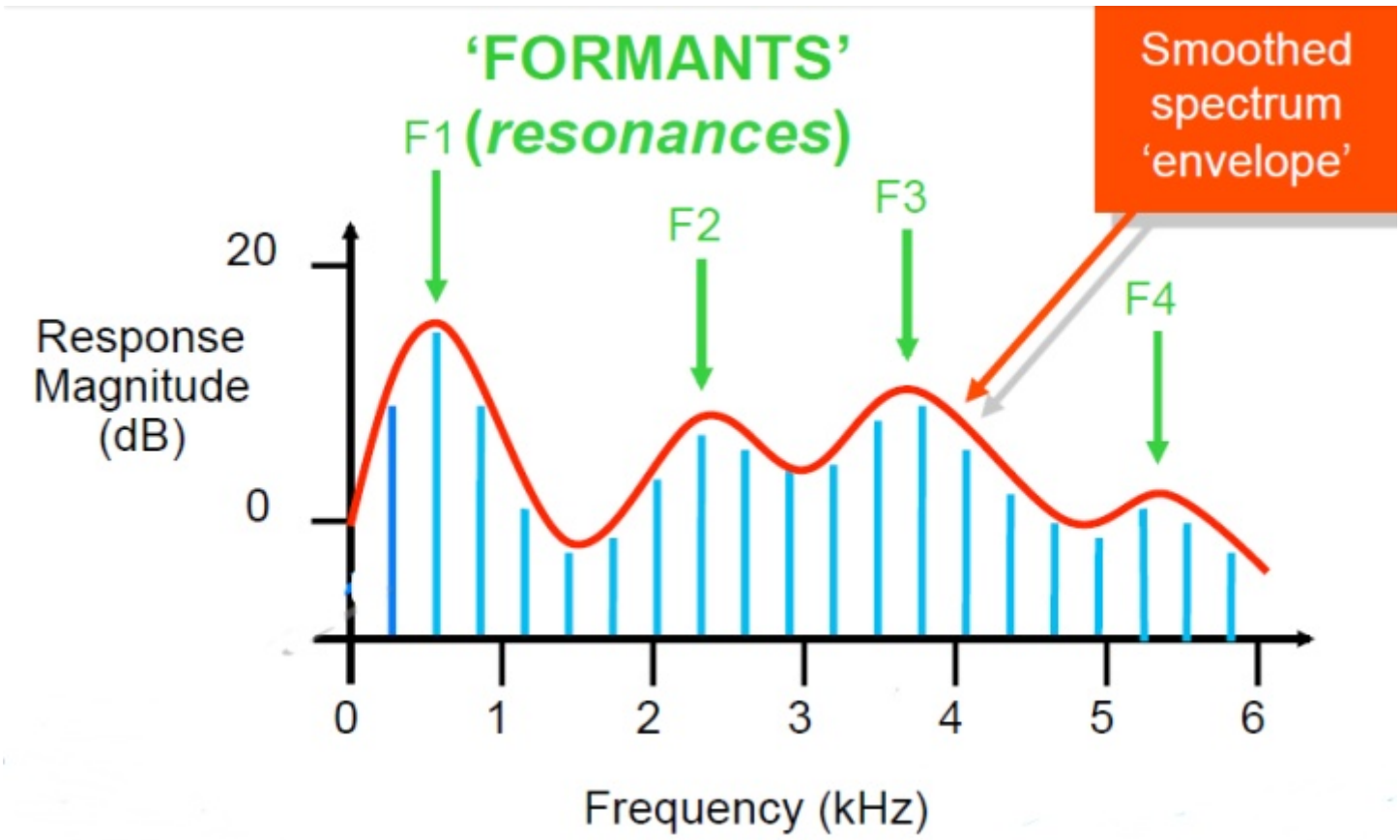
我们检测到你可能使用了 AdBlock 或 Adblock Plus，它的部分策略可能会影响到正常功能的使用（如关注）。
你可以设定特殊规则或将知乎加入白名单，以便我们更好地提供服务。（为什么？）



Pure Data教程三

Patchouli ...
移动の大图书馆

+ 关注他

11 人赞同了该文章

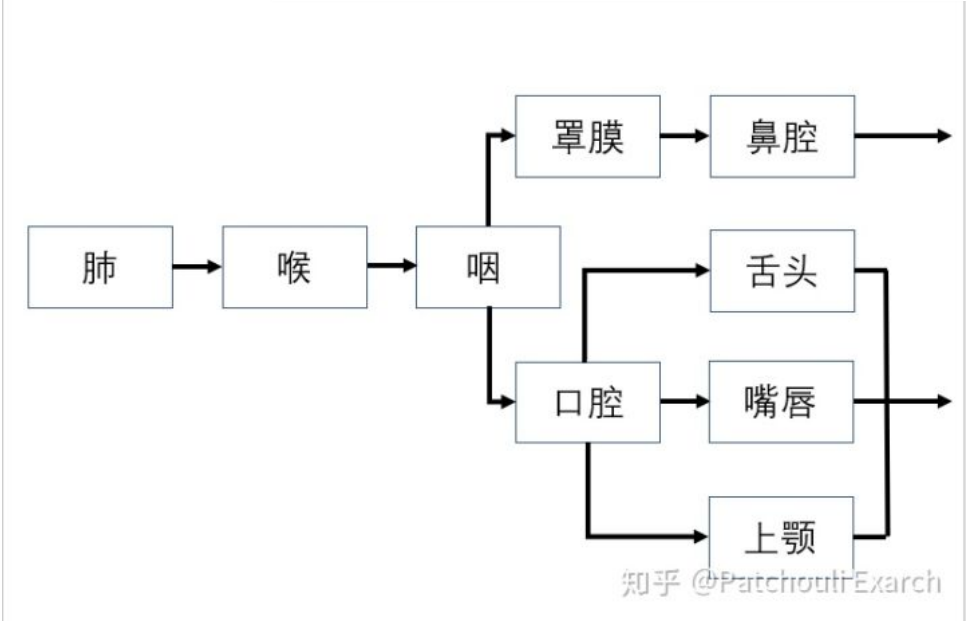
Patchouli Exarch：Pure Data教程二
zhuanlan.zhihu.com

在上一期教程中，我们讲了声学的基本原理，这期教程我们来讲一讲，说话的基本原理。

人类依靠嘴部和喉部发声，这是生物进化史上的奇迹，生物把发声器官和进食器官合并到了一起，节省了大量资源，我们的嘴喉系统不仅要负责进食，还要负责发声来和别的生物交流。但是这样的坏处也很明显，即两个功能之间会互相干扰，比如人把嗓子喊哑了的话进食就会有不舒适的感觉，亦或是需要以声音谋生的人为了保护嗓子很多东西都不能吃。比如唱曲的老师为了保护嗓子，大米拿剪子把尖儿铰了再往下吃，后来老师就饿死了；就不像捧哏的于老师，馒头俩俩往嗓子眼里扔还谁也挨不着谁。

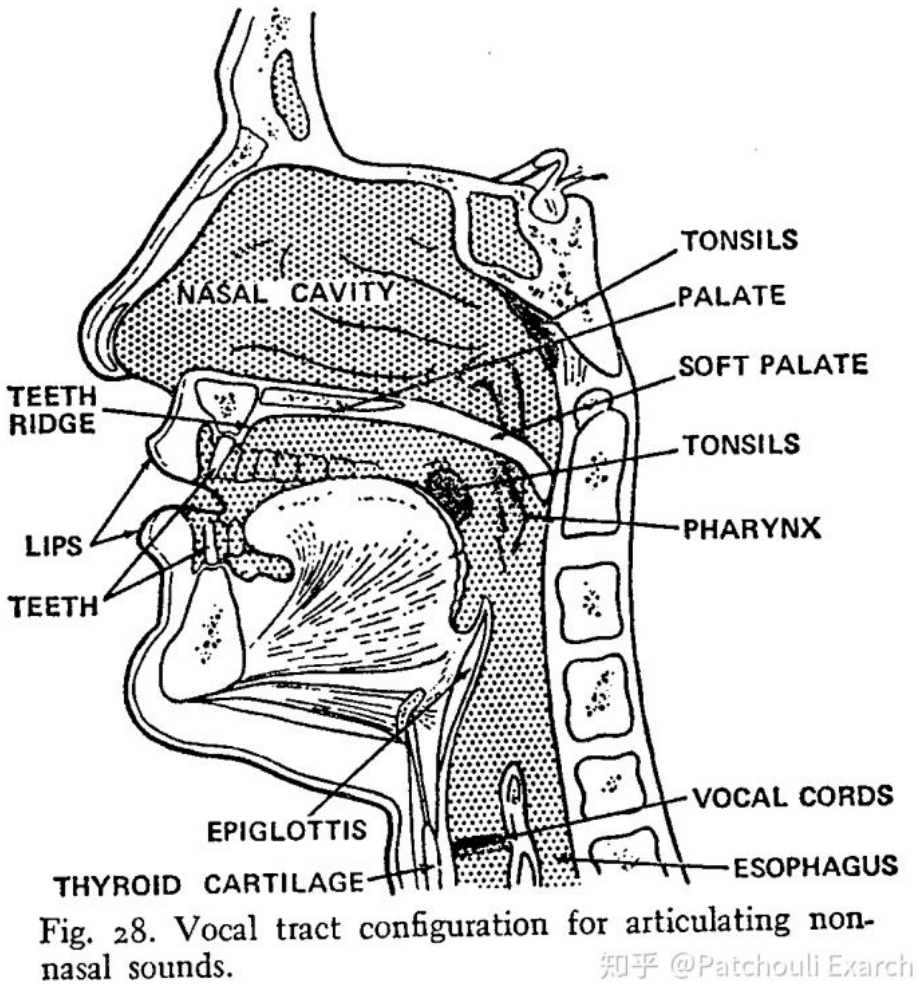
闲话不多说书归正传，我下面将用一张流程图来展示人的发声过程：

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏



人的发声过程

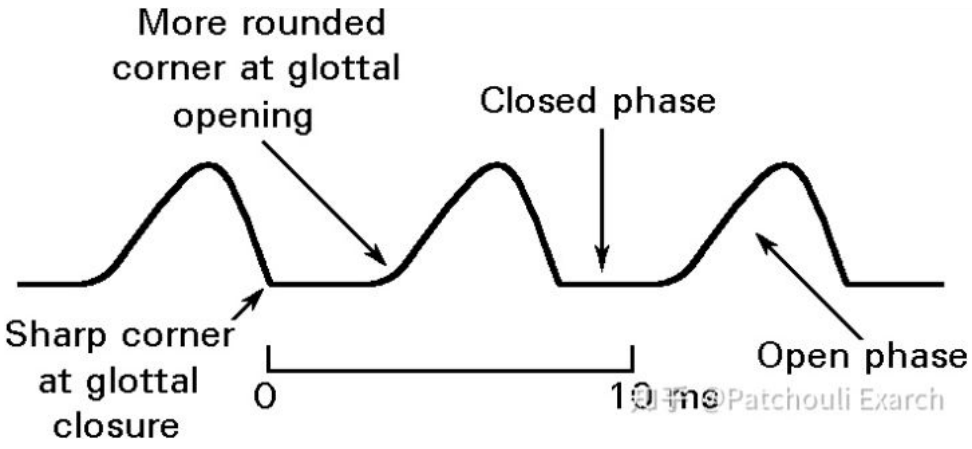
其中肺是声波的主要能量来源，能量的大小由肺活量和呼气率决定，一个瘘了的人是无论如何无法做到有底气地说话的。喉是声音的主要来源，喉咙发出的声音是包含许多频率的谐波，这个谐波的基频由人的体型决定（就是上一讲中提到的共振管）。而舌头是最重要的一环，他是提升语音清晰度的部件，所以人能不能说清楚话舌头很关键。



The Speech Chain, Denes.P.B,1973

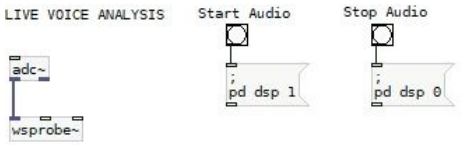
来自肺部的气压在封闭的声带下面聚积，声带反复折叠分开再折叠产生小的空气脉冲。所谓的发声，其实就是人类对气流的调制。附着在声带上的肌肉的力量决定了声带的振动速度，从而决定了每个人说话的基频（F0）。
频率时需要通过开闭声门来

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ···

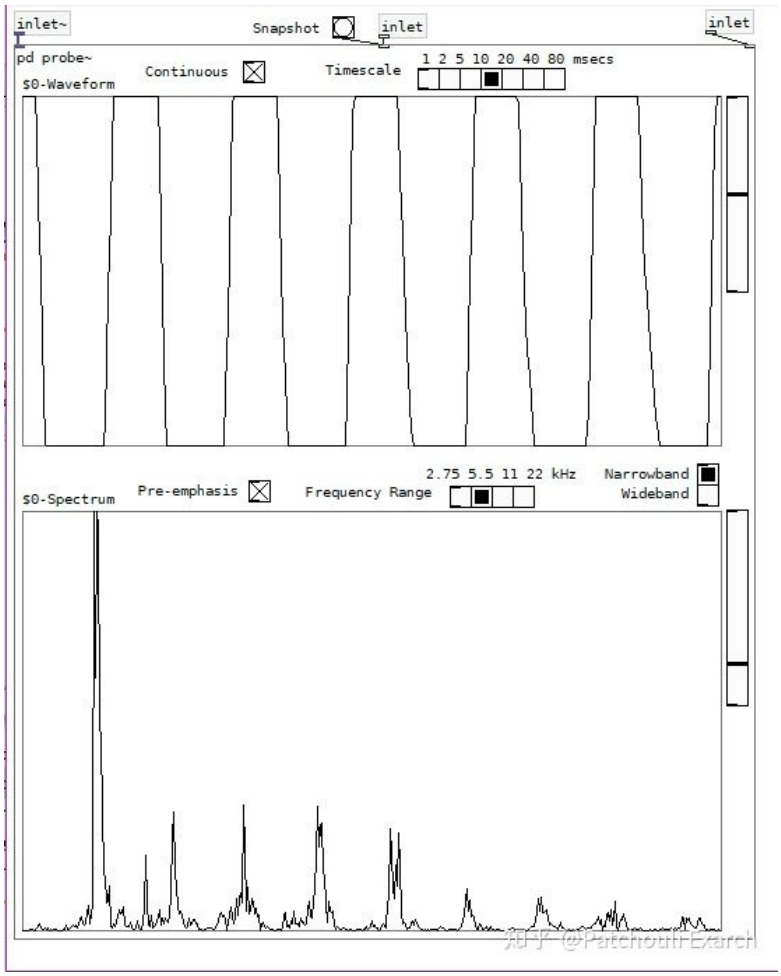


SpeechSynthesis and Recognition, Holmes.W, 2002

那么我们应该怎样用Pure data来观测人声的基频和谐振呢?

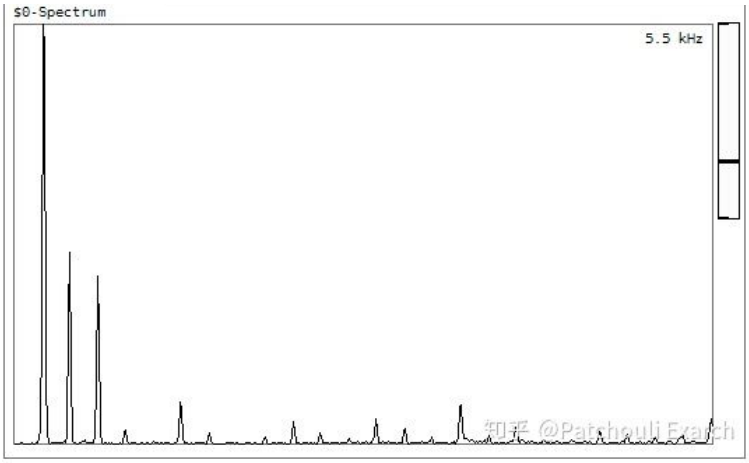


我们将Pure data做如上配置，然后右键点击wsprobe~选择内容，就会出现以下展示界面。



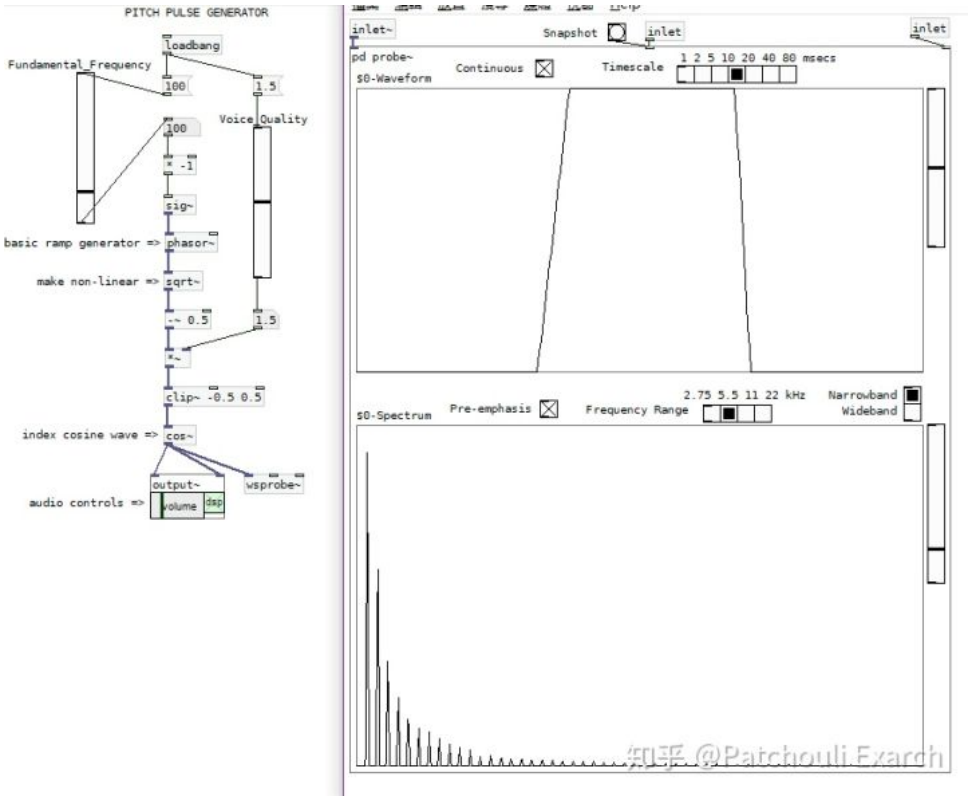
下方的频谱就是我用我的嗓子尽可能地唱一个纯粹的高音出来，最终频谱就会呈这样的形态分布，你会发现每个振幅峰的间距是差不多的。当然了，我不是专业的歌唱表演艺术家，我甚至不知道声乐有几种唱法，所以出来的频谱就这么糙，假如让专业人士来唱呢估计频谱就是这样的:

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ...



那么我们如何来用Pure data模拟出我们的基频和谐波呢?

将Pure data如此配置就能生成最清澈标准的嗓音:

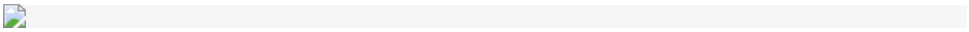


你们不要觉得这个东西没有卵用，事实上许多不能说话的人就是靠这个技术说的话。

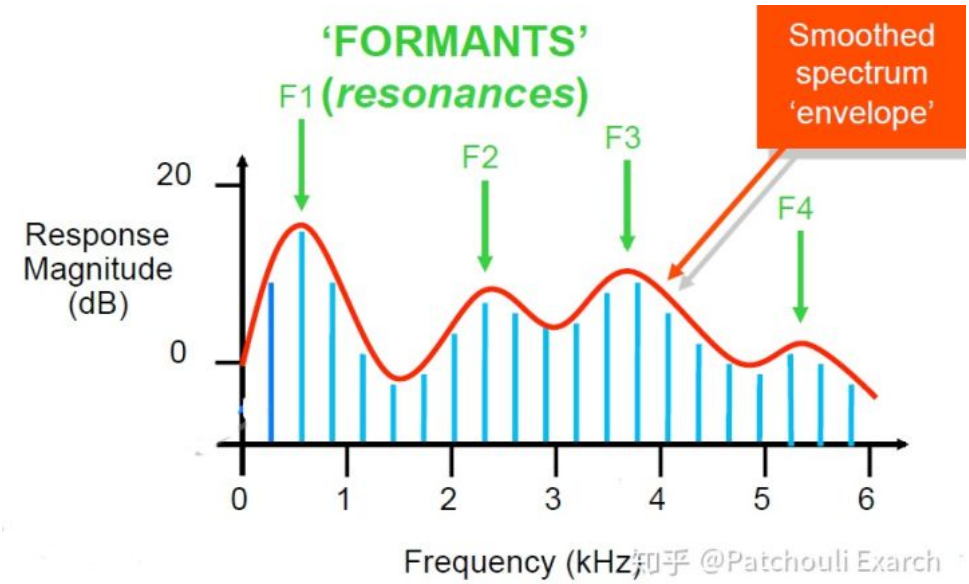
这个看起来像飞机杯一样的东西，可以被失语的人顶在喉咙上用来发声，他会检测你喉咙处的基频，然后帮你拟合出谐波以达到说话的目的，虽然很不清晰，但是已经能听懂了，这就是上面这个代码对这个社会的贡献。

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ...

上一讲中，我们收尾在了滤波器上，其实我们人的整条声道就是个滤波器，如第一张流程图所示，声音从喉出来的一瞬间，就要通过包括喉咙本身在内的各道滤波器。其中声门和嘴唇是主要的滤波器，用共振的特性声门发出来的复杂声音变成谐波并且抑制其他声音，而舌头、上颚，牙齿，软腭等等部分则是次要滤波器，功能是强调或者削弱某些特定倍数的频率来增强语音的辨析度。一份纯粹的声音从声门出来后应该是这个样子的：



经过声道的滤波处理之后就会变成这样：



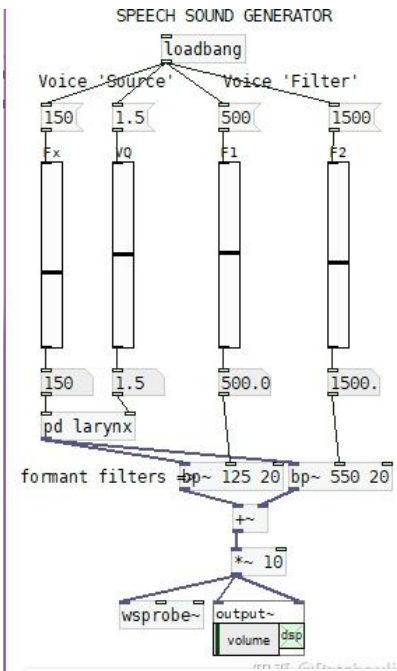
Formants

F1~4都是会被系统强调的音，是语音中主要用来辨识内容的频率点。

关于人的声道是怎么滤波的，我这里给大家推荐一个网站：Pink Trombone，这个网站可以通过调节声道中各个器官的位置来观察人在这样的声道配置下会发出什么样的声音。

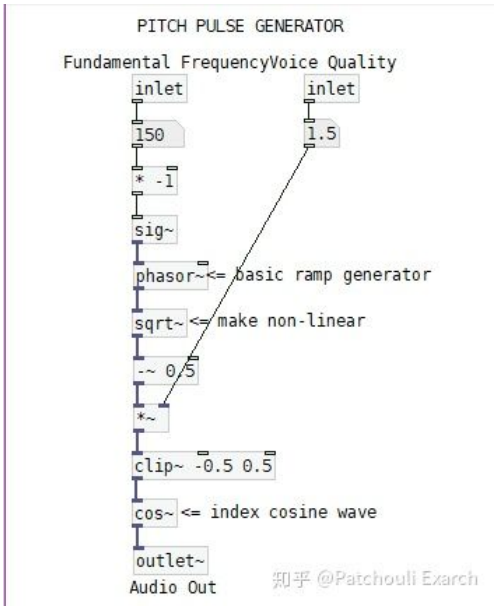
教我语音识别的大叔是个语言天才，会无数语言，而且说什么语言都能以假乱真；究其原因，他在学每门语言的时候，会去研究所有语言中每个发音的声道形态，比如舌头牙齿上颚之间的位置关系等等，以此保证自己发出一个和母语使用者完全一模一样的音，真是神奇的语言学习方式。

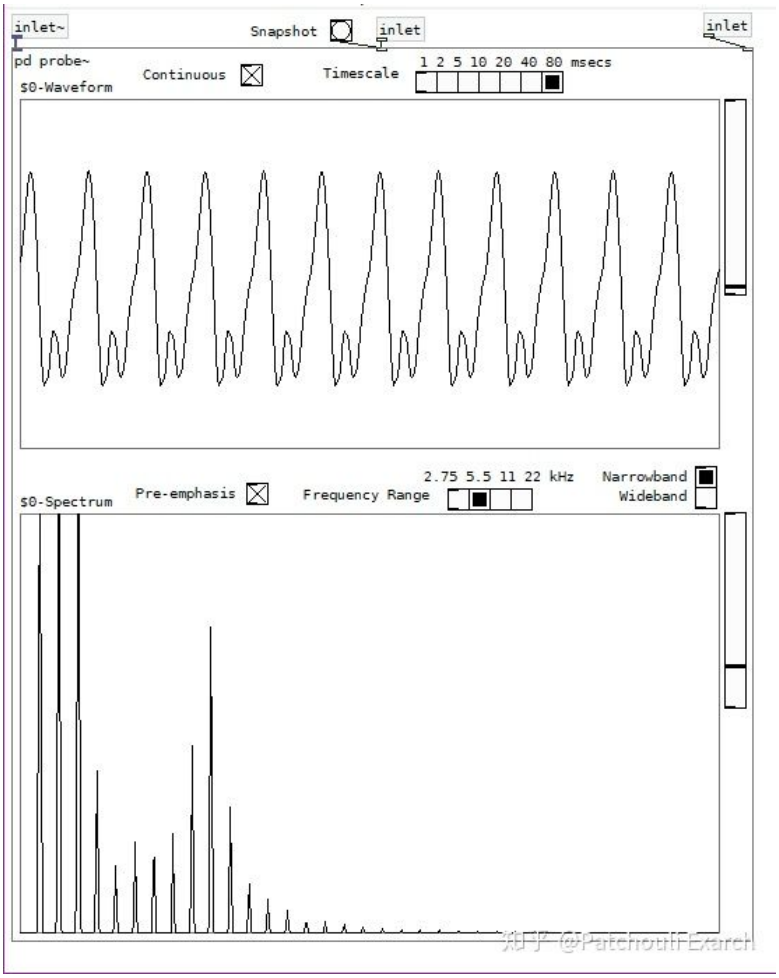
回归正题，我们同样可以设计一个Pure data程序来模拟上面的F1~F4：



▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ···

其中pd larynx是：

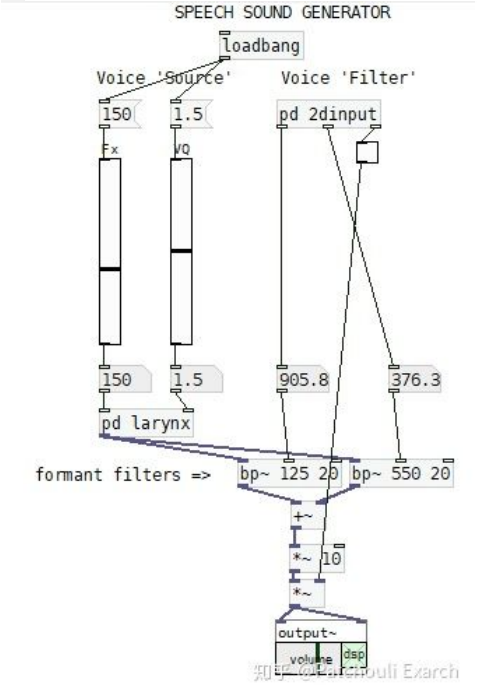




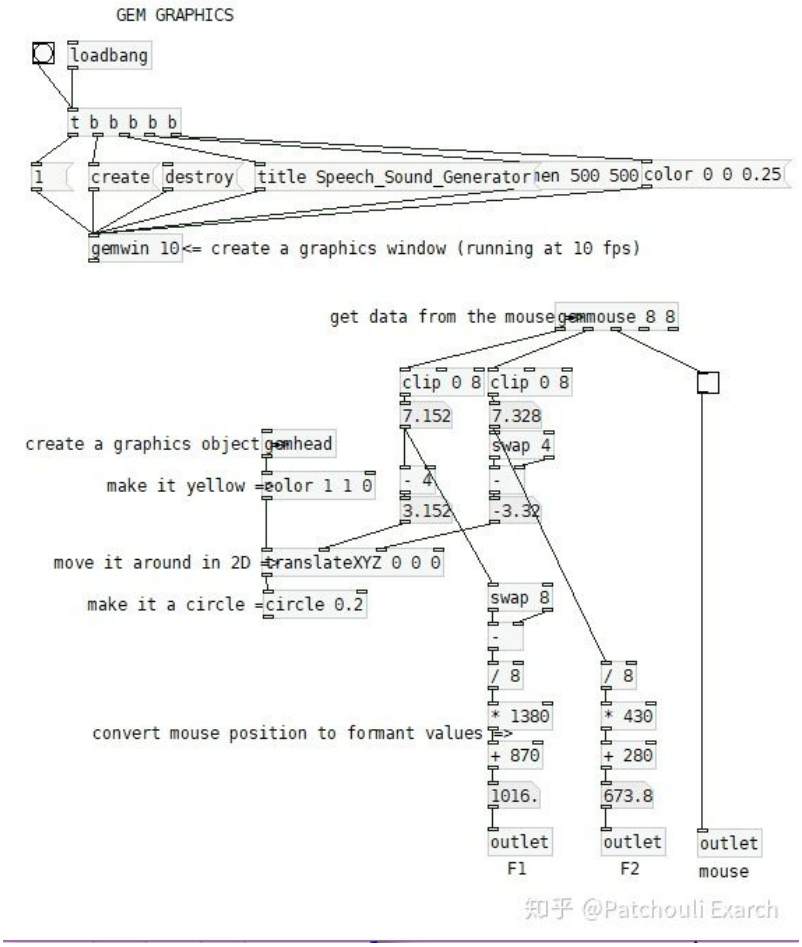
我们能狗非常明显的看到形如图“Formants”中的F1和F2两个被强调的谐波。

我们甚至可以做出一个二维球操控板来听人的不同的谐波被强调时的发声：

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ...



其中pd 2dinput是这样的：



上面的小程序就可以简单模拟人的声道发声了。

但是声音只能从人的声道产生吗?

答案是否定的，我们可以吹哨，可以弹舌头，可以让气流快速的通过牙缝，就像人疼痛的时候会咬着牙倒吸一口凉气，发出嘶嘶的声音，这气流的方向都反了，必是不能算入声道的发声中去。所以其实我们发现，人的发声过程不一样，声音的种类也就不一样，这在我们语音识别中十分重要。简单来讲，我们根据发声方式把声音分成浊音、好音、摩擦、爆破四种声音；浊音是声带震动产生的

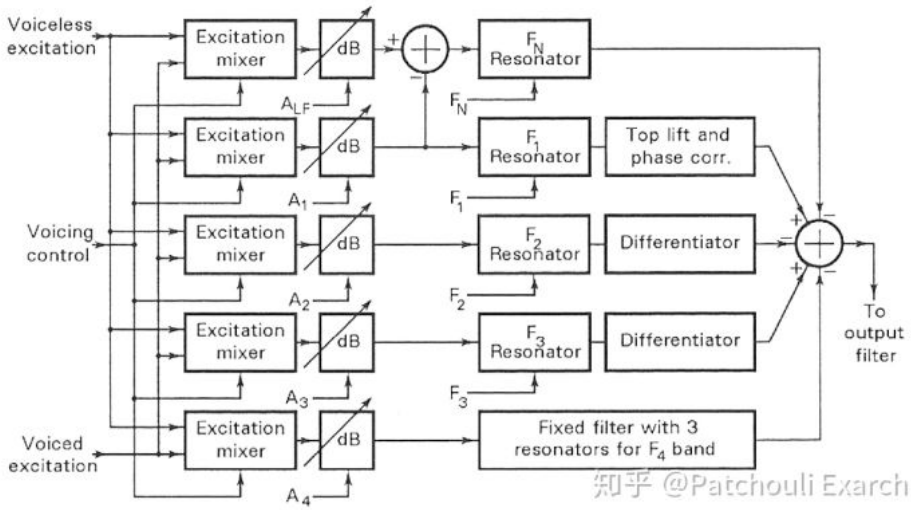
▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ...

声音，与之相对的——轻音则是声带不震动的声音，摩擦音则是由声门（各种声门，包括我刚才说的牙齿）收缩出的湍流引起的，而爆破就是喷口，是突然释放阻滞的气流发出的声音。

在我们分析的时候，一般把人的语音用12个参数标定，当使用12个参数时，帧长一般是10ms。

- FN，低频区，低于250Hz的声音都算250Hz
- ALF，低频区的声压级，也就是音量
- F1，第一共振峰的频率
- A1，第一共振峰的声压级
- F2，第二共振峰的频率
- A2，第二共振峰的声压级
- F3，第三共振峰的频率
- A3，第三共振峰的声压级
- AHF，第四共振峰的声压级，固定在3500Hz，毕竟没人用哨音说话
- V，发声程度
- F0，基频
- MS，声门脉冲标记/空间比（一半是固定的）

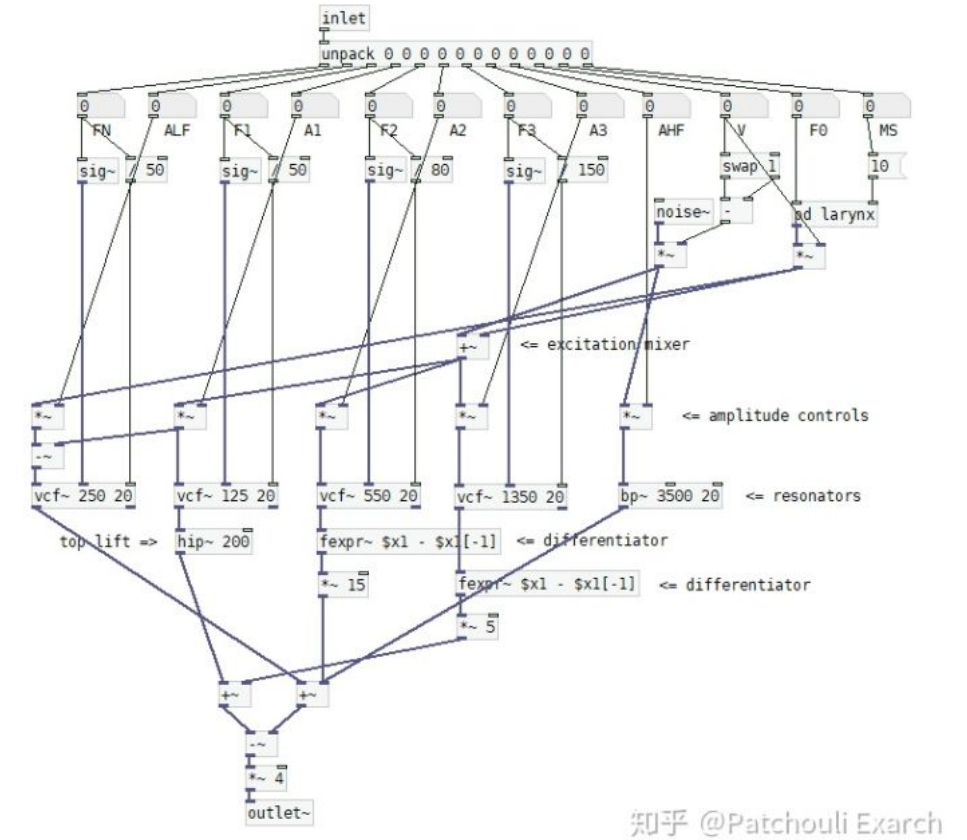
而Holmes先生设计的最早的声音合成器就是根据以上十二个参数完成的：



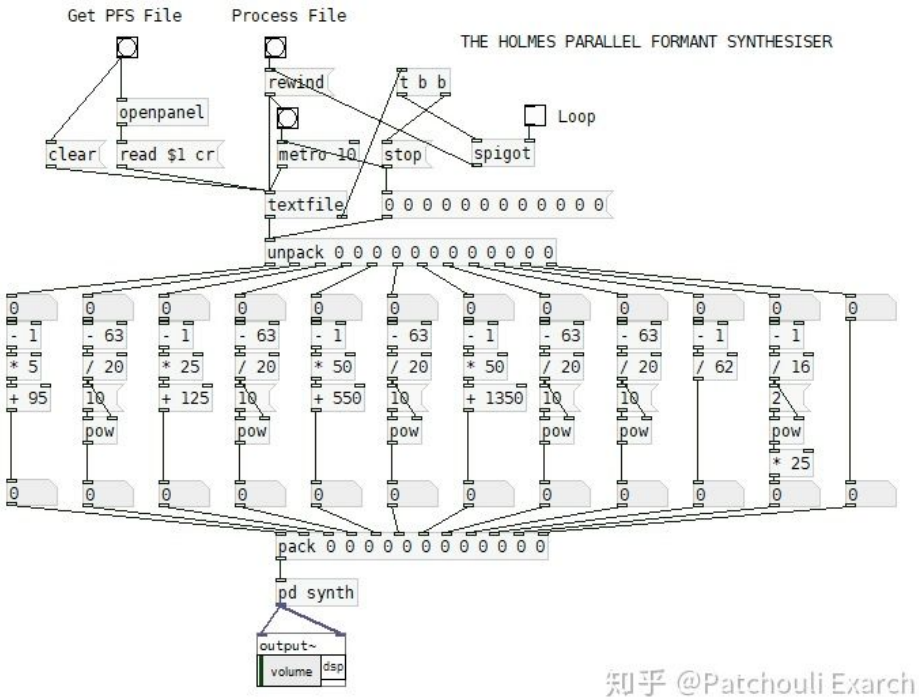
SpeechSynthesis and Recognition, Holmes.W, 2002

我们就完全可以根据Holmes设计的框架做一个我们自己的语音合成器，首先是框架实现：

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ···



然后是把函数运用到实例中去：



先点击Get PFS File，然后再Process File就可以把语音合成出来了，我这里给大家上传了一份PFS文件，大家可以用记事本打开，打开之后大家可以看到里面是密密麻麻的数字。是的，这就是那12个参数，我们用Holmes的方法成功地把复杂的人类语言压缩成了简单的文本文件，还能用语音合成器再重新播放出来。

https://pan.baidu.com/s/1wau-8WZp3TSuGbFFr62rnQ
pan.baidu.com

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ···

以上就是本期教程的全部内容了，主要是教会大家如何做一个简单的语音合成器。

发布于 2018-07-15

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

语音合成 语音学

文章被以下专栏收录

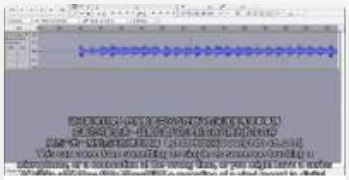
Pure Data入门教程
本专栏旨在提供人工智能方面的教程（其实就是我自己的学习笔记），包括语音文字...

关注专栏

推荐阅读

一、语音处理基础知识

Tux ZZ



Lynda中文 Audacity清理修复音频教程 Cleaning Repairing...

zhenwei009

语音合成中的Mel谱和MFCC谱无区别

语音合成目前比较流行的方案是Tacotron(2) + WaveNet(WaveRNN, LPCNet)等神经网络声码器。这些方案的流程大致相同，先由文本生成特征谱，再将特征谱重建为音频。在选择特征谱的时候，有的使...

木不shi... 发表于语音合成拾...

30个praa 语音学pr

法国猫博士

还没有评论

评论由作者筛选后显示

▲ 赞同 11 ▼ 添加评论 分享 喜欢 收藏 ...