

语音信号的时域处理： 端点检测与基音周期估计

(语音信号数字处理课程报告)

姓 名： 肖文韬
学 号： 2020214245

二〇二〇年十月十一日

目 录

目录.....	I
插图清单.....	II
第 1 章 任务一：声学参数.....	1
1.1 显示和查看 SoundEditor 中的功能	1
1.2 比较和解释不同设置参数	2
1.2.1 语谱图.....	2
1.2.2 音强.....	2
1.2.3 基音轮廓.....	3
1.2.4 共振峰.....	3
1.2.5 脉冲.....	3
1.3 解释 Praat 提取声学参数的原理	4
1.3.1 音强.....	4
1.3.2 音高.....	4
1.3.3 语谱图.....	4
1.4 共振峰和频谱图的关系	5
1.5 脉冲	5
第 2 章 任务二：发音与听觉感知.....	6
2.1 语音的谐波 harmonics.....	6
2.2 比较 EGG 和语音信号波形	6
2.3 不同情绪的基音轮廓	9
2.4 不同语气的基音轮廓	9
2.5 宽带语谱图和窄带语谱图	10
2.6 掩蔽效应	12
参考文献.....	13

插图清单

图 1.1 Praat SoundEditor 界面概览	1
图 2.1 谐波	6
图 2.2 比较 EGG 和语音信号波形.....	7
图 2.3 不同情绪的基音轮廓	8
图 2.4 不同语气的基音轮廓	10
图 2.5 宽带语谱图和窄带语谱图	11
图 2.6 一个窄带语谱图的例子	11
图 2.7 掩蔽效应	12

第1章 任务一：声学参数

1.1 显示和查看 SoundEditor 中的功能

打开 Praat, 在 Object window 中点击 Open \rightarrow Read from file..., 导入 GuoL/40004.wav 文件。然后选中该项目, 点击 View & Edit, 进入 SoundEditor window。进去之后默认只展示波形 waveform、语谱图 spectrogram 和部分其他信息。

我们可以在菜单栏依次选中 Show Pitch/Intensity/Formant/Pulses 来展示音强 intensity、基音轮廓 pitch contour、共振峰 formant 和脉冲 pulses。

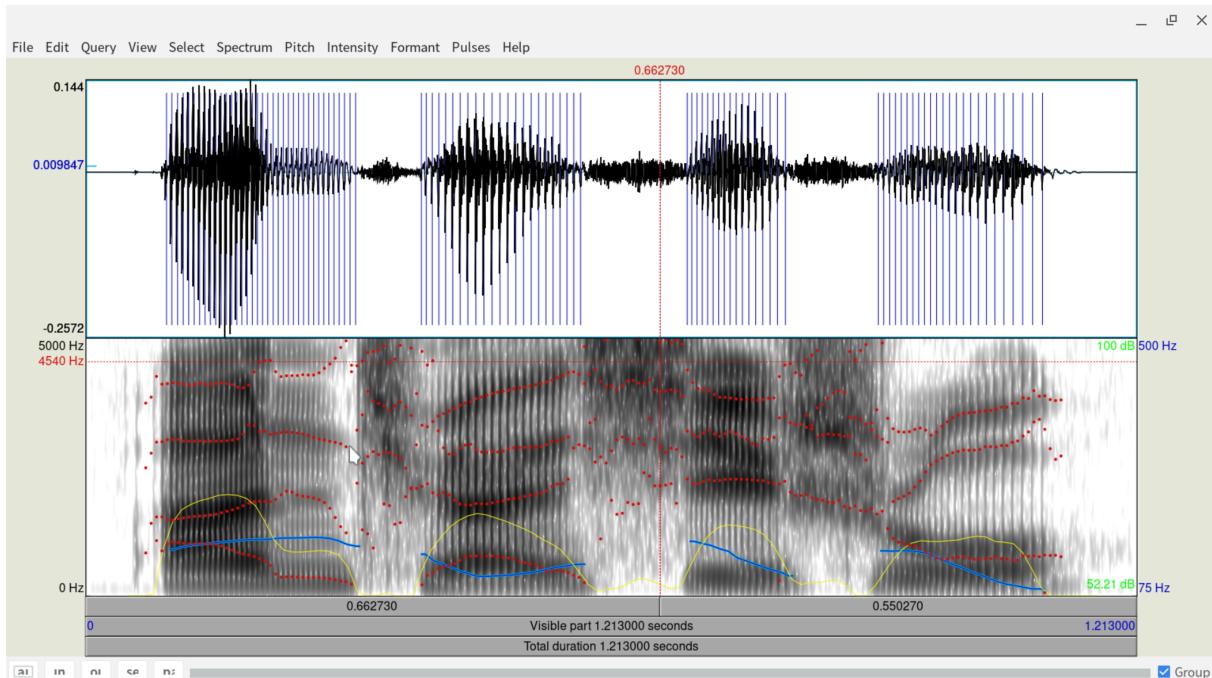


图 1.1 Praat SoundEditor 界面概览

图 1.1 展示了一个典型的 SoundEditor 界面，该界面的图形与对应的项目有：

- 波形 waveform:** 界面上半部分中黑色的内容就是波形图，可以在右下角对波形图进行缩小放大以及移动查看范围。横轴为时间，纵轴为音强。左侧上下两个黑色数字展示的数值区间。
- 语谱图 spectrogram:** 语谱图（频谱图）为界面下半部分的黑色内容。横轴跟波形图一样是时间，纵轴是频率。左侧上下两个数字展示的显示的频率区间的上下界（单位：Hz），一般下界是 0 Hz，上界是 5000 Hz。语谱图中黑色内容的灰度代表对应频率的能量密度（振幅）大小，颜色越深代表能量越大。

3. 音强 **intensity**: 下半部分中黄色曲线就代表音强，右侧绿色字体就是刻度，单位为 dB。
4. 基音轮廓 **pitch contour**: 下半部分中蓝色的点（线条）表示基音频率 F_0 ，右侧蓝色字体表示刻度，对于清音（白噪声）是没有基音频率的。男性的 F_0 一般在 [80, 200] 之间，而女性和儿童在 [150, 250] 和 [200, 500]。不过有研究表明不同声带的人之间 F_0 范围变化比性别之间的差距更大。
5. 共振峰 **formant**: 下半部分红色的点（线条）表示共振峰，默认显示五个共振峰： F_1, F_2, F_3, F_4, F_5 。
6. 脉冲 **pulses**: 上半部分的蓝色竖线就是（声门）脉冲。

1.2 比较和解释不同设置参数

1.2.1 语谱图

在语谱图的设置中，我们可以设置查看频率范围（View of range），这可以改变语谱图频率的显示范围和密度。窗口长度（Window length）用于指定傅立叶变换的时间窗口大小，长度越大的窗口的频率分辨率（不同频带的区分度）会越大，不过其时间分辨率就会减小。

Dynamic range 长度用于将语谱图中振幅低于最高振幅超过 Dynamic range 的频率用白色表示，这是一个控制语谱图中灰色部分比例的阈值。

在语谱图的高级设置中，我们可以选择不同的窗函数（Window shape）：Gaussian, Square (none, rectangular), Hamming (raised sine-squared), Bartlett (triangular), Welch (parabolic), and Hanning (sine-squared)，因为语音信号只具有短时平稳的周期性，所以我们一般对语音信号加窗截取短时信号并对其做傅立叶变换。加窗处理等于对语音特征进行低通滤波，窗函数的不同选择对应的截止频率也不同，并且不同的窗函数会导致傅立叶变换的语谱图不一样。

1.2.2 音强

音强衡量声音的大小（响度），由声音振幅 (Amplitude) 及人离声源的距离决定。在音强设置中，查看范围（view of range）控制音强（黄线）的垂直方向的刻度范围，默认是 50 dB 到 100 dB。

如果我们拖动鼠标选择一定区域的语音信号，则音强曲线右侧或者左侧的绿色数字表示该区域的平均音强。我么可以设置其采用的平均算法：median, mean energy, mean sones, mean dB.

1.2.3 基音轮廓

基音衡量声音（谐波）的基频 (F_0) 大小，基音控制着声调和语调，是语音韵律有关的一种重要特征。

在 Praat 中，我们可以设置基音范围 (Pitch range)，单位 (Hz, mel, ERB 等)。如果我们想要查看基音范围比较大的婴儿发出的声音，我们可能需要将范围调整到 [40, 2000]。基音分析的分析窗口大小一般是 3 倍基音周期。所以下界的选择也是一种权衡：如果设置太小，会丢失部分快速变化的 F_0 (因为时间分辨率下降了)，如果设置太大，则可能会丢失很小的 F_0 。

此外，还有许多高级设置：

1. **静默阈值 (Silence threshold)**：该阈值用于控制 Praat 通过振幅判断是否为声音。如果 Praat 找不到任何的音强，可以试试调整这个。
2. **浊音阈值 (Voicing threshold)**：用于判断是否是浊音，如果 Praat 将清音识别为了浊音，可以试试提高改阈值，反之亦然。
3. **八度跳跃成本 (Octave Jump cost)**：该值影响 Praat 关于 F_0 的波动是否正常的决策。大的值表示不希望 F_0 出现唐突的变化。
4. **浊音/清音成本 (Voiced/unvoiced cost)**：增大这个值将会使 Praat 更加倾向于减少清音浊音状态之间的切换。如果 Praat 分析出来的 F_0 曲线中间断部分（被识别为清音的部分）过多，可以试试增加这个值。

1.2.4 共振峰

共振峰是在声音的频谱中能量相对集中的一些区域，共振峰不但是音质的决定因素，而且反映了声道（共振腔）的物理特征。一般 F_1, F_2 就可以很好区分不同内容了。

在 Praat 中我们可以设置其在 SoundEditor 下半部分绘图的红点大小 (dot size)。还可以通过设置共振峰数量的分析数量 (number of formants)，默认是 5，即分析并展示五条共振峰 F_1, F_2, F_3, F_4, F_5 。该值是 0.5 的任意正整数倍。

因为共振峰也是通过窗函数分析估计出来的，所以我们调整窗口大小 (window length)。注意实际的分析窗口是该窗口大小的两倍程度，因为 Praat 采用的是类 Gaussian 分析窗口， -3 dB 的频带为 $\frac{2\sqrt{6 \ln(2)}}{\pi N}$ 。

1.2.5 脉冲

在 SoundEditor 上半部分的蓝色竖线就是脉冲，表示声门闭合 (glottal closure)。两根连续的蓝色竖线表示一个 (声门) 周期 (period)。抖动 (Jitter) 衡量的就是一

段时间内连续两个周期之间差值的波动情况，它有很多种衡量方式，例如局部抖动定义为连续周期差值的绝对值的平均再除以平均周期。抖动经常被用作度量声音质量。而闪动（Shimmer）类似于抖动，不过是衡量一段时间内两个周期之间的振幅的变换情况。

脉冲有关的设置有下面两个：

1. **最大周期因子（Maximum period factor）**：如果两个连续周期的比值超过该因子，那么在计算抖动的时候这两个周期将会被抛弃。在大多数情况下默认值就是最好的。
2. **最大振幅因子（Maximum amplitude factor）**：如果两个连续周期的振幅的比值超过该因子，那么计算闪动的时候就会抛弃这两个周期。

1.3 解释 Praat 提取声学参数的原理

1.3.1 音强

首先音频信号的值被平方，然后与高斯分析窗函数进行卷积。该窗函数的有效持续时间为 $3.2/p$, 其中 p 为最小音高，这保证了被分析的周期信号具有音高同步（pitch-synchronous）的音强，波动范围不超过 10^{-5} dB。

1.3.2 音高

Praat 使用一种精确自修正算法^[1] 来计算 F_0 ，这种方法比基于 cepstrum 或 combs 的方法或原始的自相关方法更精确、抗噪、鲁棒性更好。其核心思想就是将加窗后的信号上的自修正函数除以关于该窗函数的自修正函数：

$$r_x(\tau) \approx \frac{r_{xw}(\tau)}{r_w(\tau)} \quad (1-1)$$

1.3.3 语谱图

因为语音信号具有短时平滑的特点，基于该假设，Praat 采用短时傅立叶变换（short-term spectral analysis, STFT）从语音信号得到其语谱图，具体来说：

1. **短时加窗处理**：采用合理大小的窗函数，可以通俗理解为该窗函数范围内的信号是具有比较明显的周期性的信号，这样就可以方便的进行（离散）傅立叶变换。常见的窗函数有：矩形窗（窗口范围内为 1, 其他为 0, Hamming 窗（由窗中心的 1 逐渐按照余弦递减到窗边界的 0）。
2. **傅立叶变换**：将（具有周期性的）时域信号变换为频域信号。

短时傅立叶变换的公式可以表示为：

$$\text{SFTF}(x(n)) := \sum_{m=-\infty}^{\infty} x(m)w(n-m) \exp(-jwm) \quad (1-2)$$

其中时域信号和窗函数分别为 $x(\cdot), w(\cdot)$ 。

1.4 共振峰和频谱图的关系

声音从声带发出时形成的声门波是标准的谐波信号，基频就是音高 F_0 。而后声门波警告过声道，声道作为一种谐振腔，相当于一个滤波器，不同声音发声时，声道中的发音器官（舌头、软腭等）处于不同的位置，而谐振腔（咽腔、口腔、鼻腔）形成不同的形状，造成不同声音具有不同的谐振特性。也就是说，谐波信号经过声道后，有些频率区域能量被大，而有些区域能量变小。共振峰是在声音的频谱中能量相对集中的一些区域，共振峰也就是被声道特别放大的频带，与声道的谐振频率相关。

要想获得频谱切片，只需要先用鼠标点击或者拖动选中目标时间点或者时间段，然后在菜单中找到 **Spectrum → View spectral slice** 即可。

1.5 脉冲

脉冲指的是声门脉冲，每一条蓝色竖线就代表一次声门闭合。脉冲的开始在波形图中表现为局部最低点。

脉冲直接反映的就是音高轮廓，也就是基音频率 F_0 。有了基频，我们就可以得到谐波。通过脉冲可以计算出抖动（Jitter）和闪动（Shimmer），进而用于计算谐波噪声比（Harmonics-to-Noise Ratio, HNR）。谐波噪声比可以用于衡量声音质量。

第 2 章 任务二：发音与听觉感知

2.1 语音的谐波 harmonics

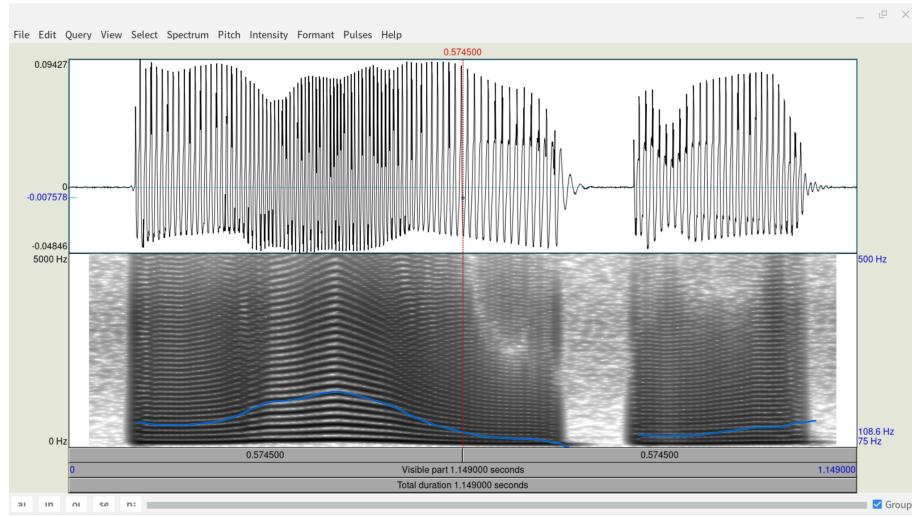


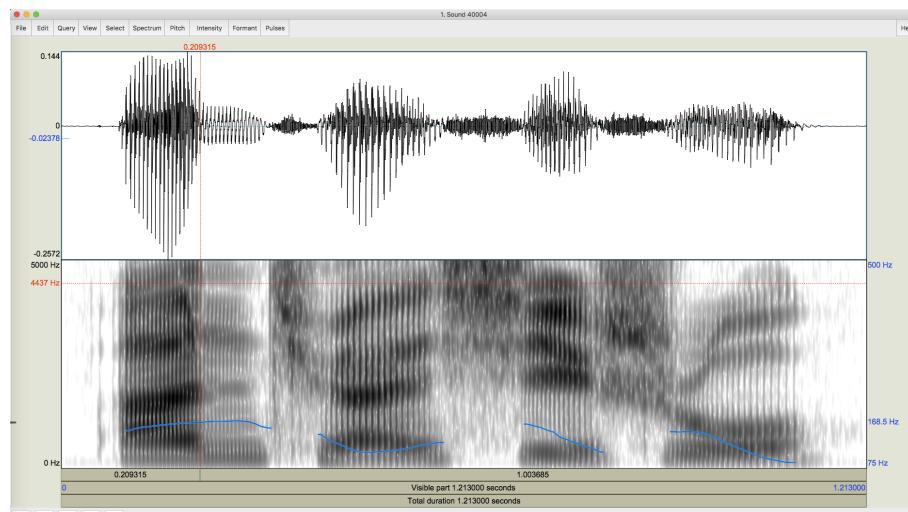
图 2.1 谐波

从波形图中我们很难看出语音的谐波分量，不过我们可以通过语谱图轻松地看出语音的谐波及其能量。不过 Praat 默认的参数设置 ($N = 5 \text{ ms}$ 的宽带语谱图) 使得我们依然不能轻松地从语谱图中找出各个谐波分量。我们只需要更改语谱图设置，把窗长调整到 30 ms ，这时候，我们会发现语谱图中有非常多的细条纹。当我们固定到某个时间刻度 t 时，我们可以看出其频谱离散的，且是等距的，这就是谐波信号。从频率最低的基频 F_0 开始，所有谐波分量都是基频的整数倍。并且谐波分量随着频率的增大，能量是逐渐减小的。

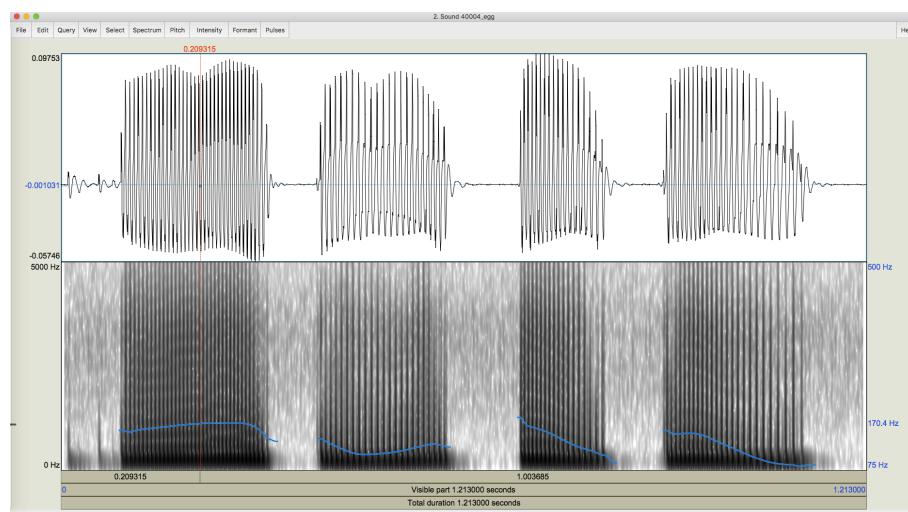
2.2 比较 EGG 和语音信号波形

在图 2.2 中展示了语音信号波形和 EGG 信号波形的示例。首先，EGG 和语音信号的时间长度是一样的，并且他们的清音和浊音的区域是一致的。其次，音高（基频 F_0 ）是一致的，虽然从波形图中无法直接看出来。

不同点在于，语音信号是 EGG 经过声道调制 (Filter) 后的输出。EGG 信号在同一段浊音内表现出来的波形是比较规整的谐波信号。具有较明显的周期性，也就是说其振幅的局部最大值的抖动较小。并且对于清音部分，其能量很小，而且波动幅度也很小，说明噪声比较微弱。其语谱图中的能量也是从基频到谐波分量，



(a) 语音信号波形



(b) EGG 信号波形

图 2.2 比较 EGG 和语音信号波形

随着频率增加，能量逐渐下降。

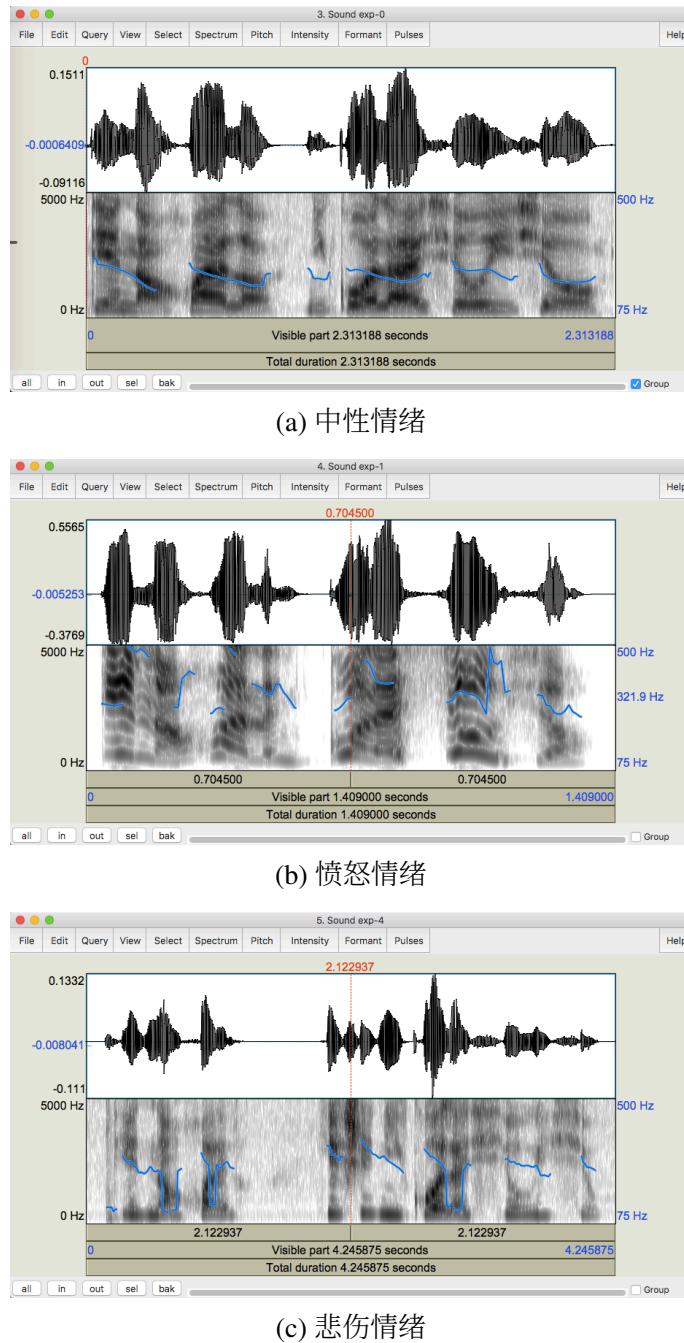


图 2.3 不同情绪的基音轮廓

而语音信号表现为其抖动更加明显，并且在大时间尺度上没有明显的周期性，而在很小一段时间内其信号具有比较明显的周期性。并且对于清音部分，可以明显看出其能量和波动都比 EGG 明显，说明噪声比较大。从语谱图中我们可以看到，各个频率的能量分布与频率大小关系不大，不过具有比较明显的峰值特性，也就是说，语谱图中某一部分区域的能量比较大，而其他地方能量比较小，看起来有点

像山峰山谷。

2.3 不同情绪的基音轮廓

可以从图2.3中看出：

1. 中性情绪的基音轮廓比较平滑，变化幅度不会很大。浊音和清音（没有基音轮廓的部分）的识别较好，间断点较少。
2. 愤怒情绪的基音轮廓抖动比较大，有些部分浊音被错误识别为了清音，所以基音轮廓有更加多的间断点。同时，可以看出，其平均的基音频率都要高于中性情绪，基音轮廓中有许多上升的部分。听该段音频的时候明显感觉声音起伏落差明显，而且很急促。
3. 悲伤情绪的基音轮廓抖动同样也比较大，也有一部分浊音被错误识别为清音的情况，不过间断点略少于愤怒情绪。同样的，其平均基音频率要低于中性情绪，其基音轮廓中有许多下降的部分。听的时候也明显感觉声音要低沉很多。

从图中我们还可以看出，除了三种情绪的语音信号的基音轮廓不一样之外，他们的语谱图也有着明显的不一样。也就是说这三种语音在发音的时候，他们不仅声门波不一样，他们的声道对语音信号的调制作用也是不一样的。语谱图不一样同样也代表了共振峰是不一样的。此外，因为基音轮廓不一样，所以脉冲是不一样的，并且由基频决定的谐波也是不一样的。基本上，这三种语音信号，除了清音浊音的分布差不多（因为说的内容都是一样的几个字），其他许多声学特征都有明显差别，比如它们之间的振幅（能量）的差别很挺大的。不过因为语速会有区别，所以它们的浊音持续时间也会有所区别。

2.4 不同语气的基音轮廓

因为不同情绪的声音同样语气也会有区别，所以不同语气的基音轮廓的区别也类似于不同情绪。参考图2.4，具体来说有以下一些差异：

1. **陈述语气**：对每一个浊音单元，我们可以看出基音轮廓都是下降的，属于降调语气（忽略第一个浊音单元后面突然上升的那一部分，那部分是识别错误）。
2. **疑问语气**：对前半部分的浊音单元，我们可以看出基音轮廓都是下降的，因为前面几个字基本上是陈述语气。而后半部分的浊音单元，尤其是最后一个字——书——的时候，它的基音轮廓是明显的升调的，因为此时是疑问语气。

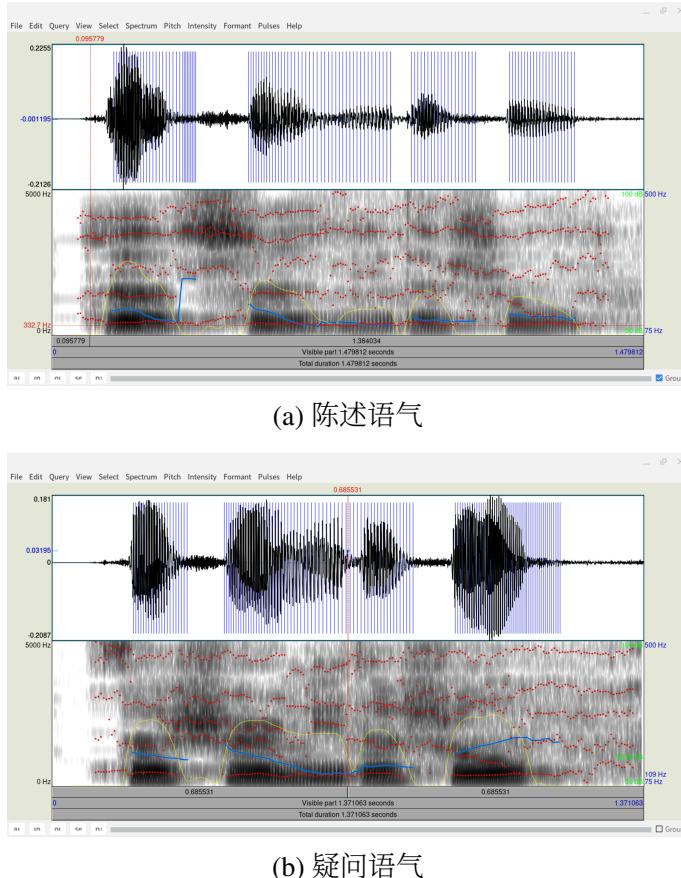


图 2.4 不同语气的基音轮廓

2.5 宽带语谱图和窄带语谱图

参考 Praat 官方手册^①中对宽带语谱图和窄带语谱图的介绍，我们分别设置窗长度为 5 ms 和 30 ms。宽带语谱图和窄带语谱图的对比在图 2.5 中展示。两者的差异可概括为：

- 宽带语谱图：**采用较小的窗口（以及较小的窗移动），这样得出的语谱图的频率分辨率较小，不过时间分辨率较大。频率分辨率较小意味着频带较宽，频率间的区分度较小，具体来说就是语谱图中的条纹之间的分隔不是很明显。不过时间分辨率大，说明不同时间之间的频谱变化可以更加容易被发现。
- 窄带语谱图：**采用较大的窗口（以及较大的窗移动），得到的语谱图的频率分辨率较大，而时间分辨率较小。从图中可以看出窄带语谱图的条纹（频率的能量集中的区域）效果很明显，很容易区分。

可能从 $N = 30$ ms 的窄带语谱图中对于时间分辨率的感知还不够明显，加入我们把窗长调整为 100 ms，这时候从图 2.6 中就可以明显感受到，时间方向上，许

^① https://www.fon.hum.uva.nl/praat/manual/Intro_3_2__Configuring_the_spectrogram.html

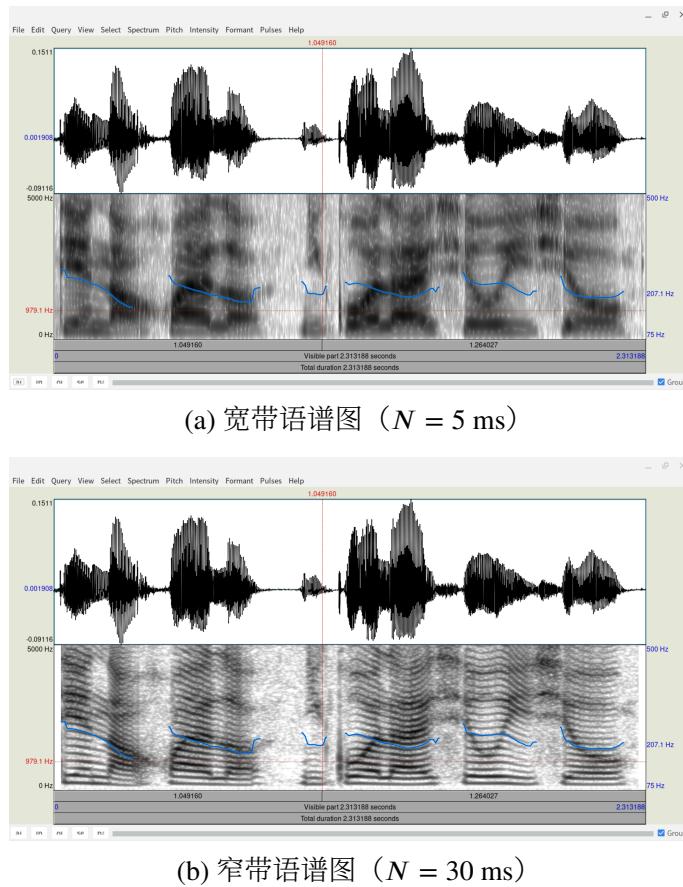


图 2.5 宽带语谱图和窄带语谱图

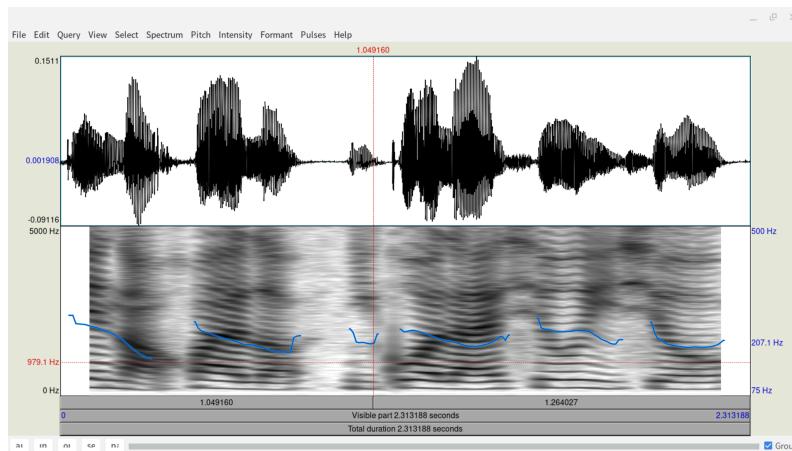


图 2.6 一个窄带语谱图的例子

多频谱变换已经不见了，也就是说过于平滑了以至于丢失了许多细节。

2.6 掩蔽效应

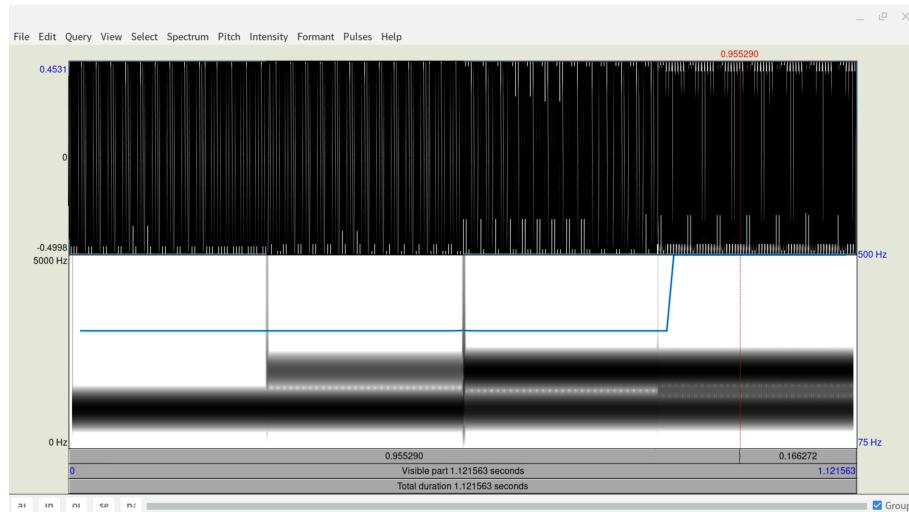


图 2.7 掩蔽效应

掩蔽效应是一种心理学现象，是由人耳对声音频率分辨机制决定的。是指一个较强声音的附件，相对较弱的声音不易被人耳察觉，即被强音所掩蔽。

图 2.7 展示了一段用以说明掩蔽效应的声音的波形图和语谱图。该音频共有四个阶段，第一阶段只有一个频率的声音；第二阶段有两个频率的声音，不过低频能量更大；第三阶段有两个频率，并且他们的能量差不多一样；第四个阶段有三个频率的声音，最低频和最高频的能量相同且较大，而中间频率的能量较小。对于这四个阶段，我们人耳只能感受出两种声音，也就是第一阶段和第二阶段一样的声音，第三阶段和第四阶段一样的声音。这是因为低能量的频率被高能量（强音）所掩蔽，这被称为**同时掩蔽**。同样的，在时间上相邻的声音之间也有掩蔽现象，被称为**异时掩蔽**。

参考文献

- [1] Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound[C]// IFA Proceedings 17. 1993: 97-110.