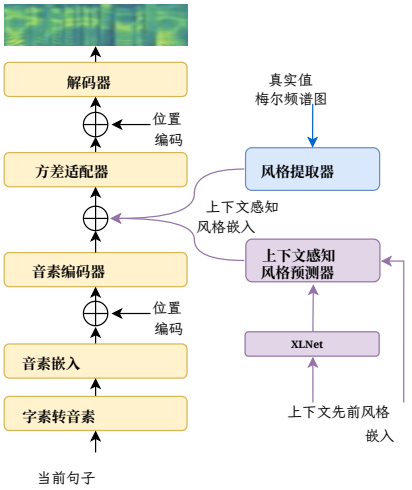

图 1。所提方法的整体架构 ø¢·ɔḿ el

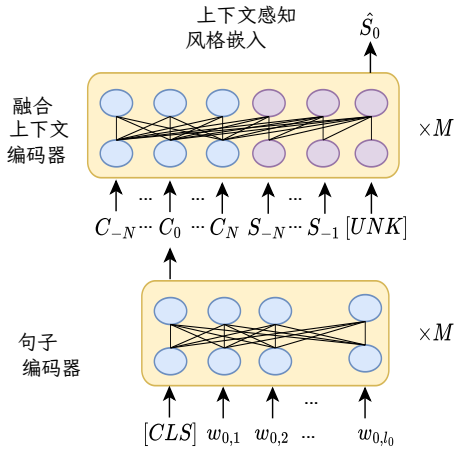

图 2。上下文感知风格预测器

Speech 2 [2] 的序列到序列声学模型。风格提取器用于从语音中提取风格嵌入。上下文感知风格预测器用于预测每个句子的风格嵌入，其输入包括该句子的上下文以及先前语音的风格嵌入。随后，将提取器或预测器提供的风格嵌入复制到音素级，并添加到声学模型中音素编码器的输出上，以合成富有表现力且连贯的语音。各个组成部分的详细介绍如下。

2.1. 风格提取器

受全局风格标记（GST） [18]模型在风格迁移方面成功的启发，我们引入了一种风格提取器，以无监督的方式从语音中提取话语级风格嵌入。该风格提取器由参考编码器和风格词元层组成。这些模块中的每一个都采用与原始GST相同的架构和超参数。梅尔频谱图被输入风格提取器，其输出被视为该话语的风格嵌入。

2.2. 上下文感知风格预测器

为了更好地利用文本侧的上下文信息和语音侧的风格信息，我们提出了一种基于 HCE 的上下文感知风格预测器。上下文信息表示为当前句子之外固定数量的过去和未来相邻句子。风格信息表示为从过去句子的语音中提取的风格嵌入，因为在推理过程中，模型只能获得之前的语音。

设 N 为过去或未来所考虑的句子数量。我们首先将所有 $2N+1$ 句子连接起来，形成一个长文本序列作为上下文。由于 XLNet [19] 可以直接处理更长的文本，甚至是不受长度限制的段落，因此我们使用 XLNet 从该序列中提取词级文本嵌入，记为 ($w_{-N,\,1}$，$w_{-N,\,2}$，…，$w_{N,\,l_N}$)，其中 l_i 是句子 i中的词数。为了捕捉句子之间固有的结构信息，我们引入了一个由句子编码器和融合上下文编码器组成的分层框架，如图 2 所示。这两个编码器具有相似的架构，均由一组 Transformer 模块组成。首先，将每个句子的词嵌入序列输入句子编码器，其中前面添加一个特殊的 [CLS]词元。然后，句子编码器通过探索句子中的低级词义，将输入转换为隐藏状态序列。与 [CLS] 词元对应的第一个隐藏状态被视为句子 i 的上下文词元，记为 C_i。之后，上下文词元序列 $C_{-N}, C_{1-N}, \cdots, C_N$ 被传递至融合上下文编码器，以学习句子间的高级上下文语义，用于风格预测。

为进一步提升风格的连贯性，对应于前文句子的风格嵌入序列 S_{-N}，S_{1-N}，$\cdots$，S_{-1} 也会传入融合上下文编码器。我们将文本侧和语音侧信息合并，形成新的词元序列，并添加一个特殊的未知词元（[UNK]），用于标记待预测的风格嵌入。词元序列 e 表示为：

$$e = (C_{-N}, C_{1-N}, \cdots, C_N, S_{-N}, S_{1-N}, \cdots, S_{-1}, [UNK]) \tag{1}$$

为了更好地学习超句子韵律模式，对于每个词元，其输入表示由相应的词元、类别、位置和片段嵌入相加构成。当 $N = 2$ 时，这种构造的可视化结果见图3。类别嵌入用于区分文本侧词元和语音侧词元。位置嵌入用于提供句子的相对位置信息。而片段嵌入用于提供句子在其段落中的位置。不同于句子编码器，我们专门为融合上下文编码器设计了一种混合注意力掩码，灵感来自 [20], ，以更好地利用上下文信息和先前风格信息。如图2所示，上下文词元只能从两个方向关注彼此的其他上下文词元，而风格标记可以关注先前的风格标记和自身，以及所有上下文词元。这与人类感知的过程一致。在对这两类信息进行编码后，上下文编码器的输出是一系列上下文感知表示，最后一个表示（即与特殊 [UNK] 词元相对应的表示）被用作当前句子的预测上下文感知风格嵌入。

2.3. Training Strategy and Inference Procedure

在训练阶段，为了促使上下文感知风格预测器学习风格表征，我们训练了所提方法。

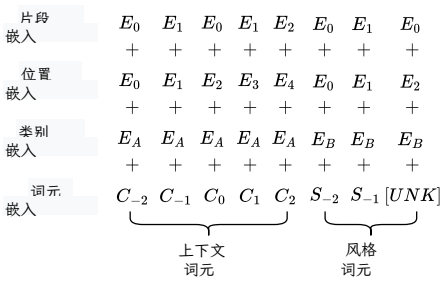

图3。上下文感知风格预测器输入表示

采用知识蒸馏策略，分三个步骤进行。第一步，以无监督的方式联合训练声学模型和风格提取器，以获得训练良好的风格提取器。第二步，我们使用从当前句子的真实语音中提取的风格嵌入作为目标，指导预测器的训练，同时也从真实语音中提取之前的风格嵌入。此外，当过去或未来句子的数量少于 L 时，我们会在文本侧和语音侧都应用零填充方法。最后，我们以较低的学习率联合训练声学模型和预测器，以进一步提升合成语音的自然度。

在有声读物的多句推理过程中，整段长篇语音会逐句生成并按顺序合并。由于此阶段没有真实标注，提取器可以从先前合成的语音或其他人工确定的语音中提取之前的风格嵌入。这样，我们的方法便能生成说话风格恰当且连贯的长篇语音。

3. 实验

3.1. 实验设置

实验中，我们使用了一个内部的单说话人有声读物语料库。该语料库包含约 30 小时的录音，由一位专业的男性普通话母语者朗读一部富有表现力的小说中的约 87 个章节录制而成。数据集共包含 14500 段音频片段，其中 90% 用于训练，10% 用于验证和测试。

对于特征提取，我们将原始波形转换为 80 维梅尔频谱图，采样率为 24 kHz。我们的实验使用了预训练的中文 XLNet-base 模型[2]。同时，我们使用自动语音识别工具进行强制对齐，以获取音素时长。上下文感知风格预测器所考虑的上下文由当前句子、前面两个句子和后面两个句子组成。

我们使用一块 NVIDIA A10 GPU 训练模型，批量大小为 16。第一阶段训练至 180k 次迭代，第二阶段为 20k 次迭代，最后阶段为 30k 次迭代，并使用 Adam 优化器及 $\beta_1 = 0.9$，$\beta_2 = 0.98$。话语编码器和上下文编码器均包含 3 个 Transformer 模块。此外，使用训练良好的 HIFI-GAN [21] 作为声码器生成波形。

在我们的评估中，我们实现了两个基于 FastSpeech 2 的基线系统进行比较:

FastSpeech 2： 一个开源的 FastSpeech 2[3] [2]。

HCE： 分层上下文编码器（HCE） [13] 方法，仅根据上下文预测风格嵌入。

3.2. 主观评估

由于所提方法面向有声读物语音合成，我们进行了两种平均意见分（MOS）测试，以评估生成语音的表现力和自然度：1）单句平均意见得分（S-MOS）：评估单个句子的合成语音；2）多句平均意见得分（M-MOS）：评估按顺序组合多个句子的合成语音而成的整段长篇语音。我们邀请了20名中文母语者听取15个样本，并按1至5分、间隔1分的量表进行评分。如表1所示，所提方法的单句平均意见得分（S-MOS）最高，为3.990；多句平均意见得分（M-MOS）最高，为3.917。HCE的单句平均意见得分（S-MOS）接近所提方法，但其多句平均意见得分（M-MOS）与所提方法相差显著，差距超过0.15。此外，所提方法在这两项指标上均大幅优于 FastSpeech 2。

我们还分别对所提方法与两个基线系统进行了 ABX偏好测试。同样，我们进行了单句ABX（S-ABX）和多句ABX（M-ABX）测试。两个系统之间的偏好测试与平均意见得分测试的视角相同，即选择更受偏好的一方。结果如图4所示。单句情况下，所提方法的偏好率比 FastSpeech 2 高 26.3%；多句情况下高 11.1%。与 HCE 相比，单句情况下高 6.7%，多句情况下高 24.3%。

表1。不同模型的单句平均意见得分（S-MOS）和多句平均意见得分（M-MOS）及其 95%置信区间。

模型	单句平均意见得分	多句平均意见得分
真实值	4.803 ± 0.057	4.826 ± 0.060
FastSpeech 2	3.737 ± 0.084	3.792 ± 0.081
HCE	3.923 ± 0.083	3.761 ± 0.078
所提方法	$\mathbf{3.990 \pm 0.072}$	$\mathbf{3.917 \pm 0.078}$

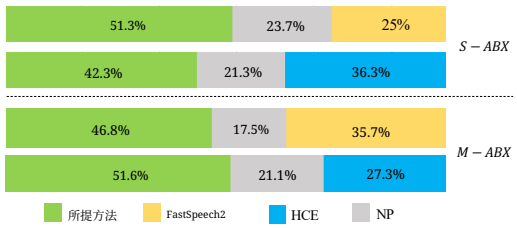

图 4。S-ABX 和 M-ABX 偏好测试的结果。NP 表示无偏好。

平均意见分（MOS）和ABX偏好测试均表明，我们的所提方法不仅能够提升多句语音的表现力和自然度，也能提升单句语音的表现力和自然度。对于单句语音，与仅使用音素序列作为输入的基础FastSpeech 2相比，我们的所提方法和HCE 表现更好，这表明考虑上下文信息确实有助于TTS 模型合成具有更多风格变化的语音。与HCE相比，我们的所提模型性能进一步提升，这是因为考虑了超句层面的韵律模式。此外，我们观察到，所合成的语音

[2]https://github.com/ymcui/Chinese-XLNet [3] https://github.com/ming024/FastSpeech2

FastSpeech 2 生成的语音在句子之间风格过渡相对平滑，但缺乏表现力；而 HCE 的表现力丰富，但句与句之间变化很大。这两者在多句测试中的表现都不佳。这表明，除了上下文之外，考虑语音侧先前的风格信息，有助于为有声读物语音合成生成富有表现力且连贯的语音。

3.3. 客观评估

为客观衡量合成性能，我们按照先前工作的做法 [2, 13]，计算基频的均方根误差（RMSE）、时长的均方误差（MSE）以及梅尔倒谱失真（MCD），作为客观评估指标。在计算基频 RMSE 之前，我们首先使用动态时间规整（DTW）构建合成梅尔频谱图与真实值之间的对齐路径。随后，按照 DTW 路径，将从合成语音中提取的基频序列与真实值对齐。我们还利用 DTW 对齐两幅梅尔频谱图，以计算最小 MCD。对于时长，我们计算预测时长与真实时长之间的 MSE。此外，为衡量风格预测的准确性，我们计算预测风格嵌入与真实风格嵌入之间的 MSE。结果如表 2 所示。所提方法在所有评估指标上均优于两个基线，表明其合成语音与真实语音更为接近。此外，我们还发现，与 HCE 相比，所提方法通过考虑语音连贯性，进一步提升了风格预测效果。

表 2。不同模型的客观评估。

	FastSpeech 2	HCE	所提方法
基频均方根误差	65.1719	63.6806	**62.1702**
时长均方误差	0.2174	0.2067	**0.1974**
MCD	5.0608	5.0376	**4.8988**
风格均方误差	–	0.3148	**0.2524**

3.4. 消融实验

我们进行了三项消融研究，以进一步探究所提方法中所采用的若干技术的有效性，包括利用混合注意力掩码、先前风格信息以及层次化变换器架构。我们采用平均意见得分差异（CMOS）来比较多句文本合成语音的自然度和表现力，并计算风格均方误差，以衡量风格预测的准确性。结果如表3所示。与所提方法相比，移除这些技术后，性能分别在不同程度上有所下降，
这表明，所有这些组件都会对我们所提模型产生显著影响。与 HCE 类似，分层架构对风格预测贡献很大。此外，结果表明，引入先前风格信息有助于提升性能，而我们所提的混合注意力掩码还能进一步改善性能。

3.5. 案例研究

如图5所示，我们通过比较 HCE 和所提方法生成的语音的梅尔频谱图与基频轮廓，并将其与拼接后的真实的前序语音进行对比，开展了一项案例研究。具体而言，红色虚线表示句子边界，其后为合成部分。它是

表3。消融研究结果。

模型	CMOS	风格均方误差
所提方法	–	**0.2524**
- 混合注意力掩码	-0.148	0.2670
- 先前风格信息	-0.172	0.2805
- 分层架构	-0.220	0.3255

可以观察到，HCE 合成的语音比所提方法合成的语音具有更大的基频波动。但由于缺少先前风格信息，从前一个话语到当前话语的基频变化显著，因此将这些话语拼接成长篇语音时会显得有些不自然。与 HCE 相比，我们所提方法合成的语音在语调趋势以及前后话语之间的风格变化方面，更接近真实语音。此外，我们尝试使用所提模型合成同一个句子，同时将前一段语音的 F_0 调整 $+50$ Hz。可以观察到，合成语音的 F_0 也随之改变，这表明所提模型成功学会了根据先前语音生成连贯风格。因此，同一个句子可以通过多种方式合成，从而实现更加自然且富有表现力的语音合成。

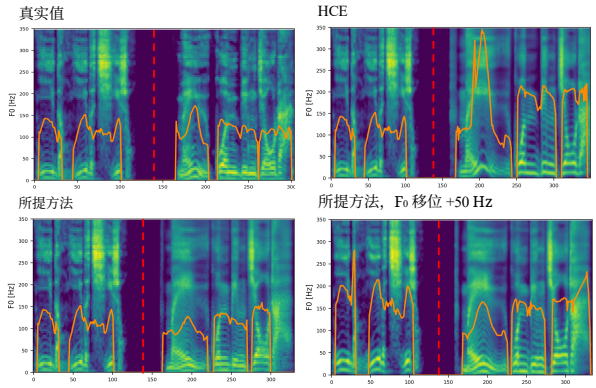

图5。测试集中一个示例话语由不同模型合成的语音的梅尔频谱图和基频轮廓。

4. 结论

在本文中，我们提出了一种用于有声读物语音合成的上下文感知连贯说话风格预测方法。引入了一种基于层次化Transformer的上下文感知风格预测器，并结合混合注意力掩码，同时考虑先前语音的上下文信息和风格信息。实验结果表明，对于单句和多句测试，所提方法均能显著提升合成语音的表现力和连贯性。

致谢：本工作得到国家自然科学基金委员会（62076144）、鹏城实验室重大重点项目（PCL2021A06，PCL2022D01）、深圳市下一代互动媒体创新技术重点实验室（ZDSYS20210623092001004）和深圳市科技计划（WDZC20220816140515001，JCYJ20220818101014030）的支持。

5. 参考文献

[1] Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan 等人，“通过以梅尔频谱预测为条件的 WaveNet 实现自然语音合成”，载于*2018 年IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2018年，第 4779–4783 页。[2] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，“FastSpeech 2：快速高质量的端到端文本转语音”，载于国际学习表征会议，2020年。[3] Daisy Stanton、Yuxuan Wang 和 RJ Skerry-Ryan，“在端到端语音合成中从文本预测富有表现力的说话风格”，载于*2018*年*IEEE*口语技术研讨会 *(SLT)*。IEEE，2018年，第 595–602 页。[4] Yuxuan Wang、RJ Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio 等人，“Tacotron：迈向端到端语音合成”，*Interspeech 2017*论文集，第 4006–4010 页，2017年。[5] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，“BERT：用于语言理解的深度双向 Transformer 预训练”，载于*NAACL-HLT*论文集，2019年，第 4171–4186 页。[6] Ya-Jie Zhang 和 Zhen-Hua Ling，“提取并预测语音合成中的词级风格变化”，*IEEE/ACM*音频、语音与语言处理汇刊，第 29 卷，第 1582–1593 页，2021年。[7] Yi Ren、Ming Lei、Zhiying Huang、Shiliang Zhang、Qian Chen、Zhijie Yan 和 Zhou Zhao，“Prosospeech：通过文本转语音中的量化向量预训练增强韵律”，载于*ICASSP 2022-2022*年*IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2022年，第 7577–7581 页。[8] Zhaoci Liu、Ningqian Wu、Yajie Zhang 和 Zhenhua Ling，“将离散词级风格变化融入用于语音合成的非自回归声学模型”，载于*Interspeech 2022*论文集，2022年，第 5508–5512 页。

[9] Xu Tan、Tao Qin、Frank Soong 和 Tie-Yan Liu，“神经语音合成综述”，*arXiv* 预印本 *arXiv:2106 15561*，2021年。。

[10] Jennifer Cole，《语境中的韵律：综述》，《语言、认知与神经科学》，第 30 卷，第 1–2 期，第 1–31 页，2015年。[11] Guanghui Xu、Wei Song、Zhengchen Zhang、Chao Zhang、Xi- aodong He 和 Bowen Zhou，“利用跨话语 BERT 嵌入改进端到端语音合成的韵律建模”，载于 *ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第 6079–6083 页。

[12] Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi、Naoko Tanji、Yusuke Ijima、Ryo Masumura 和 Hiroshi Saruwatari，《基于跨句子上下文感知词嵌入的有声读物语音合成》，发表于第 *11* 届 *ISCA* 语音合成研讨会（*SSW 11*），2021年，第 211–215 页。

[13] Shun Lei、Yixuan Zhou、Liyang Chen、Zhiyong Wu、Shiyin Kang 和 Helen Meng，《面向普通话的结合分层上下文信息的表现力说话风格建模》语音合成”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第 7922–7926 页。[14]Mireia Farr´us、Catherine Lai 和 Johanna D Moore，“面向语音合成应用的基于段落的韵律线索”，载于*Barnes J*、*Brugos A*、*Shattuck-Hufnagel S*、*Veilleux N* 编。*Speech Prosody 2016*；*2016*年*5*月*31*日—*6*月*3*日；波士顿（美国马萨诸塞州）。*[*地点未知*]*：国际语音通信协会；*2016*年。第 *1143-7* 页。，2016年。[15] `Alex Peir´o Lilja 和 Mireia Farr´us，“通过段落韵律模式提升文本转语音自然度”，载于*Klessa K*、*Bachan J*、*Wagner A*、*Karpi´nski M*，´*Sledzi´nski D* 编。《第九届国际语音韵律会议论文集》；*2018*年*6*月*13—16*日；波兰波兹南´。*[Lous Tourils]*：*ISCA*；*2018*年。第 *512-6* 页。国际语音通信协会（ISCA），2018年。[16] Peter Makarov、Syed Ammar Abbas、Mateusz Lajszczak、Arnaud Joly、Sri Karlapati、Alexis Moinet、Thomas Drugman 和 Penny Karanasou，“采用富有表现力且连贯的韵律的简单有效的多句文本转语音”，载于论文集：*Interspeech 2022*，2022年，第 3368–3372 页。[17] Liumeng Xue、Frank K Soong、Shaofei Zhang 和 Lei Xie，“Paratts：在基于段落的文本转语音中学习跨句子的语言与韵律信息”，*IEEE/ACM*音频、语音与语言处理汇刊，2022年。[18] Yuxuan Wang、Daisy Stanton、Yu Zhang、RJ-Skerry Ryan、Eric Battenberg、Joel Shor、Ying Xiao、Ye Jia、Fei Ren 和 Rif A Saurous，“风格标记：端到端语音合成中的无监督风格建模、控制与迁移”，载于国际机器学习会议。PMLR，2018年，第 5180–5189 页。[19] Zhilin Yang、Zihang Dai、Yiming Yang、Jaime Carbonell、Russ R Salakhutdinov 和 Quoc V Le，“XLNet：面向语言理解的广义自回归预训练”，《神经信息处理系统进展》，第 32 卷，2019年。[20] Li Dong、Nan Yang、Wenhui Wang、Furu Wei、Xiaodong Liu、Yu Wang、Jianfeng Gao、Ming Zhou 和 Hsiao-Wuen Hon，“面向自然语言理解与生成的统一语言模型预训练”，《神经信息处理系统进展》，第 32 卷，2019年。[21] Jungil Kong、Jaehyeon Kim 和 Jaekyoung 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 17022–17033 页，2020年。