1. 声音特征提取

1.1 常用声音特征

1.1.1 MFCC

MFCC（Mel-frequency cepstral coefficients）:梅尔频率倒谱系数。梅尔频率是基于人耳听觉特性提出来的， 它与Hz频率成非线性对应关系。梅尔频率倒谱系数(MFCC)则是利用它们之间的这种关系，计算得到的Hz频谱特征。主要用于语音数据特征提取和降低运算维度。对fbank做离散余弦变换（DCT）即可获得mfcc特征。

1.1.2 Fbank

Fbank（FilterBank）：人耳对声音频谱的响应是非线性的，Fbank就是一种前端处理算法，以类似于人耳的方式对音频进行处理，可以提高语音识别的性能。获得语音信号的fbank特征的一般步骤是：预加重、分帧、加窗、短时傅里叶变换（STFT）、mel滤波等。

1.2 声音特征提取

1.2.1 传统方式

1.2.2 基于深度学习方式

传统声音特征提取

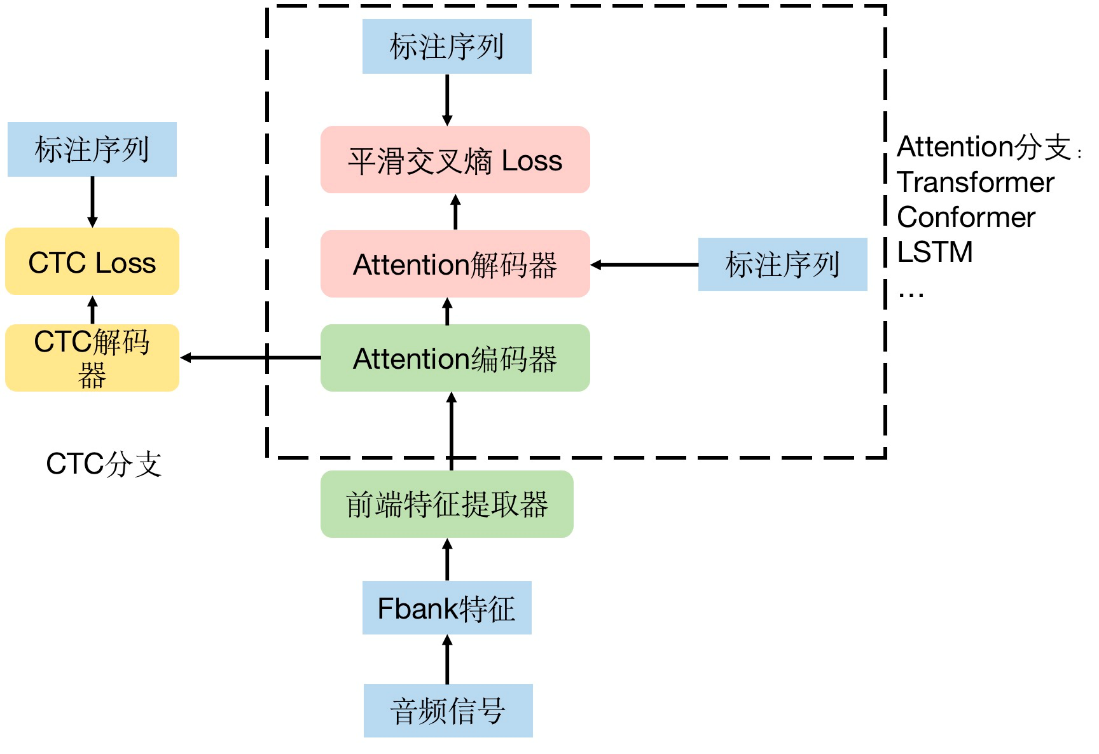
基于深度学习的声音特征提取

2. 基于深度学习的ASR

ASR模型可分为

2.1 端到端的ASR策略

端到端语音识别是区别传统语音识别的一种框架，并逐渐成为一种趋势。**传统语音识别一般分为声学模型与语言模型**，声学模型负责将音频序列转化为音素序列，常见的音素比如汉语拼音、英文音标等，语言模型则负责将这些音素序列转化成文字序列。声学模型和语言模型在训练时并不需要耦合，可以独立训练，传统语音模型的劣势在于需要有发音字典，需要有音素的标注。**端到端模型**就是克服了这一点，**直接将音频序列转化为文字序列**。



一般的端到端语音模型结构中，大致可以分成三个部分，分别是前端特征提取器、CTC分支和Attention分支：

参考文献

1. 一文入门端到端语音识别：<https://zhuanlan.zhihu.com/p/536647952>
2. 语音识别中的MFCC 与 Fbank特征的基本原理和python实现: https://zhuanlan.zhihu.com/p/363153781