arXiv:2301.08810v1 [cs.CL] 20 Jan 2023

# 利用字素预测增强文本转语音韵律的音素级 BERT

*Yinghao Aaron Li*、*Cong Han*、*Xilin Jiang*、*Nima Mesgarani*

美国哥伦比亚大学电气工程系

**摘要**

研究表明，大规模预训练语言模型能够生成更加自然的韵律模式，从而有助于提升文本转语音（TTS）模型的自然度。然而，这些模型通常基于词级或子音素级建模，并与音素联合训练，因此在下游文本转语音任务中效率较低，因为该任务只需要音素。本文提出一种音素级 BERT（PL-BERT），在常规掩码音素预测之外，增加了预测对应字素的前文本任务。主观评估表明，与最先进水平（SOTA）的 StyleTTS 基线相比，我们的音素级 BERT 编码器显著提高了分布外（OOD）文本上合成语音自然度的平均意见分（MOS）。

索引词— 文本转语音、预训练、BERT、迁移学习

## 1. 引言

近年来，文本转语音（TTS）取得了显著进展。最新研究表明，对于分布内文本，合成语音在人工评审者的主观评价中已与自然人类语音难以区分 [1]。尽管近来取得了诸多进展，但由于人类语音的韵律和情感包含丰富的信息，合成自然且富有表现力的语音仍然是一项挑战 [2]。许多 TTS 模型难以捕捉的一个关键方面，是语音的语调，也就是韵律 [3]。训练 TTS 模型就像从头开始学习一门语言。要学会外语正确的语调和情感，数百小时的输入至关重要。即便有这么多小时的输入，非母语者仍可能因其语调和韵律而轻易被辨认出来。另一方面，由于需要进行数据标注，文本转语音数据集通常远不到数百小时。数据仅有数小时，若以单纯的音素作为输入，训练出的模型预计难以捕捉自然的韵律模式。因此，需要大规模预训练模型来缓解这一问题。尤其是 BERT，已证明无论在词级[4], 还是字符级 [5, 6],，都能有效提升 TTS 模型的性能

还是句子级 [7]。

尽管这些 BERT 模型在改善语音合成的韵律和自然度方面取得了成功，但它们并未在音素级进行训练，尽管下游文本转语音任务的输入仅由音素组成。PnG-BERT[8] 尝试通过联合使用音素和字素词元作为输入，并预测音素和字素的掩码词元来解决这一问题。这种方法从字素和音素中学习到了更丰富的表征，但它仅适用于固定的字素词元集合，并且在训练过程中可能无法处理未见词。此外，字素词元的数量大得难以承受，导致模型训练和推理速度缓慢。近期的一项工作——混合音素 BERT（MP-BERT） [9], ——通过训练一个仅以音素作为输入的 BERT 模型，避免了对字素的需求。由于音素在语言表达能力上不如字素，MP-BERT 还利用字节对编码（BPE） [10] 学习了一组超音素单元，以增强所学习表征的语义内容。尽管无需字素输入，MP-BERT 在下游 TTS 任务中的性能仍可与 PnG-BERT 相媲美。然而，无法保证通过 BPE 学习到的超音素单元能够携带与字素同等丰富的语言信息。此外，超音素单元所需的词元数量在 [9],中多达 30,000 个，严重限制了训练和推理的速度。

在此，我们提出一种用于文本转语音合成的音素级 BERT 模型。通过结合整词掩码音素预测与字素预测，我们获得了一种音素级语言模型；它比 MP-BERT 更高效，且无需将字素或超音素单元作为输入。我们的贡献在于新增了一项前文本任务：预测每个音素对应的字素（音素到字素，P2G）。通过直接在音素级学习语言模型，该模型能够生成深刻掌握音素、词语与语义之间动态关系的表征，从而提升下游 TTS 任务的性能。主观评估表明，对于分布外（OOD）文本，在语音感知自然度方面，我们的音素级 BERT 显著优于当前最先进的基线 StyleTTS 模型 [3]。我们还证明，与 OOD 文本相比，针对分布内文本的评估效果并不理想，并提出了一个未来研究方向：

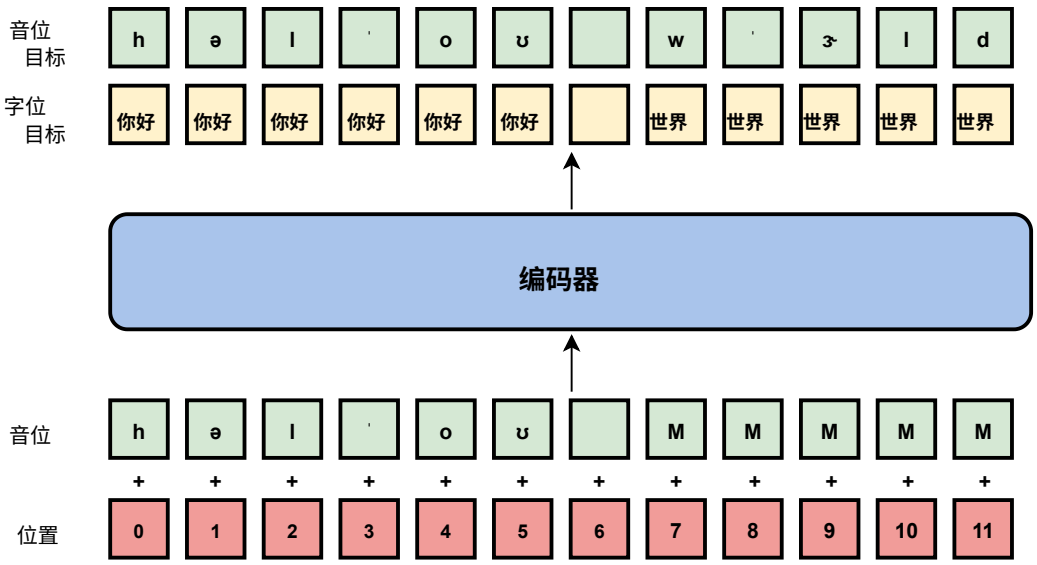


**图 1**。音素级 BERT 的预训练方案。 $M$ 表示输入词元已被掩码。Transformer 编码器将音素词元及其位置编码作为输入，并通过训练来预测被掩码的音素词元及其对应的字位词元。

通过 TTS 研究缩小分布内文本与 OOD 文本之间的性能差距。音频样例可在 https://pl-bert.github.io/ 收听。

## 2. 音素级 BERT

### 2.1. 音素表示

在音素级 BERT 中，我们仅将音素作为输入，因为音素是实现可懂语音合成所需的唯一信息。我们不使用 [8] 和 [9] 中采用的字位或超音素表示，这是因为其巨大的词汇量会降低训练和推理速度。此外，超出音素之外的额外表示还会产生未登录词问题，即在推理过程中可能出现未见过的词语或超音素单元。仅使用音素表示可以解决这些问题，使预训练编码器能够直接替代任何文本转语音系统中的文本编码器。

### 2.2. 预训练

与原始 BERT 类似，只要语料库中能够成对获得音素和字素，就可以采用自监督方式训练音素级 BERT。音素及其对应的字素可以使用外部字位到音素（G2P）工具预先生成。字素的范围可以从字符、子词单元到完整词语。由于字符或子词单元的发音可能会发生变化，因此可能需要通过动态规划算法进一步进行字素—音素对齐具体取决于许多语言中的语境，例如日语。为简单起见，我们仅使用完整词语作为词元，以避免额外的字素—音素对齐。

#### *2.2.1.* 训练目标

预训练有两个目标：掩码音素词元预测（掩码语言模型）和音素到字素（P2G）预测。与原始 BERT 模型一样，我们使用线性投影和 Softmax 函数，根据最后一层的隐藏状态预测被掩码的输入音素词元。损失函数采用多类预测中常用的交叉熵损失。对于每个音素词元，我们也会用相同的流程将其隐藏状态映射为对应字位的预测。我们仅对被掩码的词元计算掩码语言模型损失值，而对所有输入词元计算 P2G 损失。正如第 4.2 节所示，这一目标对于学习有意义的音位级语言表示至关重要，能够显著提升文本转语音任务的效果。训练目标可写为如下形式：

$$\mathcal{L}_{MLM} = \mathbb{E}_{\boldsymbol{x},I,\boldsymbol{y}_p}\left[\sum_{i\in I}\mathbf{CE}(P_{MLM}\left(E(\boldsymbol{x})\right)_i, \boldsymbol{y}_{\boldsymbol{p}_i})\right], \quad (1)$$

$$\mathcal{L}_{P2G} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_g}\left[\sum_{i=1}^{N}\mathbf{CE}(P_{P2G}\left(E(\boldsymbol{x})\right)_i, \boldsymbol{y}_{\boldsymbol{g}_i})\right], \quad (2)$$

其中， $\boldsymbol{x}$ 是经过掩码的音素词元， $\boldsymbol{y}_p$ 是原始未掩码的音素词元标签， $\boldsymbol{y}_g$ 是相应的

字素标记标签， $I$ 是掩码索引， $E$ 是我们的音素级BERT编码器， $P_{MLM}$ 是用于掩码语言模型任务的线性投影， $P_{P2G}$ 是用于 P2G 任务的线性投影， $N$ 是文本的总长度，**CE**(·)表示交叉熵损失函数。

#### *2.2.2.* 掩码策略

由于我们的目标是学习音素级语言模型，因此需要在词级进行掩码，使模型能够学习有意义的语义表示。这种掩码策略称为整词掩码 [11]，研究表明，对于以音素为输入的 BERT 模型而言，这是最有效的掩码策略 [8, 9]。我们采用整词掩码，并遵循先前的研究 [8, 9, 12]：随机选取每个序列中 15% 字素对应的音素词元进行掩码。选中某个字素后，其音素词元有 80% 的概率被替换为 MSK 标记，10% 的概率被替换为随机音素词元，还有 10% 的概率保持不变。

## 3. 实验

### 3.1. 数据集

#### *3.1.1.* 文本预训练数据

我们在英语维基百科语料库上预训练音素级 BERT 模型。该语料库包含 6,280,802 篇文章和约 7,400 万个句子。我们将数据集划分为若干部分：600 万篇文章用于训练，14 万篇用于验证，其余用于测试。我们使用 NeMo [13] 对文本进行规范化，使其与每个单词的发音相匹配。我们使用 Phonemizer [14]获取音素；它通过 eSpeak 后端将文本序列转换为国际音标（IPA）。

#### *3.1.2.* 文本转语音微调数据

我们使用 LJSpeech 数据集 [15] 评估下游 TTS 任务的性能。LJSpeech 数据集包含 13,100 个短音频片段，总时长约为 24 小时。数据集划分为训练集、验证集和测试集，分别包含 12,500、100 和 500 个样本。我们使用快速傅里叶变换大小为 2048、帧移为 300、窗长为 1200，并设有 80 个梅尔频带的参数提取梅尔频谱图。我们使用 Hifi-GAN 声码器 [16]根据梅尔频谱图合成波形。

### 3.2. 训练细节

我们的音素级 BERT 是一个 12 层 ALBERT [17] 模型，隐藏层大小为 768，中间层大小为 2,048，并设有 12 个注意力头。训练使用 3 块 Nvidia A40 GPU 进行，最大长度为 512 个词元，批次大小为 192 个样本。对于 MP-BERT 基线，我们使用了 [9] 中的 BPE 基础词典，其中包含 30,000 个超音素单元。模型训练了 1M 步，约为 10 个 epoch。

**表 1**. LJSpeech 数据集上评估所得平均意见分（含 95%置信区间）以及使用 PL-BERT 得到的比较平均意见分。

| 模型 | MOS | CMOS |
|---|---|---|
| 真实值 | 4.34（± 0.09） | 0.25 |
| 采用 PL-BERT 的 StyleTTS | 4.24（± 0.10） | 0.00 |
| StyleTTS | 4.19（± 0.10） | – 0.02 |

**表 2**。比较了分布外（OOD）数据集上 StyleTTS 搭配 MP-BERT 的平均意见分（MOS）及其 95% 置信区间（CI），以及比较平均意见分（CMOS）。此外，我们将 [3]中的分布内文本的平均意见分（MOS）（LJ）纳入比较。

| 模型 | 平均意见分（LJ）[3] | 平均意见分（OOD） | CMOS |
|---|---|---|---|
| StyleTTS 搭配 PL-BERT | – | **3.64** (± **0.09**) | **0.16** |
| StyleTTS 搭配 MP-BERT | – | 3.55 (± 0.10) | 0.00 |
| StyleTTS | **4.01** (± **0.05**) | 3.49 (± 0.09) | – 0.39 |
| VITS | 3.78 (± 0.06) | 3.37 (±0.10) | — |
| FastSpeech 2 | 2.97 (± 0.06) | 2.86 (±0.11) | — |
| Tacotron 2 | 3.01 (± 0.06) | 2.67 (±0.11) | — |

我们在 StyleTTS 训练的第二阶段对 PL-BERT 进行了 100 个周期的微调。前 50 个周期冻结 PL-BERT 的权重，之后再对其进行 50 个周期的微调，以使训练更加稳定。

### 3.3. 评估

我们通过主观评估自然度的平均意见分（MOS），来衡量合成语音的自然度。我们招募了居住在美国的英语母语者，邀请他们在 Amazon Mechanical Turk 上参与评估。在每项实验中，我们从 LJSpeech 数据集（分布内）和古腾堡图书数据集 [18]（分布外）的测试集中随机选取 30 个句子。后者被视为分布外（OOD），因为用于测试的书籍在预训练 BERT 模型和微调后的 TTS 模型的训练过程中都未曾出现过。相反，由于 LJSpeech 数据集由七本有声读物组成，测试集中文本所属的书籍在训练期间已经出现过。

对于每段文本，我们分别使用经音素级 BERT 模型微调的 StyleTTS、经 MB-BERT 微调的 StyleTTS，以及不含 BERT 的基线 StyleTTS 模型来合成语音。参考音频从训练集中选取，所选音频与目标文本之间具有最高的句子嵌入相似度，该相似度使用 sentence-BERT 计算。 [19] 每组语音由十位评分者评分，分值范围为 1 至 5 分，间隔为 0.5 分。评估每组语音时，我们随机打乱模型顺序，并

要求受试者听取语音并进行评分，但不告知他们模型标签。 [20, 21]我们加入失真语音作为注意力检测项；如果失真语音没有被评为最低分，则该受试者的所有评分都会从分析中剔除。除了 StyleTTS，我们还将 Tacorton 2 [22],FastSpeech 2 [23],和 VITS [24]作为基线模型，用于比较分布外文本的表现。为了检验我们的 PL-BERT 模型是否有帮助，以及其性能是否优于 MP-BERT，我们还开展了几项比较平均意见分（CMOS）研究，要求评分者只听取两个样本，并评判第二个样本是否优于第一个。样本顺序经过打乱，评分范围为 -6 至 6 分，间隔为 1 分。我们还开展了消融研究，以验证模型中各组成部分的有效性。我们要求受试者将我们提出的模型与移除一个组成部分的模型进行比较。为获得更显著的结果，消融研究针对分布外文本开展。此外，我们在维基百科语料库上训练了一个逻辑回归 P2G 预测器，从音素预测字素，以检验学得的表示是否包含上下文中的字素信息。

## 4. 结果

### 4.1. 文本转语音性能

如表 1 所示，对于分布内文本，使用 PL-BERT 并未带来显著提升。然而，从表 2 可以看出，在分布外文本的平均意见分和比较平均意见分两项指标上，我们的模型均显著优于未使用预训练 BERT 的基线模型。具体而言，我们的模型显著优于 MP-BERT（Wilcoxon检验，$p < 0.05$），比较平均意见分高出 0.16。这表明，与使用亚音位单元相比，使用音位预测进行训练能使文本转语音模型更好地泛化到未见过的文本。我们还注意到，分布内文本与分布外文本之间的平均意见分差距很大。StyleTTS 配合 PL-BERT 的结果与真实值之间的平均意见分差异在统计上并不显著（$p > 0.05$），不过，比较平均意见分显示，评分者略微倾向于真实值，而非我们的模型。然而，当输入文本属于分布外文本时，性能会大幅下降。这种性能差距并非 StyleTTS 模型所独有，而是许多文本转语音模型普遍存在的问题，如表 2 所示。这些平均意见分比 [3] 报告的结果低约 0.4 到 0.5 分，且差异显著。结果表明，未来的研究应更加重视对分布外文本的评估。此外，由于我们的模型无需处理亚音位词元，因此在单块 NVIDIA A40 GPU 上的速度比 MX-BERT 快 1.05 倍。

### 4.2. 消融研究

表 3 显示，在训练过程中移除 $\mathcal{L}_{P2G}$ 时，性能略有下降。然而，CMOS 仍高于不使用 BERT 的基线 StyleTTS 模型。这表明

**表 3**。消融研究：与采用 PL-BERT 的 StyleTTS 相比，验证 MLM、P2G 和 BERT 的有效性。

| 模型 | CMOS | 准确率（Top-1） | 准确率（Top-5） |
|---|---|---|---|
| 本文提出的方法 | 0.00 | 67.48% | 90.33% |
| 不含 $\mathcal{L}_{P2G}$ | −0.11 | 13.45 % | 24.97% |
| 不含 $\mathcal{L}_{MLM}$ | −4.57 | 68.73 % | 90.24 % |
| 不含 PL-BERT | −0.44 | — | — |

使用预训练音素级 BERT，即使在不使用 $\mathcal{L}_{P2G}$ 的情况下进行训练，也能改善下游 TTS 任务。这可以归因于我们的掩码策略：其中，整个字素音素都会被掩码，因此仅凭 $\mathcal{L}_{MLM}$ 就能学习到足够丰富的语言表征，从而帮助完成下游 TTS 任务。然而，当移除 $\mathcal{L}_{MLM}$ 时，CMOS 会大幅下降，表明仅使用 $\mathcal{L}_{P2G}$ 训练的模型无法保留输入音素信息，因此不能用于下游 TTS 任务。我们注意到，从训练目标中移除 $\mathcal{L}_{P2G}$ 后，P2G 预测准确率会大幅下降。这表明，即使采用整词掩码，$\mathcal{L}_{MLM}$ 也不能保证模型学习到词级语言表征。这在一定程度上解释了为什么我们的模型优于 MP-BERT，因为 MP-BERT 缺少 $\mathcal{L}_{P2G}$ 目标，而该目标能够使模型在音素级别学习语言表征。

## 5. 结论

在本研究中，我们提出了音素级 BERT，这是一种音素级语言模型，可生成上下文化嵌入，从而提升下游 TTS 任务的自然度和韵律。与以往的研究不同，我们的模型仅以音素作为输入，大幅减少了训练和推理所需的资源。我们证明，与未针对文本转语音模型微调 BERT 编码器的基线 StyleTTS 模型相比，我们的模型性能显著更优；同时也证明，对于分布外（OOD）文本，我们的预训练策略优于 MP-BERT。我们发现，许多现有 TTS 模型在分布内文本与 OOD 文本之间存在性能差距。由于分布内文本在实际应用中很少使用，我们主张未来的研究应更加关注面向 OOD 文本的 TTS 开发。

## 6. 致谢

我们感谢 Gavin Mischler 在本项目开发阶段对模型质量提供反馈。本项目由美国国立卫生研究院（NIH- 美国国家耳聋及其他沟通障碍研究所）资助，并获得 Marie-Josee 和 Henry R. Kravis 的资助。

## 7. 参考文献


[1] X. Tan、J. Chen、H. Liu、J. Cong、C. Zhang、Y. Liu、X. Wang、Y. Leng、Y. Yi、L. He等人，《Naturalspeech：具有人类水平质量的端到端文本转语音合成》，*arXiv*预印本 *arXiv:2205.04421*，2022。

[2] X.Tan、T. Qin、F. Soong 和 T.-Y. Liu，《神经语音合成综述》，*arXiv* 预印本 *arXiv:2106.15561*，2021。

[3] Y. A. Li、C. Han 和 N. Mesgarani，《Styletts：一种用于自然、多样文本转语音合成的风格生成模型》，*arXiv* 预印本 *arXiv:2205.15439*，2022。

[4] T. Kenter、M. K. Sharma 和 R. Clark，《通过融入 BERT 模型改进基于循环神经网络的英语文本转语音合成韵律》，2020。

[5] Y. Xiao、L. He、H. Ming 和 F. K. Soong，《在多说话人普通话神经文本转语音中利用语言学特征和 BERT 派生特征改进韵律》，载于*ICASSP 2020—2020 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2020，第 6704–6708 页。

[6] Y. Zhang、L. Deng 和 Y. Wang，“基于蒸馏BERT模型的统一普通话文本转语音前端”，*arXiv* 预印本 *arXiv:2012.15404*，2020。

[7] G. Xu、W. Song、Z. Zhang、C. Zhang、X. He 和 B. 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，第 6079–6083 页。

[8] Y. Jia、H. Zen、J. Shen、Y. Zhang 和 Y. Wu，“PNG BERT：面向神经文本转语音的音素和字素增强型 BERT”，*arXiv* 预印本 *arXiv:2103.15060*，2021。

[9] G. Zhang、K. Song、X. Tan、D. Tan、Y. Yan、Y. Liu、G. Wang、W. Zhou、T. Qin、T. Lee 等，“混合音素 BERT：通过混合音素和超音素表示改进用于文本转语音的 BERT”，*arXiv* 预印本 *arXiv:2203.17190*，2022。

[10] R. Sennrich、B. Haddow 和 A. Birch，“使用子词单元对稀有词进行神经机器翻译”，*arXiv* 预印本 *arXiv:1508.07909*，2015。

[11] Y. Cui、W. Che、T. Liu、B. Qin 和 Z. Yang，“使用全词掩码对中文 BERT 进行预训练”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，页 3504–3514，2021。

[12] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，《Roberta：一种稳健优化的 BERT 预训练方法》，*arXiv*预印本 *arXiv:1907.11692*，2019。

[13] O. Kuchaiev、J. Li、H. Nguyen、O. Hrinchuk、R. Leary、B. Ginsburg、S. Kriman、S. Beliaev、V. Lavrukhin、J. Cook 等，《NeMo：使用神经模块构建人工智能应用的工具包》，*arXiv* 预印本 *arXiv:1909.09577*，2019。[14] M. Bernard 和 H. Titeux，《Phonemizer：Python 中多语言文本到音素的转写》，《开放源代码软件期刊》，第 6 卷第 68 期，第 3958 页，2021 年。[在线]。可访问：https://doi.org/10.21105/joss.03958[15] K. Ito 和 L. Johnson，《LJ 语音数据集》，https: // keithito.com/LJ-Speech-Dataset/，2017 年。

[16] 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，2020。

[17] Z. Lan、M. Chen、S. Goodman、K. Gimpel、P. Sharma 和 R. Soricut，“ALBERT：一种用于自监督语言表示学习的轻量级 BERT”，*arXiv* 预印本 *arXiv:1909.11942*，2019。

[18] M. Gerlach 和 F. Font-Clos，《用于自然语言统计分析和计量语言学的标准化古腾堡计划语料库》，《*Entropy*》，第 22 卷，第 1 期，126 页，2020 年。

[19] N. Reimers 和 I. Gurevych，《利用知识蒸馏将单语句子嵌入转换为多语言嵌入》，*arXiv* 预印本 *arXiv:2004.09813*，2020 年 4 月。[在线]。网址：http://arxiv.org/abs/2004.09813

[20]Y. A. Li、A. Zare 和 N. Mesgarani，《StarGANv2-VC：一种多样化、无监督、非平行的自然语音转换框架》，*arXiv* 预印本 *arXiv:2107.10394*，2021 年。

[21] Y. A. Li、C. Han 和 N. Mesgarani，《StyleTTS-VC：通过从基于风格的文本转语音模型进行知识迁移实现的一次性语音转换》，*arXiv* 预印本 *arXiv:2212.14227*，2022 年。

[22] 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的 WaveNet 实现自然文本转语音合成”，载于*2018 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018，第 4779–4783 页。

[23] 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”，载于国际学习表征会议，2021 年。 [在线]。可访问：https://openreview.net/forum?id=piLPYqxtWuA

[24] J. Kim、J. Kong 和 J. Son，“用于端到端文本转语音的对抗学习条件变分自编码器”，载于第 *38* 届机器学习国际会议论文集，系列：机器学习研究论文集，M. Meila 和 T. Zhang 编，第 139 卷，2021年7月18日至24日，第 5530–5540 页。