基于 GMM 符号化和置信判别的汉语方言自动辨识研究*)

沈兆勇[1] 顾明亮[1,2] 杨亦鸣[1]

(徐州师范大学语言研究所语言科学与神经认识工程江苏省重点实验室 徐州 221116)[1]

(徐州师范大学物理系 徐州 221116)[2]

摘 要 近年来汉语方言自动辨识研究有了初步进展,但由于缺乏带有语音标注的方言音库,性能优越的并行音素识别-语言模型(PPRLM)方法尚未得到研究和运用。本文借助高斯混合模型(GMM)符号化器把 PPRLM 的思想方法引入到汉语方言辨识中,并通过融合置信判别使系统能够用于开集辨识。仿真实验表明,本文方法具有很高的稳定性和可靠性,综合性能较为优越。

关键词 汉语方言自动辨识,PPRLM 方法,GMM 符号化,置信判别

A Study about Chinese Dialect Identification Based on GMM Tokenization & Confidence Measure

SHEN Zhao-Yong[1] GU Ming-Liang[1,2] YANG Yi-Ming[1]

(Institute of Linguistics, Xuzhou Normal University, Jiangsu College Key Lab of Linguistic Science and Neuro-cognitive Engineering, Xuzhou 221116)[1] (Department of Physics, Xuzhou Normal University, Xuzhou 221116)[2]

Abstract Lately the study of Chinese dialect identification (CDI) shows some progress. Yet the excellent method-parallel phone recognizers followed by language modeling(PPRLM)-has not be study in CDI field due to the lack of dialect corpus with annotation. In this paper, we study CDI using a method like PPRLM by virtue of GMM tokenizer, further we study the combination of a confidence measure to use the method in open-set task. Simulation results show that this CDI method is an excellent method with high stability and reliability.

Keywords Chinese dialect identification, PPRLM method, GMM tokenization, Confidence measure

1 引言

汉语方言自动辨识是计算机自动辨识一段汉语语音的方言种类的技术,在语音识别、信息检索、旅游服务、刑侦及军事监听等领域有着重要的应用价值。目前该研究尚处于起步阶段。2002 年以来,我国台湾的蔡伟和、新加坡的 B. P. Lim 等先后进行了基于高斯混合二元模型、融合全局特征等方法的相关研究[1,2]。由于缺乏带有语音标注的汉语方言音库,语种辨识领域中性能优越的并行音素识别-语言模型方法(PPRLM)[3,4]尚未得到应有的研究。另外,目前研究都是基于闭集强制判别的,在应用中缺乏稳定性和可靠性。本文通过引入高斯混合模型(GMM)符号化[5]避免了对语音库的标注要求,实现了 PPRLM 思想方法在汉语方言辨识中的应用。同时为增强系统的可靠性,本文比较选取了置信判别方法,使其能够应用到开集判别领域。仿真实验表明,基于 GMM 符号化和置信判别的汉语方言辨识方法具有较快的辨识速度和良好的辨识率,综合性能优于以往方法。

2 高斯混合模型(GMM)符号化

本文的符号化过程类似于 PPRLM 中的音素识别功能,只不过这里的训练语音数据不再需要标注。基于 GMM 符号化和置信判别方法的辨识系统结构如图 1 所示。它由语音信号预处理、特征提取、GMM 符号化、语言建模、分类器识别和置信判别 6 个部分组成。其中,语音信号预处理主要包括采样、量化、去噪、端点检测、预加重、分帧和加窗等操作。本文所用采样频率是 11kHz,量化级是 16bit,帧长为 256 点、帧移 128 点。特征提取则利用美尔倒谱系数(MFCC)计算办法得到,这里取 12 维 MFCC 加 12 维一阶差分美尔倒谱系数组成特征向量。

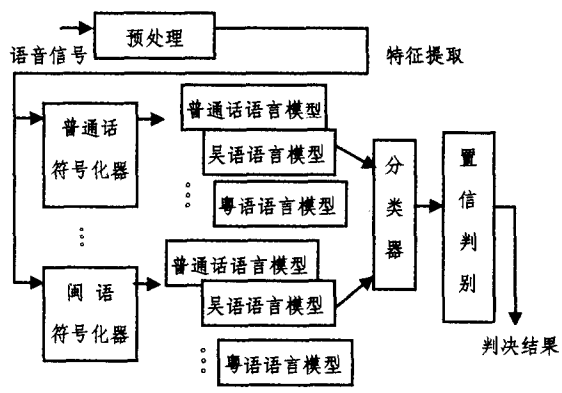

图 1 本文系统结构简图

2.1 GMM 符号化原理

符号化通过高斯混合模型(GMM)(也可以称作 GMM 符号化器)各子分布对语音帧的打分实现。GMM 在系统中主要用作对不同方言语音进行声学建模。利用 EM 算法[6]及每

*)本文得到了国家社会科学基金重点项目(01AYY004),江苏省社会科学基金项目(06J5BYY006),江苏省"十五"社科基金项目(K3-013),徐州师范大学人文社会科学基金项目(06XWB28)。**沈兆勇** 硕士,主要研究方向:语言信号处理方言自动辨识等;**顾明亮** 副教授,主要研究方向:信号处理、神经网络等;**杨亦鸣** 教授、博士生导师,主要研究方向:理论语言学、神经语言等。

一种方言的训练语音数据，可以训练得到各方言的模型参数：$\Omega_k=\{w_i,\lambda_i\}$，$\lambda_i=\{\vec{\mu_i},\sum_i\}$，这里 Ω_k 表示第 k 类方言的模型参数，w_i 为各高斯混合元的权重，它满足：$\sum_{i=1}^{M}w_i=1$，λ_i 表示第 i 个高斯分量参数集，其中的 $\vec{\mu_i}$ 是均值向量，$\sum_i$ 是协方差矩阵，$i=1,2,3,\cdots,M$，M 是高斯混合元的总数。在测试阶段，语音信号被逐帧解码：对于每一帧，GMM 依据公式 $v_i=\arg\max_{1<=j<=M}\{b(x_i\backslash\lambda_i)\}$ 输出其对应的符号（符号可以用数字或字母表示）。也就是说，GMM 对于输入的每一帧语音（特征向量）按照似然度指定一个声学空间的子空间，这里是一个高斯子分布，并输出相应子分布的标号。式中，x_i 为第 i 帧向量，λ_i 为第 j 个高斯分量（子分布）的参数（包括均值 $\vec{\mu}$、方差 Σ 和权重 λ），M 是高斯混合数，v_i 为第 i 帧向量所对应的符号。显然，符号化是通过上述最大后验概率把特征向量 x_i 和高斯分量 j 联系起来并输出符号 v_i。在系统中，这种符号化在无需语音标注的情况下实现了类似音素识别的功能。

2.2 语言模型

PPRLM 系统通过对音素识别得到的音素流进行语言建模，大大提高了系统的辨识正确率。本系统仿照 PPRLM 系统中的语言建模过程，对得到的符号串进行语言建模。本文语言模型部分是一个插值二元模型[7]，其概率计算可以由公式 $\hat{p}(v_t\backslash v_{t-1})=\lambda_0+\lambda_1\hat{p}(v_t)+\lambda_2\hat{p}(v_t\backslash v_{t-1})$ 给出，其中权重 λ_i 表示对每一部分概率的置信度，λ_0 为被估计概率的最小值阈值，概率 $\hat{p}(v_t)$ 对应于训练数据中符号 v_t 出现的概率，概率 $\hat{p}(v_t\backslash v_{t-1})$ 是 v_t 紧随 v_{t-1} 出现的二元概率。实验中，本文把权值设置为：$\lambda_0=0.001$，$\lambda_1=0.333$，$\lambda_2=0.666$。

2.3 后端分类器

后端分类器在基于音素识别（Phone Recognition，PR）方法的系统中被广泛采用，其目的是从上阶段语言模型组的打分中进一步获取区分信息。本文比较了两类较为常用的分类器：高斯分类器和人工神经网络（ANN，Artifical Neural Network）分类器。高斯分类器是把各语言模型的打分看成特征向量的一维分量，然后用一个多维均值和协方差的高斯模型来获取各类别对语言模型组打分的统计分布特征。ANN 分类器的设计则相对复杂和多样。本系统采用的是前馈型两层 BP 神经网络（NN）分类器。在设计方面，为了在网络学习中获得较快的收敛速度和克服类别之间的耦合，选用了单输出型神经网络[8,9]。具体做法是：首先为每个类建立一个单输出神经网络，然后对每一类进行分别训练，并将属于这一类样本的期望输出设为 1，而把属于其它类的期望输出设为 0。特别地，训练中必须用所有类所别训练数据对每一个网络模型进行有监督训练。在分类阶段，将未知类别的打分样本输入到每一个网络，然后把输出最接近 1 的网络的类别判定为决策类别。

3 置信判别

置信判别是对分类器的分类结果在一定的置信度下进行最后的接受或拒绝处理的一种方法，对于系统开集辨识条件下拒绝集外语音（非目标类别）、信噪比过低的语音或者非语音输入等具有重要作用。当输入语音为集外语音或带有较多的噪音等而难以做出可靠的判别时，拒绝判别是一个更为理想的选择。为融入有效的置信判别，提高系统的稳定性和可靠性，本文比较了两类共 3 种不同的置信判别方法。第一类是用所有非目标方言训练语音训练一个综合背景 CBG（Composite Background）语言模型[10]，一者这是因为各背景方言没有足够的数据来训练各自的背景语言模型，二者出于系统运算效率的考虑。第二类是不建立 CBG 而使用直接的在线置信判别，本文又分别采用了 GMM 声学打分的置信判别和语言模型（Language Model，LM）打分的置信判别两种方法。在融入 CBG 置信判别的系统中，置信判别是通过比较最优假设和 CBG 的打分差别实现的，而第二类不包括 CBG 系统是通过在不同阶段比较最优假设打分和训练数据平均打分的差别实现的。

4 实验

4.1 使用的汉语方言语音库

语音库的建库目的和用途不同，其选取的语料种类和数量也各不相同。在方言语音库方面，我国在上世纪 90 年代中后期建立了面向传统方言学研究的现代汉语方言语音库[11]，但专门面向方言辨识的方言语音库尚属空白。为此，参照 OGI 国际多语种语音库的设计建立方法[12]，本文建立了一个面向方言辨识的小型多说话人、非特定文本、连续语音、独白模式汉语方言语音库。语料包含个人介绍（姓名、性别、年龄、毕业学校、家庭住址、家乡特产、特色家乡话等）、方言调查词句和故事讲述 3 部分。语音库在广播系统播音室录制，以 11025Hz 采样、16bit 量化。该语音库包括现代汉语的 7 大主要方言[13]：北方方言、吴方言、粤方言、闽方言、湘方言、赣方言和客家话，分别以普通话、苏州话、广州话、厦门话、长沙话、南昌话和广东梅县话为代表，每种方言说话人设计为 10～12 人，男女比例为 1∶1。本文以前 4 种方言为辨识目标方言，其它方言及其次方言为集外方言。语音库分为 3 部分：训练集、开发集和测试集，分别用于训练 GMM 符号化器和语言模型、系统性能提升比如训练后端分类器以及测试整个系统的性能。训练集中前 4 种方言和作为整体的集外方言每种各有一个约 60 分钟的训练语料，测试集和开发集为 15 秒的语音段的集合，测试集中上述 5 种各有 60 段时长 15 秒的测试语音。同样地，开发集中也各有 60 段时长 15 秒的语音。以上 3 个语音集语音数据互不交叉重叠。

4.2 基线闭集强制判别系统实验

本实验主要是在闭集条件下测试 GMM（符号化器）的阶数、并行 GMM 符号化器的路数以及不同后端分类器对辨识率的影响。

首先考察 GMM 符号化器的阶数对系统辨识率的影响。这里 GMM 阶数分别取 16、32、64、128。由于语料规模的限制，128 阶以上本文暂未进行研究。本实验同时还对高斯分类器和 ANN 分类器在系统条件下做了比较。图 2 显示了 GMM 阶数和系统平均辨识率之间的关系，这里的平均辨识率在 4 路并行 GMM 符号化器系统下计算。实验表明，总体上系统性能有随着 GMM 阶数的增加而提升的趋势，但 GMM 阶数达到 32 以后提升速度明显放缓。这或许是因为增加 GMM 的阶数，能够对语音进行更为细致的分析和描绘，而增加到一定程度后，这种效果便趋于弱化。同时还可以看到，本系统中 ANN 分类器在 GMM 取 32 阶后分类效果开始优于高斯分类器，特别在 128 阶时高出 10 多个百分点。这表明在本系统，ANN 分类器有更强的分类能力和容错能力。

考察 GMM 符号化器并行路数对系统辨识率的影响，实验结果见表 1。和 PPRLM 系统一样，每一个符号化器都只

（下转第 236 页）

复杂情况，抗干扰性能力很强，并且此系统还具有成本低、监控范围广、响应时间快、可以使用传统的二总线网络，不需更换网络等多种优点。但是系统还是有一定误报，可以在此系统中加入传统感温、感烟等传感器，有望进一步减少误报率，提高稳定性。

参 考 文 献

1 易继锴，张蔚蔚. 模糊神经网络技术及其在火灾探测中的应用. 北京工业大学学报，2001，27(3)：337

2 卢瑞祥，牟轩沁，等. 一种基于红外图像识别的自动消防监控系统. 电子技术应用，1998，2：8

3 季萍，卢结成. 一种总线制火灾图像探测系统的设计与实现. 计算机工程与应用，2004，26：212

4 宋卫国，范维澄，吴龙标. 基于人工神经网络的火灾图像探测方法. 火灾科学，1999，8(3)：49

(上接第 211 页)

用一种方言语音训练，符号化器数目采取从吴方言到粤方言、普通话、闽方言逐一增加的方式。实验中 GMM 符号化器阶数均取为 128，分类器采用 ANN 分类器。表 1 表明，本系统 4 路并行时系统稳定性和辨识率最好，而 1 路单符号化器时性能最差。这表明增加符号化器数目能够更好地刻画语音中的音素和类音素搭配规律。

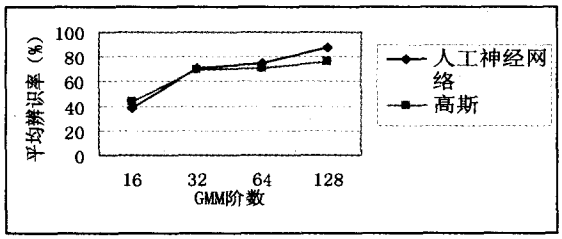

图 2 不同 GMM 阶数及两类分类器的比较

表 1 不同路数 GMM 并行时系统性能的比较

性能波动	不同路数 GMM 并行辨识率(%)			
	单路	2 路	3 路	4 路
最低	15.2	35.5	57.0	70.2
平均	32.5	60.2	67.5	87.3
最高	55.0	82.1	88.3	95.2

4.3 基于置信判别的开集实验

这部分开集条件下的实验用来测试和比较上文提到的两类 3 种置信判别方法。实验中系统采用 128 阶 GMM4 路并行方式，置信判别的模型和阈值用开发集的数据训练。由于 CBG 的加入，使用 CBG 的高斯分类器和人工神经网络分类器用 20 维的打分向量训练，而不用 CBG 的两类分类器都用 16 维的打分向量训练。实验结果如表 2 所示。表 2 表明，系统使用 CBG 置信判别时的性能明显优于使用一般置信判别的系统，这说明背景模型在有置信判别的使用 PPRLM 方法的系统中是非常重要的模块。在分类器方面，本实验再次验证了该系统中 ANN 分类器效果要优于高斯分类器。

表 2 两类分类器下不同置信方法的比较

置信方法 分 类		高斯分类器辨识率(%)		ANN 分类器辨识率(%)	
		总体	置信	总体	置信
不用 CBG	GMM	47.7	58.6	62.0	69.4
	LM	61.7	72.5	69.8	76.7
CBG		71.5	82.5	75.4	86.2

结束语 本文通过引入 GMM 符号化研究了语种辨识的 PPRLM 思想方法在汉语方言自动辨识中的应用，并在此基础上研究比较了开集条件下的置信判别。实验结果表明基于 GMM 符号化的类 PPRLM 汉语方言自动辨识方法辨识率高、运算效率突出而且对训练语音库没有标注要求，是一种具有很强扩展性和移植性的综合性能优越的方法。置信判别的比较实验表明，基于 CBG 的置信判别是类 PPRLM 方法在开集应用中的重要组成部分，它增强了系统的稳定性和可靠性，使其能够应用到开集辨识领域。

参 考 文 献

1 Tsai Wuei-He, Chang Wen-Whe. Discrimination Training of Guassian Mixture Bigram Models with Application to Chinese Dialect Identification [J]. Speech Communication 2002, 36: 317～326

2 Lim Boon Pang, Li Haizhou, Ma Bin. Using Local & Global Phonotactic Features in Chinese Dialect Identification [C]. In: Proc. of ICASSP'05, 2005, 1: 577 ～580

3 Muthusamy Y K, Barnard E, Cole R A. Reviewing automatic language identification [C]. IEEE Signal Processing Mag, 1994, 11 (4): 33～3

4 Zissman M A. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech [C]. IEEE Trans. Speech and Audio Pro, 1996, 4(1): 31～34

5 Torres-Carrasquillo P A, Reynolds D A, Deller J R Jr. Language identification using Gaussian mixture model tokenization [C]. In: Proc. of ICASSP 2002, 12002: 757～760

6 Laird N M, Lange N, Stram D. Maximum Likelihood Computations with Repeated Measures: Applications of the EM algorithm [J]. Journal of the American Statistical Association, 1987, 82: 97 ～105

7 Jelinek F. Statistical Methords for Speech Recognition [M]. Cambridge, Massachusetts, MIT Press, 1999

8 Biederman D C, Ososanya E. Capacity of several neural networks with respect to digital adder and multiplier System Theory [C]. In: Proc. of the Twenty-Seventh Southeastern Symposium on Neural Network, 1995. 305～308

9 赵 力. 语音信号处理[M]. 北京：机械工业出版社，2001

10 Gleason T P, Zissman M A. Composite background models and score standardization for language identification system [C]. ICASSP, 2001(1): 529～532

11 侯精一主编. 现代汉语方言音库[M]. 上海教育出版社，1994～1999

12 Muthusamy Y K. A segmental approach to automatic language identification: [Doctor thesis]. Hyderabad, Indian: Jawaharlal Nehru Technological University, 1993

13 黄伯荣，廖序东主编. 现代汉语[M]. 增订二版. 北京：高等教育出版社，1997