1. 百度的deep voice已经更新到第三代了， 去年十月的时候已经有了deep voice3
2. 我看了一些review，第三代和第一第二代差距特别大 几乎是一个全新的东西，DV3借鉴了17年的Facebook CNN-seq2seq的结构convolutional sequence-to-sequence model
3. 前端
4. G2P目标是得到音素。手写体转音素使用的是encoder-decoder结构
5. 音频切分目标是得到音素和音频的对齐信息，知道每句话中每个音素在对应音频中的起点和终点。使用的方法是语音识别论文[Deep Speech 2](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1512.02595)中的对齐方法
6. 时长预测和基频预测模型为同一个模型多任务预测方式。模型输入为带有重音标注的音素序列，输出为时长、是否发音概率、基频F0
7. 全部输入到声学模型中合成语音
8. 声学模型

百度对WaveNet改进方向主要是：改变网络层数、残差通道数、用矩阵乘代替上采样卷积、CPU优化、GPU优化等

1. Tacotron的出现
2. 减少特征工程，只需要输入文本即可，其他特征模型自己学习
3. 各种条件方便添加，例如语种、说话人、情感等
4. 避免了多个子模型的误差传递和积累

**评价**

1） 模型复杂。用一个模型端到端方式合成语音，虽然省去了中间步骤，但模型复杂，不好调试，不好训练

1. 模型除错难。训练再好的模型，也可能对某些文本发音错误，这时，想要纠正这些错误，很难；需要重新准备数据、再次训练，再次训练也不一定能克服那些问题，代价很大。
2. 人为干预能力差。参数合成往往可以人为指定语速、重音、断句、停顿、韵律等信息，进行个性化合成。然而，端到端合成这些信息全部由模型自己学习，很难加入人为控制。因此，很难产品化。
3. 端到端不彻底。严格意义上讲，Tacotron也并非端到端，模型输出的是梅尔频谱(mel-scale spectrogram)，再用CBHG结构转为频谱幅度(spectral magnitude)，最后使用了Griffin-Lim这样的vocoder转为最终的音频。对此，好处是Tacotron中的seq2seq结构部分更容易训练，后处理部分可以单独训练，缺点是，后处理网络和Griffin-Lim本身的局限会影响到音质。

我这段时间看了很多关于tts的博客和论文，发现deep voice的内核仍然是基于wavenet，我仔细看了前两代的deep voice，无非是在wavenet的基础上进行的改善，改进方向主要是：改变网络层数、残差通道数、用矩阵乘代替上采样卷积、CPU优化、GPU优化等

虽然如此 但是网上很少找到前两代的实现代码，目前只有看到