****

**本科毕业设计（论文）外文翻译**

|  |  |
| --- | --- |
| **学 院：** | 计算机科学与工程学院 |
| **专 业：** | 计算机科学与技术 |
| **学生姓名：** | 彭郑威 |
| **学生学号：** | 201930343087 |
| **指导教师：** | 李粤 |

|  |  |  |
| --- | --- | --- |
| **中文译名** | **FASTPITCH：具有音高预测的并行文本到语音转换** | |
| **外文原文名** | **FASTPITCH: PARALLEL TEXT-TO-SPEECH WITH PITCH PREDICTION** | |
| **外文原文版出处** | | **Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6588-6592.**  **链接：<https://arxiv.org/pdf/2006.06873.pdf>** |
| **译 文：**  **摘要**  我们提出了FastPitch，一个基于FastSpeech的完全并行的文本到语音模型，此模型以基频轮廓为条件。该模型在推理期间预测音高轮廓。通过改变这些预测，生成的语音可以更加具有表现力，更好地匹配语音的语义，并且最终更加地吸引听众。使用FastPitch均匀地增加或减少音调产生类似于语音的自主控制的语音。对频，率轮廓的调节提高了合成语音的整体质量，使其与最先进的技术相媲美。它不会引入额外的开销，并且FastPitch保留了有利的，全并行的Transformer架构，具有超过900倍的时间因子用于典型话语的梅尔频谱图合成。  **关键词**：文本转语音，语音合成，基频   1. **介绍**   神经文本到语音（TTS）的最新进展使得能够实时合成听起来自然的，类似人类的语音。并行模型能够数量级地快于自回归模型来合成梅尔频谱图，要么是依赖于外部的对齐，要么是自身对齐。TTS模型可以以语音的质量为条件，例如语言特征和基频。后者已经多次被证明可以提高神经模型的质量，但也可以提高级联模型的质量。对F0的调节是增加歌唱能力或与其他扬声器相适应的常用方法。  在本文中，我们提出了FastPitch，一个提高了合成语音质量的基于FastSpeech的前馈模型。对于每个输入符号的估计的基频，我们简单地将其称为音高轮廓，通过对它们进行调节，FastPitch可以和最先进的自回归TTS模型相匹配。我们表明，这种音高轮廓的确定性建模解决了普通前馈Transformer架构的质量缺陷。这些最有可能是由于在文本输入中缺乏足够的语言信息的情况下，相同语音单位的不同发音被折叠。对极品的调节也提高了收敛性，并且消除了在FastSpeech中使用的梅尔频谱图目标的知识提取的需要。我们需要注意的是，同时开发的FastSpeech 2 描述了类似的方法。  结合WaveGlow，FastPitch能够以比实时快60倍的速度合成梅尔频谱图，而无需求助于内核级优化。因为该模型学习来预测，以及使用每个输入符号的一个值的低分辨率的音高，这使得它可以很容易地交互式地调整音高，从而能够在音高编辑中实现实际应用。FastPitch中F0的恒定偏移产生自然发生的低声调和高声调的音高变化，保持对说话人身份的感知。我们的结论是，该模型学习模仿声带的运动，它发生在声音的自主控制。   1. **模型描述**   FastPitch的架构如图1所示。它基于FastSpeech，主要由两个前馈Transformer（Feed-Forward Transformer, FFTr）堆栈组成。第一个Transformer在输入词元（tokens）的分辨率下工作，第二个Transformer在输出帧的分辨率下工作。    图1 **FastPitch的架构**遵循FastSpeech，针对每个时间位置预测单个音调值  令 为输入词汇单元的序列， 为目标的梅尔刻度频谱图的帧。第一个FFTr堆栈产生隐式表示 ，隐式表示 用于通过 1-D CNN网络预测每个字符的持续时间和平均音高：  ，  这里，，。接下来，音高被投影以匹配隐式表示 的维度，并被添加到 。得到的和 被离散地上采样并被传输到输出FFTr上，其生成输出的梅尔频谱图序列。  上述的 序列中的 含有 个。  在训练过程中使用基本事实 和 ，在推理过程中使用预测的 和 。该模型优化了预测模态和真实模态之间的均方误差（MSE）：   * 1. **输入符号的持续时间**   输入信号的持续时间由一个基于LJSpeech-1.1训练的Tacotron 2 模型来估计。令 为最终的Tacotron 2 注意力矩阵，第 个输入符号的持续时间 。因为Tacotron 2 又一个单一的注意力矩阵，我们不需要在注意力头之间进行选择，因为这是多头Transformer模型所必需的。  FastPitch在对齐质量上来说是鲁棒的。我们观察到由不同的Tacotron 2 模型所提取的持续时间是不同的（见图2），持续时间最长的大约出现在相同的位置，但是可能分配给不同的字符。令人惊讶的是，这些不同的对齐模型产生的不同的FastPitch模型，其合成相似质量的语音。    图2 使用不同的Tacotron 2 模型提取的不同字符的持续时间，允许训练相似质量的FastPitch   * 1. **输入符号的音高**   我们使用精确的自相关方法通过声学周期性检测获得真实的音高值。加窗的信号由汉斯窗（Hann）计算得到。此算法找到归一化自相关函数的最大值的阵列，其成为候选频率。通过候选阵列的最低成本路径由维特比算法（Viterbi）计算得到。该路径使候选频率之间的转变最小化。我们设置窗口大小以匹配训练梅尔频谱图的分辨率，以获得每帧的F0值。  使用提取的持续时间 对每个输入符号的F0值进行平均（见图3），无声的值从计算中排除。对于训练，将值标准化为平均值为0以及标准偏差为1。如果对于特定的符号没有有声的F0估计，则将其音高设置为0。我们还没有看到在对数域对F0建模的任何改进。  我们尝试将每个符号平均三个音高值，希望捕获到每个符号的开始、中间和结束的音高，然而，该模型被判定为劣质的（见4.1章节）    图3 对话语*In being comparatively* 的基频估计，原始值以蓝色显示，输入字符的平均值以绿色显示   1. **相关工作**   FastSpeech模型与我们的模型同时开发，它有一种不同的方法来调节F0并使用音素（phoneme）输入。预测的轮廓对于每个梅尔频谱图帧具有一个值的分辨率，离散为256个频率值。此外，此模型以能量为条件。在FastPitch中，预测的轮廓对每个输入符号都有一个值，在实验中，这种较低的分辨率使得模型更容易预测轮廓，并且用户稍后可以交互式地修改音高。我们发现，这种分辨率使得模型在训练过程中区分不同的字形发音方式。此外，对较高共振峰的调节可能会略微提高质量，可以通过成对比较捕获（见4.1.1章节）。  文本到语音的主要范例是两阶段合成：首先从文本产生梅尔刻度频谱图，然后从声码器模型产生实际声波。在尝试加速生成中，并行模型已经得到探索。除了在这项工作中研究的Transformer模型之外，卷积GAN-TTS还能合成具有最先进质量的原始音频波形，它取决于语言和音高特征。  并行化现有模型的努力包括类似于FastSpeech的持续时间预测，应用于Tacotron，WaveRNN或者基于流的模型。持续时间的显式建模通常使用与HMM的推断和训练相关联的动态编程算法。Viterbi路径对齐的Glow-TTS，FastPitch已经通过前向-后向算法的变体进行了改进。  音高的显式神经建模与神经TTS语音转换模型一起引入，该模型与IBM Research的其他模型具有相似性。具有LSTM的变分自动编码器生成网络对韵律进行建模，并且在训练之前使用单独的工具计算音高。韵律信息被编码在四个值的向量中：对数持续时间（log-duration）、起始对数音高（start log-pitch）、结束对数音高（end log-pitch）和对数能量（log-energy）。   1. **实验**   带有[预训练检查点](https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/FastPitch)和[合成样本](https://fastpitch.github.io/)的源代码可在线获得，我们使用预训练的WaveGlow合成用于评估的波形。  该模型在公开可用的LJSpeech 1.1 数据集上进行训练，该数据集包含以22050Hz记录的大约24小时的单说话人语音。我们使用语音识别模型手动纠正数据检查期间发现的样本LJ034-0138和LJ031-0175的错误转录。我们谨慎地使用与WaveGlow模型相同的训练/测试/验证数据集分割。在使用预训练模型时，这个细节很容易被忽略，并且可能在评估过程中泄露训练数据并夸大结果。  模型的参数大多数遵循FastSpeech。每个FFTr层由具有核大小3和384/1536的输入/输出通道的1-D卷积、ReLU激活、具有核大小3和1536/384输入/输出滤波器的1-D卷积，其后的Dropout和LayerNorm组成。持续时长预测器和音高预测器有相同的架构：具有核大小3和384/256通道的1-D卷积、具有256/256通道的1-D卷积、每个通道之后是ReLU，LayerNorm和Dropout层，最后一层将每个256通道的向量投影到标量。Dropout率是0.1，在注意力头上的也是如此。  所有被描述的模型都在字素（graphemes）上进行训练。在字素上进行训练导致模型具有相似的质量，不论是Tacotron 2还是Montreal Forced Aligner的持续时间。不过，音素（phonemes）和字素（graphemes）混合训练的方法引入了令人不快的伪像。  FastPitch已经在8张NVIDIA V100 GPU上进行了训练，每张GPU有32个例子和自动混合精度。训练在两个小时后收敛，完全训练需要5.5个小时。我们使用LAMB优化器，其学习率为0.1，，，。学习率在1000个warmup步中被增加，然后根据Transformer的时间表进行衰减。我们使用的权重衰减。   * 1. **评估**   我们将FastPitch模型和Tacotron 2 模型进行了比较（见表1）。这些样本已经在Amazon Turk使用众包音频质量评估工具包（CrowdSourced Audio Quality Evaluation Toolkit）进行评分。我们已经从LJSpeech-1.1的开发子集的前30个样本中生成了语音。每个模型至少收集了250个分数，总共有60个独特的Turkers参与研究。为了获得资格，Turkers 被要求通过听力测试。    表1 95%置信区间的平均意见得分，这两个模型都是在字素输入上训练的   * + 1. **成对比较**   生成模型给超参数调整带来困难。生成的样品的质量是主观的，并且运行大规模研究耗时且成本高。为了有效地对多个模型进行排名，并避免开发人员在长时间内对样本进行评分时出现评分偏差，我们研究了盲比较样本对的方法。成对比较允许建立一个全球排名，假设技能评级是可传递的。  在内部研究随机选择50多个参与者得分对样本。Glicko-2评级系统，从人类玩家下棋，体育和在线游戏，以及生成模型自动评分的上下文中获取知识，被用来建立一个基于这些得分的排名（见图4）。FastPitch用1、2、4个注意力头、6和10个Transformer 层的变种，以及在每个输入符号的1个和3个预测的音高值的分辨率进行比较。此外，这种评价方法在跟踪多个超参数设置的开发过程中已经被证明是有用的，即使只有少数的评价者。   * + 1. **多说话者**   FastPitch很容易扩展到多说话者。我们在LJSpeech-1.1数据集上训练了一个模型，其中包含来自两位女性说话者的额外内部训练数据：说话者1（8330个样本，共计13.6小时），说话者2（18995个样本，共计17.3小时）。我们通过向输入词元（token） 添加全局说话人嵌入来对说话人模型进行调节。为了进行比较，我们选择了多说话人的Tacotron 2 和Flowtron，后者是一个自回归的基于流的模型。所有的模型都是在相同的数据上进行训练的，多说话人的Tacotron 2 已经被用于提取FastPitch的训练对齐。结果总结在表2中。    表2 在LJSpeech dev样本上评估的多说话者的95%置信区间的平均意见得分   * 1. **音高条件化和推理性能**   可以在推理期间修改预测的音高轮廓以控制所生成的语音的某些感知质量。它可以用来增加或减少F0，提高音高的表现力和多样性。本文所附的音频样本演示了增加、减少或反转单个语音的平均值周围的频率，以及在多说话人模型的说话人之间插值的效果。我们鼓励读者去听。  图5展示了将频率偏移50Hz的示例。与频域中的简单移位相比，FastPitch保留了说话者的感知身份，并对语音的自动调制期间发生的声带动作进行建模。    图5 使用FastPitch通过在推理期间将常数添加到预测的音高 上来移动F0。成对的显示为：移位频谱图以及移位和未移位频谱图之间的绝对差。  推理性能测量是在FP16精度和TorchScript序列化模型的NVIDIA A100 GPU上进行的。在批量大小（batch size）为1的情况下，来自LJSpeech-1.1训练集的前2048个话语的平均实时因子（Real Time Factor，RTF）为 （见表3）。使用WaveGlow，完整音频合成的RTF降低到 。在Intel Xeon Gold 6240 CPU上测得的RTF为 。FastPitch 适用于合成样本的实时编辑，每个输入符号的单个音调值很容易被人类解释，我们在上述网站的视频剪辑中演示了样本。    表3 来自LJSpeech训练子集的前2048个话语上的梅尔频谱图生成的平均延迟和实时因子   1. **结论**   我们已经提出了FastPitch，一个基于FastSpeech的并行的文本到语音模型，能够快速合成高保真的梅尔尺度的频谱图，在韵律上具有高度的控制度。该模型展示了如何调节韵律信息可以显著提高前馈模型中合成语音的收敛性和质量，使其独立输出的语音更加连贯，并产生最先进的结果。我们的音高调节方法比许多文献中已知的方法更简单。它不引入额外的开销，并在交互式地调整韵律方面开辟了实际应用的可能性，因为该模型是快速的，高度表达的，并提出了潜在的多说话者的情况。   1. **致谢**   作者要感谢Dabi Ahn，Alvaro Garcia和Grzegorz Karch在实验和模型评估方面的帮助，以及Jan Chorowski，Jo˜ao Felipe Santos，Przemek Strzelczyk和Rafael Valle在论文准备期间提供的有用的讨论和支持。 | | |

说明：

1.外文原文出处包括出版社、出版时间、期刊的刊名、刊号、刊期；如翻译内容为外文著作，需写明翻译的著作名称、章节及页码；如翻译多篇文献，需将各篇文献分别列出，翻译外文文献数量最多不超过3篇。

2.外文翻译内容字体为小四号，宋体（英文需用Times New Roman字体），行距为固定值20磅。

3.外文原文用A4纸打印，附在译文后面，与译文一起装订。