Texture Synthesis: Explore Representations of Mind

http://freemind.pluskid.org/misc/texture-synthesis-explore-representations-of-mind

posted on Free Mind on March 2, 2013
generated with pandoc on December 3, 2015
category: MISC

tags: Fun, Talk

很久很久没有更新了，其实有好多想写的东西，草草的开头也有三个，只是一直非常忙完全没有时间好好把他们写出来。于是写点闲聊的东西来凑凑数吧！:D 其实是上周 Speech Recognition 课上的一个 Guest Lecture 上听到一些好玩的东西。

前来做客的是新近加入 MIT Brain and Cognitive Science faculty 的 Josh McDermott（发现 MIT 有很多 faculty member 都是曾在本校求学的呢），他讲的内容主要是声音的 Texture Synthesis。一般说到 Texture Synthesis 的话，最先想到的都会是 Vision 里的东西吧，那声音的 Texture 究竟又是什么呢？

实际上就连 Vision 里面的 texture 本身也是一个比较难以定义的东西吧？我都不知道中文应该称作什么好，说“纹理”似乎意义太狭隘了。实际上一个“东西”究竟算不算 texture，也完全取决于你观看的视角、尺度之类的各种因素。总之，这种只可意会的东西，我就直接盗用一下他们 Vision 课上的一页 slides 好了：

Figure 1: .....

不过呢，且不论实际中的 texture 究竟是什么，当人们要开始动手研究的时候，就不得不做一些更详细的限制和描述，比如通常要求是“stochastic”、“stationary”等等。在 Vision 里的典型例子就是像地板啊、墙壁啊这种**背景**里的东西。其实到声音里也是类似的，这里主要研究的就是所谓的“Background Sounds”，典型的例子是流水啊、风吹啊之类的，

或者比如酒吧里一堆一堆的人一起说话所形成的闹哄哄的声音也可以算在里面。

然后这里要做的就是人工合成背景音，或者说 sound texture。当然我觉得做 texture synthesis 应该并不是真正的目的，真正的目的其实是理解人脑对声音是采用如何的一种形式进行表达的，而 texture synthesis 只是作为验证理论模型的一种手段。因为从各种方面来说 representation 都是一个更为基本和重要的问题，这一点在之前介绍 Marr 的那本书的时候也已经提到过了。

这里提出的模型是，人脑是以记录一些统计量的方式来实现 sound texture 的 representation 的。首先，声音信号大致会经过如下图所示的一些预处理：

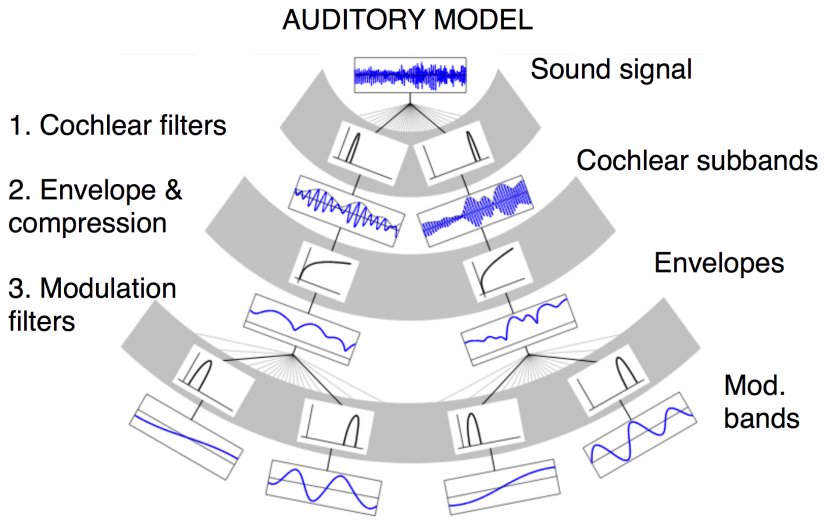

Figure 2: .....

前面的这些机制是经过多年的研究差不多已经成为公认的事实了。不过三言两语也解释不清楚，总之可以看成一个特征抽取的黑盒子，从最初的音频信号得到一堆的特征。然后从得到的这些特征可以进行统计分析，计算一些统计量（诸如一阶矩、二阶矩之类的），motivation 当然是因为 texture 本身就有 stochastic 和 stationary 的特性，所以采用统计量作为 representation，一方面可以允许同一类型的 texture（例如都是流水的声音）有不同的 instance，另一方面也能区分不同的 texture 类别。

不过，虽然这种方法看起来还是蛮自然的，但这却不能直接作为人脑就是使用这种 representation 的证据。证明的方法是：如果人脑确实是使用这种 representaion 的话，那么如果我 synthesis 出来不一样的随机的声音片段，但是如果保证它的这些统计量都和某一个 texture 类别匹配上的话，人将会无法区别出 “真正的” 自然产生的 sound texture 和合成的 texture。

感兴趣的同学可以去 Josh McDermott 的主页[1]上听一下他们得到的结果。他们采用的方法简单来说是以 random 信号出发，采用 gradient

[1] 他的主页应该不久会搬家吧，如果一段时间以后这个链接无法访问了，可以直接 Google 他的名字。

descend 的方式修改信号使得统计量匹配上目标 texture 的统计量特征。可以说许多情况都还是相当逼真的，他在上课的时候演示出来还是让我小吃了一惊，特别是最后给的一个合成酒吧嘈杂的人声的例子，可以说是相当困难的吧。

当然也有一些本来似乎应该比较简单的但是结果却能明显感觉到合成的声音的不真实性的情况，诸如鼓掌、敲鼓之类的声音。这样的情况在 Vision 中也有，典型的例子就是砖墙的 texture synthesis，并且在 Vision 里显得更加直观一点。如下图所示的那样，Naive 地去做随机 synthesis 的话，很可能由于边界匹配、整体模式等各种问题而很明显地显得不自然了。

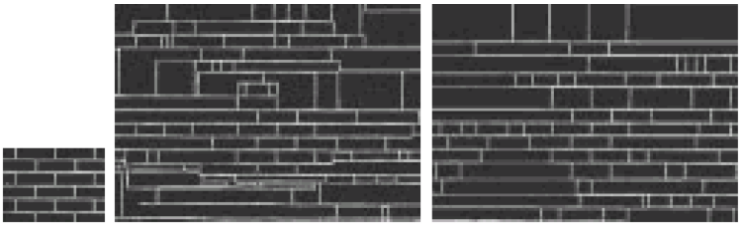

Figure 3: .....

当然，也不能因此全盘否定这种基于统计量的表达模型，因为人脑对声音也好，图像也好，肯定是有多种不同的表达方式的，分别会针对不同类型的数据。比如，很明显地，对于 Speech 来说，肯定就不会是基于统计量的方式来表达。对此他们也有做实验来进行验证，人在面对于背景音这种 texture 是主要特征的声音和像 speech 之类的有明显结构的声音的时候会有很不相同的行为。我想像砖墙这样的例子的困难之处大概也是在于他不小心具有了一些对于人类来说非常容易抽取的全局模式的缘故吧。这样的实验似乎也是 Vision 的版本比较直观一点，想象一下让你区分两张图片是不是一样的——比如，每张图片只给你显示一下下。如果给你的是两张内容不同但是类型相同的 texture 的话，你是很难区分开的，但是如果是其他类型的图片，比如说两张人脸，那么大多数人都会有非常高的 discrimination 能力。

这让我突然想起了刚入学的时候看到实验室一个师兄在 2011 年 NIPS 发的一篇叫做 Why The Brain Separates Face Recognition From Object Recognition 的文章，当时觉得标题很好玩，闲聊的时候问他是怎么回事，他说就是提出了一个模型，有可能是对的有可能是错的，然后用一些实验去验证一下，具体是怎么回事你还是该去看看论文本身的。比较惭愧的是后来一直没有去看.....不过当时对于 “提出一个模型，然后用一些实验去验证” 这件事情不太有概念，现在似乎也开始明白了一些。:D

嗯，回到 Speech Recognition 课的 Guest Talk 上，我在想原本请 Josh 过来应该是主要是给大家介绍人们对人自己的声音处理系统的一些认识和理解吧。虽然前面有一部分确实是在讲这个，不过后面就完全开始讲 Background Sounds Synthesis 了，我个人倒是觉得是非常有趣的一堂课呢。实际上从小学到大学再到 graduate student 的课，能感觉到的一种明显的变化就是讲授的内容越来越 “有特点” 了。比如，低年级的时候学的

各种东西，基本上都是标准化了的，全国甚至全球都是统一的，到后来就开始接触一些比较前沿的内容，相对还没有经历过历史的沉淀，于是各种 notation 呀、细节之类的也会随着学校、老师的不同而有所差别。再到 graduate 课程的时候，有许多课甚至从课的名字上就散发着 Instructor 自己的研究的气息。

一开始我还不太习惯这一点，想着这些老师都不好好上课，讲一门课就偷懒拣和自己的研究相关的部分重点讲，其他的部分就一笔带过。当然虽然也不排除教授们偷懒的嫌疑:P，不过这结果其实是一个很大的好处：因为教授们实际上是在讲自己的研究，所以你看到的是一些实打实的内容，而不是照着教科书照本宣科，同时也有更多的机会接触到最新的最前沿的一些工作。当然，双刃剑的弊端是大部分情况是没有合适的教科书，不同的参考书也会因为各种细节不同而导致你只会越看越晕；甚至可能是根本没有书籍形式的参考资料存在。然而话说回来，做研究的时候，不也就是这样的吗？这样一想的话，倒是觉得是挺必要的锻炼了。

就 Speech 这门课而言也有它很独特的地方，比如说 Spectrogram Reading，据说全世界只有咱这一家教这个，因为这个东西就是 Instructor Victor Zue 他自己发明的啊……

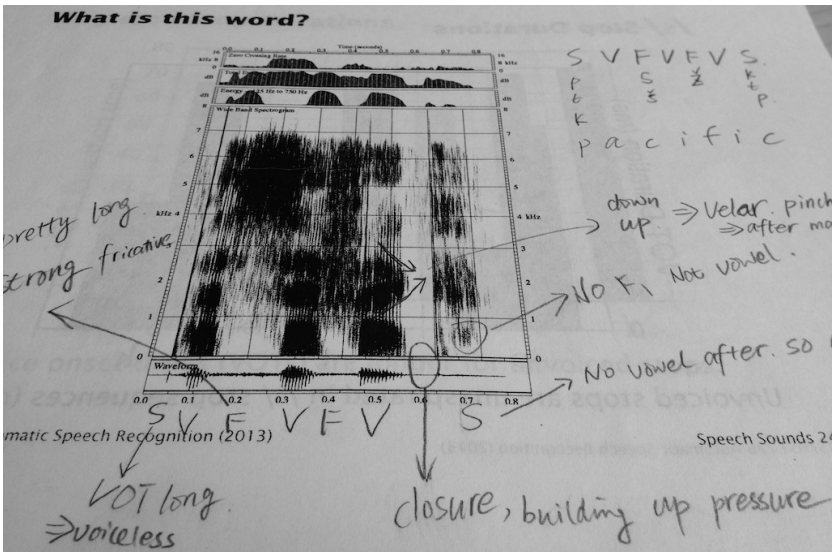

Figure 4: .....

嘛，比起读唇术读心术之类的绝世武功来说似乎是没有那么有用，因为你必须得先有声音信号，然后你又不要直接去听那个信号，而是转换成 wideband 谱图，用眼睛去看。但是不管怎么说还是非常有趣的。据说他们 Speech Group 的人组内还有定期的 “读谱 party”，不过他们读的是更复杂的一个句子的谱。怎么说呢，人肉读谱这件事，与其说是一门技术大概更多的还是一门艺术吧，即便是 Victor 本人也不能保证总是能读出来吧；另一方面人类的语言其实有很多规则和限制，比如三个辅音同时出现的时候只可能会是哪些情况，等等。Victor 说他有一次在外面出差，学生给他发邮件就直接给的是 spectrogram，不过正讲故事兴起的时

候突然想起来还是先上课吧……他还有一点好玩的地方是每次上课会拿一堆红包来，每个里面有五美元，谁要是在课堂上指出了他哪里有讲错的地方，就奖励红包一个。虽然目前为止因为各种原因发出去了好几个红包（比如第一个读出谱的人），但是由于错误被指出还只给过一次呢。

哈，总之是一门很有意思的课，只是每次 pset 都给 50 道左右的题，真是让人求生不得求死不能啊，特别是在除了 pset 还有 quiz 还有 term project 的情况下；特别是这学期其他的课也各种 tough 的情况下。感觉新学期的强度和上学期比起来完全就是好像从桃白白变到了短笛大魔王一样，特别让人心有余悸的是后面还有弗利薩、沙魯、魔人布歐什么的在等着的感觉……>_< 希望能平安度过……阿门！