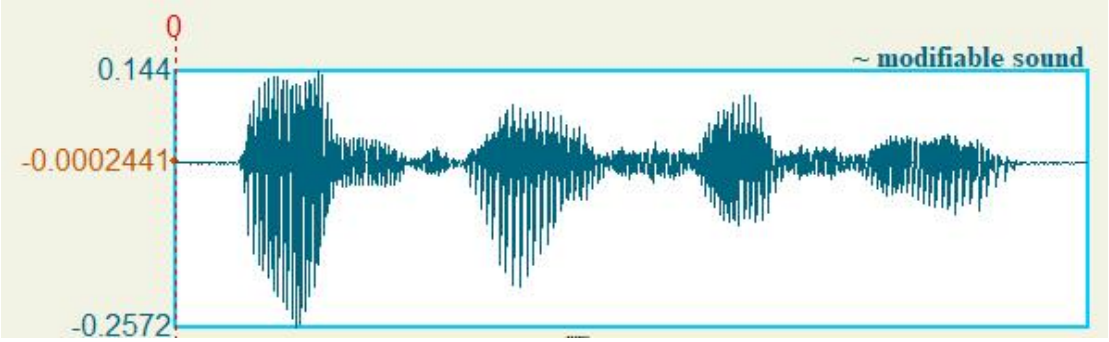
实验一：Praat 使用及语音信号处理算法基础

任务一：声学参数（20 分）

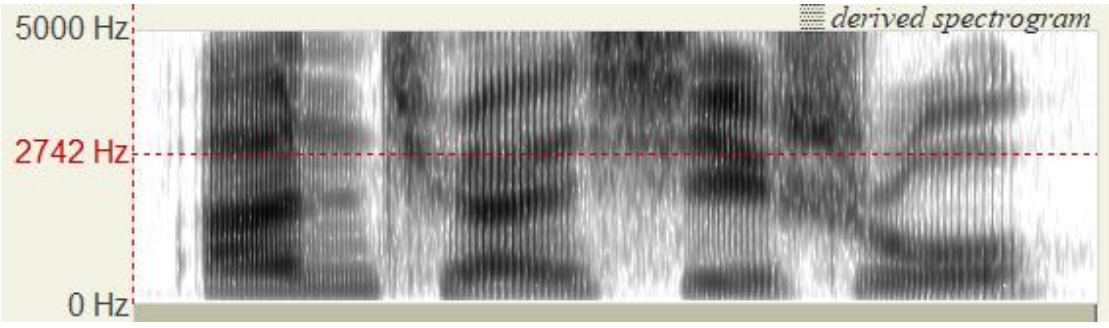
加载“GuoL/40004.wav”音频，在此基础上进行以下操作并回答如下问题：

1) 显示和查看波形 waveform、语谱图 spectrogram、音强 intensity、基音轮廓 pitch contour、共振峰 formant 和脉冲 pulses。（2 分）

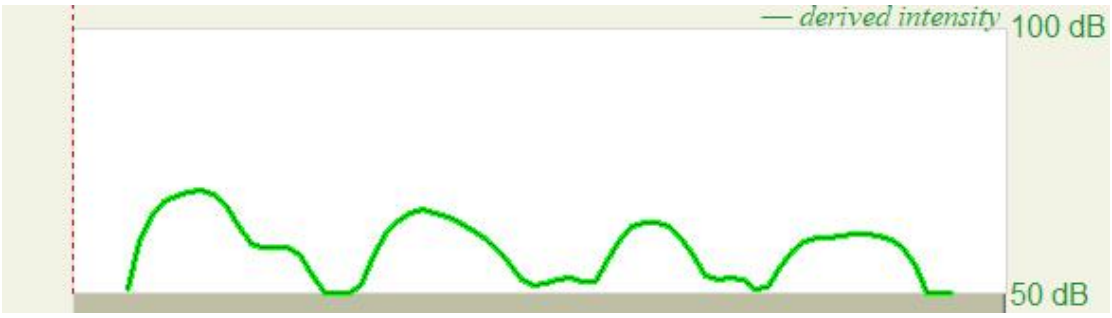
解：波形 waveform 如下图所示：



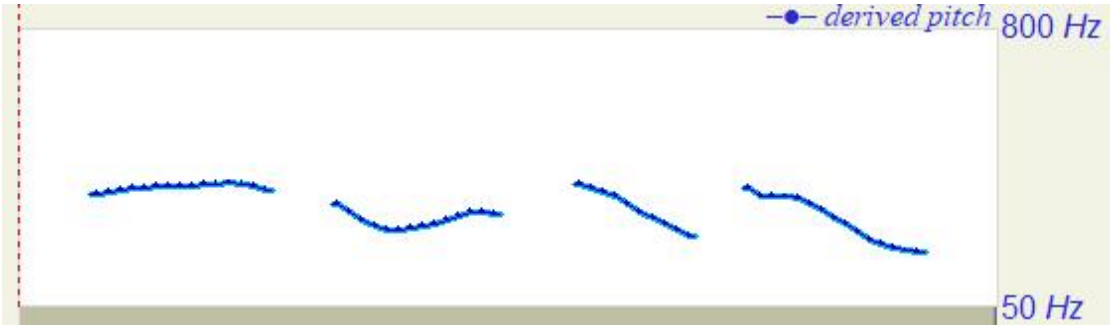
语谱图 spectrogram 如下图所示：



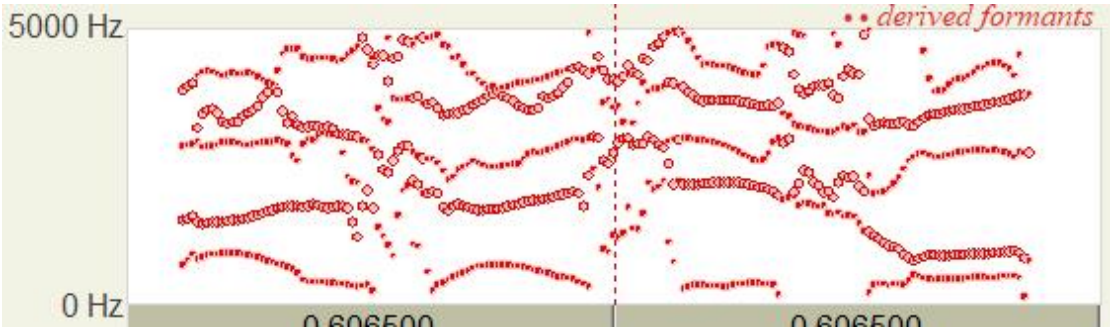
音强 intensity 如下图所示：



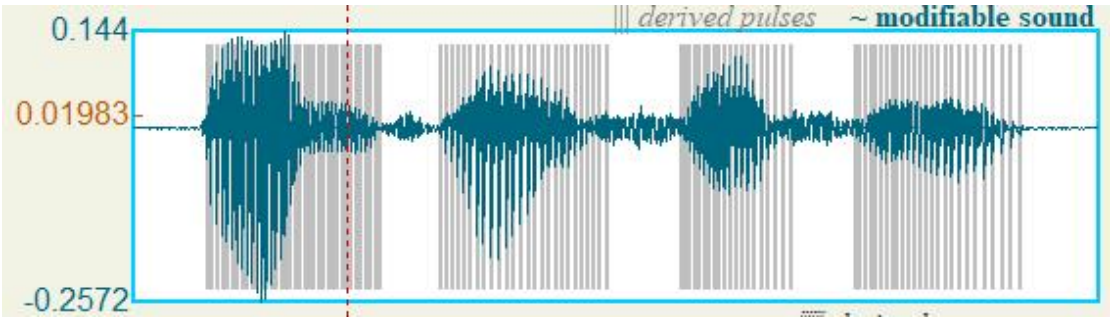
基音轮廓 pitch contour：



共振峰 formant：



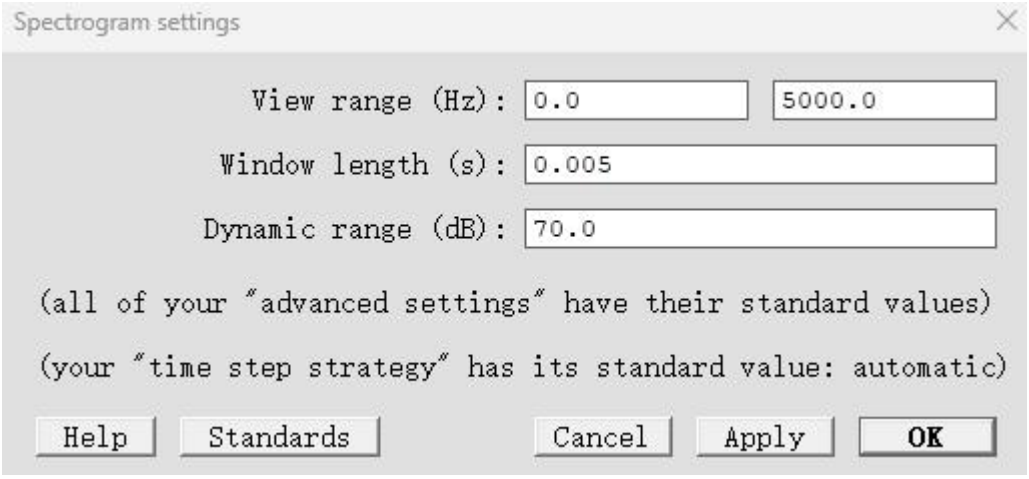
脉冲 pulses：



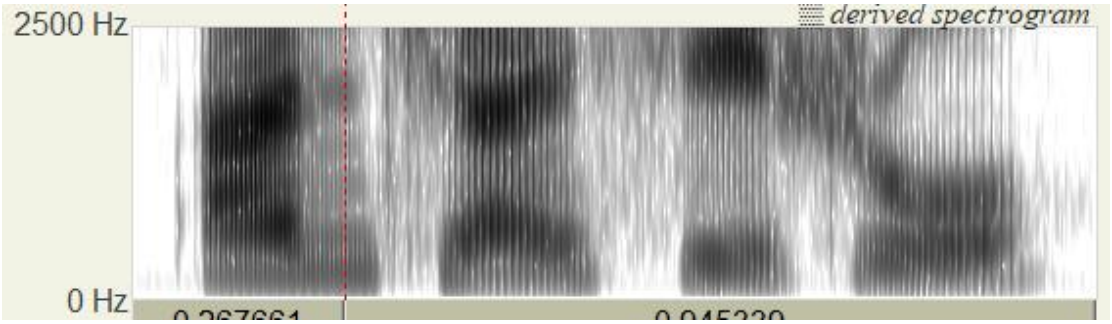
2) 更改每个声学参数的计算/提取的设置参数，并观察比较不同设置参数对应的结果差别。（2 分）

1. 更改语谱图 spectrogram 参数： （每次更改均在前一次基础上）

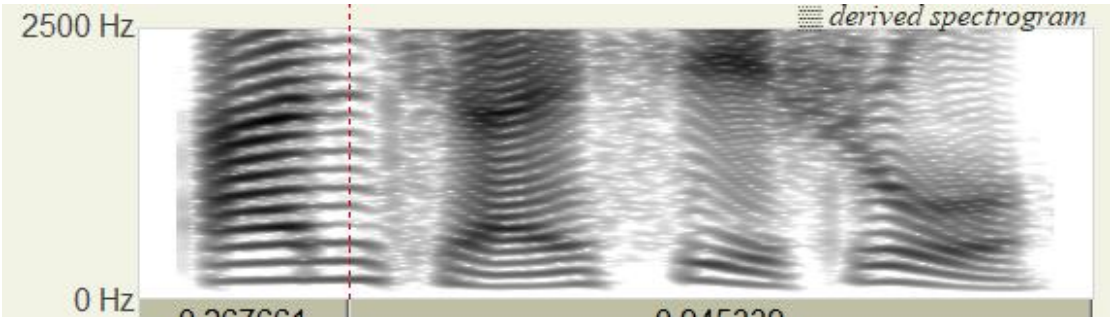
- 参数标准值/原参数如下所示：



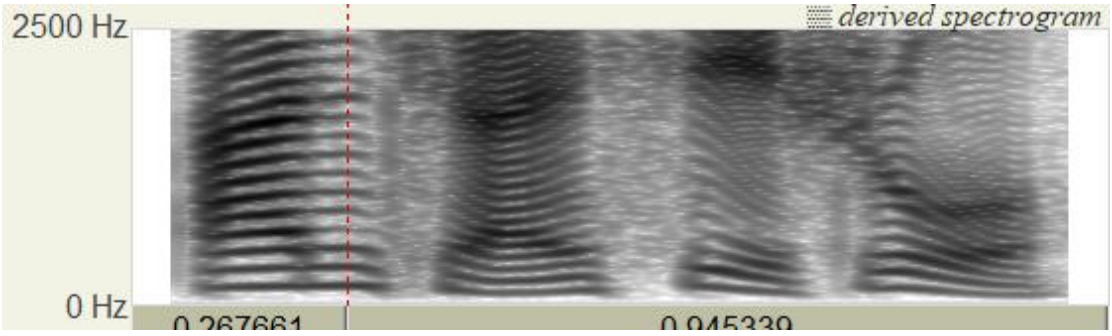
- 将 View range 上限由 5000 改为 2500：



- 将 Window length 由 0.005 改为 0.05：

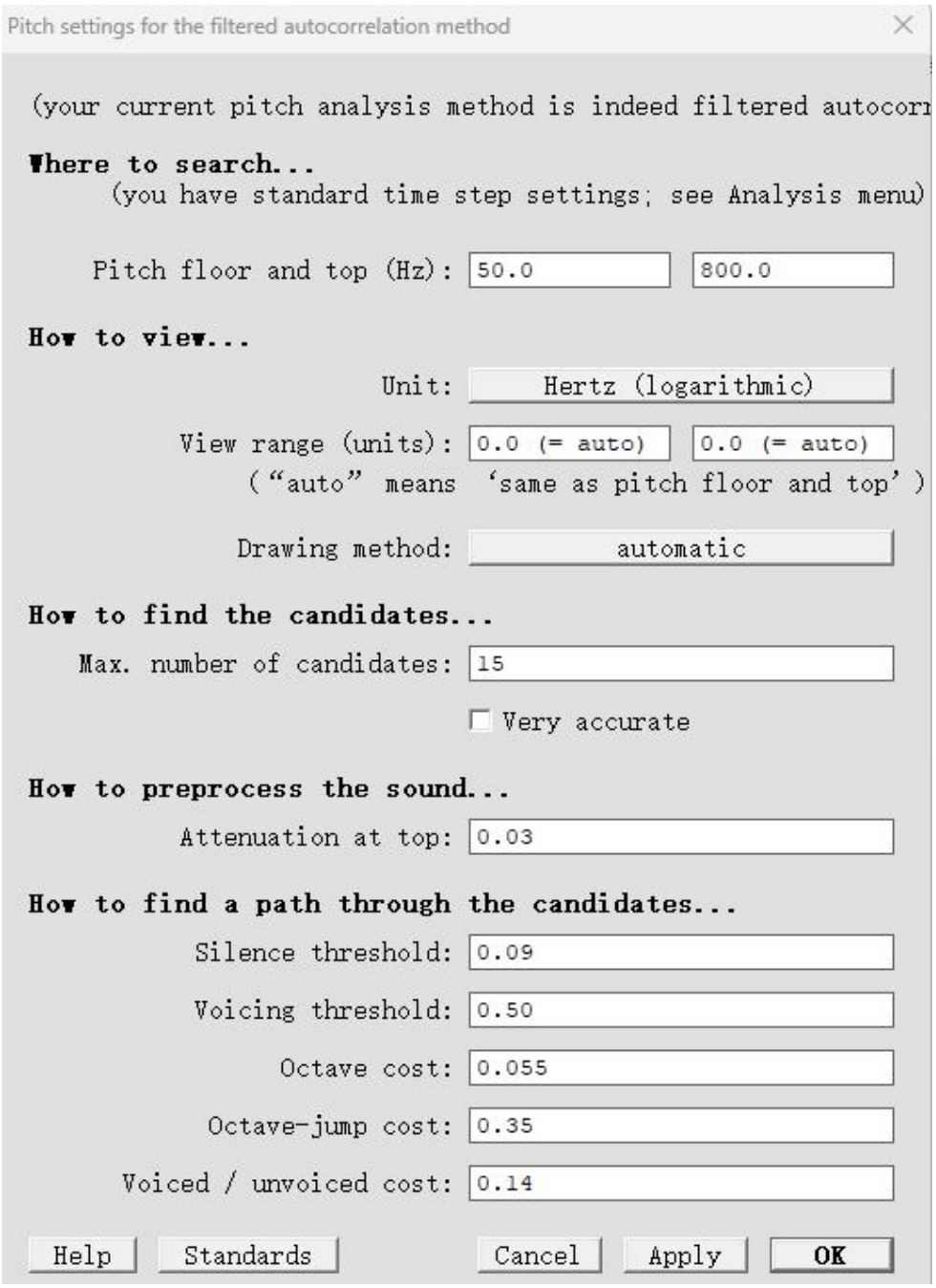


- 将 Dynamic range 由 70 改为 100：

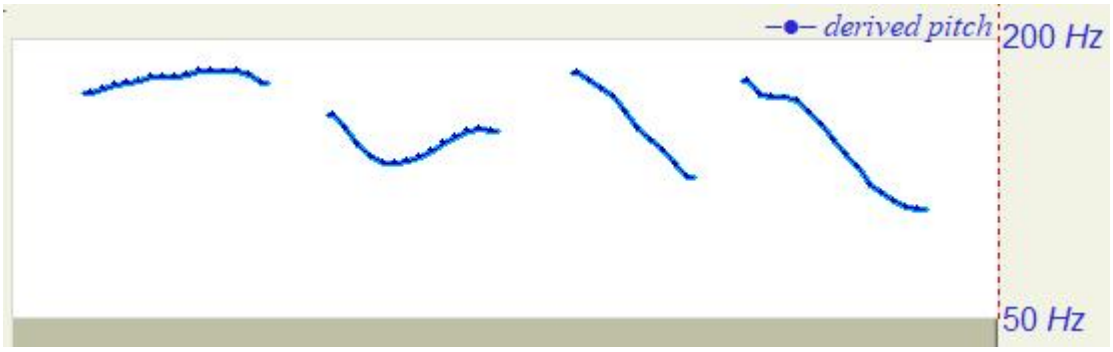


2. 更改基音轮廓 pitch contour 参数：

- 参数标准值/原参数如下所示：



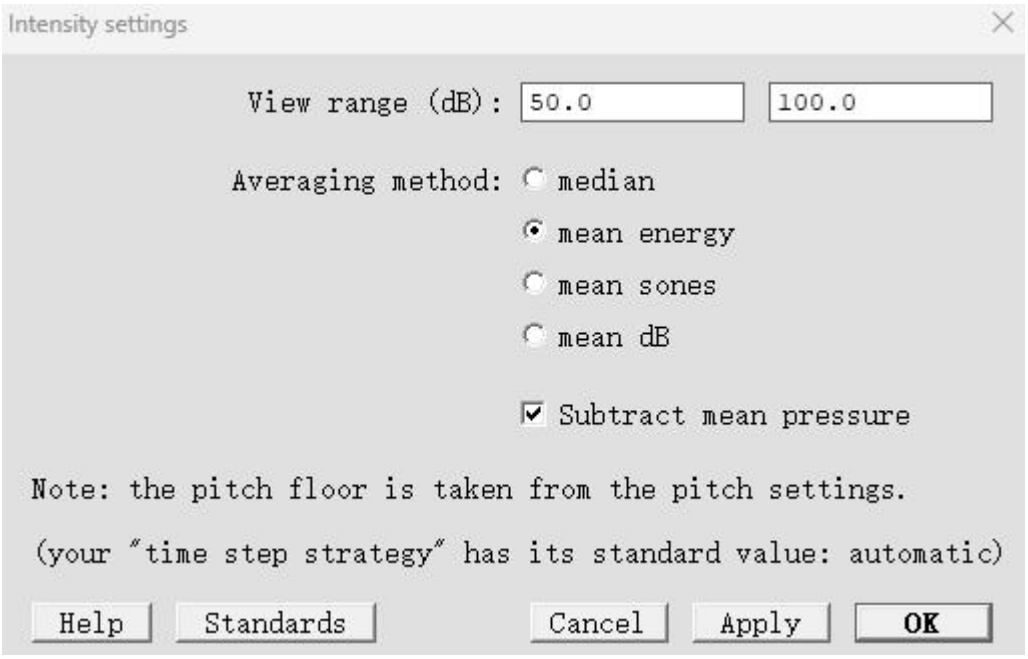
- 将 View range 上限由 800 修改到 200：



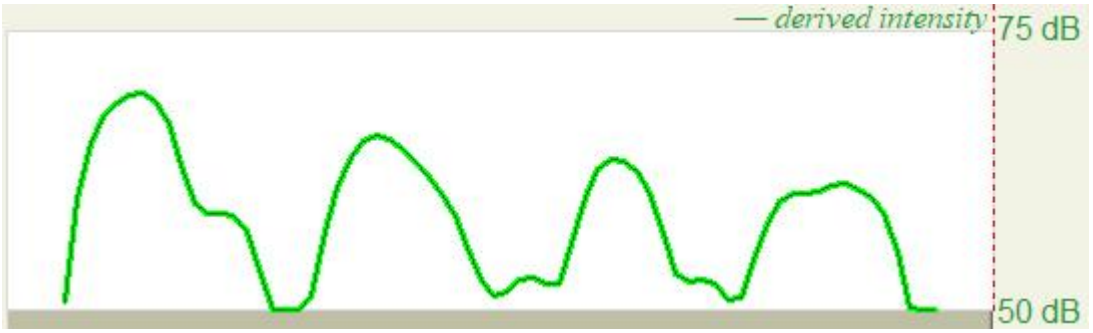
下面参数实在太多了 QAQ，就改这一个参数吧

3. 更改音强 intensity 参数：

- 参数标准值/原参数如下所示：

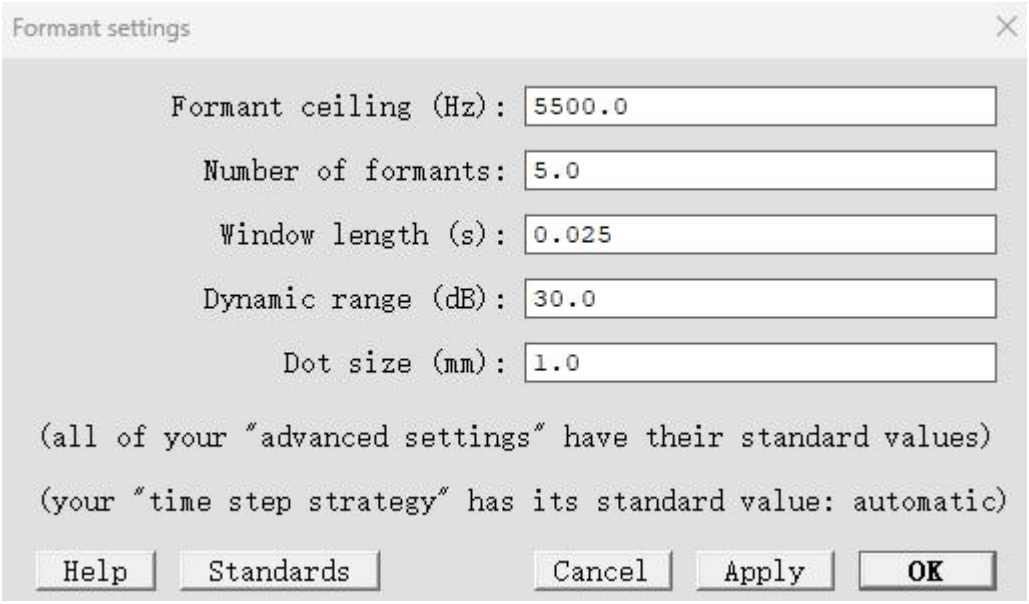


- 将 View range 上限由 100 修改到 75：

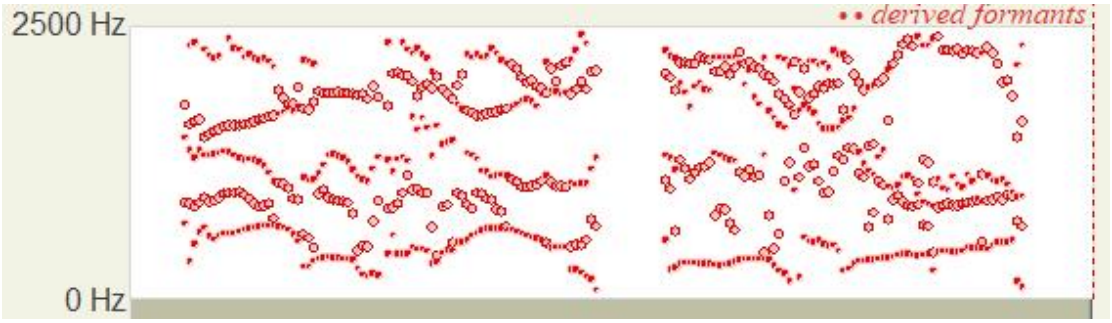


4. 更改共振峰 formant 参数：

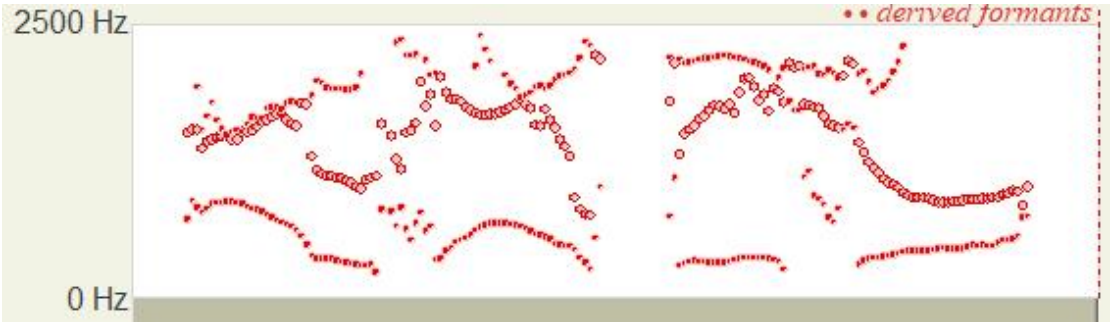
- 参数标准值/原参数如下所示：



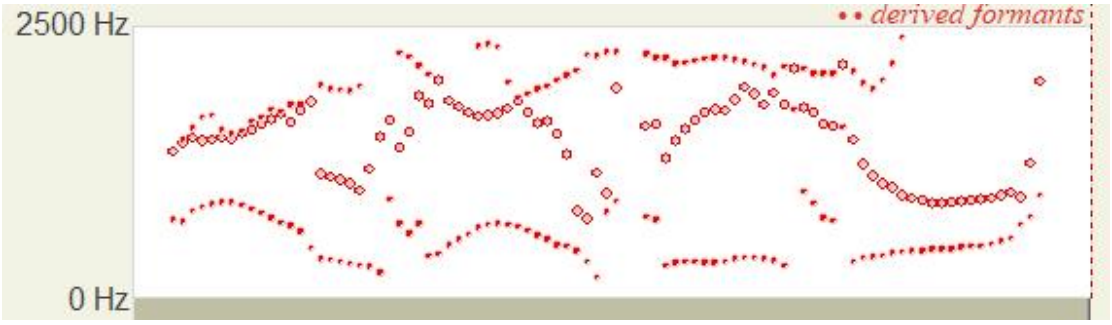
- 将 Fromant ceiling 由 5500 修改到 2500：



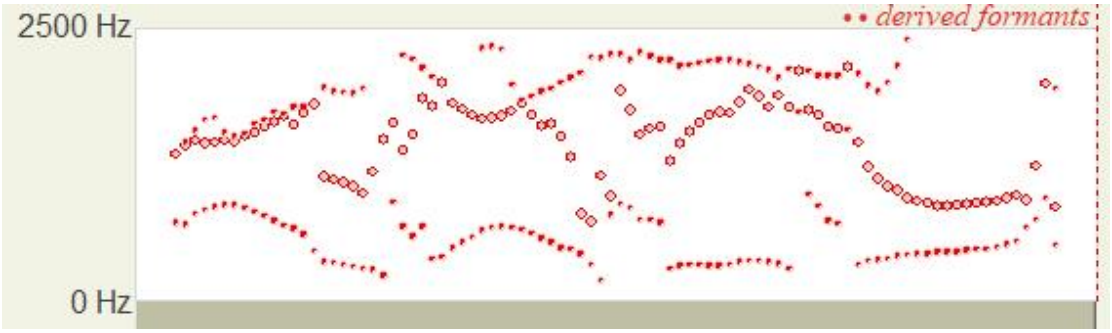
- 将 Number of formants 由 5 修改到 3：



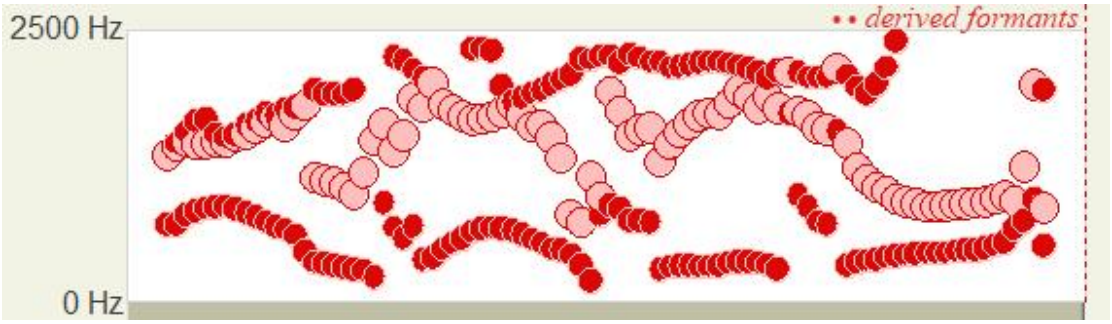
- 将 Window length 由 0.025 修改到 0.05：



- 将 Dynamic range 由 30 修改到 60：



- 将 dot size 由 1 修改到 3：



3) 解释上述设置参数的含义。（4 分）

1. 语谱图 spectrogram 设置参数含义：

- **View range(Hz)**：查看范围，即窗口中显示出来的频率范围，纵坐标的上下限即为参数设置的值。
- **Window length(s)**：窗口长度，即频谱分析窗口的持续时间，值越大，语谱图中的每条竖线越宽；反之，值越小，语谱图中的每条竖线越细。

- **Dynamic range(dB)**：动态范围，值越小，语谱图中的白色部分面积越大；反之，值越大，语谱图中的每条竖线越小。

2. **基音轮廓 pitch contour 设置参数含义：**

- **pitch range(Hz)**：基音范围，即进行音高分析时设置的范围，即分析图右边的纵坐标区间。

3. **音强 intensity 设置参数含义：**

- **View range(dB)**：查看范围，即窗口中显示出来的音强范围，分析图中纵坐标的上下限即为参数设置的值。

4. **共振峰 formant 设置参数含义：**

- **Formant ceiling(Hz)**：最高分析基频到哪个频率，画点的纵坐标上限不会高于设置的值。
- **Number of formants(1)**：共振峰的数量，即告诉算法到底要追踪几条共振峰，粉红色线和红色线的全部数量即为共振峰的数量，且两条线交错绘制。
- **Window length(s)**：窗口长度，即频谱分析窗口的持续时间，W 值越大，绘制共振峰的点越稀疏；反之，值越小，绘制共振峰的点越密集。
- **Dynamic range(dB)**：动态范围，值越大，绘制共振峰低能量区域的点越稀疏；反之，值越小，绘制共振峰低能量区域的点越密集。
- **Dot size(mm)**：圆点半径，值越大图中圆点越大。

4) 解释 Praat 提取音强 intensity、音高 pitch 和语谱图 spectrogram 的原理与算法。（4 分）

经查阅资料得知，Praat 提取音强 intensity、音高 pitch 和语谱图 spectrogram 的原理与算法如下：

1. 音强 Intensity 的提取原理

- Praat 通过计算音频信号的平方值来估算声音的能量。这是因为信号的能量与其振幅平方成正比。
- 为了获得随时间变化的音强，Praat 使用固定长度的窗函数（如汉明窗或矩形窗）对信号进行分割，在每个时间窗口内计算能量。通常，窗口长度在 30 到 50 毫秒之间。
- 为了更符合人耳对响度的感知，Praat 通常会对能量值取对数。这个对数值就代表了音强。

2. 音高 Pitch 的提取原理

Praat 提取音高使用的是自相关法（Autocorrelation），这是一种常用的音高跟踪算法。

- 首先对音频信号进行预处理，去除直流偏移并施加高通滤波，以减少低频噪声的影响。
- 将音频信号划分为短时帧，通常是 10 到 40 毫秒的窗口。音高是基于每个短时帧计算的。
- 对每一帧信号，计算自相关函数，即将信号与其延迟版本相乘并累加。通过寻找自相关函数的最大值位置来确定基频周期。
- 根据找到的自相关峰值位置，可以推导出该帧的基频。峰值对

应的时间延迟是基频周期，然后通过取其倒数得出基频。

3. 语谱图 Spectrogram 的生成原理

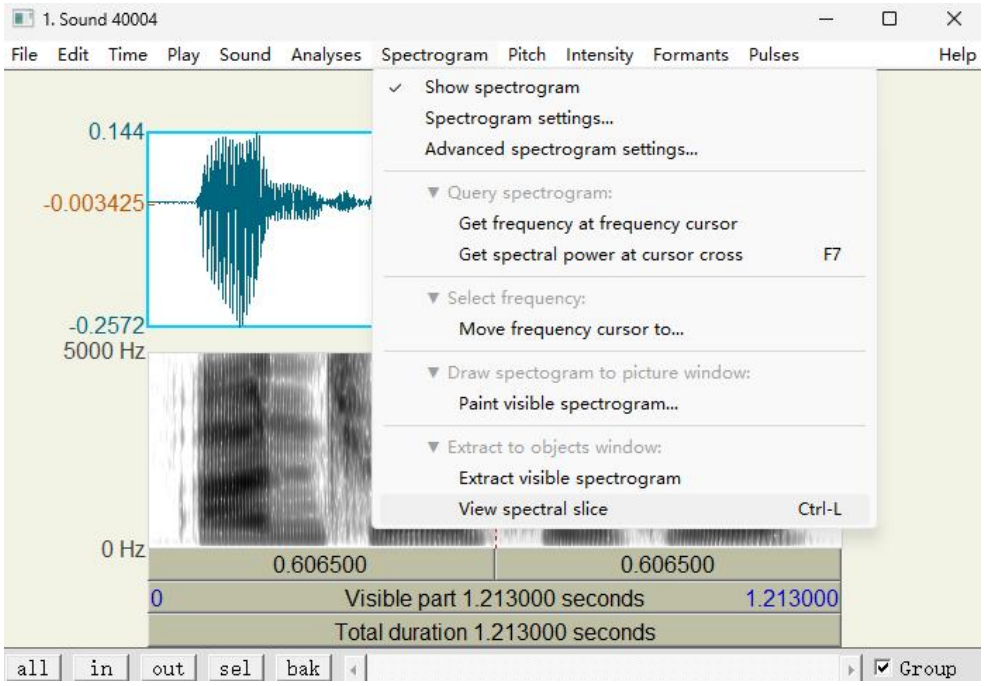
Praat 使用短时傅里叶变换 (Short-Time Fourier Transform, STFT) 来生成语谱图。

- **时窗分割**：与音高提取类似，音频信号被分割成小的时窗（通常 5 到 30 毫秒），每个窗口内的信号被视为一个独立的片段。
- **窗函数加权**：为了减少窗口边界效应，Praat 使用窗函数（如汉明窗、矩形窗等）对每个窗口内的信号进行加权。
- **傅里叶变换**：对每个窗口内的信号应用快速傅里叶变换 (Fast Fourier Transform, FFT)，将时域信号转换到频域，得到该窗口内不同频率分量的振幅。
- **频谱显示**：将每个窗口的频率分量的强度（通常取其对数值）作为纵坐标，时间作为横坐标，生成一幅随时间变化的频率强度图。
- **时频分辨率平衡**：语谱图的分辨率取决于窗口大小。较短的窗口提供较高的时间分辨率，较长的窗口则提供较好的频率分辨率。Praat 允许用户调节这些参数以适应不同的分析需求。

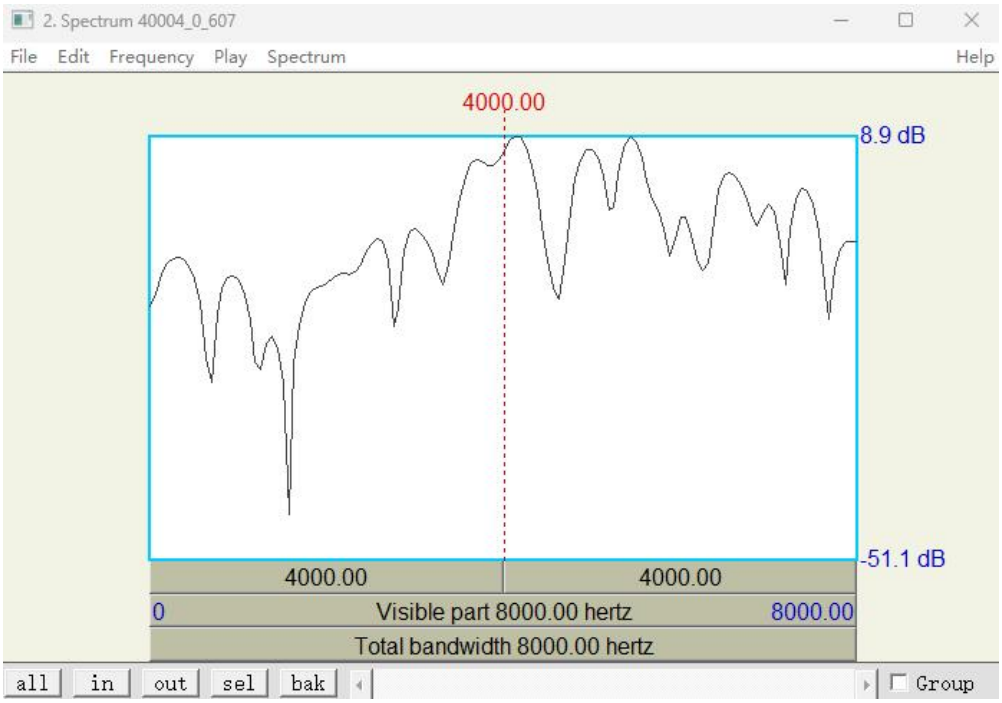
5) 共振峰 formant 和频谱图 spectrogram 之间的关系是什么？如何用 Praat 获得频谱切片 spectral slice？ (4 分)

共振峰是被声道特别放大的频带，是指在声音的频谱中能量相对集中的一些区域（语谱峰值）。在频谱图上，共振峰通常表现为频率上的明亮或高能量的频段。这种高能量的频率带反映了语音信号中哪些频率成分被声道增强了，从而形成了共振峰。因此，共振峰实际上是频谱图中的局部频率增强区域。

在 Praat 中点击下图所示按钮即可获得频谱切片 spectral slice：



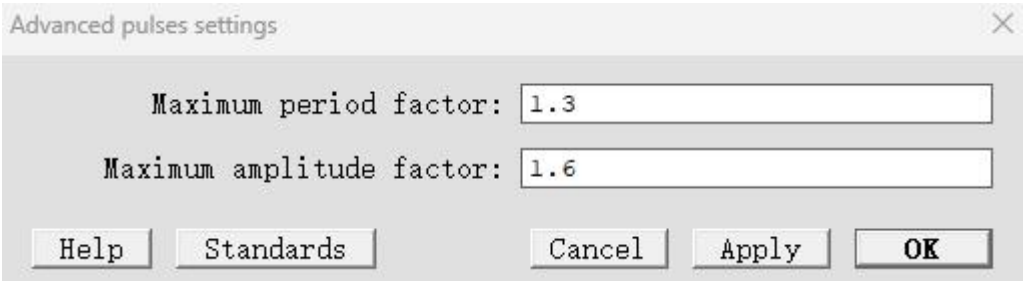
得到结果如下图所示：



6) 什么是脉冲 pulses？哪些声学参数与脉冲有关？ （4 分）

脉冲 Pulses 是指一种短时持续的信号，通常在物理、声学、电磁等领域中广泛使用。声学中的脉冲通常是指时间上短暂的声音信号，可能表现为单一的声音峰值或短时间内的声音爆发。这种信号的特点是持续时间较短，并且能量集中在某个较短的时间段内。

Praat 中与脉冲相关的可设置参数如下所示：



其中，Maximum period factor 是指相邻两个脉冲之间的周期变化允许的最大比例变化。Praat 的脉冲检测算法默认相邻两个脉冲的周期应该比较稳定，但在现实语音中，声带振动的周期会有一定波动。

该参数允许一些波动的存在，但超出设定值的变化会被认为是噪声或异常信号，从而不会被识别为脉冲。

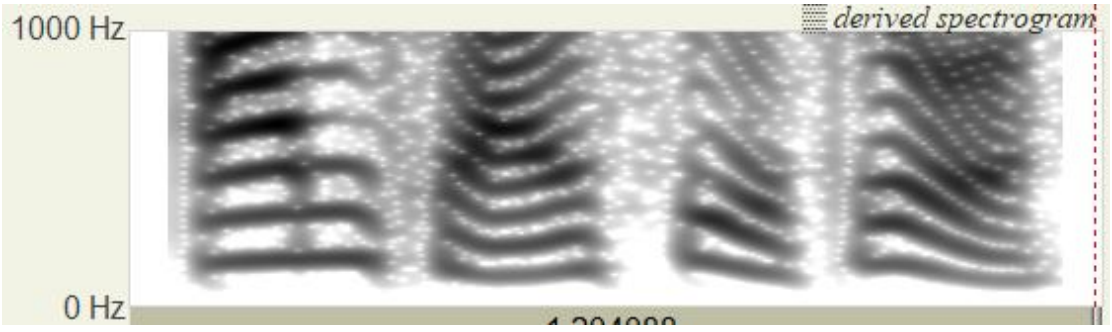
Maximum amplitude factor 是指相邻两个脉冲的幅度变化允许的最大比例变化。Praat 假定在正常情况下，声带振动产生的相邻脉冲幅度（即声音的强度）应该比较一致。这个参数允许一些幅度变化，但如果相邻脉冲的幅度变化超过了设定值，Praat 将认为这是异常的幅度变化，可能是由于噪声或发声不稳定导致的，因此不会标记为脉冲。

此外，脉冲持续时间、脉冲重复频率、脉冲幅度等声学参数也与脉冲有关。

任务二：发音与听觉感知（20 分）

1. 从文件夹“GuoL”中加载所提供的语音波形，观察并解释语音的谐波 harmonics。（4 分）

本题以任务一中的“GuoL/40004.wav”音频为例，对其进行操作：将 View range 调整为 0 到 1000， Window length 调整为 0.05，便于观察谐波 harmonics，得到频谱图如下：



频谱图最下方的线即为基频，基频是语音信号中最低的频率成分，

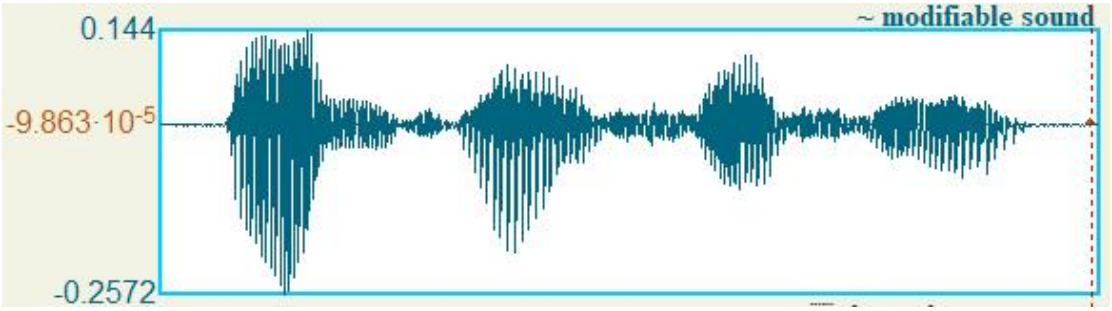
通常由声带的振动产生。在语音信号中，基音决定了声调的高度，即音高。

基频上面的平行线即为谐波，谐波的频率为基频的整数倍，频率为基频 2 倍的谐波为二次谐波，频率为基频 n 倍的谐波为 n 次谐波。谐波在语音信号中产生共振和音色。

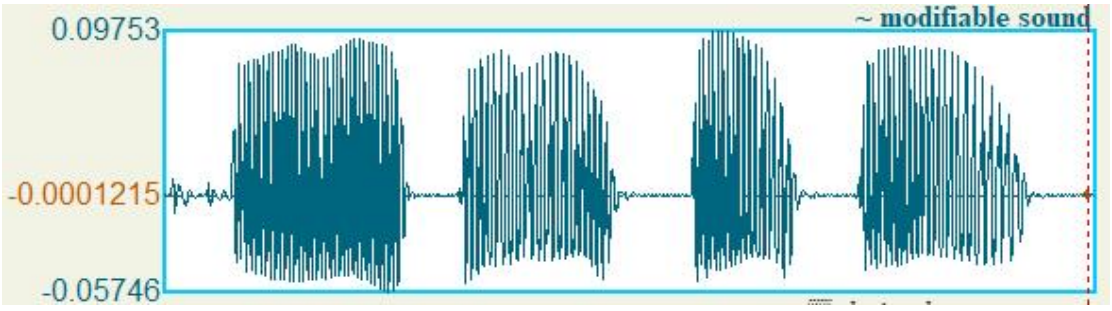
2. 从文件夹“GuoL”中加载所提供的语音波形，比较 EGG 信号和语音信号的波形 waveforms 差别。（2 分）

本题以“GuoL/40004.wav”和“GuoL/40004.egg.wav”音频为例。

“GuoL/40004.wav”的波形如下：



“GuoL/40004.egg.wav”的波形如下：



这两者具有显著的差异，主要体现在以下几个方面：

（1）波形的复杂性

语音波形较为复杂，包含高频细节（如振幅的快速波动）。这

种复杂性来自声带振动后在声道中的调制，以及声学环境对声音的影响。

EGG 波形相比语音信号更加规整，通常呈现为较为规则的波动。每个波峰和波谷代表声带的开合周期，因为 EGG 主要测量声带接触的电阻变化，不包含声道共振的影响。

（2）波形的周期性

虽然语音信号中也有一定的周期性，特别是在发声部分可以看到相对稳定的重复模式，但由于语音信号包含复杂的谐波和共振，周期性可能并不如 EGG 信号明显。

由于 EGG 主要反映声带的周期性振动，其波形的周期性非常明显，特别是在连续发声的段落，可以清楚地看到均匀的周期波动。

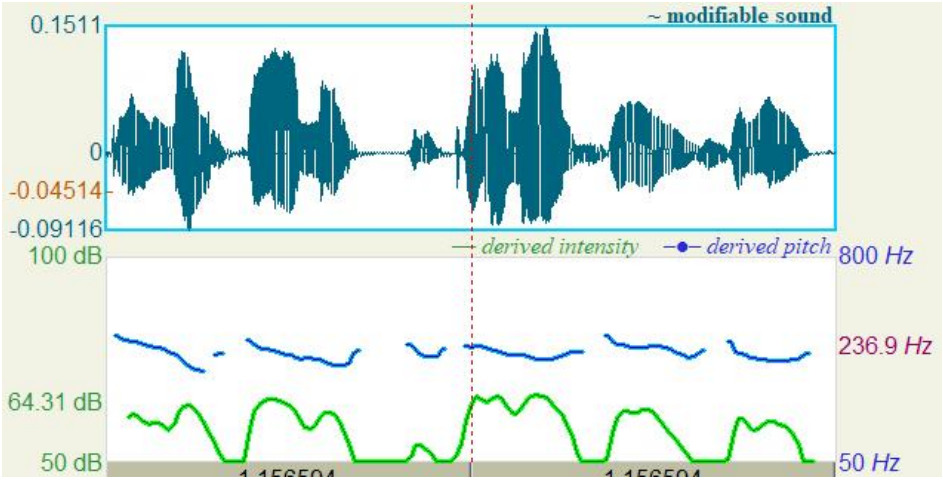
（3）波形的振幅

语音信号的振幅变化较大，尤其是由于语音中的语音强度和不同的声音产生机制（例如元音、辅音），因此波形的振幅会呈现出显著的变化。

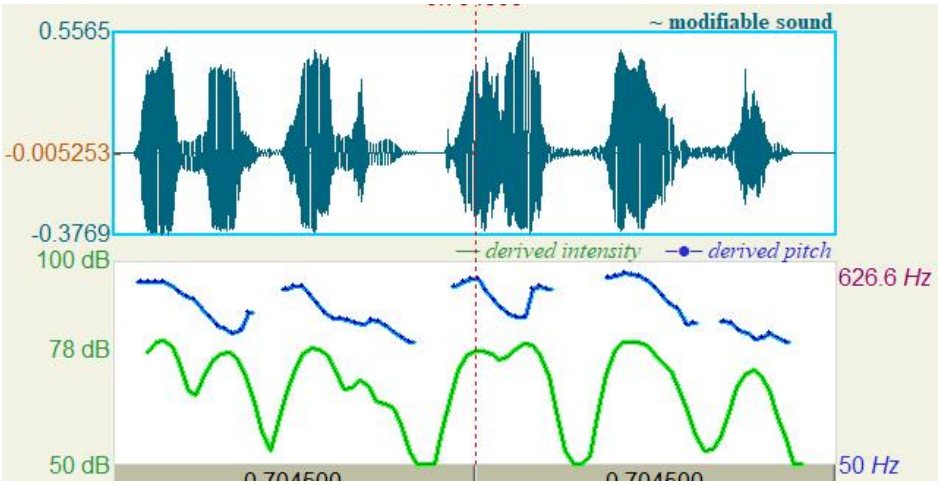
EGG 信号的振幅较为恒定，主要反映声带的接触面积变化，因此在稳定发声时，其波形振幅变化较小。

3. 绘制文件夹“Emotion”中“exp-0.wav”（中性）、“exp-1.wav”（愤怒）、“exp-4.wav”（悲伤）的基音轮廓 pitch contour（参考第二周讲义），比较三者之间的差异。思考：除了 pitch contour 外，三个语音还有什么其他声学特征的差别。（6 分）

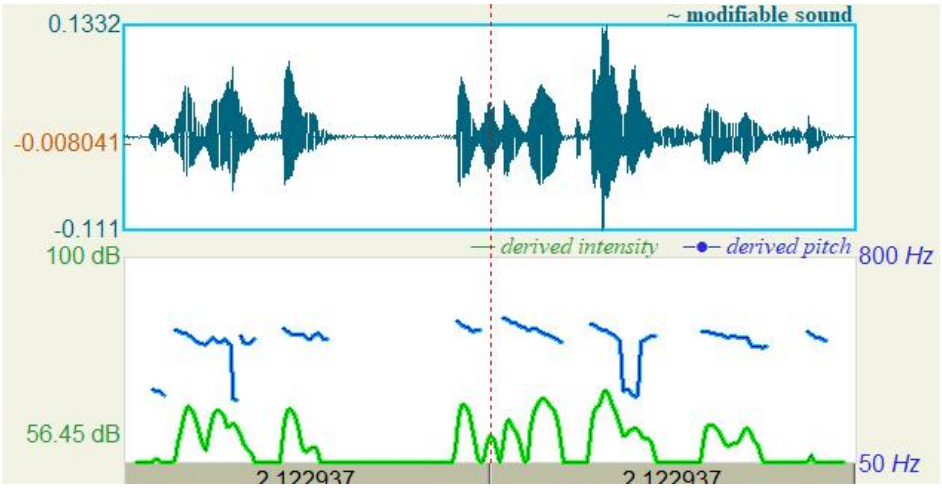
“exp-0.wav”（中性）的基音轮廓等特征如下图所示：（基音轮廓用蓝色线绘制，音强用绿色线绘制）



“exp-1.wav”（愤怒）的基音轮廓等特征如下图所示：



“exp-4.wav”（悲伤）的基音轮廓等特征如下图所示：



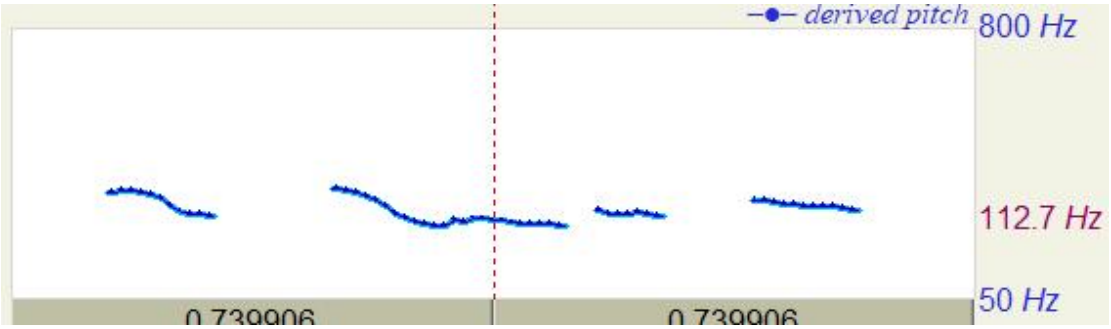
上面三图对比可以看出，“exp-0.wav”（中性）的基音轮廓主要分布在中间，音调相对适中；且比较连续平稳；“exp-1.wav”（愤怒）的基音轮廓分布比较离散，不连续；且音调较高；“exp-4.wav”（悲伤）的基音轮廓分布也比较连续，但音调偏低。

除了基音轮廓外, 三个语音的音强也有所不同: 相比于“exp-0.wav”（中性）的音强，“exp-1.wav”（愤怒）的音强较高，“exp-4.wav”（悲伤）的较低。

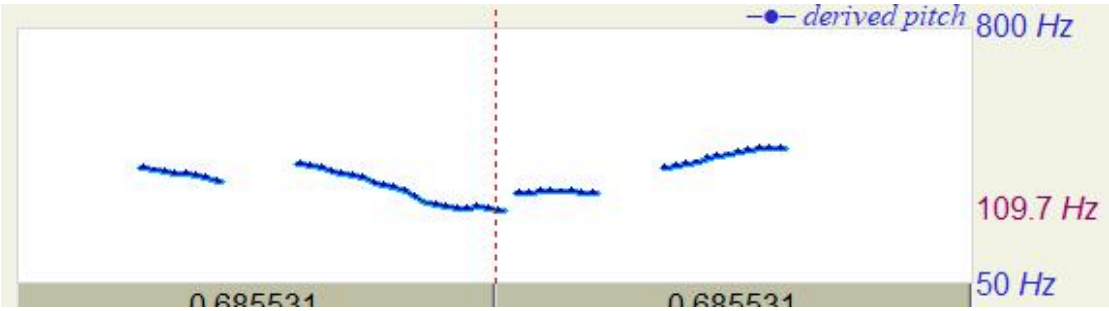
三个语音的波形差异主要体现在振幅上，相比于“exp-0.wav”（中性）的波形，“exp-1.wav”（愤怒）的平均振幅较大，“exp-4.wav”（悲伤）的平均振幅较小。

4. 绘制文件夹“Book”中“book_declaration.wav”、“book_question.wav”的基音轮廓 pitchcontour（参考第二周讲义），比较二者之间的差异。（2 分）

“book_declaration.wav”的基音轮廓如下图所示：



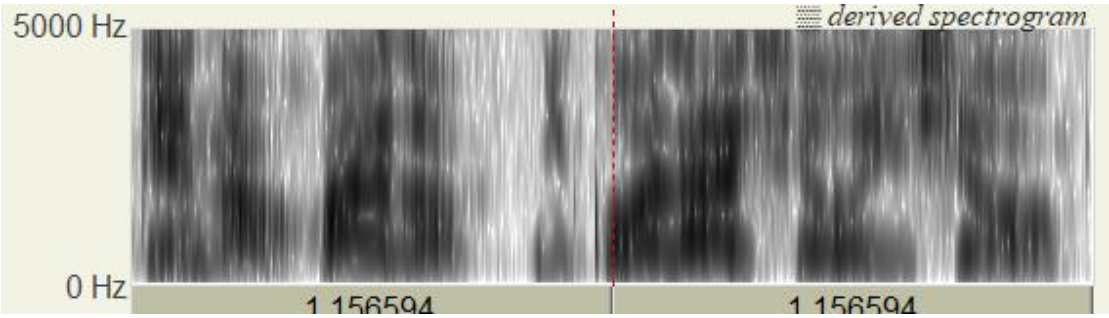
“book_question.wav”的基音轮廓如下图所示：



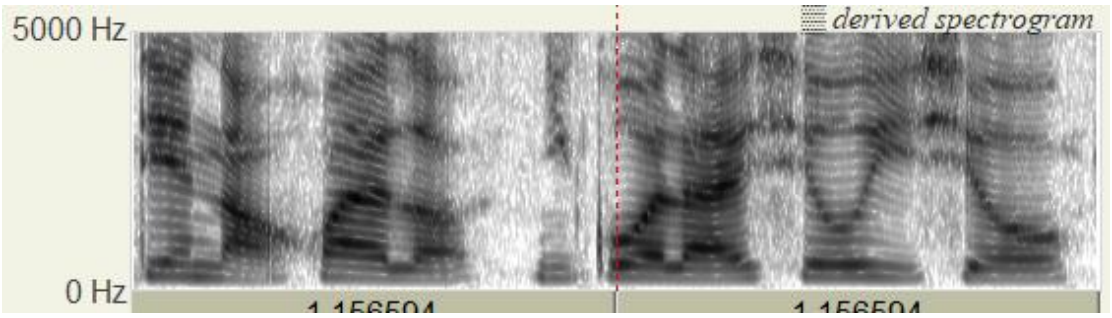
由上图可以看出，“book_declaration.wav”的基音轮廓呈下降趋势，可能因为 declaration 为陈述语气；“book_question.wav”的基音轮廓呈上升趋势，可能因为 question 为疑问语气。

5. 绘制文件夹“Emotion”中“exp-0.wav”的宽带语谱图和窄带语谱图（参考第二周讲义），比较二者之间的差异。（4 分）

由第二周讲义可知，频率分辨率取 300-400Hz，时间分辨率 2-5ms 时得到宽带语谱图，故取 Window length 为 0.002，得“exp-0.wav”的宽带语谱图如下所示：



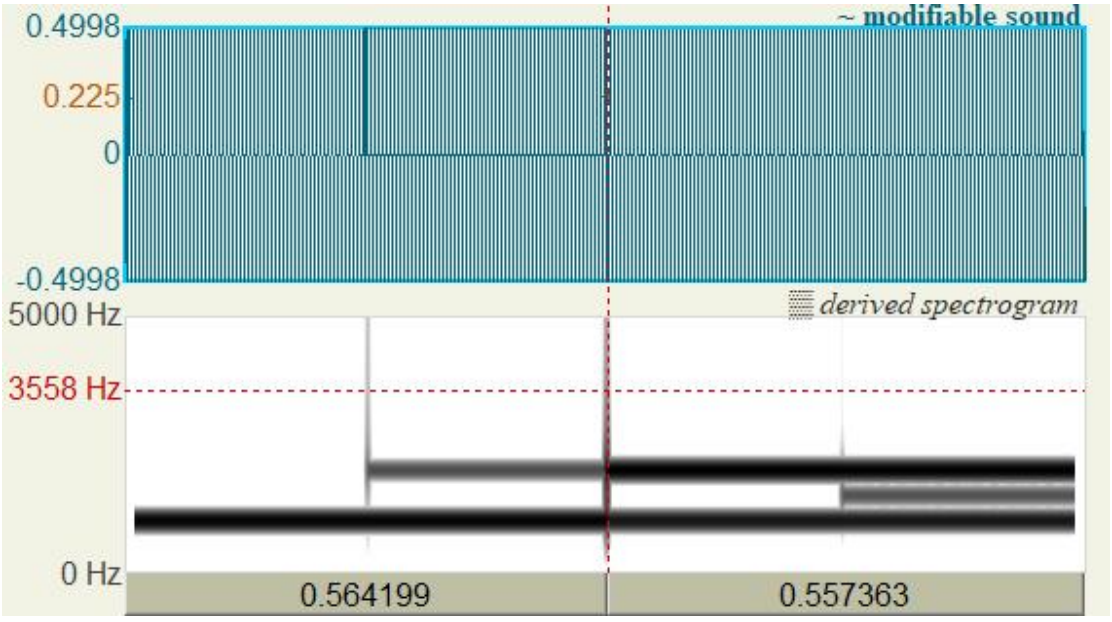
频率分辨率取 50-100Hz，时间分辨率 5-10ms 时得到窄带语谱图，故取 Window length 为 0.01 ，得“exp-0.wav”的窄带语谱图如下所示：



可以看出，宽带语谱图中出现的频谱线较宽，难以分辨出较为密集的谐波，具有良好的时间分辨率，但频率分辨率较差；窄带语谱图中出现的频谱线较细，能够清晰地看到谐波结构，具有良好的频率分辨率，但时间分辨率较差。

6. 加载文件夹“Mask”中“abc.wav”，显示该音频的波形 waveform 和频谱图 spectrogram，听不同的部分，体会掩蔽效应 mask 的效果，说明该段音频不同时间段分别包括哪些频率的声音，并说明在不同时间段你听到的声音情况，解释掩蔽效应。 (2 分)

“abc.wav”的波形和频谱图如下图所示：



由图可知，频谱图大概可分为四个阶段，大致分别为 0s-0.28s、

0.28s-0.56s、0.56s-0.84s 和 0.84s-1.12s。其中第一阶段只有 600-1400Hz 的低频分量, 后三阶段多了 1600Hz-2400Hz 的高频分量, 且第四阶段还有一段中频分量。

但在实际听取时, 只能明显区分出三段声音, 分别为 0s-0.28s 段、0.28s-0.56s 段及 0.56s-1.12s 段。这三段声音逐段变尖，音调逐段变高。第二段比第一段高显然是因为多了一段高频分量；第三段比第二段高是因为高频分量有所加强。

但有意思的是 0.56s-0.84s 和 0.84s-1.12s 这两段时间的声音听不出差别。如果用掩蔽效应来解释，就是高频分量太强，把新加入的中频分量完全掩蔽了，所以加入中频分量对声音几乎没有影响。