上周我用五天时间仔细的阅读了老师给我交代的阅读3篇论文的任务，并对Yiteng Huang的SUPERVISED NOISE REDUCTION FOR MULTICHANNEL KEYWORD SPOTTING和来自洛桑联邦理工学院的Dual-Microphone Noise Removal for Keyword Spotting有着自己比较深入的理解。

这两篇论文的主要目的都是为双麦克风提出一种自适应消除或智能语音增强以使用户发出的唤醒词即使在一些声音不利的环境下也可以将关键词看起来更清晰易懂的一种算法（KWS）。Yiteng Huang的论文更是提出了一种基于人类听觉系系统研究的远场精细的关键词发现的算法，通过引入从神经网络分类器到信号预处理前端的反馈路径，进而将麦克风阵列语音增强和机器学习相结合，从而使前段降噪从中受益更好的为后端服务。 这种结合的思想可以很好使机器适应不同环境强度的噪声，极大的提高了KWS的性能。

一．来自洛桑联邦理工学院的Dual-Microphone Noise Removal for Keyword Spotting

先提出了一些关于日常生活中的VUI（是一种与虚拟助手互动的主要方式）的具体例子例如Google Assistant等，并希望解决如何在嘈杂的环境中实现背景噪声的消除，以增强关键字识别的效果。通过现有的基于相干性的技术和机器学习技术提出了一些算法的说明和结果的展示。

基于相干性的技术

它是一种用于检测两个信号之间关系的一种统计数据，通常可以用于去估计线性输入和输出之间的功率传输。该论文提出的这种基于相干性语音增强的算法依赖于两个声道中的语音信号是相关的，但噪声信号确是不相关的。所以如果噪声信号之间的相干函数接近1，则语音信号占主导，所以将无衰减通过。但如果接近0，该语音信号将不复存在。对于一些手机中的特殊应用，这种双麦克风语音增强也是基于相干方程和卡尔曼滤波器。卡尔曼滤波器是使用一系列随时间推移观察到的包含统计噪声的测量值。通过每个时间帧在变量上的联合概率分布去更准确的估计所需要的值。

机器学习技术

一些深度学习基于使用了不同类型的神经网络结构，例如深度神经网络或者卷积神经网络。

算法的说明

用两个有小段距离的麦克风去捕捉周围环境的背景噪声，分别将两个麦克风通道，噪声，语音信号给出相对应的函数关系并在短时傅里叶变换域中实现了维纳过滤器和具有延时滤波器系数的ANC滤波器。

维纳滤波器

通过估计来自双麦克风的另一个麦克风的一个信号，在给定遗忘因子α情况下，考虑到先前的帧，针对每一个帧在STFT域中计算滤波器系数。如果录音中关键字持续时间很短并基本由噪声组成时，该滤波器只能进行去噪处理并保留关键字。

首先要对两个信号做短时傅里叶变换，对于每个帧的傅里叶变换可以得到

*φxs*[*i*] = *αs*[*i*] *∗ x*[*i*]*∗* + (1 1 *α*)*φxs*[*i i* 1]

*φxx*[*i*] = *αx*[*i*] *∗ x*[*i*]*∗* + (1 1 *α*)*φxx*[*i i* 1]

当两个信号之间的自相关度很高的时候h接近1.

根据上述内容可得到误差系数。

具有延时滤波器系数的自适应噪声消除

之前我们介绍了其基本概念和使用的条件等，接下来我们提及它的基于的几个假设：关键字的前面必须是噪声而且关键字的持续时间必须要短，至少短于1s。

对于维纳滤波器和具有延时滤波器系数的自适应噪声消除这两种算法，该论文进行了四种情况下的测试：

1. 添加了语音的延时输入
2. 带有语音添加的延时输入，并且两个信号都使用不同的RIR进行滤波
3. 来自录音的原始信号加上语音输入
4. 来自录音的原始信号已经包括语音信号

Yiteng Huang的多通道KWS监督降噪

随着由语音驱动的搜索查询增多，语音检索设备正在数以百万计的家庭中使用说明语音正在成为人们当今快速变幻的数字世界中寻求帮助的重要方式。

语音接口的免提和预设关键字去召唤语音服务对检测远场，噪准确性有着很大帮助。因为一些常用设备的KWS系统需要连续收听来自麦克风的音频输入。因此其必须要具有很小的内存和功耗，为满足这些条件，我们提出了HMM，这在DNN的自动语音识别起了很大作用并基于RNN,DNN,CNN构建。虽然机器智能无法达到人类听觉系统所表现出的能力但一些多样式训练MTR，自动增益控制等还是能够提高其远场失帧等问题。

Cocktail party effect：指的是我们可以将听力集中到对话上，自动过滤掉周围其他对话和噪声的能力，它是通过大脑和耳蜗来做到这一点的

因为这种效应具有自下而上的驱动和自上而下的处理但现有的KWS系统仅仅实现了自下而上的机制，却没有自上而下的反馈，所以在该篇论文的研究中提出了一种新的构架将神经网络NN的关键字输出反馈到多通道前端预处理器中从而更有效地消除背景噪声。

新系统包括CNN分类器，关键字的筛选和STFT域的自适应噪声消除

CNN分类器：

KWS的一个框架，用于声学建模。由5层组成，输入和输出层之间有3个隐藏层

具体描述：对单通道麦克风信号分成两个不同帧的跳数。对于其每一帧都要使用PCEN计算并归一化滤波器组的能量。CNN的输入由堆叠的特征向量组成，在输入层的顶部是卷积层，卷积层可以在时间和频率上扫描308个不重叠的8\*8色块，该卷积层后面是具有32个输出的线性投影层和具有128个输出完全连接的整流线型单元层。并在输出层的神经单元上使用softmax激活功能。

关键字的筛选：监督语音增强

监督语音增强实际上监督的是立体声语音增强的一种算法，虽然KWS的核心是单声道，但我们还是希望其能利用立体声麦克风来提高性能。

具体描述以及算法概述：

当捕捉到新的立体声音频样本帧时，基于单声道CNN的KWS分类器使用左声道在上下文中去计算新帧的关键字，再进行预设阈值进行比较。比较出来的新帧将被标记为3个类别：

1. 触发（高于高阈值）
2. 接近触发（两个阈值之间）
3. 噪声（低于阈值）

所得出的数据帧会被推入固定长度的缓冲区，且控制器的设计可以处理缓冲区中的数据并分为3个情况。

1. 缓冲区中的数据帧都是噪声帧，则控制器仅从缓冲区输出一个数据。而后将其应用ANC并于右通道卷积减去左通道形成误差信号，调整系数并将该信号作为已过滤帧输出发送到CNN进一步处理。
2. 缓冲区包含一些接近触发但没有触发帧的时候将会刷新整个缓冲区，但注意此过程不会更新滤波器系数，该情况只是在不取消关键字的情况下抑制噪声
3. 当缓冲区包含至少一个触发帧时，第一个CNN已成功识别了关键字，此时需要刷新系数并将其传递到第二个卷积

特别注意：筛分缓冲器的长度。如果时间太长，则ANC过滤器会在KWS更早的停止这会使滤波器背景噪声效率降低。如果太短，则缓冲区无法容纳整个关键字实例并会产生过多的失真

STFT自适应噪声消除

它是基于短时傅里叶变换的RLS方法，可以实现ANC子系统。