语音变调实验报告 胡晋侨 2000013141

一、语音变调的基本原理

人类语言的生成可以简化为源-滤波器模型：声带作为激励源产生声音信号，经过声道的滤波效应，产生若干能量较集中的共振峰，最后再经过嘴唇辐射生成语音。基频即频率最低的共振峰，即声带振动的频率。通常认为基频决定了人类对音调的感知，而共振峰的位置、形状等则包含了语义信息。故变调即是调整基频、共振峰的频率，在不改变语义信息的基础上改变人类感知的音调。

二、语音变调的方法

方法一：LPC分析

我们假定每一时刻输出的语音只和之前若干时刻输出的语音以及当前时刻的激励源线性相关。使用最小二乘法计算出线性相关的系数，便可以解出激励源的波形，可以从中获取基频的信息。而从频谱包络中能得到共振峰的信息。故改变基频的位置就可以改变音调。为了保证基频与共振峰的相对位置，也需要改变共振峰的位置。

方法二：基于STFT的方法

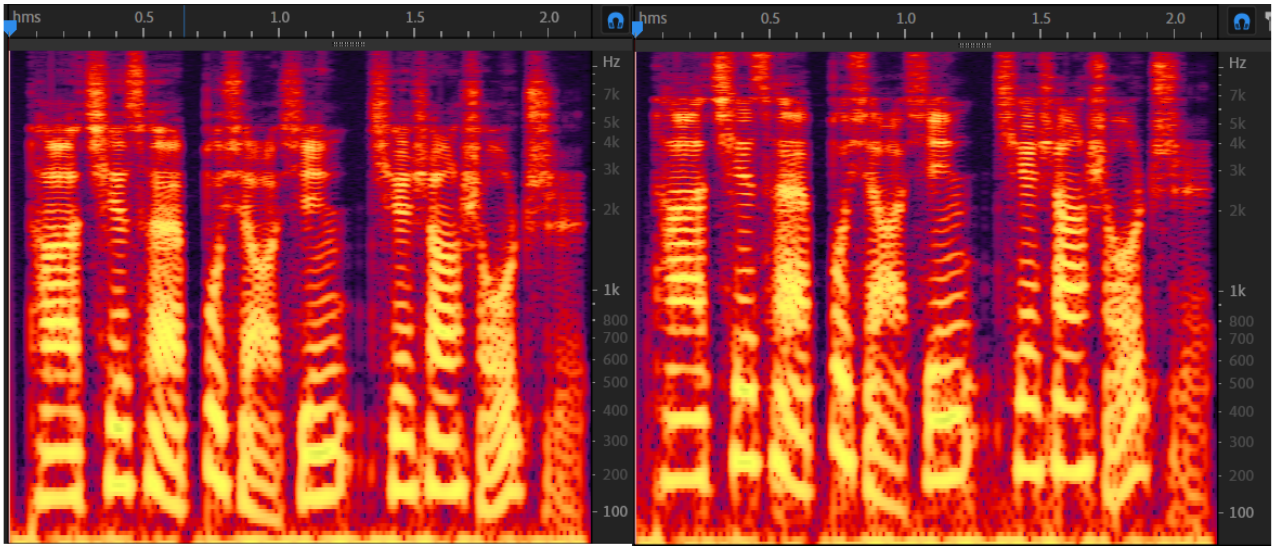
语音序列在较短时间(10ms-30ms)内可以看作平稳序列，故可以对其分帧加窗处理，对每一帧进行傅里叶变换之后在频域中提高其频率，再进行逆傅里叶变换得到时域信号，理论上来说即可完成语音变调。

三、实验结果及分析

这里均使用提供的语音文件，音调变为原来的1.5倍，并使用Adobe Audition绘制频谱图。

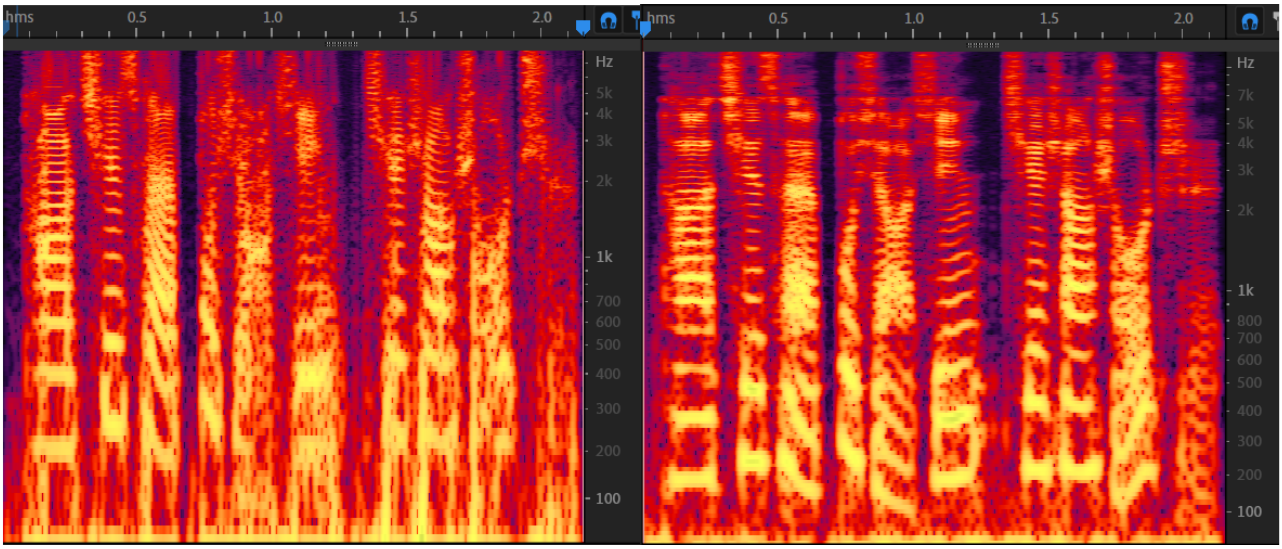
1.使用Adobe Audition 变调

使用Audition自带的变调功能进行变调，主观听感最为真实、自然，故将其作为标准参考。（下图左侧为原始频谱图，右侧为变调后频谱图）



2.使用LPC分析变调

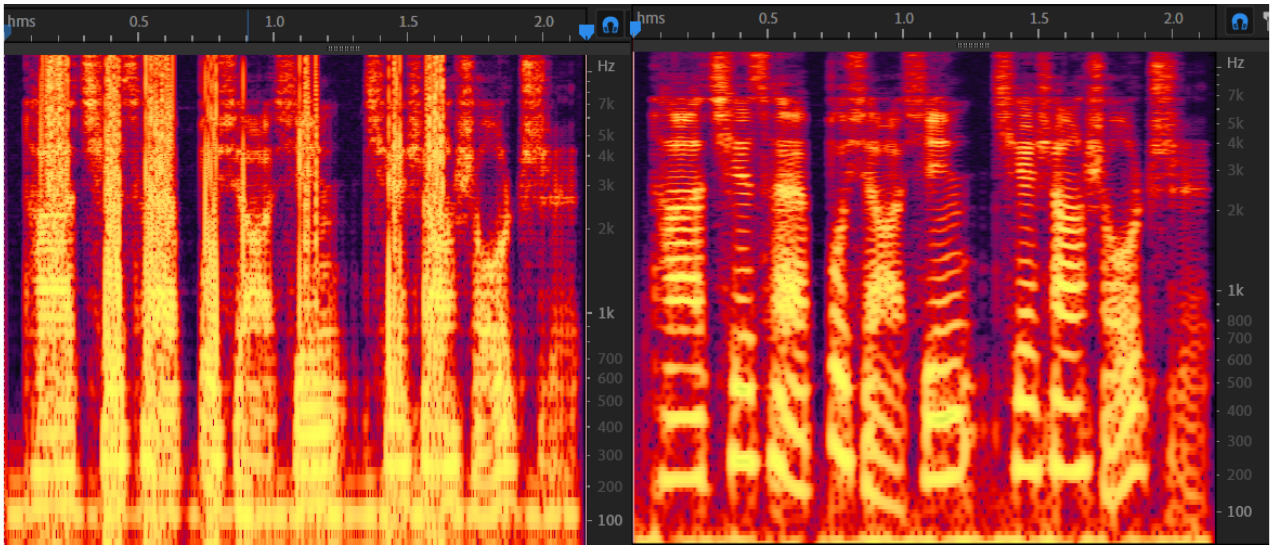
使用LPC分析方法，得到的声音主观感受有一些毛刺、“咔咔”声，也不太清晰，观察其频谱图如下（右侧为标准参考）：



可见其频谱图中出现一些明亮竖线，推测其为“咔咔”声的主要来源；同时不同时间点均存在能量泄露的情况，导致其共振峰的包络没有参考组清晰，但总体来说较为不错

3.使用STFT变调

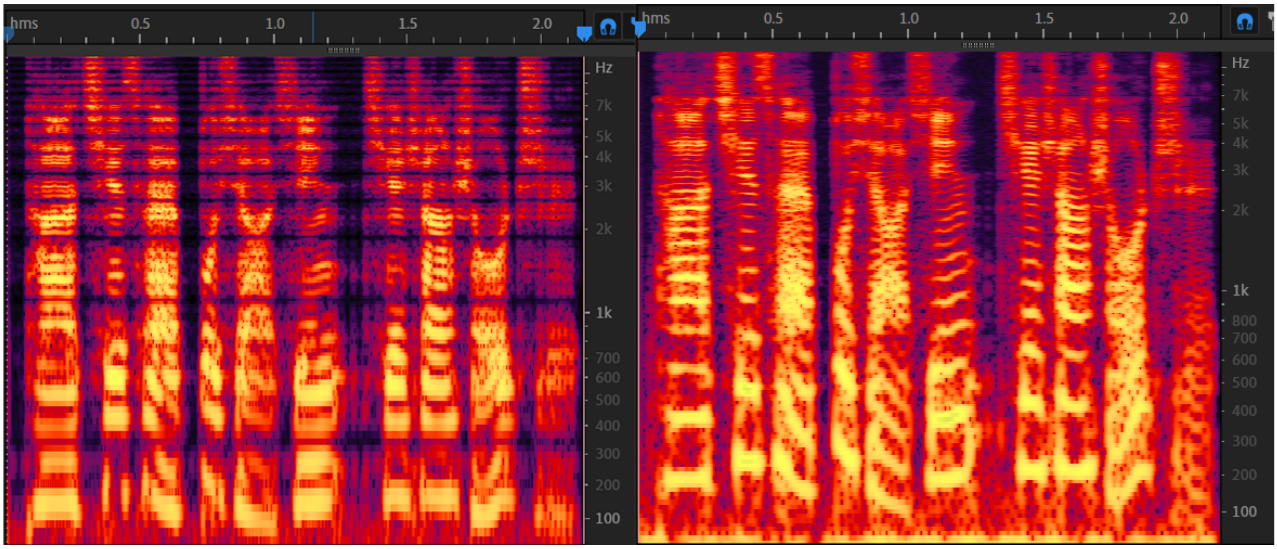
直接使用STFT得到原声音的频谱序列，将所有的频率乘以1.5再做逆变换得到序列。帧长20ms，帧重叠10ms。声音主观感受十分模糊，能听到背景中的“嗡嗡”声和“沙沙”声。



观察频谱图，可见能量严重泄漏，共振峰的包络也发生很大的变化。
上网查找相关资料后得知，虽然从理论上来讲，直接对频谱图进行操作，再进行逆变换一定能得到理想的频谱图，但是实际上对每一帧分别做逆变换再拼接时，信号的相位会有差异导致帧与帧之间不能很好拼接，相当于引入了一个随机噪声，因此会产生杂音。

4.对3的改进

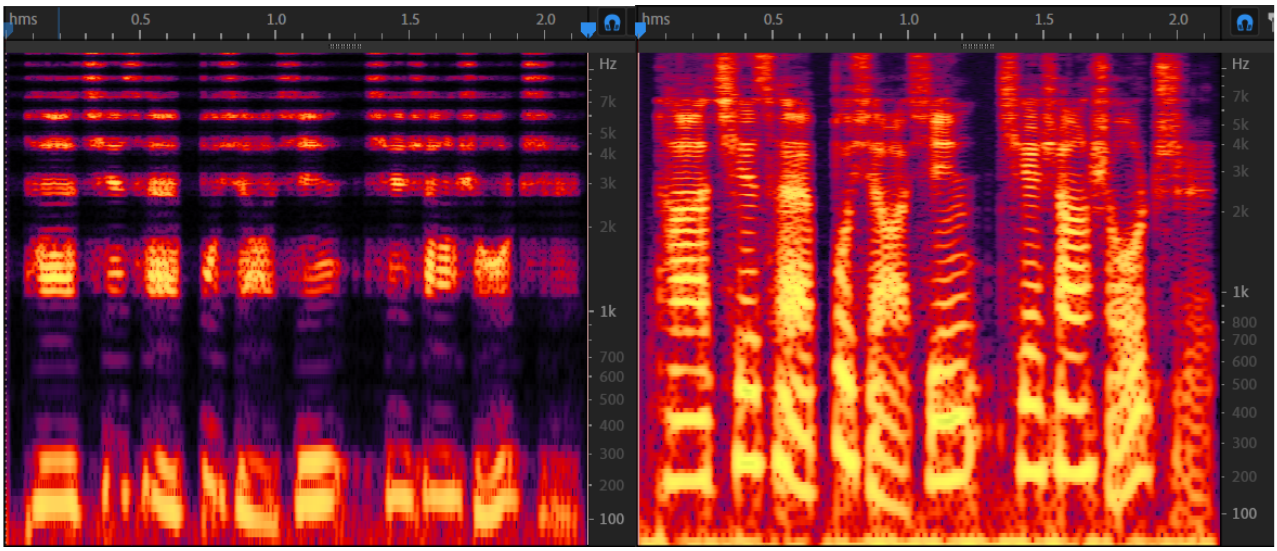
解决帧之间的相位差主要有两种方法：手动进行相位补偿或者增加帧与帧之间的重合程度，这里采用后一种。经过多次实验，将重叠部分长度由0.5*帧长改为0.8*帧长时，效果最佳，其频谱图如下：



可见共振峰的结构基本得到保持，杂音也较少

但图中明显存在很多条暗横线，怀疑其为窗函数的频谱对声音频谱造成的污染

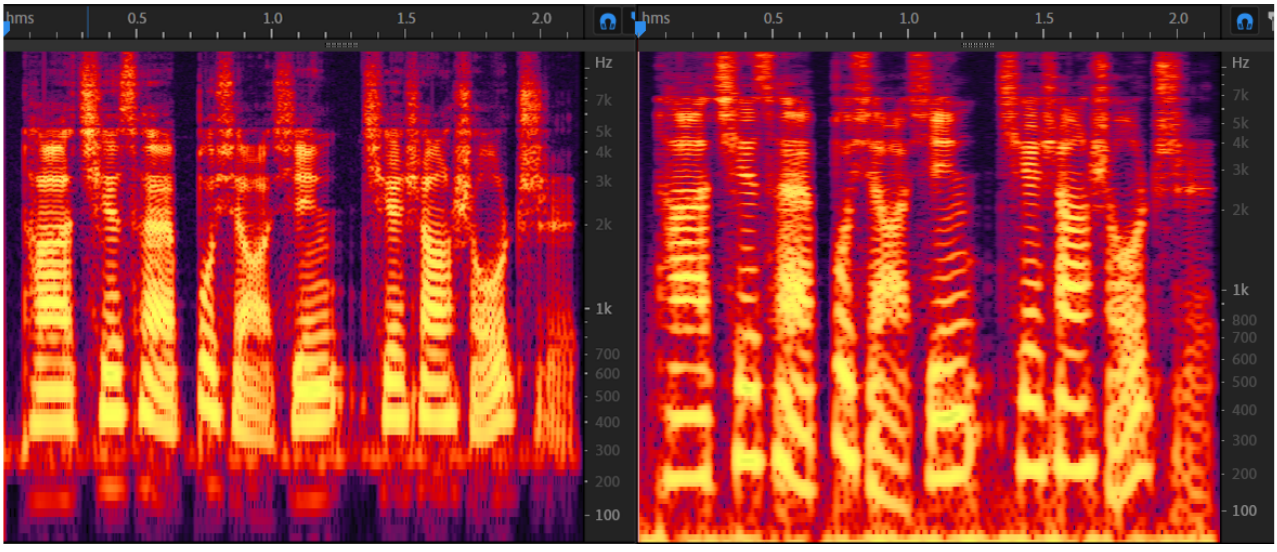
当重叠部分长度继续增加的时候，这一影响也明显加重，例如下图是重叠部分长度为0.9*帧长时的频谱图



可见很多频带的信息几乎完全丢失

5.另一种简单粗暴的尝试

尝试直接将stft后得到的频谱线性向上平移，再做逆变换。虽然基频实际上提高了许多，但主观听感是感觉音调变化很少。可能是因为共振峰之间的相对位置并没有发生变化，直接以线性方式改变频谱图并不会很大程度影响人类的听感。其频谱图如下：



四、总结

我原本以为语音变调就是简简单单在频域上修改，再做逆变换得到原序列，但实际上比这复杂得多。人类感知的音调不仅仅取决于声音的基频，还和共振峰的相对位置有关。如果仅仅是改变基频而不改变共振峰的相对位置并不会对音调的感知造成太大变化。如果想进一步更逼真地变换音调，则需要实现帧与帧之间的相位对齐，同时统计男、女声之间共振峰的差异，并基于统计数据进行变调。这大概就进入机器学习的领域了。如果使用RNN的话，应该还能够学习到上下文之间的相关性（如元辅音之间的关系，等等）。