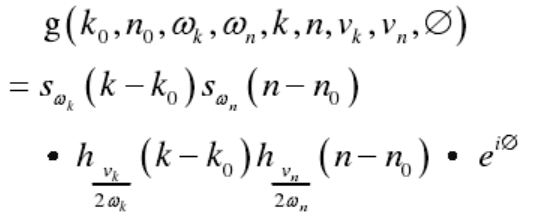
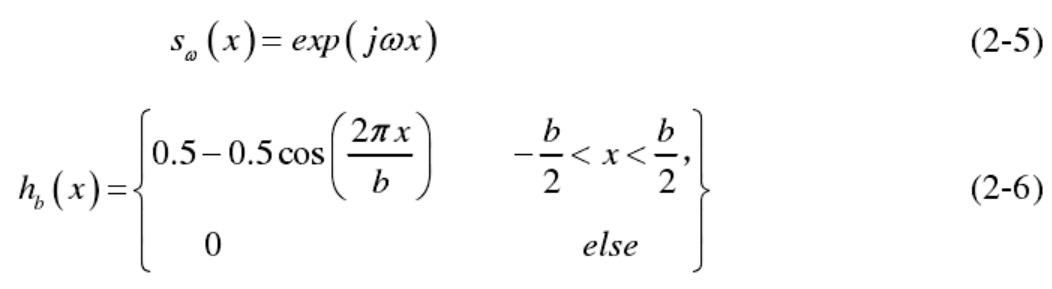
本周主要做了MATLAB对音频特征提取在语谱图中的显示，阅读了基于深度神经网络的音频事件检测 这篇论文，得知了gabor特征提取方法，并找到相应代码进行实现。

Gabor 滤波器组由一组两维的 Gabor 滤波器组成，Gabor 滤波器最初在文献中被提出，后来被用于听觉和视觉系统中生物学处理的模型。每个滤波器由时频域的包络函数和时频的载波函数定义。Gabor 滤波器函数定义如下：



其中，k为频率索引，n 为帧索引，k0表示载波频率，n0表示时间帧的中心，为谱调制频率，为时间调制频率，和为载波在频域和时域维数的半周期数， 为一个加性的全局相位。

一个 Gabor 滤波器可以被定义为复合正弦载波的输出，调制频率与和保持一致，载波函数和包络函数定义如下：



由于当 =0或=0时滤波器函数有无穷多个，因此所有滤波器的大小都被限制使用 69

个信道和 40帧的时间窗，这些限制与每个时频滤波器维数的最大值大致一致。除此之外，调制频率和载波范围线性相关，所有滤波器都是 constant-Q 滤波器，而且和也都保持一致。

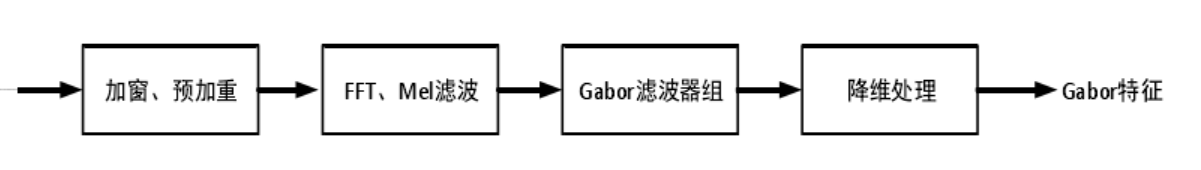
时-频调制频率的定义如公式(2-7)和(2-8)所示：

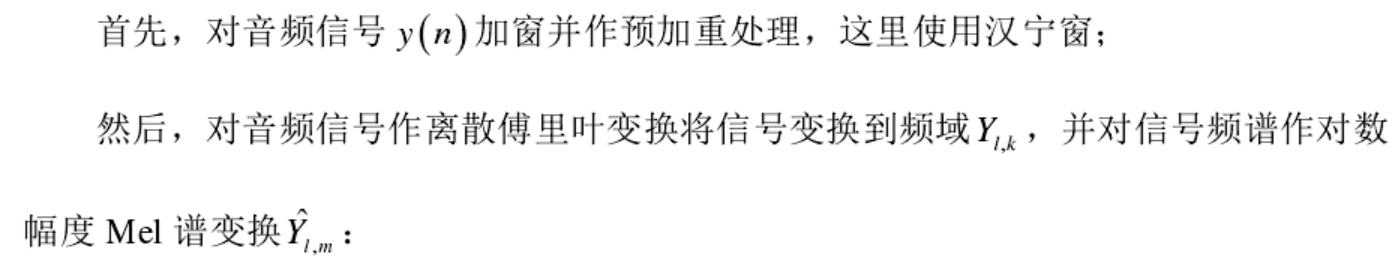


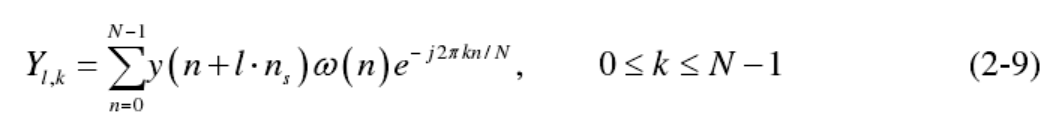
其中，dx为相邻滤波器的相对距离，dx越小滤波器之间的重叠越大。这样定义可以保证相邻滤波器之间的调制转换保持固定的重叠，也可以使每个滤波器共享不同的时-频调制频率 和，但也因此限制和其它滤波器的相关性。调制频率和可以取正负值，决定了滤波器被调谐的时-频方向。

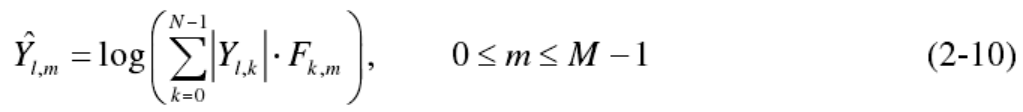
Gabor 特征提取

Gabor 特征是一种时-频特征，它的提取过程如图







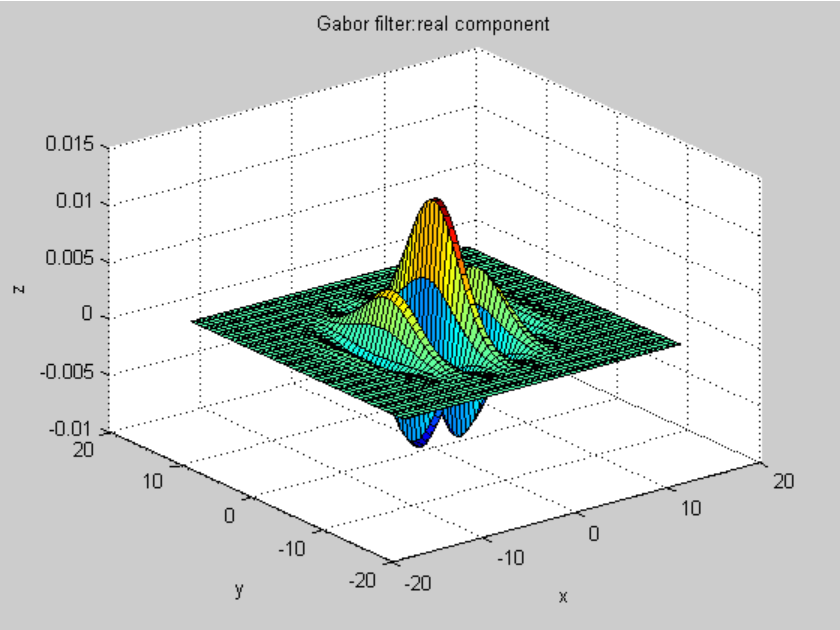


接着，将对数 Mel 谱系数 Y 送入两维的 Gabor 滤波器，取 Gabor 滤波器输出的实部作为该信号的 Gabor 特征由于将Gabor滤波器组应用到每个Mel滤波器会得到一个高维的特征表示，如本文选用23个Mel滤波器和41个Gabor滤波器，则Gabor滤波器的输出有23\*41=943维。文献指出相邻信道的滤波器输出高度相关，因此可以对Gabor滤波器的输出进行降维处理。有文献比较了两种降维方式，一种是对滤波器的输出进行主成分分析；另一种是对Gabor滤波器的输出作二次抽样。实验结果表明后者的识别效果更好，最后保留311维的Gabor特征作为最有代表的特征成分。

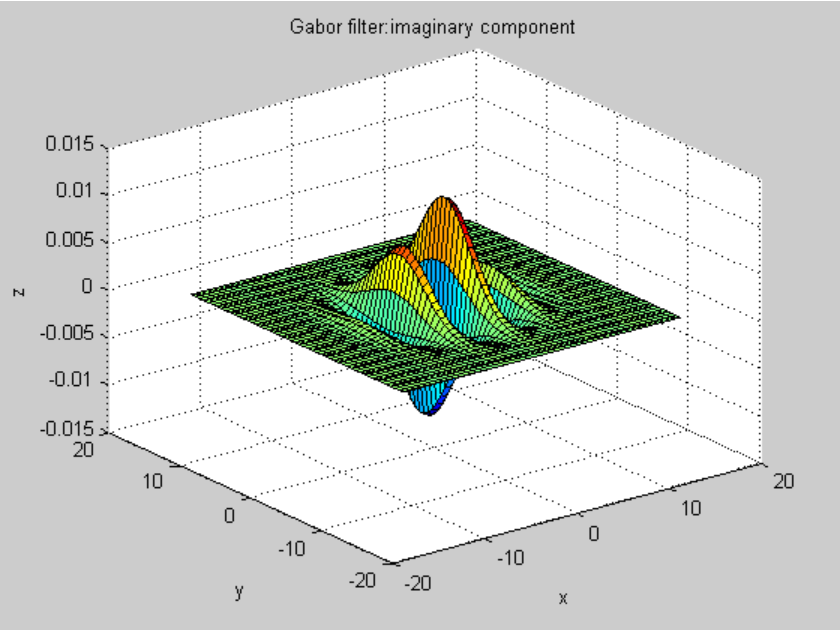
用Gabor 函数形成的二维Gabor 滤波器具有在空间域和频率域同时取得最优局部化的特性，因此能够很好地描述对应于**空间频率(尺度)、空间位置及方向选择性**的局部结构信息。Gabor滤波器的频率和方向表示接近人类视觉系统对于频率和方向的表示，并且它们常备用于纹理表示和描述。在图像处理领域，Gabor滤波器是一个用于边缘检测的线性滤波器。在空域，一个2维的Gabor滤波器是一个正弦平面波和高斯核函数的乘积。Gabor滤波器是自相似的，也就是说，所有Gabor滤波器都可以从一个母小波经过膨胀和旋转产生。实际应用中，Gabor滤波器可以在频域的不同尺度，不同方向上提取相关特征。

代码实现：

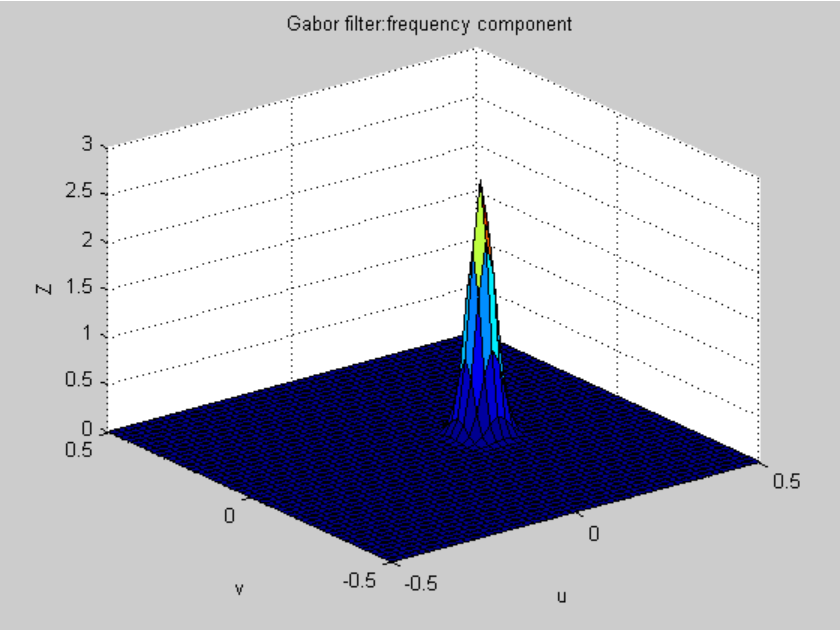
Gabor滤波器实部分量



Gabor滤波器 虚部分量



频率分量



语谱图：

语音的发音过程中, 声道通常都是处于运动状态的, 因此它的共振峰特性也是时变的。不过这个时变过程比起振动过程来说要缓慢得多, 因此一般可以假定它是短时平稳的, 每一时刻我们都可以用这时刻附近的一短段( 例如 15ms ) 语音信号分析得到一种频谱。对语音信号连续地进行频谱分析就可以得到一种二维图谱, 其横坐标表示时间, 纵坐标表示频率, 而每像素的灰度值大小反映相应时刻和相应频率的信号能量密度。这种时频图称为语谱图 ( Sonogram 或 Spec-trogam ) , 这种反映语音信号动态频谱特性的时频图在语音分析中有重要实用价值, 被视为可视语言。

语谱图中的花纹有横杠、乱纹和竖直条等。**横杠**是与时间轴平行的几条深黑色带纹, 它们是**共振峰**。 从横杠对应的频率和宽度可以确定相应的共振峰频率和带宽。在一个语音段的语谱图中, 有没有横杠出现是判断它是否是浊音的重要标志。竖直条(又叫冲直条) 是语谱图中出现与时间轴垂直的一条窄黑条。 每个竖直条相当于一个基音, 条纹的起点相当于声门 脉冲的起点, 条纹之间的距离表示基音周期。条纹越 密表示基音频率越高。

%Winsiz: 定义帧长, 一般应取2 的幂次, 其目的是适合FFT 的要求。通过对Winsiz 的选值可实现宽带频谱或窄带频谱的选择。

function sogram1(Winsiz,Shift,Base,Mode,Gray);

[x,Fs]=wavread(' a007\_110\_120.wav');

n=fix((length(x)-Winsiz)/Shift)+1;

A=zeros(1+Winsiz/2,n);

%分帧

s=enframe(x,Winsiz,Shift);% Shift: 定义帧移值。一般此值小于或等于Winsiz。

Shift 值越小, 时域分辨率越高。

%设置最大递归次数

set(0,'RecursionLimit',500)

for i=1:n

z=fft(s(i,:));%快速傅里叶变换

z=z(1:(Winsiz/2)+1);

z=z.\*conj(z);

z=10\*log10(z);

A(:,i)=z;

End

% Base: 基准电平值。注意, 此值的设定需根据实际经验, 可以通过在多次运行此程序中给出不同的Base 值, 观察所获得的频谱图的视觉和分辨率效果, 选择一个合适的Base 值, 如果没有特别要求, 可取默认值Base=0。

L0=(A>Base);

L1=(A<Base);

B=A.\*L0 +Base\*L1;

L=(B- Base)./(max(max(B))- Base);

y=[0:Winsiz/2]\*Fs/Winsiz;

x=[0:n-1]\*Shift;

% Mode: 定义显示模式。1 伪彩色映射, 0 为灰度映射

if Mode==1

colormap('jet' );

else

mymode =gray;

mymode =mymode (Gray: - 1:1,:);

colormap(mymode);

end

imagesc(x,y,L);

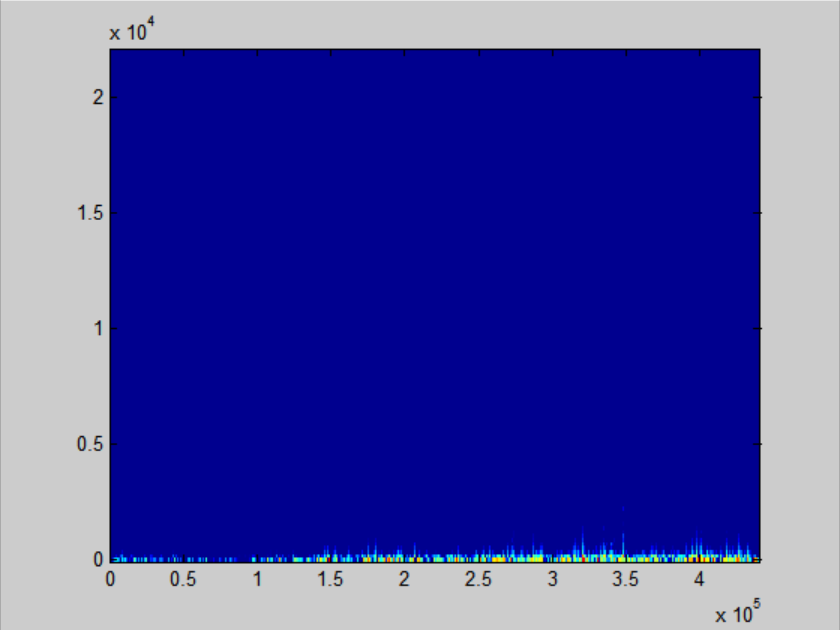
axis xy;

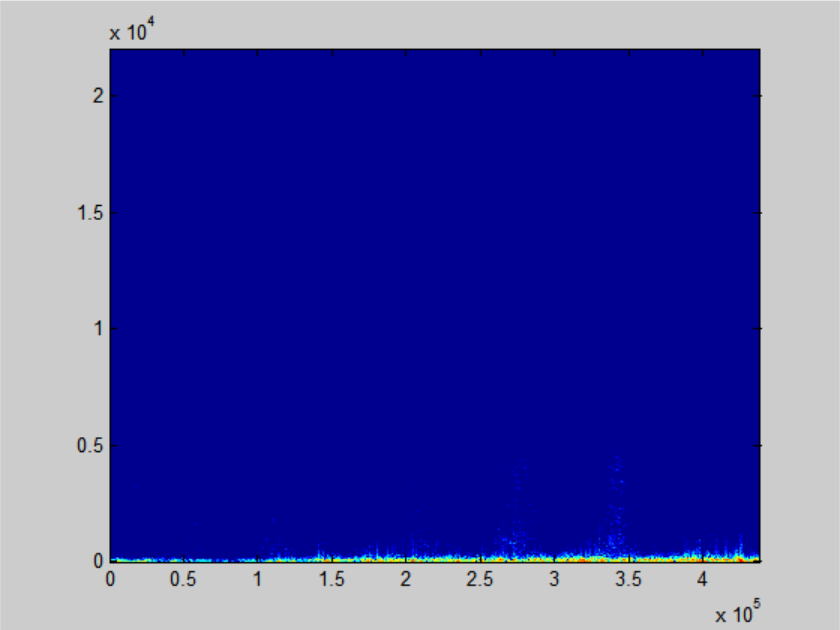
然后，编写主程序，输入相应的参数，得到语谱图：

close all;

sogram1(2048,128,0,1,64);

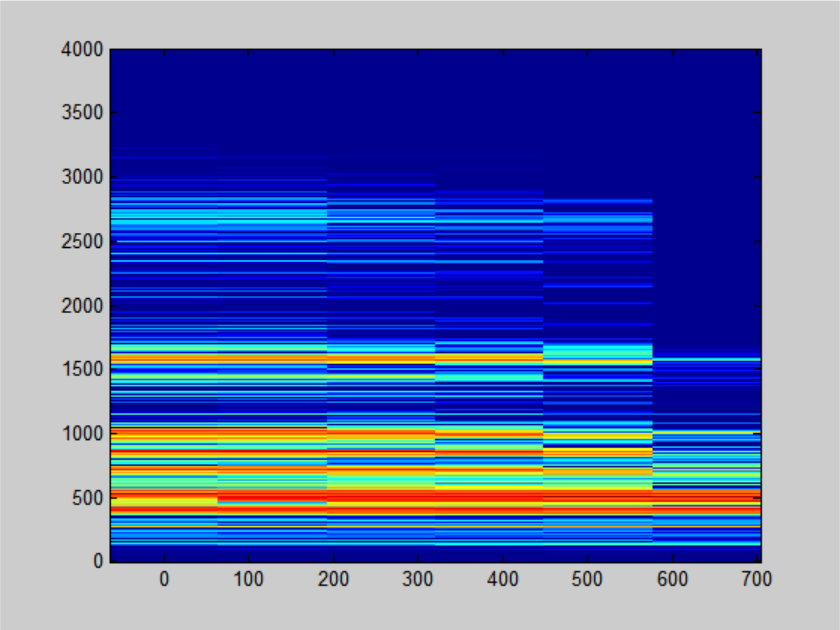
Dcase 海浪音频片段输入，

帧移位=256



帧移位=2048

由图可见，帧移位对环境音的影响不大，语谱图特征不明显



语音输入“Apple”

由图可知，Apple 这个单词中有大量浊音，条纹密集程度不高，音频频率也不高。

目前的技术难点是如何批量导入大量音频进行特征提取，并转化成语谱图，现在只是将每一部分开实现，并只是分析片段的音频特征，CNN训练也没有实现音频的批量读取，这是我需要继续深入学习的地方。