样式标记：端到端语音合成中的无监督样式建模，控制和传输

王雨轩[1]黛西斯坦顿[1]张钰[1]RJ Skerry-Ryan[1]Eric Battenberg[1]乔尔绍尔[1]英晓[1]
非人[1]叶嘉[1]Rif A. Saurous[1]

摘要

在这项工作中，我们提出了“全球风格标记”（GST），这是一个在Tacotron中共同培训的嵌入库，这是一个最先进的端到端语音合成系统。 嵌入训练没有明确的标签，但学会建模大范围的声学表现力。 消费税导致了一系列重要结果。 他们生成的软解释“标签”可用于以新颖的方式控制合成，如变化速度和讲话风格 - 独立于文本内容。 它们也可以用于样式转换，在整个长形文本语料库中复制单个音频剪辑的发言风格。 在嘈杂的，未标记的发现数据上进行训练时，GST学会将噪音和说话人身份因数分解，为高度可扩展但强大的语音合成提供了一条途径。

1. 介绍

在过去的几年中，使用深度神经网络来合成自然发音的人类语音（Zen et al。，2016; van den Oord等人，2016; Wang等人，2017a; Arik等人，2017; 泰格曼等人。，2017; 沉等人。，2017）。 随着文本到语音（TTS）模型的快速发展，许多应用程序（如有声读物解说，新闻阅读器和会话助理）的机会越来越多。 神经模型显示出有力地综合表达性长篇演讲的潜力，但该领域的研究尚处于起步阶段。

为了传递真实的类人言语，TTS系统必须学会模仿韵律。韵律是言语中的一些现象的汇合，例如辅助语言信息，语调，压力和风格。 在这项工作中我们关注

[1]Google, Inc. .. 通信地址： 王宇轩 <yxwang@google.com>.

声音演示可以在HTTPS：//google.github。 IO / tacotron /出版/ global_style_ 令牌

在风格建模上，其目标是为模型提供选择适合给定上下文的讲话风格的能力。 虽然难以精确定义，但风格包含丰富的信息，如意图和情感，并影响说话者对语调和流动的选择。 适当的风格渲染会影响整体感知（例如，参见（例如）“情感韵律”泰勒，2009）），这对于有声读物和新闻阅读器等应用程序非常重要。

风格建模提出了几个挑战。 首先，没有“正确的”韵律风格的客观测量，使得建模和评估困难。 获取大型数据集的注释可能代价昂贵，同样也存在问题，因为人们的评价者往往不同意。 其次，具有表现力的声音中的高动态范围很难建模。 许多TTS模型（包括最近的端到端系统）仅在其输入数据上学习平均韵律分布，从而产生较少表达性言语，尤其是对于长式短语。 此外，他们往往缺乏控制语音合成表达的能力。

这项工作试图通过向Tacotron引入“全球风格令牌”（GST）来解决上述问题（Wang等人，2017a; 沉等人。，2017），这是一个最先进的端到端TTS模型。 GST训练时没有任何韵律标签，但仍然发现了大量的表现风格。 内部架构本身产生可用于执行各种风格控制和传输任务的软解释“标签”，从而显着改善表现性长时间合成。 GST可以直接应用于嘈杂的，未标记的数据，为高度可扩展但强大的语音合成提供了一条途径。

2. 模型架构

我们的模型基于Tacotron（Wang等人，2017a; 沉等人。，2017），序列到序列（seq2seq）模型，直接从字形或音素输入预测mel谱图。 这些梅尔谱图通过低资源反演算法转换为波形（格里芬和林，1984）或神经声码器，如WaveNet（van den Oord等人，2016）。 我们指出，对于Tacotron而言，声码器的选择不会影响韵律，也就是说

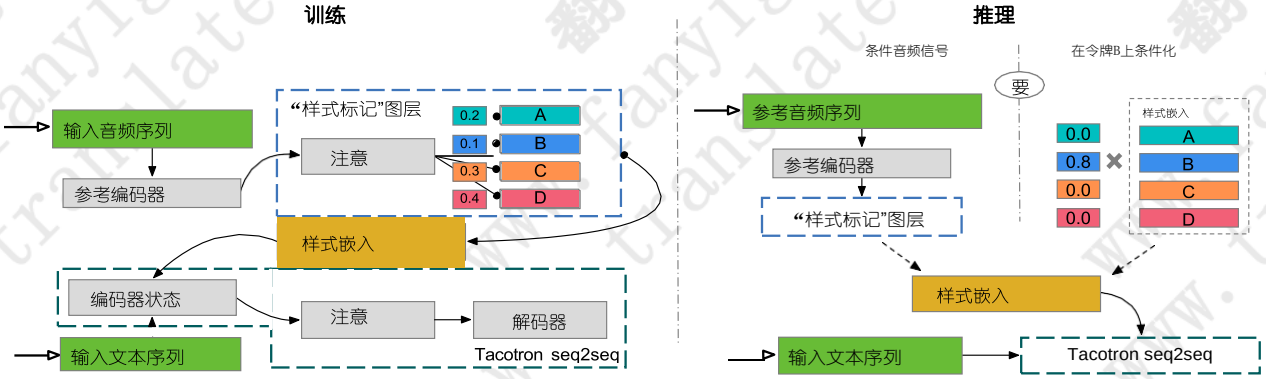


图1. 模型图。 在训练期间，训练目标的对数谱图被馈送到参考编码器，然后是样式标记层。 生成的样式嵌入用于调整Tacotron文本编码器状态。 在推断过程中，我们可以提供一个任意的参考信号来合成具有其说话风格的文本。 或者，我们可以移除参考编码器，并使用学习到的可解释令牌直接控制合成。

由seq2seq模型建模。

我们建议的GST模型，如图所示1由参考编码器，样式注意，样式嵌入和序列 - 序列（Tacotron）模型组成。

2.1. 训练

在培训期间，信息流经模型如下：

- 参考编码器，在（斯凯利瑞安 等人。，2018）将可变长度音频信号的韵律压缩成一个固定长度的向量，我们称之为参考嵌入。 在训练期间，参考信号是地面真实音频。
- 参考嵌入被传递给样式标记层，在那里它被用作关注模块的查询向量。 在这里，关注不用于学习对齐。 相反，它学习参考嵌入和随机初始化嵌入库中每个令牌之间的相似性度量。 这组嵌入，我们交替地调用全局样式标记，GST或令牌嵌入，在所有训练序列中共享。
- 注意模块输出一组组合权重，其表示每个样式标记对编码参考嵌入的贡献。 我们称之为风格嵌入的GST的加权和被传递给文本编码器，以便在每个时间步进行调节。
- 样式表征层与模型的其余部分共同训练，仅由Tacotron解码器的重建损失驱动。 GST因此不需要任何明确的样式或韵律标签。

2.2. 推理

GST架构专为推理模式下的强大而灵活的控制而设计。在这种模式下，信息可以通过以下两种方式之一流过模型：

1. 我们可以直接在某些标记上调整文本编码器，如图的推理模式图的右侧所示1 （“令牌B条件”）。 这允许在没有参考信号的情况下进行风格控制和操纵。
2. 我们可以输入不同的音频信号（其记录不需要与要合成的文本相匹配）来实现样式转换。 这在图的推理模式图的左侧描述1 （“音频信号条件”）。

这些将在第一节中详细讨论6.

3. 模型细节

3.1. Tacotron建筑

对于我们的基线和GST增强Tacotron系统，我们使用与（王 等人。，2017a）除了一些细节。 我们使用音素输入来加速训练，稍微改变解码器，用两层256单元的LSTM代替GRU单元; 这些正规化使用区域输出（克鲁格等人，2017），概率为0.1。 解码器每次输出80个通道的log-mel频谱图能量，两个帧，通过输出线性频谱图的扩张卷积网络运行。 我们通过Griffin-Lim运行这些以快速重建波形。 用WaveNet声码器代替Griffin-Lim可以直接提高音频保真度（沉等人。，2017).

基准模型达到4.0平均意见分数

（MOS），跑赢3.82 MOS报道（王 等人。，2017a）在同一套评估中。 因此它是一个非常强大的基准。

3.2. 样式令牌体系结构

3.2.1. 参考编码器

参考编码器由一个卷积堆栈和一个RNN组成。 它需要输入一个log-mel谱图，该谱图首先传递给一个由3个3×3内核，2×2步幅，批量归一化组成的6个2-D卷积层的堆栈，ReLU激活功能。 我们分别为6个卷积层使用32,32,64,64,128和128个输出通道。 然后将得到的输出张量形成3维（保留输出时间分辨率）并馈送到单层128单元单向GRU。 最后一个GRU状态用作参考嵌入，然后将其作为输入提供给样式标记层。

3.2.2. 样式代币层

样式标记层由一组样式标记嵌入和注意模块组成。 除非另有说明，否则我们的实验使用10个令牌，我们发现这些令牌足以在训练数据中表示一小部分但丰富的韵律维度。 为了匹配文本编码器状态的维度，每个令牌嵌入是256-D。 同样，文本编码器状态使用tanh激活；我们发现，在应用注意力之前对GST应用tanh激活会导致更大的令牌差异。 基于内容的tanh关注使用softmax激活来输出权标上的一组组合权重；然后将所得GST的加权组合用于调节。 我们尝试了不同的调节位置组合，发现复制样式嵌入并简单地将其添加到每个文本编码器状态表现最佳。

虽然我们在这项工作中使用基于内容的注意力作为相似性度量，但用替代方法替代它是微不足道的。 点积关注，基于位置的注意力，甚至注意机制的组合都可以学习不同类型的风格标记。 在我们的实验中，我们发现使用多头注意力（瓦斯瓦尼等人。，2017）显着提高样式转换性能，而且，比简单地增加令牌数量更有效。 当使用h注意力头时，我们将令牌嵌入尺寸设置为256 / h并且连接注意力输出，使得最终样式嵌入尺寸保持相同。

4. 模型解释

4.1. 端到端的聚类/量化

直观地说，GST模型可以被认为是将参考嵌入分解为端到端的方法一组基向量或软集群 - 即样式标记。 如上所述，每个样式标记的贡献由关注分数表示，但可以用任何期望的相似性度量来替换。 GST层在概念上与VQ-VAE编码器有些相似（范登 奥德等人。，2017），因为它学习了其输入的量化表示。 我们还尝试用离散的类似VQ的查找表层来替换GST层，但还没有看到类似的结果。

这种分解概念也可以推广到其他模型中，例如，分解变分潜在模型许 等人。，2017），它通过在分解层次图形模型中明确地表述它来利用语音信号的多尺度特性。 其依赖序列的先验是由嵌入表格制定的，该表格与GST类似，但没有基于注意力的聚类。 GST可能用于减少所需的样本以学习每个先前的嵌入。

4.2. 内存增强神经网络

GST嵌入也可以被看作是存储从训练数据中提取的样式信息的外部存储器。 参考信号在训练时指导内存写入，内存在推理时读取。 我们可能会利用内存扩展网络的最新进展（格雷夫斯 等人。，2014）进一步改善GST学习。

5. 相关工作

韵律和说话风格模型在TTS社区已经研究了数十年。 但是，大多数现有模型需要显式标签，例如情感或说话者代码（Luong等人，2017）。 尽管少量研究探索了自动标记，但仍然监督学习，需要昂贵的注释进行模型培训。 AuToBI，例如，（罗森伯格，2010）旨在生产ToBI（Silverman等人，1992）可以被其他TTS模型使用的标签。 然而，AuToBI仍需要培训注释，ToBI作为手工设计的标签系统，已知其性能有限（怀特曼，2002）.

基于群集的建模（Eyben等人。，2012；拖延，2017）与我们的工作有关。 Jauk（2017），例如，使用i向量（Dehak等人，2011）和其他声学特征来将训练组和训练模型分组在不同分区中。 然而，这些方法依赖于一套复杂的手工设计功能，并且需要在单独的步骤中训练一个中性的语音模型。

如前所述，（Skerry-Ryan等人。，2018）介绍了这项工作中使用的参考嵌入，并说明它可用于从参考信号传输韵律。 但是，这种嵌入不能启用可解释的样式控制，我们将在Section中显示6 它

在一些风格转移任务上推广不力。

我们的工作大大扩展了（Wang等人, 2017b），但有几个根本的区别。 首先，（Wang等人, 2017b）使用来自Tacotron解码器的单个帧作为查询来学习令牌。 因此它只模拟主要对应于F0的“局部”变化。 的GST

而是使用整个参考信号的摘要作为输入，并且因此能够揭示表现性合成所必需的局部和全局属性。 其次，与解码器端调节相比（Wang等人, 2017b），GST的设计允许文本输入以解开式样嵌入为条件。 我们在部分中展示了对于风格控制和转移的重要影响

6.2。 最后，GST可以用于清洁录制和嘈杂的数据。 我们在第一节详细讨论这个问题及其意义7.

6. 实验：风格控制和转移

在本节中，我们使用Section中的推理方法来衡量GST控制和转换口语风格的能力2.2.

我们使用147小时的美国英语有声读物数据来训练模型。 这些由2013年暴雪挑战赛发言人Catherine Byers以动画和情绪化的讲故事风格阅读。 一些书籍包含非常富有表现力的动态范围的人物声音，这对模型来说很具挑战性。

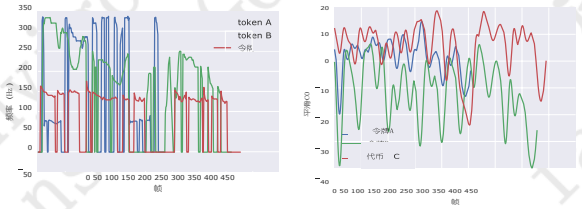
正如生成模型常见的那样，客观度量标准通常与感知关系不好（Theis等人, 2015）。 虽然我们在下面的一些实验中使用可视化，但我们强烈建议读者倾听我们提供的样本演示页面.

6.1. 样式控制

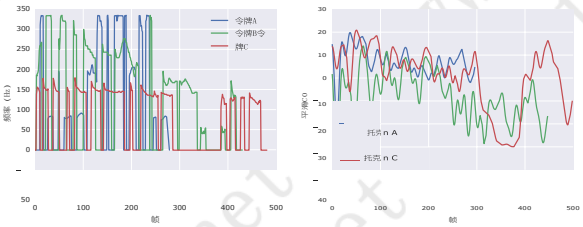
6.1.1. 风格选择

最简单的控制方法是在个人令牌上调节模型。 在推断时，我们只需用特定的，可选的缩放标记替换样式嵌入。

以这种方式调节有几个好处。 首先，它允许我们检查每个令牌编码的样式属性。 经验地说，我们发现每个标记不仅可以表示音调和强度，还可以表示其他各种属性，例如说话率和情绪。 这可以在图中看到2，它显示了从10-令牌GST模型中用三种不同风格标记（标度= 0.3）合成的两个句子。 这些图显示F0和C0（能量）曲线在不同风格的令牌中有很大不同。 然而，尽管事实上输入语句A和B是完全不同的，但由每个令牌产生的F0和C0轮廓遵循明显的相对趋势。



(a) 句子A



(b) B句

图2. 两个不同句子的F0和C0（对数刻度），用三个标记合成。与文本内容无关，相同标记相对于其他标记具有相同的F0 / C0趋势。

事实上，感性上，红色标记对应于较低音调的语音，绿色标记对应于渐减的音高，蓝色标记对应较快的语速（注意两个图中的总音频持续时间）。

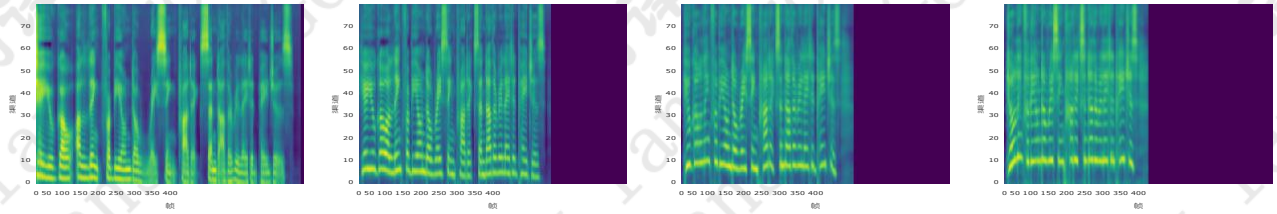
单令牌调节还揭示并非所有的令牌都能捕获单个属性：虽然一个令牌可以学习代表口语速率，但其他的可以学习反映训练数据中风格共现的属性混合（例如一个低调的令牌，例如，也可以编码较慢的讲话速率）。鼓励更独立的风格属性学习是正在进行的工作的重要焦点。

除了提供可解释性外，样式表征条件处理还可以提高合成质量。 考虑具有大量韵律变化的训练数据的长形合成问题。 许多TTS模型学习生成“平均”韵律风格，这对于表达性数据集可能是有问题的，因为表征它们的特征变化被破坏了。 这也可能导致不良的副作用，例如每个句子结束时音高不断下降。 我们发现对“活泼”的音符进行调节可以解决这两个问题，显着改善韵律变化。

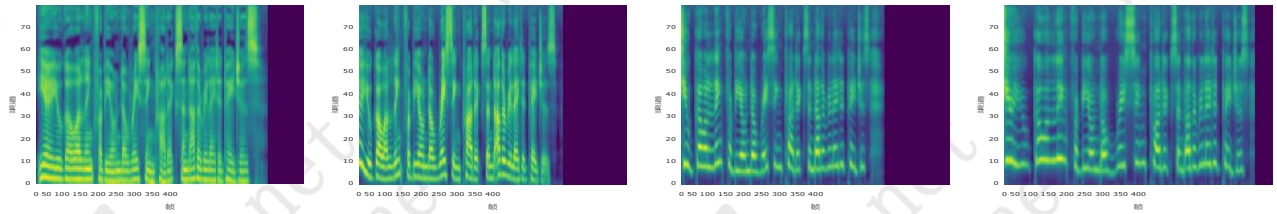
可以找到音乐选择的音频示例这里.

6.1.2. 样式缩放

另一种控制样式令牌输出的方法是通过缩放。 我们发现乘以令牌嵌入的标量值会加强其风格效果。 （请注意，较大的缩放比例值可能导致无法理解的语音，这表明未来改善稳定性的工作。）如图所示3，它显示了由两个不同的标记合成的话语的频谱图。 感知上，这些令牌编码

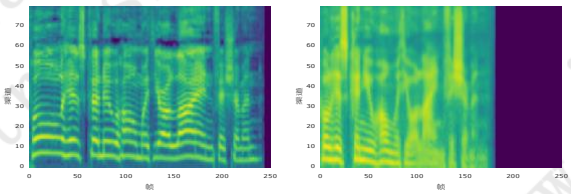


(a) 令牌A（速度）

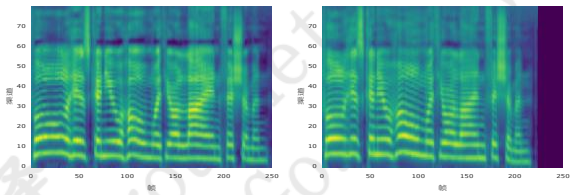


(b) 令牌B（动画）

图3. 标记缩放的影响。 从左到右，我们将这两个令牌分别按-0.3, 0.1, 0.3, 0.5进行缩放。 请注意，该模型似乎表现出反向效应（例如快速减慢或动态平稳）和负面缩放比例，这在训练过程中从未见过。



（a）原始音频　　（b）基线Tacotron



(c) 直接调节　　（d）GST调理

图4. 平行样式传输的Log-Mel谱图。

两种不同的口语风格：更快的说话率（3(a)）和更多的动画语音（3(b)）。 数字3(a) 表明增加快速语速标记的缩放因子会导致时域中的频谱图逐渐压缩。 同样，图3(b) 表明增加动画语音标记的缩放因子会导致音高变化的相应增加。 即使对于负值（说话速度变慢，言语变得平静），这些样式缩放效果也能保持，尽管模型在训练期间只看到正值（softmax）值。

音频缩放的音频示例可以找到这里.

6.1.3. 样式采样

我们还可以通过修改样式标记层内的注意模块权重来控制推理过程中的合成。 由于GST注意力产生一组组合权重，所以可以手动优化这些组合权重以产生期望的插值。我们也可以使用随机生成的softmax权重对样式空间进行采样。 采样分集可以通过调整softmax温度来控制。

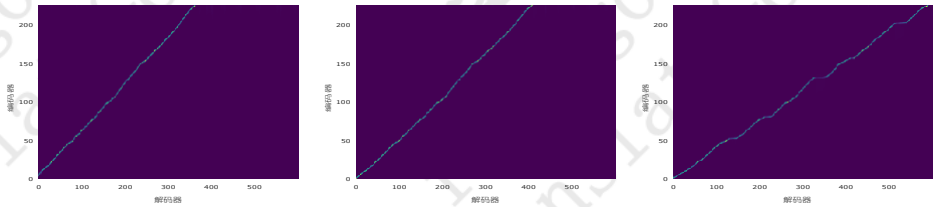
6.1.4. 文本方式控制/变形

尽管在训练期间将相同样式的嵌入添加到所有文本编码器状态，但推理模式中并不需要这种情况。 正如我们的音频样本所演示的，这允许我们通过针对输入文本的不同片段对一个或多个令牌进行调节来进行分段样式控制或变形。

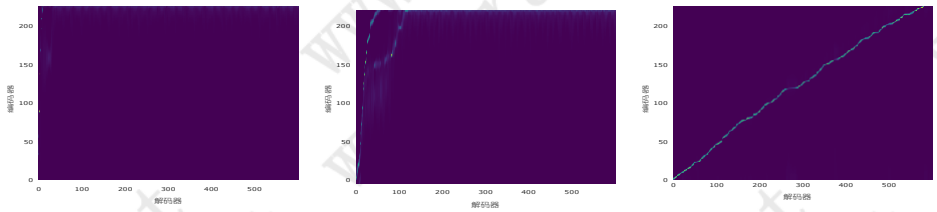
可以找到音乐变体的音频示例这里.

6.2. 风格转移

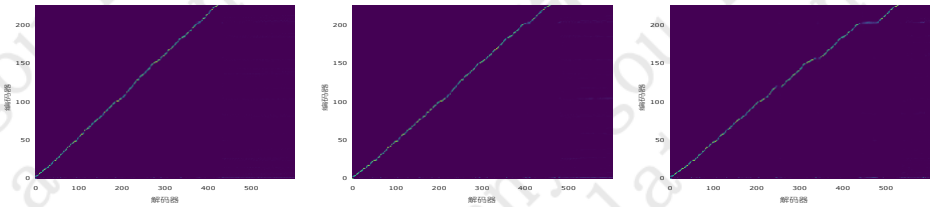
风格转移是一个活跃的研究领域，旨在综合参考信号的韵律风格中的短语（Wu等人, 2013; Nakashika等人, 2016; Kinnunen等人, 2017）。 GST模型可以以任何风格令牌的凸组合为条件的属性非常适合这一任务； 在推断时间内（方法2）2.2），我们可以简单地提供一个参考信号来指导令牌组合权重的选择。 以下实验使用4头GST注意力。



(a) 10令牌GST



(b) 直接调节（128-D）



(c) 256令牌GST

图5. 非平行样式传输的鲁棒性 从左到右：通过馈送三个文献长度分别为10, 96, 321字符的参考文献获得的注意对齐。 目标文本长度为258个字符。

6.2.1. 并行式传输

数字4 显示并行传输任务的频谱图，其中要合成的文本与参考信号的文本相匹配。 GST模型频谱图位于右下角，与其他三个基线相比：（a）地面实况输入信号（即参考）； （b）由基线Tacotron模型（仅从文本推断声学）进行的推断； 和（c）由（斯凯利瑞安 等人。，2018），Tacotron系统直接在128-D参考嵌入中调整文本编码器。

我们看到，只有文本输入，基准Tacotron模型并不与参考信号的韵律风格紧密匹配。 相反，直接调节方法（Skerry-Ryan等人。，2018）导致近乎时间对准的良好韵律转移。 GST模型介于两者之间：虽然其输出持续时间和共振峰跃迁并不精确地与参考时间相匹配，但总体光谱包络确实如此。 感知上，GST类似于参考的韵律风格。

可以找到并行式传输的音频示例这里.

6.2.2. 非平行式传输

我们接下来展示一个非并行传输任务的结果，其中TTS系统必须在其中合成任意文本参考信号的韵律风格。 我们为这项任务选择了三种不同的参考信号，并测试了在综合相同目标短语时GST模型复制每种风格的效果。 由于长格式合成可以从适当的文体渲染中获益很多，因此我们使用了长（258个字符）的目标短语。 我们选择了不同长度的源短语（分别为10, 96和321个字符）。 数字5 显示用于在每个源信号上调节的合成的对齐矩阵。

第一行显示了一个10-token GST模型。 这个模型强有力地概括了所有三个调节输入，如良好的比对图所证明的。 最下面的一行显示了一个具有相同行为的256-token GST模型；我们包括这个模型来表明即使当令牌（256）的数量大于参考嵌入维度（128）时，GST仍然保持稳健。

中间一行显示了一个带有直接参考嵌入条件的模型。注意矩阵表明，这个模型在短语源短语的条件下失败了，因为它试图将其合成压缩到与参考文献相同的时间间隔。 虽然模型在以最长输入为条件的情况下成功调整，但对于某些单词来说，可理解性较差：每个话语嵌入从源头捕获太多信息（例如时间和语音），从而伤害泛化。

表1. Gx有声书综合对Tacotron基线的SxS主观偏好（%）和p值。 每一行显示GST推理调节不同的参考信号（A和B）。 p值是针对3点和7点评级系统给出的。

	偏好（%）			P-VALUE	
	基础	中性	GST	3点	7-POINT
信号A	32.9	26.5	40.6	P = 0.0552	P = 0.0131
信号B	33.1	21.9	45.0	P=0.0038	P=0.0003

表2. 稳健MOS作为训练集中干扰百分比的函数。 总的训练集大小是相同的。

NOISE%	基准TACOTRON	消费税
50%	2.819 ± 0.269	4.080 ± 0.075
75%	1.819 ± 0.227	3.993 ± 0.074
90%	1.609 ± 0.131	4.031 ± 0.082
95%	1.353 ± 0.090	3.997 ± 0.066

为了按比例评估这种方法的质量，我们对Tacotron基线进行了非平行GST样式转移的并行主观测试。 我们使用了60个有声书句子的评估集，其中包括许多长短语。 我们通过在两种不同的叙述式参考信号上调节模型产生两组GST输出，这些信号在训练期间看不见。 并行主观测试表明，评估者优选两组GST合成对Tacotron基线的偏好，如表中所示1.

GST在非平行样式传输上的性能非常重要，因为它允许使用源信号来引导任意文本的强健文体综合。

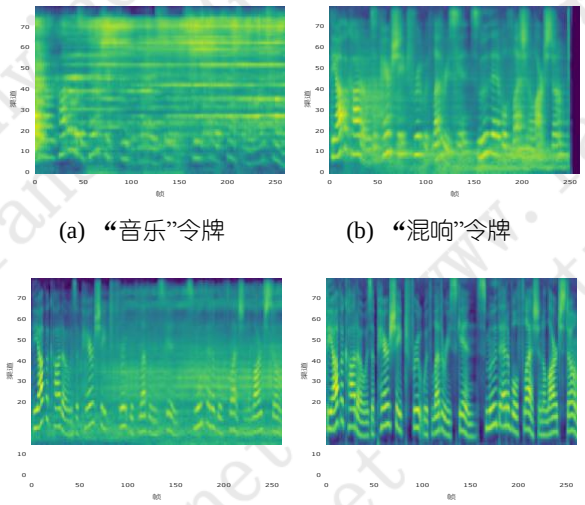
可以找到非平行样式传输的音频示例这里.

7. 实验：未标记的噪声发现数据

录音室质量的数据既经济又耗时。 尽管互联网上有大量丰富的现实生活表达言论，但它往往很嘈杂，难以标注。 在本节中，我们将演示如何使用GST直接从嘈杂的数据中直接训练健壮模型，而无需进行修改。

7.1. 人为噪声数据

作为第一个实验，我们通过向干净的语音添加噪声来人工生成训练集。 这里的动机是在执行受控实验时模拟真实的噪音数据。 为了达到这个目的，我们把单个说话者美国英语专有数据集从（Wang等人，2017a）变成房间模拟器（Kim等人，2017），增加了不同类型的背景噪音和房间混响。 该



(a) “音乐”令牌 (b) “混响”令牌

(c) “噪音”令牌 (d) 清洁令牌

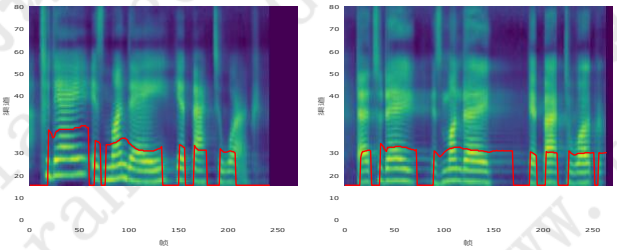
图6. 发现嘈杂和干净的令牌。

信噪比（SNR）范围为5-25 dB，房间混响的T60范围为100-900 ms。 我们分别创建了四个不同的训练集，其中50%，75%，90%和95%的输入分别被黑化。

在这些数据集上训练GST增强Tacotron之后，我们在第一节中描述的第一种模式下进行推理2.2。 我们不是提供参考信号，而是调整每个独立样式标记的模型，从而为我们提供每个标记已学习的可解释，可听的感知。有趣的是，我们发现不同的噪声被视为样式并“吸收”成不同的令牌。 我们用图中的几个标记来说明频谱图6。 我们可以看到（和听）这些令牌显然对应于不同的干扰类型，如音乐，混响和一般背景噪音。 重要的是，这种方法揭示了学习令牌的一个子集也对应于完全干净的讲话。 这意味着我们可以通过在单个干净的样式标记上调节模型来合成任意文本输入的干净语音。

为了证明这一点，我们使用手动识别的干净样式标记（缩放到0.3）运行推理，然后使用MOS自然度测试评估输出。 我们使用相同的100个词组评估集合（Wang等人，2017a），收集来自众包的母语人士的8个评分。 表2显示基线Tacotron和“干净标记”GST模型的MOS结果。虽然基线Tacotron在数据集100%清洁时达到4.0 MOS，但随着干扰增加MOS降低，降至1.353的低分。 由于该模型没有语音或噪声的先验知识，因此它会盲目地模拟训练集中的所有统计数据，从而在合成过程中产生大量噪音。

相比之下，GST模型达到约4.0 MOS



(a) 令牌A (b) 令牌B

图7. 从TED数据训练的GST模型中随机选择的两个令牌的Log-Mel谱图（覆盖有F0轨迹）。 这两个令牌揭示了两个不同的发言者。

所有的噪音条件。 请注意，令牌的数量需要随噪声的百分比一起增加才能达到此结果。 例如，一个10-令牌GST模型在50％的噪音数据集上训练时会产生干净的令牌，但噪音较大的数据集需要20令牌模型。 未来的工作可能会探索如何将令牌的数量自动适应给定的数据分布。

可以找到这些模型的音频示例这里.

7.2. 真正的多扬声器发现数据

我们的第二个实验使用真实数据 该数据集由439个官方TED YouTube频道视频挖掘的音轨组成。 轨道包含显着的声学变化，包括声道变化（近场和远场语音），噪声（如笑声）和混响。 我们使用一个endpointer将音轨分割成短片段，然后使用ASR模型创建<文本，音频>训练对。 尽管ASR模型会产生大量的转录和错位错误，但我们不会执行其他预处理。 最终的训练集约68小时，包含约439位演讲者。

在不使用任何元数据作为标签的情况下，我们为了比较而训练基线Tacotron和1024-token GST模型。 正如预期的那样，基线未能学习，因为多说话者数据太多。 GST模型结果如图所示7。 这显示了通过在两个随机选择的标记上调节模型生成的与F0轨道重叠的相同短语的光谱图。 检查训练的GST，我们发现不同的令牌对应于不同的说话者。 这意味着，为了与特定讲话者的声音合成，我们可以简单地将来自该讲话者的音频作为参考信号馈送。 参见章节7.3 进行更多的定量评估。

最后，我们利用了这样一个事实，即大部分会谈都是英文的，但一小部分是用西班牙文。 在这个实验中，我们比较了基线和GST使能噪声数据模型在交叉语言风格的传输任务。 对于基线，

表3. 西班牙语到英语无监督语言转换实验的WER。 请注意，WER低估了真正的可理解性分数；我们只关心相对差异。

模型	WER (INS / DEL / SUB)
GST	18.68 (6.13/2.37/10.18)
多扬声器	56.18 (3.75/20.27/32.14)

我们训练一个多扬声器Tacotron类似于（Ping等人，2017），使用视频ID作为扬声器标签的代理。 在西班牙语音标签上标注条件后，我们将合成100个英语短语。 对于GST系统，我们提供来自同一西班牙语发言者的参考信号并合成相同的100个英语短语。 虽然说话者的西班牙口音不被保留，但我们发现GST模型产生完全可理解的具有与说话者相似的音高范围的英语语音。 相比之下，多扬声器Tacotron输出则不太清晰。

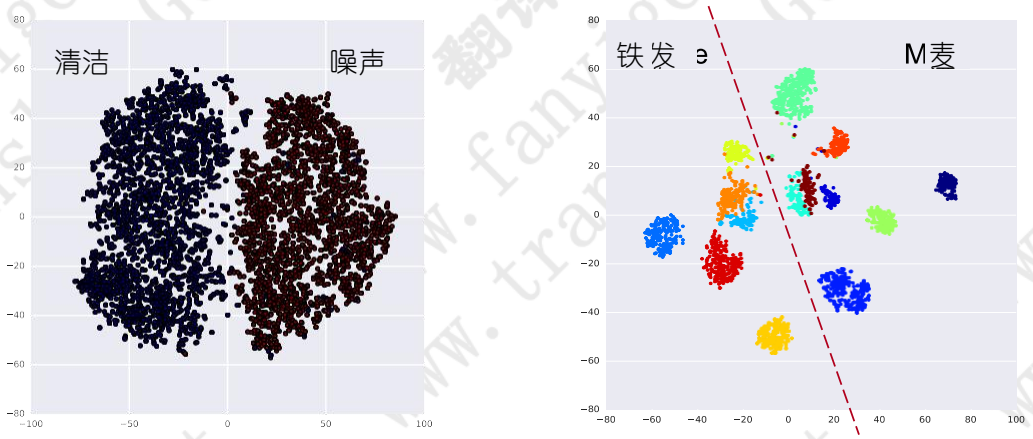
为了客观地评估这个结果，我们计算了合成语音的英语ASR模型的误码率（WER）。 如表所示3，GST话语的WER远低于多说话者模型的WER。

结果强烈地证实GST学习了从文本内容中解开的嵌入。 虽然这是一个令人激动的早期成果，但是使用GST进行韵律保存语言转换的深入研究仍在进行中。

7.3. 定量评估

我们使用t-SNE（Maaten和Hinton，2008）可视化从人造噪声和TED数据集学习的样式嵌入。 数字8(a) 显示从人造噪声数据集（50％清洁）中学习的嵌入清楚地分为两类。 数字8(b) 显示包含14个TED通话数据扬声器的2,000个随机抽取样本的样式嵌入。 我们看到样本被很好地分为14个集群，每个集群对应一个单独的说话人。 女性和男性讲话者线性可分离。

我们还使用样式嵌入作为特征来利用线性判别分析来执行噪声和说话者分类。 结果显示在表中4。 对于噪声分类，GST以99.2％的准确度揭示真实标签。 对于说话人分类，我们使用TED视频ID作为真实标签，并与i矢量方法进行比较（Dehak等人，2011），这是现代说话人验证系统中使用的标准表示。 对于这项任务，测试集包含431个扬声器。 虽然都是用短语（平均持续时间为3.75秒）进行训练和测试，但我们可以看到GST与i向量相当。 这是一个令人鼓舞的结果，因为i矢量是专门为说话人分类而设计的。 我们推测商品及服务税有可能应用于说话者diarization。



(a) 50%的噪音数据　　(b) 多扬声器TED

图8. 使用t-SNE的样式嵌入可视化。

表4. 使用GST和i向量的分类精度（噪声与干净和TED说话者ID）。 尽管在生成模型中进行了培训，但GST编码了丰富的区分信息。

嵌入算法	人工数据	TED（431演讲人）
GST	99.2%	75.0%
I-VECTOR	/	73.4%

7.4. 启示

以上结果对未来TTS对发现数据的研究具有重要意义。 首先，由于GST对声学和文本噪声的稳健性，自动数据挖掘管线的设计可能会大大简化。 例如，精确的分割和ASR模型不再是构建高质量TTS模型所必需的。 其次，风格属性（例如情感）通常很难标注为大规模噪音数据。 使用GST或权重自动生成样式注释可能会大大减少人员在回路中的工作量。

8. 结论和讨论

这项工作引入了Global Style Tokens，这是一种在端到端TTS系统中建模风格的强大方法。 消费税是直观的，易于实施，并没有明确的标签学习。 我们已经表明，在表达语音数据的训练中，GST模型产生可用于控制和传输样式的可解释的嵌入。 我们还证明，虽然GST最初被设想为模仿说话风格，但它是揭示数据中潜在变化的一种通用技术。 这通过对未标记的嘈杂发现数据的实验证实，该数据显示GST模型学习将各种噪声和扬声器因素分解成不同的样式标记。

还有很多需要调查的内容，包括改进GST的学习，以及使用GST权重作为目标来预测文本。 最后，虽然我们在本文中仅将GST应用于Tacotron，但我们相信它可以很容易地被其他类型的端到端TTS模型使用。 更一般地说，我们设想GST可以应用于受益于可解释性，可控性和健壮性的其他问题领域。 例如，GST可以类似地用于文本到图像和神经机器翻译模型。

致谢

作者感谢Aren Jansen，Rob Clark，陈志峰，Ron J. Weiss，Mike Schuster，吴永辉，Patrick Nguyen和机器听证会，Google Brain以及Google TTS团队的有益讨论和反馈。

参考

Arik，Sercan O，Chrzanowski，Mike，Coates，Adam，Di-amos，Gregory，Gibiansky，Andrew，Kang，Yongguo，Li，Xian，Miller，John，Raiman，Jonathan，Sengupta，Shubho等人。 深沉的声音：实时神经文本到语音。 ICML，2017。

Dehak，Najim，Kenny，Patrick J，Dehak，Re'da，Dumouchel，Pierre和Ouellet，Pierre。 说话人验证的前端因素分析。 IEEE Transactions on Audio，Speech，and Language Processing，19（4）：788-798，2011。

Eyben，Florian，Buchholz，Sabine和Braunschweiler，Norbert。 情感和声音风格的无监督聚类为表现型tts。 在ICASSP，第4009-4012页。 IEEE，2012。

格雷夫斯，亚历克斯，韦恩，格雷格和丹尼尔卡，伊沃。 神经图灵机。 arXiv预印本arXiv：1410.5401,2014。

格里芬，丹尼尔和林，宰。 来自改进的短时傅里叶变换的信号估计。 IEEE Transactions on Acoustics, Speech, and Signal Processing, 32 (2) : 236-243,
1984.

徐，卫宁，张宇，玻璃，詹姆斯。 无监督地学习来自顺序数据的解开和可解释的表示。 2017年神经信息处理系统进展。

Jauk, Igor。 用于表达性语音合成的无监督学习。博士论文，Universitat Polite`cnica de Catalunya, 2017。

Kim, Chanwoo, Misra, Ananya, Chin, Kean, Hughes, Thad， Narayanan， Arun， Sainath， Tara和Bacchiani, Michiel。 在虚拟房间中生成大规模模拟话语，以训练谷歌家庭中用于远场语音识别的深度神经网络。 PROC。 INTERSPEECH。 ISCA, 2017年。

Kinnunen, Tomi, Juvela, Lauri, Alku, Paavo和Yamagishi, Junichi。 使用i向量的非平行语音转换plda：致力于统一说话者验证和转换。 在ICASSP, 2017年。

Krueger, David, Maharaj, Tegan, Krama'r, Ja'nos, Pezeshki, Mohammad, Ballas, Nicolas, Ke, Nan迷迭香， Goyal， Anirudh， Bengio， Yoshua，Larochelle, Hugo, Courville, Aaron等。 防区：通过随机保存隐藏的激活来规范RNN。 在Proc。ICLR, 2017。

Luong, Hieu-Thi, Takaki, Shinji, Henter, Gustav Eje和Yamagishi, Junichi。 使用输入代码调整和控制基于dnn的语音合成。 在ICASSP中，第4905-4909页。IEEE, 2017。

Maaten, Laurens van der和Hinton, 杰弗里。 使用t-sne可视化数据。 Journal of machine learning research, 9 (11月) : 2579-2605,2008。

Nakashika, Toru, Takiguchi, Tetsuya, Minami, Yasuhiro, Nakashika, Toru, Takiguchi, Tetsuya和Minami, Yasuhiro。 使用自适应限制玻尔兹曼机器进行语音转换的非平行训练。 IEEE / ACM Trans。音频，语音和朗。 Proc。，24 (11) : 2032-2045, 2016年11月。

Ping, Wei, Peng, Kainan, Gibiansky, Andrew, Arik, Ser-can O, Kannan, Ajay, Narang, Sharan, Raiman, Jonathan和Miller, John。 Deep voice 3：2000-扬声器神经文本到语音。 arXiv预印本arXiv：1710.07654，2017。

罗森伯格，安德鲁。 AuToBI - 自动ToBI注释工具。 在Interspeech，第146-149页，2010年。URLHTTP://eniac.cs.qc.cuny.edu/andrew/autobi/.

Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerry-Ryan, RJ等人。 通过对梅尔谱图预测进行调节波网络自然合成。 arXiv预印本arXiv：1712.05884，2017。

Silverman, Kim, Beckman, Mary, Pitrelli, John, Ostendorf, Mori, Wightman, Colin, Price, Patti, Pierrehumbert, Janet和Hirschberg, Julia。 ToBI：标注英语韵律的标准。 第二届国际口语处理会议，1992年。

Skerry-Ryan, RJ, Battenberg, Eric, Xiao, Ying, Wang, Yuxuan, Stanton, Daisy, Shor, Joel, Weiss, Ron J., Clark, Rob和Saurous, Rif A.向端到端韵律转移用Tacotron表达语音合成。 arXiv预印本，2018年。

Taigman, Yaniv, Wolf, Lior, Polyak, Adam和Nachmani, Eliya。 通过语音循环对野外演讲者进行语音合成。 arXiv预印本arXiv：1707.06588，2017。

泰勒，保罗。 文本到语音合成。 剑桥大学出版社，2009年。

Theis, Lucas, Oord, Aa¨ronvan den和Bethge, Matthias。 关于生成模型评估的说明。 arXiv预印本arXiv：1511.01844,2015。

van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalch-brenner, Nal, Senior, Andrew和Kavukcuoglu, Koray。 Wavenet：原始音频的生成模型。 CoRR abs / 1609.03499,2016。

van den Oord, Aaron, Vinyals, Oriol等人。 神经离散表示学习。 In Advances in Neural Information Processing Systems，pp.6309-6318，2017。

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz和Polosukhin, Illia。 注意是你所需要的。 In Advances in Neural Information Processing Systems，pp.6000-6010,2017。

王玉轩Skerry-Ryan RJ斯坦顿Daisy Wu Yonghui Weiss Ron J. Jaitly Navdeep Yang Zongheng Xiao Ying Ying Chen Zhifeng Bengio Samy Le Le Quoc Agiomyrgiannakis, Yannis, Clark, Rob和Saurous, Rif A. Tacotron：迈向端到端的语音合成。 在Proc。Interspeech，pp.4006-4010，August 2017a。 网址 https://arxiv.org/abs/1703.10135.

王玉轩Skerry Ryan RJ肖恩Ying Stanton Daisy Shor Joel Battenberg Eric Eric Clark Rob和

Saurous, Rif A.发现表达式语音合成的潜在风格因素。 ML4Audio研讨会，NIPS，2017b。

Wightman, Colin W. Tobi还是不tobi？ 在2002年的Speech Prosody，国际会议上，2002年。

吴志正，郑，英，李，海州。 有条件的限制玻尔兹曼机器进行语音转换。 在ChinaSIP，2013年。

Zen, Heiga, Agiomyrgiannakis, Yannis, Egberts, Niels， Henderson， Fergus 和 Szczepaniak，Przemysław。 基于LSTM-RNN的快速，紧凑和高质量的统计参数语音合成器适用于移动设备。 Proceedings Interspeech，2016。