目　　录

第一章　　引言

第二章　　背景介绍

2.1　自动语音识别

2.1.1　特征提取

对于语音识别系统而言，声学信号作为用户的唯一输入，需要承载用于识别的所有信息。如果仅对声学信号在时域上的波形进行分析，很难从中提取出对识别有用的特征，因为即使同一个人说同样一段话，单从波形上看都会有很大差别。然而，一个受过训练的人，可以通过语谱图区分不同的元音，因为元音的频率成分相对固定，不同元音的频谱图会有明显的差别。根据这一特性，我们可以在频域上对信号进行分析，从声学信号中提取与频率有关的特征，用来作为识别系统的输入。目前语音识别系统常用的声学特征包括：梅尔频率倒谱系数（MFCC）、感知线性预测(PLP)、Filter-bank等。

梅尔频率倒谱系数(MFCC)作为语音识别中比较常用的声学特征参数，其原理是模仿人耳的听觉机理，将以赫兹为单位的频率变换成梅尔频率，使用在梅尔刻度上等距分布的梅尔滤波器组搜集不同频段的能量，通过逆离散傅里叶变换（IDFT）计算倒谱系数，实现声源和滤波器的分离，并降低不同维度特征之间的相关性。最后加入能量以及帧与帧之间的变换的信息。计算梅尔频率倒谱系数的详细过程如下：模拟信号经过采样和量化，转换为数字信号x[n]，n对应采样时刻。接下来对x[n]加窗，由于声学特征是用来区分语音信号中不同音素的，所以我们需要分析大致对应每个音素长度的部分的波形，这就需要对整个信号做加窗处理，窗口外部的信号全部设为零，只保留窗口内部的信号。一般情况下窗长设为25ms，每10ms 向前移动一个时间窗。这样每段音频都转化成了相互之间有重叠部分的固定长度的数字向量。MFCC 提取过程中普遍使用的窗函数为Hamming Window，其公式为:

$$f(x)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{L}) & 0\le x\le L-1\\ 0 & \text{otherwise.}\end{cases} \tag{2-1}$$

接下来是对加窗后的数字信号做离散傅里叶变换，离散傅里叶变换的目的是计算信号在不同频段所包含的能量。其公式为:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn} \tag{2-2}$$

DFT的N个输出对应N个离散频带，X[k]为复数，代表当前频率成分的幅值和相位。FFT是实现离散傅里叶变换的高效算法，但限制是N必须为2的正整数次幂。最后FFT的输出表示每个频段的能量。前文已经提到过，MFCC是基于人耳的听觉感知机理设计的，人耳对频率的感知是非线性的，超过1000Hz时，人类对频率的变换越来越不敏感。研究表明，在特征提取时，通过建模人类听觉的这种特性，能够提高识别系统的性能[Davis and Mermelstein]。MFCC通过引入梅尔刻度来模拟人耳的机制，梅尔刻度[Stevents and Volkmann]是描述声调的单位，两组信号如果在声调的感知上是等距的，那么它们的梅尔频率也是等距的。以赫兹为单位的频率与梅尔频率之间的关系可以使用如下公式表示：

$$mel(f) = 1127\ln(1 + \frac{f}{700}) \tag{2-3}$$

计算MFCC时，通过放置一组三角滤波器来收集不同频带的能量，三角滤波器在Mel刻度下是等宽均匀排布在整个频率范围内的。接着对每个梅尔滤波器的输出取对数，这样可以减弱特征对输入变化的敏感性，比如发音人与麦克风之间距离的变化。如果直接将梅尔滤波器输出取对数后的值作为特征，因为各个滤波器输出值之间相关性不为零，会造成后续训练高斯混合模型时协方差矩阵无法使用对角阵，所以需要进一步处理。倒谱（cepstrum）是对频谱取对数之后做逆傅里叶变换，倒谱可以实现声源和滤波器分离，并去除特征不同维度之间的相关性，因此取倒谱系数的前12维作为MFCC的特征。由于每一帧的能量和当前帧所属音素有关，可以将能量作为MFCC的一个维度：

$$E(m) = \sum_{t=t_1}^{t_2} x^2[t] \tag{2-4}$$

其中m表示帧的标号，t_1和t_2分别代表帧的起始时刻和终止时刻。除了能量之外，前后帧之间的变化信息也有助于识别不同的音素，所以MFCC一般还会加入倒谱系数每一维的一阶差分和二阶差分，以及能量的一阶差分和二阶差分。最终，从每一帧信号中提取出39维的MFCC向量，用于训练声学模型和识别。

2.1.2 声学模型

Dynamic Time Warping(DTW)是动态编程在语音识别中的应用，通过DTW可以搭建最简单的语音识别器。DTW可以解决计算两个音频之间距离时，时长不匹配的问题。通过引入warping function Tx(t), Ty(t),t=1,2,...T. 通过限制Warping function的限制条件：如果给定对齐信息，很容易计算两个音频特征样本之间的距离。给定样本X和Y，有很多可选的对齐方式，并且随着Tx和Ty 的增加，可选的对齐方式数量成指数级增长。这里就要用到Dynamic Programming（DP），通过保存历史信息，找到最优路径。

2.1.3 发音词典

2.1.4 语言模型

2.1.5 解码器

2.2 Kaldi工具箱

2.3 拉萨方言语音识别

第三章　　声学模型训练

3.1　传统的GMM-HMM方法

3.2　目前广泛使用的DNN-HMM方法

3.3　LSTM

第四章　　拉萨方言语音识别系统

4.1　拉萨方言数据库

4.2　发音词典

4.3　语言模型训练

4.4　声学模型训练

4.4.1　声调相关特征

4.4.2　模型融合

4.5　识别结果及分析

第五章　　总结