基于多空间概率分布的汉语连续语音声调识别研究

倪崇嘉[1,2] 刘文举[2] 徐 波[2]
(山东财政学院统计与数理学院 济南 250014)[1]
(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)[2]

摘 要 汉语是一种带声调的语言,声调信息在汉语语音识别中具有非常重要的意义。提出了embedded声调模型与explicit声调模型相结合的方法用以识别汉语连续语音的声调。该方法能够将逐帧的基频信息和较强时长的基频信息相结合来识别声调。在"863-Test"和"TestCorpus98"测试集上的实验表明,该方法分别能够达到96.12%和93.78%的声调识别正确率。

关键词 声调,基频,多空间概率分布
中图法分类号 TP319 **文献标识码** A

Research about Tone Recognition of Mandarin Continuous Speech Based on Multi-space Probability Distribution

NI Chong-jia[1,2] LIU Wen-ju[2] XU Bo[2]
(School of Statistics and Mathematics, Shandong University of Finance, Jinan 250014, China)[1]
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract Chinese Mandarin is the tonal language. Tone is important to Mandarin speech recognition. We proposed a method to recognize the tone of Mandarin continuous speech, which is the combination of embedded tone model and explicit tone model. This method can fuse the fundamental frequency information of short time and long time. The experiments in "863-Test" and "TestCorpus98" test show that our proposed method can achieve 96.12% and 93.78% tone recognition correct rate separatively.

Keywords Tone, Fundamental frequency, Multi-space probability distribution

1 引言

汉语是单音节结构的带调语言,共有5种声调,即阴平、阳平、上声、去声和轻声。声调在汉语词义辨识中担当了重要的角色,例如,"妈/ma1/","麻/ma2/","马/ma3/"和"骂/ma4/",如果声调变了,那么整个词的意义就变了。汉语中共有1300多个带调的音节,如果去掉声调,汉语中音节有400多个。通过对汉语词典的研究表明,如果去掉声调,汉语中大约有30%的词是同音词,如果考虑了声调,那么同音词的数量将大大降低。文献[1]中的实验结果表明:加入声调后同音词的数目由平均6.8个降到3.2个。因此,如果能够准确地获得声调信息,将会大大提高汉语语音识别的性能,声调的建模和识别成为汉语语音识别研究中重要的研究内容。多年来,许多学者对声调识别做了大量的研究,并取得了很多研究成果。在研究方法上可以分为两种:explicit tone recognition和embedded tone modeling[2]。explicit tone recognition就是通过训练数据,建立声学模型,然后通过声学信号单独识别声调[3-6]。声调识别的结果还可以对语音识别的结果进行后处理。而embedded tone modeling则是通过添加基频分量以及基频的一阶、二阶差分到已有的声学特征,如MFCC、PLP等,通过已有的语音识别框架对声调进行识别[7,8]。并且,在最近几年,关于汉语声调识别的研究已经扩展到其他的汉语方言,如广东话(或粤语)的声调识别[2,9,10]。在explicit tone recognition方面,建立声调模型的方法主要有决策树(Decision tree, DT)[6]、神经网络(Neural Network, NN)[11]、隐马尔可夫模型(Hidden Markov model, HMM)[4]、支持向量机(Support Vector Machine, SVM)[10]和高斯混合模型(Gaussian mixture model, GMM)[9,12]。最近,在声调识别方面取得较大进展的研究分别是基于支持向量机(Support Vector Machine, SVM)[10]和高斯混合模型(Gaussian mixture model, GMM)[9,12]。文献[10]利用SVM对连续语音的广东话声调进行识别,其采用自适应的五度制的方法对基频进行正则化以减少或消除基频的动态变化对声调识别的影响,同时上述的正则化的方法还考虑了语调的影响对基频进行正则化,该方法对广东话声调识别的正确率能够达到71.50%,较之前的对广东话的声调识别率有了较大提高。文献[9]提出了超级声调模型(supratone model),考虑了在话语中上下文的影响,对超级声调模型利用GMM进行建模,对广东话的声调进

到稿日期:2011-03-08 返修日期:2011-05-01 本文受国家自然科学基金(90820303,60675026,90820011),国家高技术研究863计划(2006010Z4073,2006AA01Z194)和国家重点基础研究发展973计划(2004CB318105)资助。

倪崇嘉(1979—),男,博士,讲师,主要研究领域为语音识别韵律模型;刘文举(1960—),男,博士,研究员,博士生导师,主要研究领域为语音识别;徐 波(1966—),男,博士,研究员,博士生导师,主要研究领域为语音识别、数字内容理解。

行识别。文献[12]则利用 TRUES(Tone Recognition Using Extended Segments)对连续语音的普通话声调进行识别。其是利用动态规划的方法对时间域上的 AMDF 函数抽取连续的基频曲线,然后建立 tri-supratone 模型,该方法能够获得85.07%的声调识别率。

embedded tone modeling 方法主要是考虑了将基频作为输入特征加入到已有的语音识别系统框架内,是逐帧的。我们知道声调是负载到音节上的,并且如果考虑到上下文影响,声调的时间跨度更广,可能是数十帧。因此,embedded tone modeling 方法在刻画长时信息方面是有缺陷的。另外,在传统的声调识别的研究中,很多研究是首先由语音识别系统获得每一个音节的时间切分边界,然后再建立 explicit tone 模型进行声调识别。explicit tone 模型也有不足,那就是一些有用的通过语音识别系统获得的信息没有得到充分的利用,如当前音节的上下文声调等。本文将结合 explicit tone modeling 和 embedded tone modeling 两种方法对连续语音中的声调进行识别。我们首先利用多空间概率分布(Multi-Space Probability Distribution,MSD)建立 embedded tone modeling,然后根据语音识别系统获得的每一个音节时间切分信息,利用 explicit 声调模型对 embedded 声调模型的结果进行修正,以获得最终的声调。图 1 给出了系统的总体框架。

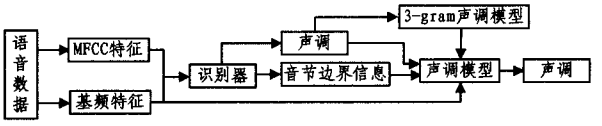

图 1 声调识别系统的总体框架

本文在第 2 节介绍基于多空间概率分布的 embedded 声调模型;第 3 节介绍 explicit 声调模型,主要包括输入的特征、建模的方法等;第 4 节给出实验结果;最后给出结论以及下一步的工作。

2 基于多空间概率分布的 embedded 声调模型

MSD 是由 K. Tokuda 等人提出的,用于建模随机的分段连续的基频轨迹,并成功地应用到基于 HMM 的语音合成中[13,14]。

在 MSD 中它考虑样本空间 Ω 由 G 个子空间构成:

$$\Omega=\bigcup_{g=1}^{G}\Omega_g \tag{1}$$

式中,Ω_g 是一个 n_g 维的实空间 R^{n_g},g 是空间索引值,每个空间都有自己的维度 n_g,其中一些空间可以有相同的维度。

对每一个空间 Ω_g,它都有一个相应的空间权重 w_g(或称为此空间出现的概率),即 $P(\Omega_g)=w_g$,其中,$\sum_{g=1}^{G}w_g=1$。如果 $n_g>0$,则空间对应有一个概率密度函数 $N_g(x)$,$x\in R^{n_g}$,其中,$\int N_g(x)\mathrm{d}x=1$。如果 $n_g=0$,我们假设 Ω_g 只有一个样本。

对于任意一个事件 E,此空间的样本 o 由两部分组成,空间索引集 X 和一个随机变量 $x\in R^n$,即 $o=(X,x)$(注意这里 X 指定的所有子空间必须是 n 维的,但并不是说 X 必须包含样本空间 Ω 中所有维度为 n 的子空间)。则观测到 o 的概率定义为:

$$b(o)=\sum_{g\in S(o)}w_gN_g(V(o)) \tag{2}$$

式中,$S(o)=X$,$V(o)=x$。需要注意的是,当 $n_g=0$ 时,$N_g(x)\equiv 1$。

我们知道,汉语是一种带调的语言,声调有重要的辩义作用。因此,将基频作为输入特征已经广泛应用到汉语的语音识别系统中。由于基频在一些清音段和静音段是不存在的,因此,基频仅仅是一个分段连续的变量。如果将基频不做修改直接作为特征用于 HMM/GMM 建模,会造成零方差。因此,很多人也给出了解决问题的方法,如将基频曲线连续化等。但是,这就有一个问题:本来静音段没有基频,而仅仅是出于计算的考虑才给它赋值,这不是很合理。

MSD 很好地解决了这一问题。对于有声的区域,基频被认为是一维来自几个一维 Gaussian 子空间的观测值序列。对于无声的区域,基频仅仅被视为一个 0/1 开关符号。在当前大多数语音识别系统中,一般采用混合高斯对输出的分布进行建模。因此,MSD 假设:在零维无声的子空间,输出的分布是 Kronecker δ 函数;在一维有声的子空间,分段连续的基频通过混合高斯建模。

3 Explicit 声调模型

第 1 节介绍了很多建模的方法,如 GMM、SVM、MLP 等。本节首先给出建立 explicit 声调模型时所用到的特征,然后介绍建模的方法以及该 explicit 声调模型与 embedded 声调模型的结合。

首先,提取下列特征:

PPTone:当前音节之前第二个音节的声调;

PTone:当前音节之前第一个音节的声调;

fPro:3-gram 声调模型的概率;

PMean1:当前音节之前音节被平均分成三份以后的第一部分音高的平均值;

PMean2:当前音节之前音节被平均分成三份以后的第二部分音高的平均值;

PMean3:当前音节之前音节被平均分成三份以后的第三部分音高的平均值;

CMean1:当前音节被平均分成三份以后的第一部分音高的平均值;

CMean2:当前音节被平均分成三份以后的第二部分音高的平均值;

CMean3:当前音节被平均分成三份以后的第三部分音高的平均值;

FMean1:当前音节之后音节被平均分成三份以后的第一部分音高的平均值;

FMean2:当前音节之后音节被平均分成三份以后的第二部分音高的平均值;

FMean3:当前音节之后音节被平均分成三份以后的第三部分音高的平均值。

总共有 12 个特征。

然后,利用多层感知机(Multiple Layer Perception,MLP),采用有监督的机器学习建立基于 MLP 的 explicit 声调模型。训练用到的数据主要是从训练语音识别系统的大规模的连续语音库中获得。

最后,利用训练好的 explicit 声调模型对 embedded 声调模型识别的结果进行修正。如果利用 explicit 声调模型识别出的声调的概率大于某个阈值,则修改该声调;否则,不对声调进行修改。

4 实验

4.1 语料库与实验设置

训练语料主要包括“863”连续语音库和 Intel 连续语音库中北京口音的语料。“863”连续语音库是近年比较权威的用于汉语语音识别开发的语音库。它包括 200 说话人(100 男声,100 女声),每人 520 到 625 句,覆盖了 2185 个连续语句。说话人来自北京等 6 省 2 市,没有明显的口音。文本取自《人民日报》,考虑了语料的声学平衡和覆盖性。语料库录音环境为安静的实验室环境。语料采样率 16kHz,16bit 量化。每句语音文件前后保留了大概 1 秒的静音段。实验中,使用了 83 男声数据作为标注和训练语料(48373 句,55.6 小时)。Intel 连续语音库包含了来自北京等 6 个城市的带有方言口音说话人的普通话语音。实验中,我们只采用带有北京口音的男声语音(45759 句,55 小时左右)。

测试语料包含两部分,一部分是“863-Test”,包括 240 句话,17.1 分钟,共 6 个男声说话人的测试语料,其录音环境和“863”连续语音库语料一致。另一部分是“Test Corpus98”,包括 779 句话,46.6 分钟,共 13 个说话人的测试语料,13 个说话人语音没有在训练语料中出现。

在基线语音识别系统 Baseline 中,建模单元采用声韵母,其中声母 24 个,韵母 37 个以及一个静音。每一个韵母有 5 个调:阴平、阳平、上声、去声和轻声。在此基础上的音素集包括静音(去除训练集中没有出现的音素)共有 205 个。模型是采用上下文相关的三音子模型。实验中采用的特征为 12 维 MFCC 参数,以及归一化能量和这些参数的一阶、二阶差分,共 39 维。

在结合 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 中,建模单元的选择、音素以及音素集中包含音素的个数与基线系统一样。结合 embedded 声调模型的语音识别系统的输入特征除了包括 39 维 MFCC 特征以外(该部分输入特征与基线系统一样),还包括了基频以及其一阶、二阶差分。因此,该语音识别系统的输入特征是 42 维。在训练该识别系统时,分两个流,一个流是 39 维 MFCC 特征,另一个流是 3 维基频特征。我们对这两个流设计了两个不同的问题集,并且在利用决策树绑定时,进行流相关状态绑定。

实验中用到的语言模型是包含 47489 个词的三元语言模型,大约 255M。

对于训练多层感知机 MLP 时,设置隐层的个数是 1,隐层中包含结点的个数是输入层结点的一半。

4.2 实验结果

表 1 列出了基线系统 Baseline 和结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 在测试集上关于汉字的识别率。

表 1 基线系统 Baseline 和结合 embedded 声调模型的系统 Tri-Pho-MSD-HMM 的汉字的识别率

系统	测试集	正确率(%)	替换率(%)	删除率(%)	插入率(%)	错误率(%)
Baseline	863-Test	89.41	10.24	0.35	0.29	10.87
	TestCorpus98	84.90	14.43	0.67	0.33	15.42
Tri-Pho-MSD-HMM	863-Test	91.48	8.36	0.16	0.35	8.87
	TestCorpus98	87.19	12.18	0.63	0.38	13.19

从表 1 可以看到:(1)在“863-Test”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的正确率比基线系统 Baseline 高 2.07%,错误率比基线系统 Baseline 降低了 2%;(2)在“TestCorpus98”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的正确率比基线系统 Baseline 高 2.29%,错误率比基线系统 Baseline 降低了 2.23%;(3)无论在“863-Test”测试集,还是在“TestCorpus98”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 要比基线系统 Baseline 好。

由识别出的汉字,可以得到两个系统的声调的识别率,表 2 列出了基线系统 Baseline 和结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的声调的识别率。

表 2 基线系统 Baseline 和结合 embedded 声调模型的系统 Tri-Pho-MSD-HMM 的声调的识别率

系统	测试集	正确率(%)	替换率(%)	删除率(%)	插入率(%)	错误率(%)
Baseline	863-Test	93.35	5.44	1.21	1.14	7.79
	TestCorpus98	91.07	6.90	2.02	1.68	10.60
Tri-Pho-MSD-HMM	863-Test	95.20	4.29	0.51	0.70	5.50
	TestCorpus98	92.82	5.92	1.25	1.01	8.18

从表 2 可以看到,(1)在“863-Test”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的声调识别的正确率比基线系统 Baseline 高 1.85%,错误率比基线系统 Baseline 降低了 2.29%;(2)在“TestCorpus98”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的声调识别的正确率比基线系统 Baseline 高 1.75%,错误率比基线系统 Baseline 降低了 2.42%;(3)无论在“863-Test”测试集,还是在“TestCorpus98”测试集上,结合了 embedded 声调模型的语音识别系统 Tri-Pho-MSD-HMM 的声调识别的正确率要比基线系统 Baseline 高,错误率比基线系统 Baseline 要低。

下面,利用基于 MLP 的 explicit 声调模型对识别出来的声调进行修正,表 3 列出了结合了 explicit 声调模型的声调的识别结果。

表 3 基线系统 Baseline 和结合 embedded 声调模型的系统 Tri-Pho-MSD-HMM 分别与 explicit 声调模型结合以后声调的识别结果

系统	测试集	正确率(%)	替换率(%)	删除率(%)	插入率(%)	错误率(%)
Baseline	863-Test	94.37	4.36	1.27	1.21	6.84
	TestCorpus98	91.59	6.31	2.10	1.76	10.17
Tri-Pho-MSD-HMM	863-Test	96.12	3.37	0.51	0.70	4.58
	TestCorpus98	93.78	4.98	1.24	1.00	7.22

从表 3 可以看到:(1)Tri-Pho-MSD-HMM 系统结合了 explicit 声调模型之后关于声调识别的正确率比基线系统 Baseline 结合了 explicit 声调模型之后的声调识别的正确率要高。在“863-Test”测试集上,正确率提高了 1.75%,在“TestCorpus98”测试集上,正确率提高了 2.19%。(2)Tri-Pho-MSD-HMM 系统结合了 explicit 声调模型之后关于声调识别的错误率比基线系统 Baseline 结合了 explicit 声调模型之后的声调识别的错误率要低。在“863-Test”测试集上,错误率降低了 2.26%,在“Test-Corpus98”测试集上,错误率降低了 2.95%。

对比表 3 和表 2,可以看到:(1)结合 explicit 声调模型的基线系统 Baseline 的声调识别的正确率比基线系统 Baseline

(下转第 241 页)

参 考 文 献

[1] Ehrlich P R, Raven P H. Butterflies and plants: a study in co-evolution[J]. Evolution, 1964, 18(4): 586-608

[2] Li Z Y, Tong T S. Research on ANN evolutionary design method based on populations evolution niche genetic algorithm[J]. Control and Desison, 2003, 18(5): 607-610

[3] Potter M A. The design and analysis of computational model of cooperative coevolution[D]. Virginia: Geroge Mason University, 1997

[4] Multi-objective cooperative coevolution of artificial neural networks[J]. Neural Networks, 2002: 1259-1278

[5] 董红斌,黄厚宽,印桂生,等. 协同演化算法研究进展[J]. 计算机研究与发展,2008,45(3):454-463

[6] 巩敦卫,孙晓燕. 变搜索区域多种群遗传算法[J]. 控制理论与应用,2006,23(2):256-260

[7] Li Bi, Yong Zheng-zheng. A high-efficient parallel genetic algorithm based on multi-level competition[J]. Acta Eleetronica Sinica, 2002, 30(12A): 2161-2162

[8] Cao X B, Li J L. Reserach on coevolutionary optimization based on ecological cooperation[J]. Journal of software, 2001, 12(4): 521-528

[9] Potter M. The Design and Analysis of a Computational Model of Cooperative Coevolution[D]. Fairfax, VA, USA: George Mason University, 1997

[10] Fan Hui-yuan, Wang Shang-jin, Xi Guang. Directional Evolution Operator Applied to Genetic Algorithm[J]. Journal of Xi'an Jiaotong University, 1999, 33(5): 45-49

[11] 陈国良. 遗传算法及其应用[M]. 北京:人民邮电出版社,1996: 1-433

[12] 王文义,秦广军,王若雨. 自适应的多种群并行遗传算法研究[J]. 计算机工程与应用,2006(15):34-36

[13] Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes[C]// Proceedings of the 15th International Conference on Data Engineering. Sydney, Australia, 1999: 1-11

[14] 赵志强,等. 基于个体适应度梯度的定向进化算法[J]. 模式识别与人工智能,2010,23(1):30-34

[15] Rosin C D, Belew R K. Methods for competitive co-evolution, finding opponents worth beating [C] // Proc of International Conference on Genetic Algorithms. San Diego: Morgan Kaufman, 1995: 256-328

[16] 苗金凤,王洪国,邵增珍,等. 基于多级搜索区域的协同进化遗传算法[J]. 计算机应用研究,2010,27(9):3345-3347

(上接第 226 页)

的声调识别的正确率在"863-Test"和"TestCor-pus98"测试集上分别高了 1.02%和 0.52%。结合 explicit 声调模型的基线系统 Baseline 的声调识别的错误率比基线系统 Baseline 的声调识别的错误率在"863-Test"和"TestCorpus98"测试集上分别低了 0.95%和 0.43%。(2)结合 explicit 声调模型的系统 Tri-Pho-MSD-HMM 的声调识别的正确率比系统 Tri-Pho-MSD-HMM 的声调识别的正确率在"863-Test"和"TestCor-pus98"测试集上分别高了 0.92%和 0.96%。结合 explicit 声调模型的系统 Tri-Pho-MSD-HMM 的声调识别的错误率比系统 Tri-Pho-MSD-HMM 的声调识别的错误率在"863-Test"和"TestCorpus98"测试集上分别低了 0.92%和 0.96%。

总之,通过结合 explicit 声调模型,使得结合了 explicit 声调模型的声调识别系统的性能比单独的 embedded 声调模型的系统要好。

结束语 本文提出了 explicit 声调模型和 embedded 声调模型结合的方法用以识别连续语音的声调。该方法不仅能够利用逐帧的短时基频信息识别声调,还结合了较长时间的基频信息识别声调。在"863-Test"和"TestCorpus98"测试集上的实验表明,该方法在"863-Test"和"TestCorpus98"测试集上声调识别的正确率分别能够达到 96.12%和 93.78%。今后,我们将探索其他的声调识别的方法以及自然口语中的声调识别问题。

参 考 文 献

[1] Xu Bo, Gao Sheng, Cao Yang, et al. Integrating tone information in continuous Mandarin recognition [C] // Proc. ISSPIS'99. Guangzhou

[2] Lee T, Lau W, Wong Y W, et al. Using tone information in Cantonese continuous speech recognition [J]. ACM Trans. on Asian Language Information Processing, 2002, 1(1): 83-102

[3] Liu J, Yu T. New tone recognition methods for Chinese continuous speech [C]// Proc. ICSLP'00. Beijing

[4] Zhang J-S, Hirose K. Anchoring hypothesis and its application to tone recognition of Chinese continuous speech [C]// Proc. ICASSP 2000. Istanbul, June 2000

[5] Peng G, Wang W S Y. An innovative prosody modeling method for Chinese speech recognition [J]. International Journal of Speech Technology, 2004, 7(4): 129-140

[6] Cao Y, Deng Y, Zhang H, et al. Decision tree based Mandarin tone model and its application to speech recognition [C]// Proc. ICASSP 2000. 2003, 3: 1759-1762

[7] Sun Y, Willett D, Brueckner R, et al. Experiments on Chinese speech recognition with tonal models and pitch estimation using the Mandarin speech data [C]// Proc. ICSLP. 2006: 1245-1248

[8] Seide F, Wang N. Two-stream modeling of Mandarin tones [C]// Proc. ICSLP. 2000: 867-870

[9] Qian Y, Song F K, Lee T. Tone-enhanced generalized character posterior probability(GCPP) for Cantonese LVCSR [J]. Computer Speech and Language, 2008, 22(4): 360-373

[10] Peng G, Wang W S Y. Tone recognition of continuous Cantonese speech based on support vector machines [J]. Speech Communication, 2005, 45(1): 49-62

[11] Chen S-H, Wang Y-R. Tone recognition of continuous Mandarin speech based on neural networks [J]. IEEE Trans. on Speech and Audio Proc, 1995, 3(2): 146-150

[12] Chen J-C, Jang J-S R. TRUES: Tone recognition using extended segments [J]. ACM Trans. on Asian Language Information Processing, 2008, 7(3): 10-33

[13] Tokuda K, Masuko T, Miyazaki N, et al. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling [C]// IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 1999

[14] Tokuda K, Masuko T, Miyazaki N, et al. Multi-Space Probability of Distribution of HMM [J]. IEICE Trans. Inf. and Sys, 2002, E85-D(3): 455-464