**[梅尔频率倒谱系数（MFCC） 学习笔记](http://www.cnblogs.com/BaroC/p/4283380.html)**

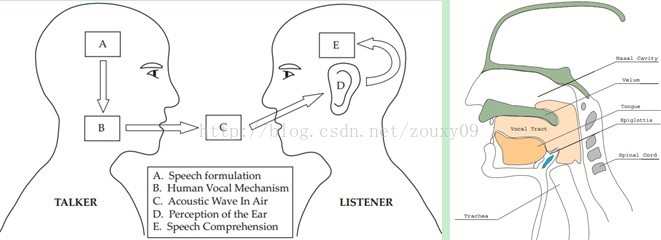
最近学习音乐自动标注的过程中，看到了有关使用MFCC提取音频特征的内容，特地在网上找到资料，学习了一下相关内容。此笔记大部分内容摘自博文 http://blog.csdn.net/zouxy09/article/details/9156785 有小部分标注和批改时我自己加上的，以便今后查阅。

**语音信号处理之（四）梅尔频率倒谱系数（MFCC）**

[zouxy09@qq.com](mailto:zouxy09@qq.com)

<http://blog.csdn.net/zouxy09>

       在任意一个Automatic speech recognition 系统中，第一步就是提取特征。换句话说，我们需要把音频信号中具有辨识性的成分提取出来，然后把其他的乱七八糟的信息扔掉，例如背景噪声啊，情绪啊等等。



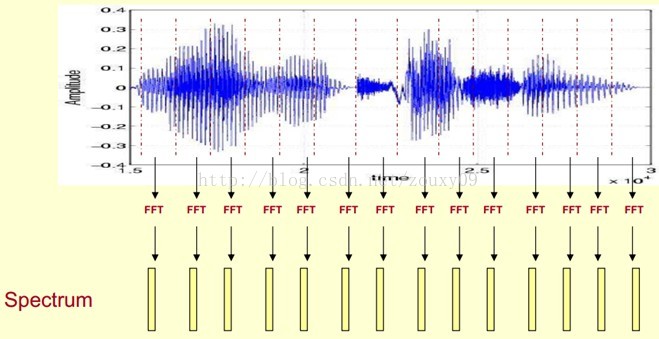
      搞清语音是怎么产生的对于我们理解语音有很大帮助。人通过声道产生声音，声道的shape（形状？）决定了发出怎样的声音。声道的shape包括舌头，牙齿等。如果我们可以准确的知道这个形状，那么我们就可以对产生的音素phoneme进行准确的描述。**声道的形状在语音短时功率谱的包络中显示出来。而MFCCs就是一种准确描述这个包络的一种特征。**

       MFCCs（Mel Frequency Cepstral Coefficents）是一种在自动语音和说话人识别中广泛使用的特征。它是在1980年由Davis和Mermelstein搞出来的。*从那时起。在语音识别领域，MFCCs在人工特征方面可谓是鹤立鸡群，一枝独秀，从未被超越啊（至于说Deep Learning的特征学习那是后话了）。*

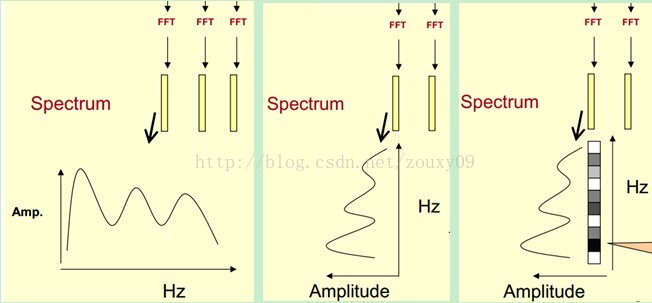
       好，到这里，我们提到了一个很重要的关键词：声道的形状，然后知道它很重要，还知道它可以在语音短时功率谱的包络中显示出来。哎，那什么是功率谱？什么是包络？什么是MFCCs？它为什么有效？如何得到？下面咱们慢慢道来。

**一、声谱图（Spectrogram）**

         我们处理的是语音信号，那么如何去描述它很重要。因为不同的描述方式放映它不同的信息。那怎样的描述方式才利于我们观测，利于我们理解呢？这里我们先来了解一个叫声谱图的东西。

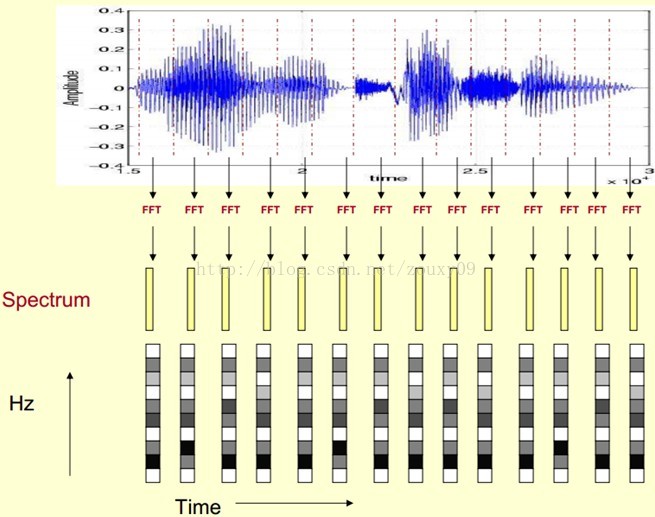


       这里，**这段语音被分为很多帧，每帧语音都对应于一个频谱（通过短时FFT计算）**，频谱表示频率与能量的关系。在实际使用中，**频谱图有三种**，即线性振幅谱、对数振幅谱、自功率谱（对数振幅谱中各谱线的振幅都作了对数计算，所以其纵坐标的单位是dB（分贝）。这个变换的目的是使那些振幅较低的成分相对高振幅成分得以拉高，以便观察掩盖在低幅噪声中的周期信号）。

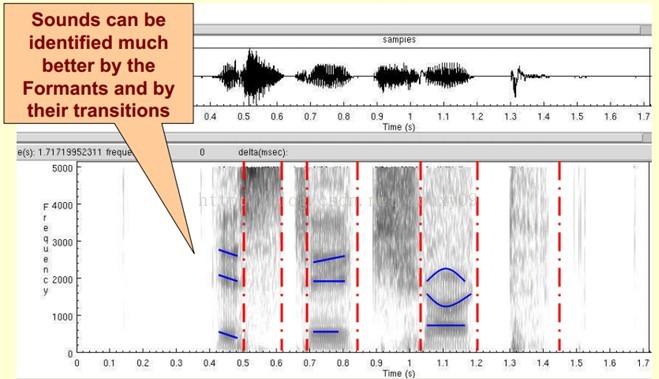


       我们先将其中一帧语音的频谱通过坐标表示出来，如上图左。现在我们将左边的频谱旋转90度。得到中间的图。然后把这些幅度映射到一个灰度级表示（也可以理解为将连续的幅度量化为256个量化值？），0表示黑，255表示白色。幅度值越大，相应的区域越黑。这样就得到了最右边的图。那为什么要这样呢？为的是增加时间这个维度，这样就可以显示一段语音而不是一帧语音的频谱，而且可以直观的看到静态和动态的信息。优点稍后呈上。

         这样我们会得到一个随着时间变化的频谱图，这个就是描述语音信号的spectrogram声谱图。



      下图是一段语音的声谱图，很黑的地方就是频谱图中的峰值（**共振峰formants**）。



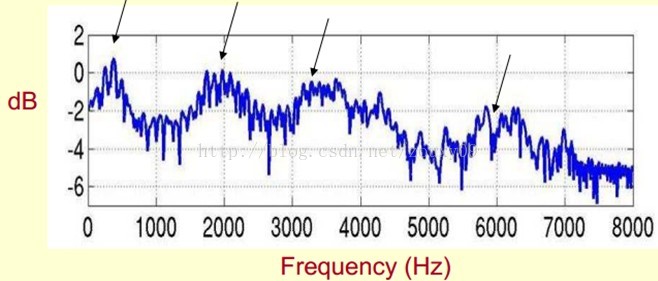
      那我们为什么要在声谱图中表示语音呢？

      首先，音素（Phones）的属性可以更好的在这里面观察出来。另外，通过观察共振峰和它们的转变可以更好的识别声音。隐马尔科夫模型（Hidden Markov Models）就是隐含地对声谱图进行建模以达到好的识别性能。还有一个作用就是它可以直观的评估TTS系统（text to speech）的好坏，直接对比合成的语音和自然的语音声谱图的匹配度即可。

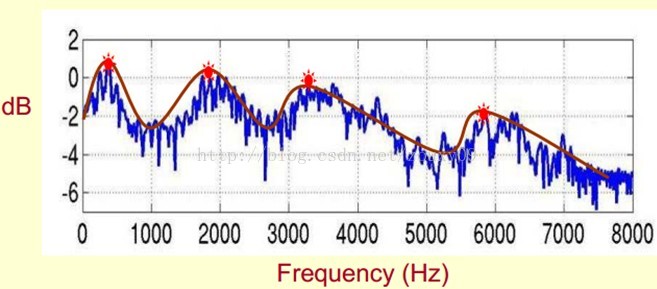
|  |
| --- |
| *通过对语音进行分帧进行时频变换，得到每一帧的FFT频谱再将各帧频谱按照时间顺序排列起来，得到时间-频率-能量分布图。很直观的表现出语音信号随时间的频率中心的变化。* |

**二、倒谱分析（Cepstrum Analysis）**

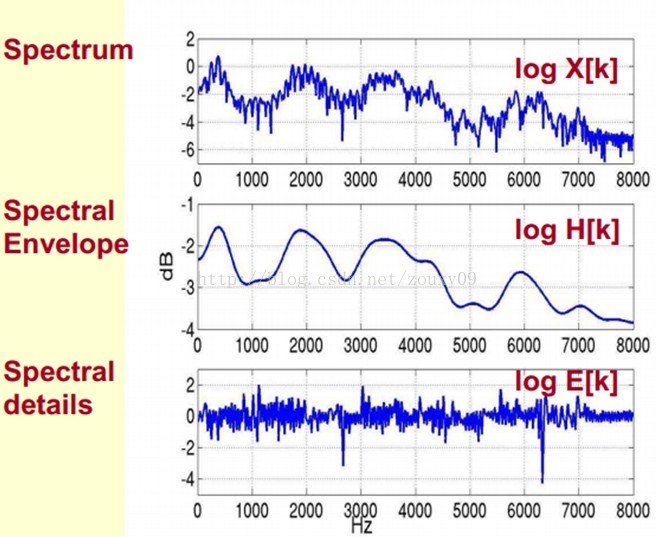
       下面是一个语音的频谱图。峰值就表示语音的主要频率成分，我们把这些峰值称为**共振峰（formants）**，而**共振峰就是携带了声音的辨识属性**（就是个人身份证一样）。所以它特别重要。用它就可以识别不同的声音。



        既然它那么重要，那我们就是需要把它提取出来！我们要提取的不仅仅是共振峰的位置，还得提取它们转变的过程。所以我们提取的是频谱的包络（Spectral Envelope）。这包络就是一条连接这些共振峰点的平滑曲线。

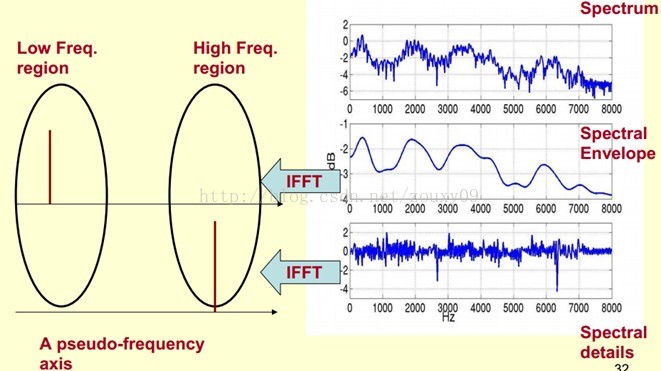


      我们可以这么理解，将原始的频谱由两部分组成：**包络和频谱的细节**。这里用到的是对数频谱，所以单位是dB。那现在我们需要把这两部分分离开，这样我们就可以得到包络了。



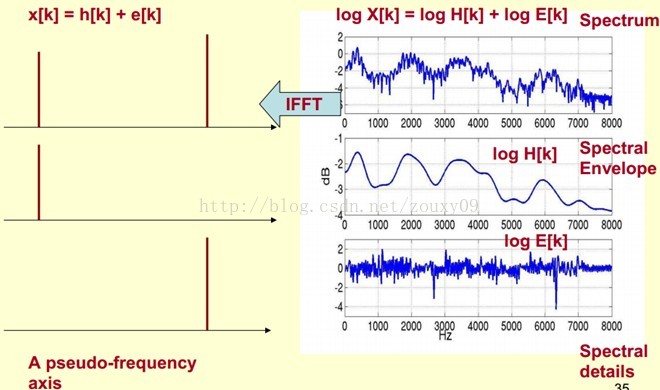
      那怎么把他们分离开呢？也就是，怎么在给定log X[k]的基础上，求得log H[k] 和 log E[k]以满足log X[k] = log H[k] + log E[k]呢？

      为了达到这个目标，我们需要Play a Mathematical Trick。这个Trick是什么呢？就是对频谱做FFT。在频谱上做傅里叶变换就相当于**逆傅里叶变换Inverse FFT (IFFT)**。需要注意的一点是，我们是在频谱的**对数域上**面处理的，这也属于Trick的一部分。这时候，在对数频谱上面做IFFT就相当于在一个**伪频率（pseudo-frequency）坐标轴**上面描述信号。



         由上面这个图我们可以看到，包络是主要是低频成分（这时候需要转变思维，这时候的横轴就不要看成是频率了，咱们可以看成时间），我们把它看成是一个每秒4个周期的正弦信号。这样我们在伪坐标轴上面的4Hz的地方给它一个峰值。而频谱的细节部分主要是高频。我们把它看成是一个每秒100个周期的正弦信号。这样我们在伪坐标轴上面的100Hz的地方给它一个峰值。

         把它俩叠加起来就是原来的频谱信号了。



      在实际中咱们已经知道log X[k]，所以我们可以得到了x[k]。那么由图可以知道，h[k]是x[k]的低频部分，那么我们将x[k]通过一个低通滤波器就可以得到h[k]了！没错，到这里咱们就可以将它们分离开了，得到了我们想要的h[k]，也就是频谱的包络。

**x[k]实际上就是倒谱Cepstrum（这个是一个新造出来的词，把频谱的单词spectrum的前面四个字母顺序倒过来就是倒谱的单词了）**。而我们所关心的h[k]就是倒谱的低频部分。h[k]描述了频谱的包络，它在语音识别中被广泛用于描述特征。

      那现在总结下**倒谱分析**，它实际上是这样一个过程：

1）将原语音信号经过傅里叶变换得到频谱：X[k]=H[k]E[k]；

　　只考虑幅度就是：|X[k] |=|H[k]||E[k] |；

2）我们在两边取对数：log||X[k] ||= log ||H[k] ||+ log ||E[k] ||。

3）再在两边取逆傅里叶变换得到：x[k]=h[k]+e[k]。

       这实际上有个专业的名字叫做**同态信号处理**。它的目的是将非线性问题转化为线性问题的处理方法。对应上面，原来的语音信号实际上是一个卷性信号（声道相当于一个线性时不变系统，声音的产生可以理解为一个激励通过这个系统），第一步通过卷积将其变成了乘性信号（时域的卷积相当于频域的乘积）。第二步通过**取对数将乘性信号转化为加性信号**，第三步进行逆变换，使其恢复为卷性信号。这时候，虽然前后均是时域序列，但它们所处的**离散时域显然不同**，所以后者称为**倒谱频域**。

      总结下，倒谱（cepstrum）就是一种信号的傅里叶变换经对数运算后再进行傅里叶反变换得到的谱。它的计算过程如下：

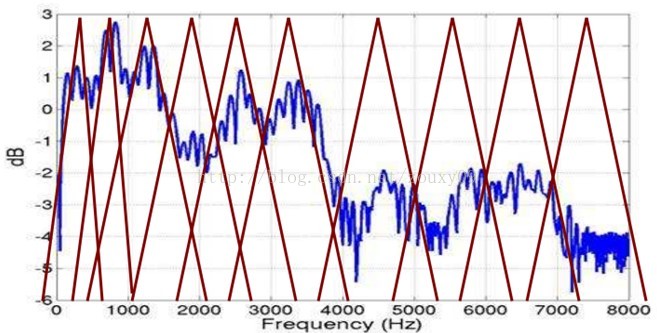


以下部分还未整理

**三、Mel频率分析（Mel-Frequency Analysis）**

         好了，到这里，我们先看看我们刚才做了什么？给我们一段语音，我们可以得到了它的频谱包络（连接所有共振峰值点的平滑曲线）了。但是，对于人类听觉感知的实验表明，人类听觉的感知只聚焦在某些特定的区域，而不是整个频谱包络。

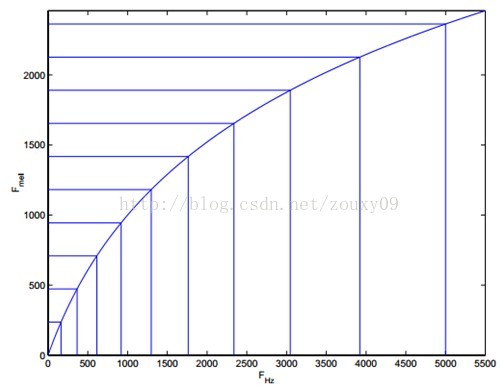
         而**Mel频率分析就是基于人类听觉感知实验**的。实验观测发现人耳就像一个滤波器组一样，它只关注某些特定的频率分量（人的听觉对频率是有选择性的）。也就说，它只让某些频率的信号通过，而压根就直接无视它不想感知的某些频率信号。但是这些滤波器在频率坐标轴上却不是统一分布的，在低频区域有很多的滤波器，他们分布比较密集，但在高频区域，滤波器的数目就变得比较少，分布很稀疏。



       人的听觉系统是一个特殊的非线性系统，它响应不同频率信号的灵敏度是不同的。在语音特征的提取上，人类听觉系统做得非常好，它不仅能提取出语义信息, 而且能提取出说话人的个人特征，这些都是现有的语音识别系统所望尘莫及的。如果在语音识别系统中能模拟人类听觉感知处理特点，就有可能提高语音的识别率。

        梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC）考虑到了人类的听觉特征，**先将线性频谱映射到基于听觉感知的Mel非线性频谱中，然后转换到倒谱上。**

        将普通频率转化到Mel频率的公式是：  
 http://img.blog.csdn.net/20130623210412187?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQvem91eHkwOQ==/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center  
      由下图可以看到，它可以将不统一的频率转化为统一的频率，也就是统一的滤波器组。



**在Mel频域内，人对音调的感知度为线性关系**。举例来说，如果两段语音的Mel频率相差两倍，则人耳听起来两者的音调也相差两倍。

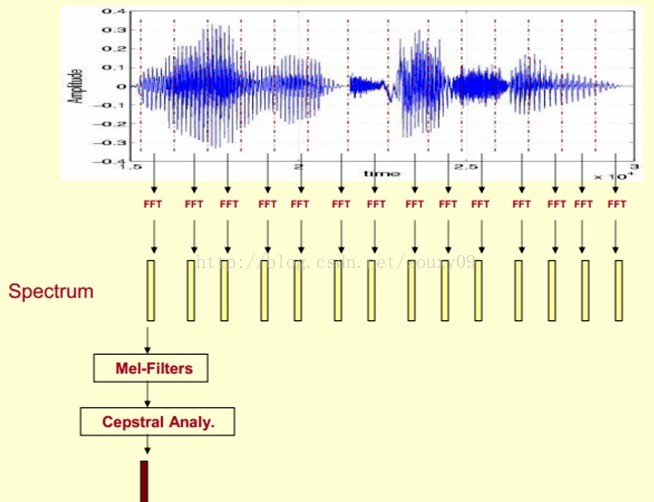
**四、Mel频率倒谱系数（Mel-Frequency Cepstral Coefficients）**

       我们将频谱通过一组Mel滤波器就得到Mel频谱。公式表述就是：log X[k] = log (Mel-Spectrum)。这时候我们在log X[k]上进行倒谱分析：

1）取对数：log X[k] = log H[k] + log E[k]。

2）进行逆变换：x[k] = h[k] + e[k]。

**在Mel频谱上面获得的倒谱系数h[k]就称为Mel频率倒谱系数，简称MFCC**。



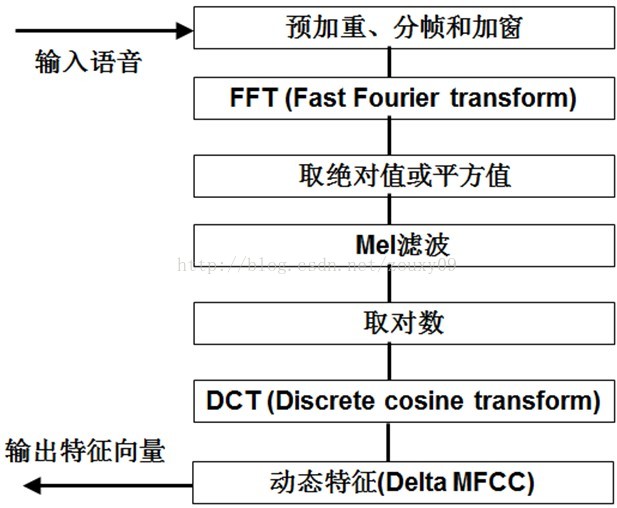
         现在咱们来总结下提取MFCC特征的过程：（具体的数学过程网上太多了，这里就不想贴了）

1）先对语音进行预加重、分帧和加窗；**（加强语音信号性能（信噪比，处理精度等）的一些预处理）**

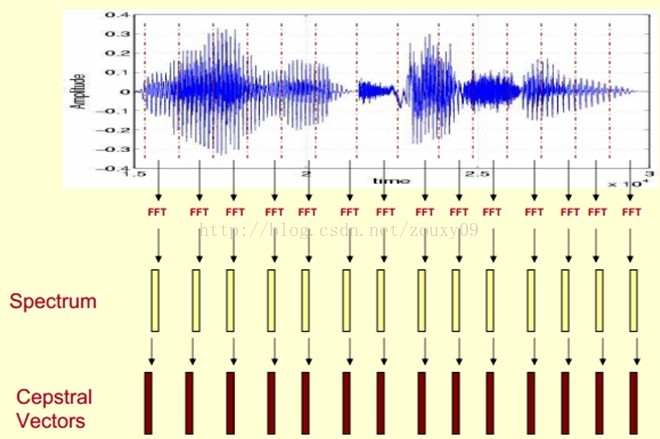
2）对每一个短时分析窗，通过FFT得到对应的频谱；**（获得分布在时间轴上不同时间窗内的频谱）**

3）将上面的频谱通过Mel滤波器组得到Mel频谱；**（通过Mel频谱，将线形的自然频谱转换为体现人类听觉特性的Mel频谱）**

4）在Mel频谱上面进行倒谱分析（取对数，做逆变换，实际逆变换一般是通过DCT离散余弦变换来实现，取DCT后的第2个到第13个系数作为MFCC系数），获得Mel频率倒谱系数MFCC，这个MFCC就是这帧语音的特征；**（倒谱分析，获得MFCC作为语音特征）**



       这时候，语音就可以通过一系列的倒谱向量来描述了，每个向量就是每帧的MFCC特征向量。



     这样就可以通过这些倒谱向量对语音分类器进行训练和识别了。

**五、参考文献**

[1]这里面还有一个比较好的教程：

<http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>

[2]本文主要参考：cmu的教程：

<http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf>

[3] C library for computing Mel Frequency Cepstral Coefficients (MFCC)

<http://code.google.com/p/libmfcc/>