本周我阅读了《基于深层神经网络的音频特征提取及场景识别研究》

文章首先介绍了声学特征分析研究现状，音频场景识别研究现状，深度神经网络的研究现状。其中我在音频场景识别现状中找到了许多前人尝试过的场景分类方法。

接下来作者声明了该论文的创新之处：

首先是特征提取方面，然后是音频场景识别方面。大体思路为：首先把特征分析角度从时域或者频域特征分析转变为基于频域的长时分析；深度神经网络在图像分析领域有着非常好的识别性能，所以借助其特征分析的能力对语谱图进行深层特征分析提取；然后使用分类器进行最后的分类。具体研究内容章节安排大致如下：

第二章介绍了本课题用作结果对比的基线系统，即基于梅尔频率倒谱系数和 k 近邻分类器的实验系统。在这一章，我们首先介绍了基线系统的基本流程，然后对流程中的主要步骤进行展开，依次详细介绍了数据预处理工作、特征提取工作，最后介绍了根据提取到的特征向量进行音频场景分类的工作。实验过程中我们根据通过调节影响实验结果的参数得到的对应结果选出最佳的参数值。

第三章介绍了基于卷积神经网络的音频特征提取及场景识别算法的研究。在这一章里，首先分析了现有音频特征的不足之处和基于频域的长时分析的好处；然后进行卷积神经网络的适用性分析，以及卷积神经网络在语谱图上进行深层特征分析提取的有效性分析；紧接着介绍了卷积神经网络的计算方式和学习算法；最后进行实验系统的算法设计以及基于此算法进行了实验验证，并把实验系统结果和基线系统结果进行对比分析。

第四章介绍了基于解卷积神经网络的音频特征分析、提取以及最后的音频场景分类。本章首先介绍了解卷积神经网络的计算方式和学习算法，然后分析了解卷积神经网络对场景特征分析的适用性，最后根据解卷积神经网络的结构特性设计了本实验系统的算法。按照设计的算法进行实验，把最终得到的实验结果和基线系统的结果进行对比分析。

# 音频场景识别现状

**音频场景**可以看做是由一个或几个特定类别的声学事件构成的音频段。它往往通过对特定声学事件的检测来对整个音频场景所属类别做出判断。例如浪花拍打沙滩的声音和海面上各种海鸟的叫声等一起构成了海滩这个特定的音频场景，要识别这一特定场景首先要检测出其中的特定声学事件如海鸟的叫声，然后才能对其场景类别做出判断。

到目前为止，音频场景识别领域已有识别的方法都是按照模式识别框架进行的。

即首先进行音频场景数据的**特征分析、提取**，

然后进行场景的**分类**。（常用到的识别器有支持向量机、高斯混合模型、k 近邻模型和隐马尔科夫模型等。）此外还有人尝试使用模型组合的方式进行音频场景的分类，Wei-Ta chu等人通过组合支持向量机和高斯混合模型来进行音频场景的分类，Khin Myo chit把隐马尔科夫模型和支持向量机进行级联来进行场景识别，最终都取得了较好的效果。

Toni Heittola等人提出基于关键声学事件直方图的场景识别方法。大致思路如下：首先进行关键声学事件的检测，然后根据检测出的声学事件对场景进行建模，最后根据得到的场景模板进行识别。Virtanen提出了使用声源分离的方式进行关键声学事件的检测，大致思路为：利用非负矩阵分解把声源信号进行分解，得到几个独立的音轨信号；然后在每个独立的信号上面按照上面的处理方式进行事件的检测，最终进行音频场景的分类，取得了较好的识别效果。由于声学事件之间存在重叠且其先验知识相对匮乏，这种情况下的声学事件检测相对就会困难很多，为此Mesaros等人提出了基于潜在语义分析的声学事件检测的方法来进行重叠事件的相关性分析，最终很好地解决了由于声学事件之间存在重叠而造成场景识别性能低的问题。

Chu等人提出使用匹配追踪的方式直接对音频场景信号进行分析，得到对于背景噪声来说更加鲁棒的特征。匹配追踪的目标是找寻一组最小数目的基底来对信号进行稀疏且有效的分解。经过匹配追踪分析，原始信号就可以用一组基底向量进行表示，即得到了新的特征向量，然后使用这种新的特征参数直接对音频场景进行分类。Peltonen等人通过使用MFCC 和子带能量比作为音频场景的声学特征，以高斯混合模型和KNN 分类器作为场景的分类器进行最终的分类，此外Guo在前人研究的基础上，通过使用频域特征和支持向量机进行音频场景的识别。Ajmera则是把基于音频场景类后验概率得到的动态特征用作场景声学特征，使用隐马尔科夫模型进行最终的场景识别。

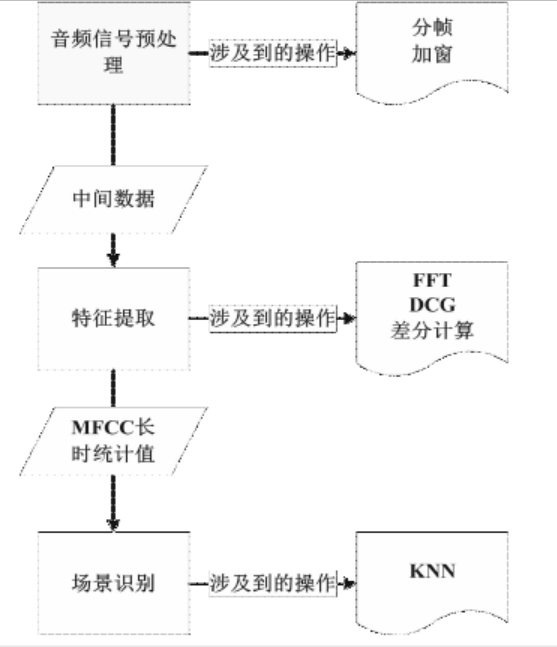
Kyuwoong等人使用 MFCC 作为场景识别的声学特征，然后通过使用高斯混合模型构建语义模型，得到语义模型直方图，最后使用KNN 分类器进行最后的场景识别。Srini Vasan等人借助音频场景识别的结果对语音信号进行分离来提高最终的语音识别性能；Akinori等人通过使用多阶 GMM 来对声学事件进行检测，然后进行音频场景的识别。

# 基于 MFCC 和 KNN 的场景识别基线系统

流程：音频信号的预处理工作 对预处理得到的中间结果进行特征分析

在得到的特征向量上进行音频场景分类。

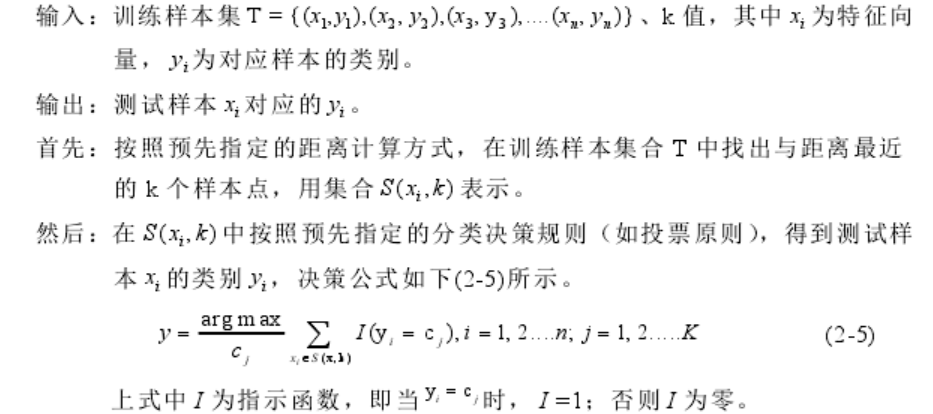
音频信号预处理工作主要是对由音频设备采集到的原始模拟信号进行处理，常规操作有：预滤波、模数转换、预加重、分帧、加窗等；特征提取的目的是提取出相对来说最能体现出当前音频信号的内容数据，我们采用的特征是MFCC,所以涉 及到 的工作有： 快速傅里叶变换 (Fast Fourier Transformation,FFT)、三角滤波、离散余弦变换(Discrete Cosine Transformation, DCT)、倒谱均值减、差分计算等；最后一步是对特征数据进行分类， KNN 分类器无须进行训练，可以直接对特征向量进行分类操作。本片论文的的基线系统基本流程如图所示



这篇论文研究的是多类音频场景识别问题，选择的分类器是 k 近邻模型。给出训练数据集，如果要判断测试样本的类别首先得在训练集中找出与该样本最近的 k 个样本，然后统计这 k 个样本对应的所有类别中每个类别包含的样本数量，最后通过多数表决方式选出包含样本数量最多的类别做为测试样本的类别归属。KNN 具有简易、灵活和有效性，最终选择其作为场景识别的分类模型。（KNN即临近算法）

k近邻算法中三个重要的基本因素为：k值的选择、样本向量之间距离的度量方式以及最后的分类决策规则

以下是K临近算法（KNN）的基本流程



在这个基线系统中，通过调整影响最终分类性能的参数即 knn 中的 k 值，来进行音频场景识别性能的分析，最终得到基线系统的整体识别结果。

# 基于卷积神经网络（CNN）的音频特征提取及场景识别

使用卷积神经网络(Convolution Neural Network, CNN)对音频场景进行时频特征分

析，提取能够反映结构性的长时特征来进行音频场景识别。这要比现有的单一从时域或频域特征更加准确，更能中和平衡各自的缺点。时域信号和频域信号都是一维信号，如果联合这两者进行分析，把时间和频率同时作为自变量，而把对应的能量值当做因变量，这就形成了时间频率平面上的二维信号。

滤波器的设计和参数调整是此部分的关键，特征语谱图的数量调整和结果分析，激活函数参数的调整和结果分析也同等重要

此部分的创新之处在于由于神经网络在图像处理方面特别是对其结构性特征分析方面取得了很好的效果，所以应用卷积神经网络在语谱图上进行长时结构性特征分析。通过设计合理的网络结构和调整影响网络性能的多个参数，甚至是不断调整网络的局部结构，最终得到了好于基线系统的识别效果。