ISSN 1000-0054 清华大学学报 (自然科学版) 2005 年 第 45 卷 第 7 期 13/38
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. 7 912-915

快速口音自适应的动态说话人选择性训练

董 明, 刘 加, 刘润生

(清华大学 电子工程系, 北京 100084)

摘 要: 为解决语音识别系统实用中的说话人口音快速自适应问题, 提出了一种动态说话人选择性训练方法。基于说话人选择性训练方法, 采用基于 Gauss 混合模型似然分数计算的置信测度选择训练用说话人, 改变训练用说话人的绝对数目选取方式, 提高了选取的效能并拓展了选取标准的推广性。根据各个训练用说话人同被适应说话人的不同似然程度, 加权地合成动态说话人选择性训练的语音模型, 提高了自适应训练的效果。实验表明: 该方法使识别率从 80.16% 提高到 84.12%, 相对误识率降低了 19.96%, 在实用中提高了基线系统的识别性能。

关键词: 语音识别; 说话人快速自适应; 置信测度

中图分类号: TN 912.34 **文献标识码**: A

文章编号: 1000-0054(2005)07-0912-04

Dynamic speaker selected training for rapid speaker adaptation

DONG Ming, LIU Jia, LIU Runsheng

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Practical speech recognition systems need rapid speaker adaptation to be effective with a wide variety of speakers. A dynamic speaker selected training method developed for rapid speaker adaptation improves the basic speaker selected training method by replacing the absolute number selection method used in the basic method with a confidence measure calculated from the Gaussian mixture model likelihood. The new method enhances both the training speaker selecting efficiency and the selecting adaptability. The dynamic acoustic model, which uses different weightings for each training speaker so that they resemble the adapted speaker, further increases the recognition accuracy rate. Simulation show that the dynamic method improves the baseline recognition accuracy rate from 80.1% to 84.1%, with a decrease of 19.96% in the relative error rate. Thus, the dynamic method rapidly increases practical speech recognition system performance.

Key words: speech recognition; rapid speaker adaptation; confidence measure

统计方法是目前大词汇量非特定人连续语音识别系统的最主流方法。为了使一个非特定人(speaker independent, SI)语音识别系统能够有好的稳健性, 用于进行系统训练的语音样本中应该尽量包含各种声学特点不同的说话人, 这样的语音库在采集上成本非常高, 而且也很难完整建立。一般认为针对某特定说话人时, 同该说话人相对应的说话人相关(speaker dependent, SD)识别系统相比 SI 系统, 相对误识率可以降低一半[1], 因此研究快速说话人自适应算法具有重要意义与实用价值。

说话人口音自适应技术(speaker adaptation, SA)是改善 SI 系统性能的一个有效手段。经典的 SA 算法主要包括最大后验概率方法(maximum a posteriori, MAP)和最大似然线性回归方法(maximum likelyhood linear regression, MLLR)[2,3]。这两种方法都需要一定规模的自适应语料数据量, 标准MAP 方法需要上百句的自适应语音才能达到较好的效果, MLLR 使用的音素模型归类可以有效减少其对自适应数据的需求, 大约在数十句自适应语音上可以取得较好效果。然而在实际的运行中, 应用需求希望 SA 算法能够在 3～5 句甚至只有被识别的当前句这样少的自适应语料状况下工作起效, 达到所谓的"快速自适应"(指系统需要的自适应数据量小)。

传统的MAP 和MLLR 方法对于快速自适应问题都几乎无能为力。目前对于快速自适应问题的解决方案主要集中在分类的思想上, 通过被适应说话人的自适应语料, 找到他所归属的类别或与之相近的其他表征说话人, 进一步使用类别中的附加信息量达到快速自适应的效果。

本文对一种基于分类思想的快速自适应方法,

收稿日期: 2004-06-09
基金项目: 国家自然科学基金资助项目 (60272016)
作者简介: 董明(1978-), 男(汉), 辽宁, 博士研究生。
通讯联系人: 刘润生, 教授, E-mail: lrs-dee@tsinghua.edu.cn

说话人选择性训练方法(speaker selection training, SST)[4,5]进行了更深一步的拓展, 将置信测度全面引入到SST 方法中而发展出动态说话人选择性训练方法(dynamic speaker selection training, DSST)。可以从性能和实用性两个方面综合地提高快速自适应的效果。

1 说话人选择性训练的基本系统

各种基于分类思想的快速自适应方法, 都出自于一个基本事实: 对于特定的被适应说话人来说, 训练集中的多个说话人总是有一些同他比较相近(发音方式、说话方式、语气语速等), 而另一些同他比较不同。SST 方法是依据自适应语料直接从训练集(N 人)中挑选出与之最相近似的M 人($M < N$), 进而使用这M 个选出的训练用说话人的统计信息合成出被适应说话人的语音模型。我们具体使用的基本SST 方法流程见图1。

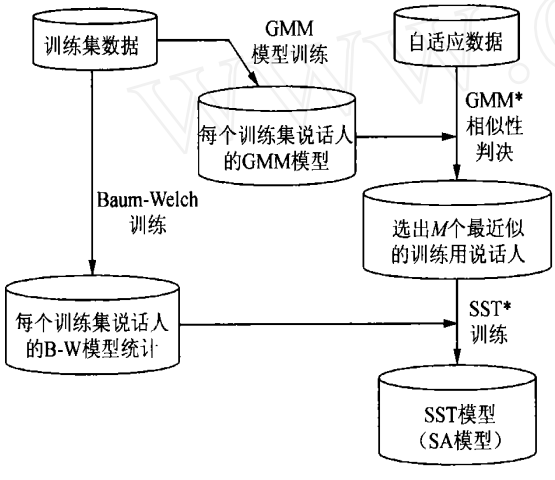

图1 基本SST 方法的流程

针对一个实际系统来分析, SST 方法仅仅将原来系统中使用的SI模型替换成SST 训练得到的模型, 对于系统其他环节完全没有影响。用于原系统的前端特征处理, 后端识别模型调整的方法仍旧可以使用。进一步地, SST 模型还可以在自适应语料增加后继续结合进行MLLR 等其他SA 的方法。SST 方法基本上可以没有任何副作用地对于系统性能产生提升, 这是SST 方法最突出的优点。

对于SST 算法而言, 有两个关键性的步骤。第一是如何依据自适应训练数据从训练集中选出同被适应说话人最相近似的训练用说话人, 第二是如何根据已经选定的训练用说话人训练出(合成出)被适应说话人的语音模型参数。这两个步骤在图1 中以“*”符号标出。在训练用说话人的选择上, 常采用计算被适应说话人语音数据同每个训练集说话人的隐马尔可夫模型(hidden markov model, HMM)或高斯混合模型(gaussian mixture model, GMM)似然分数的方法, 因为这两种模型从性能上很相近[6], 但是GMM 方法的速度和存储消耗要优于HMM 方法, 本文的基线系统使用GMM 模型选出训练用说话人。在合成被适应说话人的SST 语音模型上, 可以通过选出的训练用说话人的数据直接经过Baum-Welch 算法合成被适应说话人的SST 模型, 也可以从全部训练集的SI模型, 使用训练用说话人的语音数据进一步做MAP 或者MLLR 自适应得到被适应说话人的SST 模型。考虑到后一种方法的计算量较大, 系统实用性很差, 同时也缺少对训练用说话人数据灵活加权的余地, 本文的基线系统采取直接利用训练用说话人数据通过Baum-Welch 重估合成被适应人的SST 模型。采用的公式为:

$$\hat{\mu} = \frac{\sum_{n=1}^{N}\sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)\, o^{n,r}(t)}{\sum_{n=1}^{N}\sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)} = \frac{\sum_{n=1}^{N} Q_m^n}{\sum_{n=1}^{N} L_m^n}, \quad (1)$$

$$\hat{\Sigma}_m = \frac{\sum_{n=1}^{N}\sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)\,(o^{n,r}(t)-\mu_m)(o^{n,r}(t)-\mu_m)'}{\sum_{n=1}^{N}\sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)} = \sum_{n=1}^{N} P_m^n \Big/ \sum_{n=1}^{N} L_m^n. \quad (2)$$

其中: $\hat{\mu}_m$ 和 $\hat{\Sigma}_m$ 分别是混合Gauss 概率分布的均值和方差矢量, N 是训练用说话人数目, R_n 是某训练用说话人的训练句子数, T_r 是某句中的观察矢量数。$o^{n,r}(t)$ 是某观察矢量, $L_m^{n,r}(t)$ 是从某混合高斯激发出该观察矢量的比例,

$$L_m^n = \sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t), \quad Q_m^n = \sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)\, o^{n,r}(t),$$

$$P_m^n = \sum_{r=1}^{R_n}\sum_{t=1}^{T_r} L_m^{n,r}(t)\,(o^{n,r}(t)-\mu_m)(o^{n,r}(t)-\mu_m)'.$$

而事实上L_m^n、Q_m^n、P_m^n 都是可以预先计算好的, 因此整个SST 模型的重估可以比较快捷。

对于每个训练用说话人, 存储对其预先计算的 L_m^n、Q_m^n 和 P_m^n 所需要的存储空间大约是语音模型存储量的1.5 倍。

2 动态说话人选择性训练算法

基线系统中的说话人选择性训练方法可以在只有很少量自适应训练数据的情况下起效, 然而进一步详细推敲, 基本的SST 方法仍存在着不足, 还有改善的余地。首先, 针对自适应数据从训练集合选出M 个训练用说话人, 其数量M 是依经验确定的一

个常数。这样存在以下弊端: 一是常数M 是较强烈地依赖于训练集的。第二点在于对不同的自适应说话人, 训练集中与其相似的适合选作训练用说话人的人数, 从直观推断也应该是不尽相同的, 采用固定的经验常数M 失却选择的准确性, 也容易因为多选了人数而造成浪费的计算。另外在进行SST 模型合成的时候, 丢掉了每一个训练用说话人同被适应说话人不同的相似程度(每个GMM 模型的输出分数)这个信息, 简单地等同对待了每个训练用说话人。

基于上述的一些分析, 将置信测度理论中的置信度引入到基本SST 方法中, 发展成动态说话人选择性训练(dynamic SST, DSST)方法。对每一个预先训练好的训练集说话人的GMM 模型, 计算其与被适应说话人语音的对数似然分数:

$$\ln p_{GMM}(X/\lambda_n) = \frac{1}{T}\sum_{t=1}^{T}\ln\sum_{m=1}^{M}\left\{\frac{c_m^n(2\pi)^{-D/2}}{|\Sigma_m^n|^{1/2}}\cdot \exp\left[-\frac{1}{2}(X_t-\mu_m^n)'(\Sigma_m^n)^{-1}(X_t-\mu_m^n)\right]\right\}, \quad (3)$$

其中: $(c_m^n, \mu_m^n, \Sigma_m^n)$ 是模型λ_n 的M 个Gauss 混合概率密度分布的参数值, X 是观察矢量序列, T 是观察矢量的总数目, D 是观察矢量的维数。进一步参照"N-Avg-Best"置信测度计算出GMM 似然分数对应的置信度分数:

$$\delta_1(n) = \ln p_{GMM}(X/\lambda_n) - \frac{1}{N_{Best}}\sum_{i=1}^{N_{Best}}\ln p_{GMM}(X/\lambda_i), \quad (4)$$

$$\delta_2(n) = \frac{\delta_1(n)}{\frac{1}{N_{Best}}\sum_{i=1}^{N_{Best}}\ln p_{GMM}(X/\lambda_i) - \frac{1}{N}\sum_{i=1}^{N}\ln p_{GMM}(X/\lambda_i)}. \quad (5)$$

N_{Best}可以取值在3 到5, N 是训练用说话人的总数。通过分别选用置信度$\delta_1(n)$或$\delta_2(n)$, 对其设定经验阈值δ_T, 选择出训练用说话人,

$$f(n) = \begin{cases} 1, & \delta(n) \geq \delta_T, \\ 0, & \delta(n) < \delta_T. \end{cases} \quad (6)$$

实际中为避免有时选出的训练用说话人数目太少而无法进行Baum-Welch 重估, 再设定一个最小数目阈值M_{min}= 10～15, 确保对被适应说话人的模型估计比较稳定。

进而再根据每个训练用说话人的置信度分数, 计算其应取的合成权重。将选出的各个训练用说话人, 根据其δ_2 置信测度排序, 然后简单地线性赋予从0.5 到1.0 的权重值。并在模型合成中考虑权重因素:

$$\hat{\mu}_m = \frac{\sum_{n=1}^{N} w_n Q_m^n}{\sum_{n=1}^{N} w_n L_m^n} \quad 和 \quad \hat{\Sigma}_m = \frac{\sum_{n=1}^{N} w_n P_m^n}{\sum_{n=1}^{N} w_n L_m^n}. \quad (7)$$

3 实验结果和分析

实验基于汉语大词汇量连续语音识别系统, 在声学层上进行。语音特征参数采用归一化能量和12 阶梅尔频标倒谱参数(Mel-frequency cepstral coefficient, MFCC), 以及相应的一阶二阶差分参数, 共39 维。声学模型建立有调的三元音子模型, 识别测试每句的无调单字识别正确性。

实验系统所用到的训练语音库有2 套: 一套是863 训练语音库, 其中男性说话人和女性说话人各有83 位; 另一套是由微软亚洲研究院赠送的训练库, 包括100 位男性说话人。相应的测试语音库分别包含男女说话人各10 名和20 位男性说话人, 每个说话人都是20 句测试语音。每个测试语音库都同其对应的训练语音库是环境匹配的, 这样可以排除因为环境差别带来的自适应性能提升, 更加准确地反映说话人口音自适应的算法效果。全部实验的自适应训练语音都只有1 句话(约3 s), 识别率统计无调的单字(全音节)识别正确率。

图2 给出了采用不同的训练语音库时对两种训练用说话人选择方法的影响。数据反映各种方法在不同阈值选取下识别率的变化情况, 方块数据点对应863 库, 圆圈数据点对应微软库, 未采用动态权重合成DSST 模型。可以看到, 采用固定选择M 个训练用说话人的方法, 对不同的训练集合其经验常数的一致性是较差的。另外还可发现, 对于N 较大的库, M 反而可能较小。在使用置信测度选择训练用说话人的方法时, 对不同的训练集合, 则反映出更好的参数一致性(反映在选取对另外一个库最优的δ 时, 识别率相比于对本库最优的δ时的变化)。从实验的结果中可以看出, 在采用δ_2 置信测度选取时, DSST 方法的经验常数其外推广性比较高。

表1 的实验数据是按照两种训练用说话人选择方法, 对两个训练语音库, 分别在达到最佳自适应效果时, 所平均选出的训练用说话人个数。其中DSST 方法分别使用δ_1 和δ_2 置信测度进行选取。从数据结果中看到, DSST 方法的训练用说话人选择方法

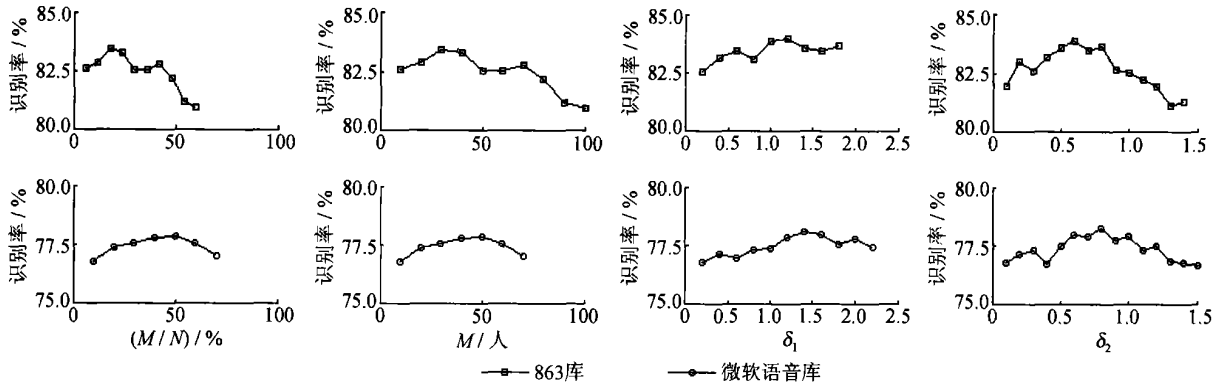

图 2　不同选择方法的一致性比较

需要使用的平均训练用说话人数要少于基本 SST 方法, 这样 DSST 方法就可以在自适应训练的过程中进行更少的计算和更少的数据吞吐量, 具有更灵活的动态适应性。

表 1　各选择方法在最佳性能时的训练人数

库名	SST /人	DSST (δ_1) /人	DSST (δ_2) /人
863 库	30	26. 2	26. 2
微软库	50	35. 5	37. 2

在表 2 中给出了不同方法进行快速自适应的识别结果(训练用说话人的选取基于 δ_2 置信测度)。对每个训练测试集合, 上面一行为识别正确率, 下面一行为相对基线系统的误识率下降比率。通过实验可以看到, SST 方法的确是一种行之有效的快速自适应方法。在仅改进了训练说话人选择策略的 DSST 方法上虽然平均选出的训练用说话人数目变少但是效果反而提高, 而在进一步应用了各个训练用说话人不同相似权重的 DSST 方法中, 自适应效果可以更一步地提高。

表 2　各种算法的自适应性能

库名	基线系统/%	SST /%	DSST (未采用动态加权) /%	DSST (采用动态加权) /%
863 库	80. 16	83. 40	83. 86	84. 12
	—	16. 33	18. 65	19. 96
微软库	75. 22	77. 80	78. 26	78. 74
	—	10. 41	12. 27	14. 20

4　结　论

本文通过对基本 SST 方法的研究, 细致分析其不足, 将置信测度全面地引入到 SST 方法之中从而发展出更加实用和高性能的 DSST 快速自适应方法。改变基本 SST 方法中训练用说话人选择上的绝对数目选取, 采用基于说话人 GMM 相似度计算的置信测度进行训练用说话人的选取, 降低了所需要的训练用说话人数目的同时也提高了识别率, 提高了选取的效能, 并拓展了选取标准的推广性。不再等同对待各个训练用说话人, 而是依据他们同被适应说话人的近似程度, 加权地合成 DSST 语音模型, 进一步提高了自适应的效果。

最后的实验数据表明系统识别率从 80. 16% 提高到 84. 12%, 相对误识率降低了 19. 96%。

参考文献　(References)

[1] Hazen Timothy J. A comparison of novel techniques for rapid speaker adaptation [J]. *Speech Communication*, 2000, **31**: 15 - 33.

[2] Gauvain Jean-Luc, Lee Chin-Hui. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains [J]. *IEEE Trans SAP*, 1994, **2**: 291 - 298.

[3] Leggetter C J, Woodland P C. Maximum likely- hood linear regression for speaker adaptation of continuous density hidden Markov models [J]. *Computer Speech and Language*, 1995, **9**(2): 171 - 185.

[4] Padmanabhan M, Bahl L R, Nahamoo D, et al. Speaker clustering and transformation for speaker adaptation in speech recognition systems [J]. *IEEE Trans on Speech and Audio Processing*, 1998, **6**(1): 71 - 77.

[5] WU Jian, CHANG Eric. Cohorts based custom models for rapid speaker and dialect adaptation [A]. Proc Eurospeech [C]. Aalborg, Denmark: ISCA Press, 2001, **2**: 1261 - 1264.

[6] HUANG Chao, CHEN Tao, CHANG Eric. Speaker selection training for large vocabulary continuous speech recognition [A]. Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing [C]. Orlando, Florida: IEEE Press, 2002 **1**: 609 - 612.