藏语声学建模方法研究

A Study on Subspace Clustering Algorithm for High-dimensional Data

专　　业：计算机科学与技术
学生姓名：李 健
指导教师：　教授

天津大学计算机科学与技术学院
二〇一六年十一月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**天津大学**或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名:　　签字日期:　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解**天津大学**有关保留、使用学位论文的规定。特授权**天津大学**可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名:　　导师签名:

签字日期:　　年　　月　　日　签字日期:　　年　　月　　日

摘　　要

中文摘要应将学位论文的内容要点简短明了地表达出来，约500 800字左右（限一页），字体为宋体小四号。内容应包括工作目的、研究方法、成果和结论。要突出本论文的创新点，语言力求精炼。为了便于文献检索，应在本页下方另起一行注明论文的关键词（3-7个）。

关键词：　关键词1　关键词2　关键词3　关键词4

ABSTRACT

Externally pressurized gas bearing has been widely used in the field of aviation, semiconductor, weave, and measurement apparatus because of its advantage of high accuracy, little friction, low heat distortion, long life-span, and no pollution. In this thesis, based on the domestic and overseas researching……

Key words： Key Word 1, Key Word 2, Key Word, Key Word 4

目　　录

第一章 绪论

1.1 自动语音识别简介

语音交互是人类社会最直接、最自然的沟通交流方式，而机器作为辅助人类生产及日常生活的工具，目前人类与各种机器交互的方式更多的还是依赖于键盘、鼠标、显示器等输入输出设备。如何摆脱鼠标、键盘，使得人与机器之间的沟通像人与人之间的沟通那样自然，是智能时代人类面临的重大挑战。想要实现人机对话，需要涉及语音识别、自然语言理解、语音合成等关键技术。其中，语音识别作为关键部分之一，就像机器的耳朵一样，机器需要依靠它来辨别人类到底在说些什么。

近些年来，随着科技的发展，在安静环境下使用近距离麦克风的应用场合，语音识别已达到实用阶段。纵观语音识别的发展史，20世纪50年代，贝尔实验室成功研制出世界上第一个语音识别系统Audrey，方法是基于元音的共振峰的测量，虽然该系统为针对特定说话人的孤立词识别，且只能识别十个英文数字的发音，但这意味着语音识别的时代开启了。20世纪60年代至70年代之间，语音识别领域取得了突破性进展。线性预测编码（Linear Predictive Coding）被应用在声学特征的提取上；动态时间规整（Dynamic Time Warping）技术用来解决模板匹配时非线性时间对齐的问题。这些关键性的突破使得特定说话人的孤立词识别成为可能。20世纪80年代，语音识别的任务开始从孤立词识别转向连续语音识别，比如识别连续朗读的数字串等。这一时期的重大进展是语音识别方法从模板匹配转为基于统计模型方法，其中最突出的是隐马尔科夫模型（Hidden Markov Model），该模型基于马尔科夫假设，实现了对时间序列结构的建模。该方法从80年代中期开始逐渐被世界范围内的研究机构广泛接受并成为主流的语音识别方法，直到今天，很多成熟的大规模连续语音识别系统依然没有脱离HMM的方法框架。20世纪90年代出现了很多判别训练方法，包括最小识别误差MCE(Minimum Classification Error)和最大互信息MMI(Maximum Mutual Information)等。相比于最大似然估计的训练方法，这些判别训练方法能够提供更好的识别性能。自2006年Hinton等人提出有效的训练深度神经网络算法开始，深度学习技术逐渐流行并在多个领域取得显著成果。在语音识别领域，深度学习用来进行声学模型建模并获得巨大成功，尤其是对于大规模的识别任务而言。

这要得益于反向传播算法的使用，以及越来越多的计算资源和训练数据。

1.2 本文研究内容及各章节安排

自20世纪60年代开始，近60年的技术积累使得语音识别性能已达到实用阶段，在某些特定的语音识别任务上，机器甚至已经超过人类。尤其是近几年深度神经网络取代传统的GMM（Gaussian Mixture Model）模型，使得识别率得到历史性突破。然而，这些性能上的突破大多都是针对英语、普通话等语料充足且已经被广泛研究理解的语种。对于许多语料匮乏语种来说，语言识别还停留在很初级的阶段。比如以藏语拉萨方言为例，目前还没有公开的比较成熟的语料库，同时也缺乏相应的拉萨方言的语音学知识，且由于藏语本身语言特性复杂，训练一个实用的语言模型十分困难，这些问题导致现阶段几乎还没有实用的藏语大规模连续语音识别系统。目前现有的关于藏语语音识别的研究主要集中在特征提取以及使用动态贝叶斯网络构建声学模型上。应用深度学习技术来对藏语声学模型建模的研究还非常少，藏语识别的研究也处在刚刚起步的阶段。即使是用于训练藏语声学模型的音素集合都还没有一个统一的参考标准。本研究从录制拉萨方言平衡语料库开始，设计了拉萨方言发音字典，尝试使用GMM-HMM、DNN-HMM及Tandem等方法训练了声学模型，通过爬取网络上的藏语文本数据训练得到语言模型，搭建了离线的拉萨方言语音识别系统。并且首次探索了如何利用拉萨方言的声调信息提高识别准确率。

本论文的章节安排如下：第一张为绪论部分，简要介绍了语音识别的任务及发展史；第二章为背景介绍，主要讲述了语音识别系统的各个组成部分及评价指标，并且介绍了本工作涉及到的语音识别工具箱。第三章总结了训练声学模型的各种方法，包括传统的GMM-HMM、目前广泛使用的DNN-HMM方法以及逐渐兴起的LSTM方法；第四章详细描述了搭建拉萨方言语音识别系统的过程，包括声学模型和语言模型的训练，以及声调特征提取的相关实验；第五章为总结和展望。

第二章　　背景介绍

本章主要介绍语音识别的一些基本概念，包括前端的特征提取、声学模型、语言模型、解码器、性能评价指标，另外简单介绍了本研究涉及到的语音识别工具箱——Kaldi。

2.1　自动语音识别

语音识别，顾名思义，是要把人的声音转化成文本，目标是要尽可能多地识别出正确的词语。语音识别系统的结构可以用下图表示：

补充语音识别系统结构图

从图中可以看出，语音识别系统总共包含五个部分，输入的音频首先经过语音识别前端的特征提取部分，现假设语音长度为T，那么经过特征提取会得到一系列固定长度的频谱特征向量X_t，t=1,2,...T。语音识别的输出是一连串字符ω_k，k=1,2,...K。ω是系统认为最能匹配输入音频的文字序列。因此语音识别的目标可以表示为找到对应的ω使之满足

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmax}} \frac{P(X|\omega) * P(\omega)}{P(X)} \tag{2-1}$$

$P(\omega|X)$ 很难直接计算，根据贝叶斯公式

$$P(\omega|X) = \frac{P(X|\omega) * P(\omega)}{P(X)} \tag{2-2}$$

对于给定的输入X，P(X)对所有ω均为定值，因此

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmax}} P(X|\omega) * P(\omega) \tag{2-3}$$

其中，$P(X|\omega)$代表声学模型，$P(\omega)$代表语言模型。解码器根据声学模型和语言模型对ω的评分，搜索所有可能的ω，得到最优解。以下是语音识别系统各部分的详细介绍。

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。下面以语音识别中常用的MFCC参数为例，详细介绍声学参数的提取过程。

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:

$$f(x)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{L}) & 0\le x\le L-1\\ 0 & \text{otherwise.}\end{cases} \tag{2-4}$$

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:

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi}{N}kn} \tag{2-5}$$

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：

$$mel(f) = 1127\ln(1 + \frac{f}{700}) \tag{2-6}$$

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：

$$E(m) = \sum_{t=t_1}^{t_2} x^2[t] \tag{2-7}$$

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

2.1.6 评价指标

2.2 Kaldi工具箱

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　拉萨方言语音识别研究现状

4.2　拉萨方言数据库

4.3　发音词典

4.4　语言模型训练

4.5　声学模型训练

4.5.1　声调相关特征

4.5.2　模型融合

4.6　识别结果及分析

第五章　　总结与展望

参考文献

[1] 胡伟. LaTeX 2ε 完全学习手册 [M]. 北京: 清华大学出版社, 书号: 978-7-302-24159-1, 2011.

[2] 邓建松, 彭冉冉, 陈长松. LaTeX 2ε 科技排版指南 [M]. 北京: 科学出版社, 书号: 7-03-009239-2/TP.1516, 2001.

[3] Lamport L. LaTeX — A Document Preparation System: User's Guide and Reference Manual [M]. 2nd ed. Reading, Massachusetts: Addison-Wesley, 1985.

[4] Knuth D E. The TeXbook [M]. Reading, Massachusetts: Addison-Wesley, 1986.

[5] Knuth D E. Computer Modern Typefaces [M]. Reading, Massachusetts: Addison-Wesley, 1986.

[6] Bezos J. The titlesec and titletoc Packages [M]. 2nd ed. Cityname: University of SomeName, 2002.

[7] P Oostrum, ifuleyou@bbsctexorg 译. LaTeX下的页面布局 [M]. 天津: 某某大学出版社, 2001.

[8] Shell M. How to Use the IEEEtran LaTeX Class [J]. Journal of LaTeX Class Files, 2002, 1 (11): 10–20.

[9] TeXGuru. LaTeX2ε用户手册 [M]. 天津: 某某大学出版社, 1999.

[10] K Reckdahl 原著, 王磊 译. Using Import graphics in LaTeX2ε, LaTeX2ε插图指南 [M]. 天津: 某某大学出版社, 2000.

[11] McDonnell J R, Wagen D. Evolving Recurrent Perceptions for Time-Series Modeling [J]. IEEE Trans. on Neural Networks, 1994, 5 (1): 24–38.

[12] XYao. Evolutionary Artifitial Neural Networks [J]. J. Of Neural Systems, 1993 (4): 203–222.

[13] 宋乐. 异源图像融合及其评价方法的研究 [D]. 天津: 天津大学, 2008.

[14] Agrawal A, Raskar R. Resolving objects at higher resolution from a single motion-blurred image [C]. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007: 1–8.

[15] Zhang J, Li X, Chen J, et al. A tree parent storage based on hashtable for XML construction [C]. In Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference on, 2010: 325–328.

[16] SNiwa, Suzuki M, Kimura K. Electrical Shock Absorber for Docking System Space [C]. In IEEE International Workshop on Intelligent Motion Control, Istenbul, 1990: 825–830.