基于高斯相似度分析的插值自适应算法

吕 萍, 王作英, 陆大
(清华大学电子工程系, 北京 100084)

摘 要: 快速说话人自适应算法在非特定人连续语音识别的应用中有重要意义. 现在流行的自适应算法多数只考虑均值的自适应. 本文提出的自适应算法可以快速的对协方差矩阵进行自适应. 该算法是用高斯相似度度量协方差矩阵间的距离, 并由此测度建立了反映协方差矩阵结构关系的二叉决策树. 树的每个中间节点包含一个类质心. 在决策树基础上, 训练多个与特定人模型相关的类质心. 自适应时, 通过对这些类质心进行线性插值得到自适应的协方差矩阵. 实验结果表明, 该方法能够在仅有一句自适应数据的情况下, 使系统误识率由 29.49% 下降到 27.55%.

关键词: 连续语音识别; 快速说话人自适应; 高斯相似度分析
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2001) 12A-1759-03

Interpolation Adaptation Algorithm Based on Gaussian Similarity Analysis

LV Ping, WANG Zuo-ying, LU Da-jin
(Department of Electronic Engineering, Tsinghua University, BeiJing 100084, China)

Abstract: Fast speaker adaptation is very important in application of speaker independent continuous speech recognition. Recently speaker adaptation methods almost just update mean vector. A new adaptation algorithm proposed in this paper can adapt covariance matrixes rapidly. Gaussian similarity is used for measuring the distance of different covariance. A binary decision tree is constructed with this measure. And each middle node includes a covariance matrix of cluster center. Lots of covariance matrixes corresponding to speaker dependent model are trained based on this tree. During adaptation, covariance matrixes are updated through linear interpolating those covariances of cluster center. It can be shown from the experiments that error rate is descended from 29.49% to 27.55% in case of just one adapted sentence.

Key words: continuous speech recognition; rapid speaker adaptation; Gaussian similarity analysis

1 引言

说话人自适应是一种提高非特定人(speaker-independent, SI)连续语音识别系统性能的有效手段[1]. 随着语音识别技术的实用化, 快速收敛的自适应技术越来越受到人们的重视. 目前最为流行的自适应方法有最大后验概率(maximum a posteriori, MAP)[2]和最大似然线性回归(maximum likelihood linear regression, MLLR)[3], 由于需要的自适应数据量较大, 不能满足快速自适应的要求.

大词汇连续语音识别系统的声学模型一般采用连续概率密度隐含马尔可夫模型(continuous density hidden Markov model, CDHMM), 而观测矢量的概率分布一般为高斯分布. 在此基础上, 一些新的快速自适应方法被提出来. 较成功的有: 说话人预分类[4], 最大似然模型插值(maximum likelihood model interpolation, MLMI)[5]和均值偏移量的相关建模法(dependency modeling of biases, DMB)[6]等. 这些方法的共同点是, 只对 HMM 模型参数中的均值向量进行自适应.

在高斯分布中, 均值向量表示的是分布的中心位置. 而协方差矩阵描述的是分布形状, 且其含有的参数比均值要多得多. 这使得协方差矩阵的快速自适应较均值困难. 文献[7]提出了基于高斯相似度分析的决策树自适应算法. 该算法以高斯相似度为协方差矩阵间的距离测度, 建立了二叉决策树. 每个中间节点对应一个类, 而该类包含一系列空间分布形状比较相近的高斯随机变量(即 HMM 状态的观测矢量). 树建好后, 可计算类质心与类中各变量间的变换关系. 自适应过程中, 由测试者提供的数据确定类的数目及自适应的类质心. 最后, 通过前面计算的变换关系导出自适应的协方差矩阵. 决策树算法取得了较好的自适应效果. 但是在自适应数据特别少的情况下(仅有几句话), 自适应效果变差. 本文提出了一种新算法, 称之为线性插值算法, 可解决快速自适应问题. 插值算法利用一些特定人(speaker dependent, SD)模型训练出特定人的类质心, 自适应时通过对这些类质心进行线性插值得到测试者的类质心. 该插值算法弥补了自适应数据过少时不能稳定的估计一个协方差矩阵的缺陷. 实验结果表明, 该算法有较好的收敛性. 当只有一句自适应数据时, 系统的误识率便由 29.49% 下降到 27.55%.

本文的安排如下, 先简要介绍高斯相似度分析和决策树自适应算法. 然后介绍线性插值自适应的基本思想和具体算法, 及其实验结果. 最后给出结论.

收稿日期: 2001-07-18; 修回日期: 2001-09-06
基金项目: "985" 重大项目(985 校-22-攻关-06)

2 高斯相似度分析

高斯相似度(Gaussian similarity)表示的是两个零均值高斯分布之间的相似程度. 基于这种度量的算法称为高斯相似度分析(*Gaussian similarity analysis*, GSA). 从空间的(几何的)角度来考虑两分布间的距离. 若 x 和 y 是分别服从 $N(0, \Sigma_x)$ 和 $N(0, \Sigma_y)$ 的零均值高斯分布, 不难得到它们的空间采样点集 $\{x_i\}$ 和 $\{x_j\}$. 我们可以在这两个点集之间找到一一映像关系 $A:\{x_i\}\rightarrow\{y_j\}$. 当然这种映像 A 必须满足 $\Sigma_y = A\,\Sigma_x A^T$. 于是分布 x 和 y 间的距离可写成:

$$d(x, y) = E(\|x - y\|^2) = E(\|x - Ax\|^2) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\|x_i - Ax_i\|^2 \quad (1)$$

因为满足条件的映像 A 不唯一, 所以 $d(x, y)$ 也不唯一. 希望找到使得 $d(x, y)$ 最小的 A, 并将此时的距离定义为两高斯分布间的高斯相似度. 由 Lagrange 乘子定理可得高斯相似度及相应的映像 A:

$$A = \sum_x^{-1/2}\left[\sum_x^{1/2}\sum_y\sum_x^{1/2}\right]^{1/2}\sum_x^{-1/2} \quad (2)$$

$$d(x, y) = tr\left(\sum_x + \sum_y - 2\left[\sum_x^{1/2}\sum_y\sum_x^{1/2}\right]^{1/2}\right) \quad (3)$$

若分布 x 和 y 间的高斯相似度小, 则这两分布间的相似程度高; 反之, 分布间的相似程度低. 显然, A 是正定对称矩阵.

3 决策树自适应算法

自适应过程由两步组成. 首先离线建立二叉决策树. 第二步为在线过程. 当有自适应数据输入时, 根据该数据更新 HMM 状态的协方差矩阵.

二叉决策树采用自顶向下的 K 均值方法建立. 先将非特定人(SI)模型的所有状态放入根节点, 接着根据式(3)用 K 均值算法将根节点分成两个子节点. 重复分裂过程, 直至叶子节点. 一个叶子节点对应一个 HMM 状态. 树建立好后, 将计算各中间节点(非叶子节点)代表的类质心 C_{Φ_j}(其中 Φ_j 表示该中间节点 j 包含的所有叶子节点集合), 及质心 C_{Φ_j} 与各叶子节点间的映像关系 $A_{i,j}$. 该决策树刻画了 HMM 状态的协方差矩阵结构关系. 属于同一类的 HMM 状态有着相似的空间分布形状. 这种相似性是比较稳定的. 即可以假设在非特定人(SI)模型中属于同一类的状态, 在特定人(SD)模型和自适应(SA)模型中仍属于同一类.

在线自适应时, 根据最大似然(ML)准则由自适应数据得到自适应类的类质心 C_{Φ_j}. 其中自适应类的数目由自适应数据动态确定. 最后通过映像关系 $A_{i,j}$ 更新各 HMM 状态的协方差矩阵 $\Sigma_j^{(SA)} = A_{i,j}C_{\Phi_j}A_{i,j}$.

4 基于 GSA 的插值自适应

人类发音器官的惯性使得 HMM 各语音单元间存在一定的相关性. 这种相关性是快速自适应算法的基础. 决策树算法中建立的二叉决策树, 简单的刻画了各状态协方差矩阵间的相关性. 不同的说话人虽然有不同的发音特点, 但是人类的发音器官结构基本是相同的. 于是不同说话人间也有一定的相关性. 因此, 我们可以将一些特定说话人的 HMM 模型进行线性插值模拟测试者的模型. 其中, 与测试者相近的特定说话人的插值系数较大, 反之则较小. 这便是本文提出的线性插值快速自适应算法的基本思想.

具体而言, 利用上面建立的二叉决策树, 测试者自适应模型(SA)中的协方差矩阵由一组的特定人模型(SD)的协方差矩阵插值得到. 这里我们说的用于插值的协方差矩阵, 并不仅仅是 HMM 状态的协方差矩阵, 还包含二叉树的各节点类中心的协方差阵. 中间节点代表了其对应的所有叶子节点, 故对其进行插值后, 其对应的所有叶子协方差矩阵都会被自适应. 利用中间节点的类质心进行插值的另一个优点是, 可以根据自适应数据量的多少动态决定用于插值中间节点的数目. 这样在保证快速自适应的同时, 改善了算法的渐进性.

下面, 具体介绍基于 GSA 的插值自适应算法.

首先根据非特定人(SI)模型, 建立起二叉决策树[7]. 然后, 按照该决策树计算在特定人(SD)模型下中间节点对应的类质心 $C_{\Phi_j}^{(s)}$, $(s = 1, \cdots, S; j = 1, \cdots, J)$. 接着, 根据测试者的自适应数据确定各中间节点的插值系数. 而插值表达式为:

$$C_{\Phi_j}^{(SA)} = \sum_{s=1}^{S}\alpha_{s,j}\cdot C_{\Phi_j}^{(s)} \qquad (j = 1, \cdots, J) \quad (4)$$

其中: j 表示中间节点; J 为中间节点的总数, 即总的自适应类数, 由自适应数据动态确定; Φ_j 表示节点 j 对应的叶子节点(即状态)的集合; $C_{\Phi_j}^{(s)}$ 表示第 s 个 SD 的第 j 个中间节点; $s = 1, 2, \cdots, S$, S 为总的 SD 模型数; $\alpha_j = \{\alpha_{s,j} \mid s = 1, 2, \cdots, S\}$ 代表第 j 个中间节点对应的线性插值系数, 且其满足下列的约束:

$$\sum_{s=1}^{S}\alpha_{s,j} = 1,\ \alpha_{s,j} \geq 0,\ s = 1, \cdots, S \quad (5)$$

该算法的待估参数是插值系数 α. 这里以高斯相似度最小作为求解参数的优化准则. 目标函数为泛函:

$$J(\alpha) = \sum_{j=1}^{J}tr\left(R_{\Phi_j}(O_t) + \sum_{s=1}^{S}\alpha_s C_{\Phi_j}^{(s)}\right) - \sum_{j=1}^{J}tr\left(2\left[\left(\sum_{s=1}^{S}\alpha_s C_{\Phi_j}^{(s)}\right)^{1/2}\cdot\left(R_{\Phi_j}(O_t)\right)\left(\sum_{s=1}^{S}\alpha_s C_{\Phi_j}^{(s)}\right)^{1/2}\right]^{1/2}\right) \quad (6)$$

其中: $R_{\Phi_j}(O_t)$ 表示属于自适应类 j 的观测特征的二阶统计量. (注: 在自适应数据极少时, 不足以稳定的估计一个协方差矩阵. 此时 $R_{\Phi_j}(O_t)$ 不是协方差矩阵, 称之为二阶统计量.)

定义 $\Omega = \{\alpha: \sum_{s=1}^{S}\alpha_s = 1, \alpha_s \geq 0, s = 1, \cdots, S\}$, 求解最优的 α 相当于:

$$\alpha^* = \arg\min_{\alpha\in\Omega} J(\alpha) \quad (7)$$

求解上式可用任何一种有约束最优化算法, 例如梯度投影法. 用式(4)得到插值的类质心后, 根据映像关系 $A_{i,j}$ 可推导出自适应的协方差矩阵:

$$\Sigma_i^{(SA)} = A_{i,j}C_{\Phi_j}^{(SA)}A_{i,j} \qquad (i \in \Phi_j, j = 1, \cdots, J) \quad (8)$$

本文实现了该算法最简单的情况, 即所有 HMM 模型状态聚成一类, 用二叉树的根节点进行插值的情况. 我们用这种最简单的情况来验证该算法的有效性. 自适应步骤具体如下:

(1)训练一个 SI 模型, 并根据 SI 模型计算所有状态聚成一类时的质心 C_Φ 及各状态与质心间的映像关系 $A_{i,\Phi}$;

(2)训练多个 SD 模型, 并计算各 SD 模型的类质心 $C_\Phi^{(s)}$, $(s = 1, \cdots, S)$;

(3) 用 SI 模型对自适应数据进行 Viterbi 分割;

(4) 自适应数据的二阶统计量 $R_\Phi(O_t)$;

(5)用梯度投影法得到最优的组合系数 $\alpha^* = \arg\min_{\alpha\in\Omega} J(\alpha)$;

(6) 计算自适应后的类质心: $C_\Phi^{(SA)} = \sum_{s=1}^{S} \alpha_s C_\Phi^{(s)}$;

(7) 计算自适应的协方差矩阵: $\Sigma_i^{(SA)} = A_{i,\alpha} C_\Phi^{(SA)} A_{i,\Phi}$ (i= 1, 2, …, N), 其中 N 为总状态数.

5　实验结果

我们将研究基于 GSA 的插值自适应算法的性能, 并给出声学层的识别结果(给出不进行自适应的基线系统与自适应后系统的对比结果). 识别系统的声学模型是基于段长分布的隐马尔可夫模型(Duration Distribution Based Hidden Markov Model, DDBHMM)[8]. 它共包含 856 个状态, 每个状态由一个均值和一个全协方差矩阵来描述. 系统中的特征参数共 45 维, 包括 14 维的倒谱, 1 维能量, 以及倒谱和能量的一阶、二阶差分. 训练集和测试集均采用国家 863 高科技计划提供的男生数据. 训练集由 70 个说话人数据组成. 集外的 13 个说话人数据作测试. 每个人的测试数据为 100 句.

首先我们研究自适应数据量极少时算法的自适应效果. 用于插值的 SD 模型数目为 70 个. 表 1 给出了自适应数据为 1、3 句(约 3 秒、9 秒) 时的系统误识率. 只有一句自适应数据时, 误识率便从 29.49% 下降到 27.55%, 识别性能相对提高了 6.6%. 当自适应数据进一步增加时, 误识率将进一步下降. 但从表中可以看出, 5 句自适应数据时该算法的自适应效果已趋于饱和. 这是因为本实验实现的是, 插值算法的最简单情况.

表 1　基于 GSA 的插值自适应算法(误识率%)

测试文件	基线系统	自适应数据量		
		一句	三句	五句
M01	35.46	30.25	30.00	30.25
M80	32.29	31.55	31.19	31.46
M81	22.98	20.29	20.29	20.05
M82	33.61	30.08	30.25	30.34
M83	28.74	26.78	26.62	26.78
M84	32.11	31.09	31.00	30.82
M93	22.98	22.98	22.66	22.98
M94	31.60	30.42	30.67	30.59
M95	26.55	25.06	25.06	25.22
M96	39.47	37.26	36.71	36.80
M97	33.82	32.93	32.52	31.30
M98	22.94	22.18	22.69	22.52
M99	20.83	17.23	17.38	17.38
平均	29.49	27.55	27.46	27.42

6　结论

本文首先简要介绍了度量两个高斯分布间相似度的方法 - 高斯相似度分析(GSA) 及决策树自适应算法. 根据快速自适应的需要, 在决策树算法的基础上提出了线性插值的自适应算法. 即以高斯相似度作为协方差矩阵间距离的度量, 由 SI 模型训练出二叉决策树, 并训练相关的 SD 模型参数. 当有少量自适应数据时, 以高斯相似度最小作为优化目标得到插值系数. 插值后的类质心作相应变换后, 便可得到各 HMM 状态自适应的协方差矩阵. 为了验证该算法, 我们给出了该算法最简单情况—所有 HMM 状态聚成一类下的实验结果. 结果表明, 该算法能实现说话人的快速自适应, 即当自适应数据仅为 1 句时系统的误识率便有相对 6.6% 的下降. 进一步的工作包括实现该算法的一般情况, 研究算法的收敛性和渐进性.

参考文献:

[1] Padmanabhan M, Bahl LR, Picheny MA. Speaker clustering and transformation for speaker adaptation in large vocabulary speech recognition systems [C]. Proceedings of ICASSP. 1996, (2): 701– 704.

[2] Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate gaussian observations of markov chains [J]. IEEE Transaction. Audio Speech Processing, 1994, 2(2): 291– 298.

[3] Legetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of continuous density HMM's [J]. Computer Speech and Language, 1995, 9(2): 171– 186.

[4] Gao Yuqing, Padmanabhan M, Picheny M. Speaker adaptation based on pre-clustering training speakers [C]. Procedings of Eurospeech. 1997: 2091– 2094.

[5] Wang Zuoying, Liu Feng. Speaker adaptation using maximum likelihood model interpolation [C]. Proceedings of ICASSP. 1999, (2): 1368– 1372.

[6] V Digalakis, S Berkowitz, et al. Rapid speech recognizer adaptation to new Speakers [C]. Proceedings of ICASSP. 1999, (2): 2102– 1206.

[7] Wu Ji, Wang Zouying. A decision tree-structured algorithm of speaker adaptation based on gaussian similarity analysis [J]. Chinese Journal of Electronics. 2001, 10(2): 166– 169.

[8] 王作英. 基于段长分布的 HMM 语音识别模型 [C]. 中文信息学会, 第二届全国汉字语音识别会议. 1989.

作者简介:

吕　萍　女. 1974 年出生于湖北省江汉. 1996 年毕业于电子科技大学自动化系, 1999 年在中国航天工业总公司第二研究院获硕士学位. 1999 年 5 月至今在清华大学电子工程系攻读博士学位, 感兴趣的研究领域有: 语音信号处理, 模式识别和智能计算机界面.

王作英　男. 1935 年出生于江西省赣县. 1959 年毕业于清华大学, 1963 年毕业于苏联莫斯科鲍曼高等工业学校制造系, 获博士学位. 自 1963 年至今在清华大学电子工程系任教. 现为该系教授, 博士生导师, 中国通信学会通信理论委员会副主任, 获国务院特殊津贴专家. 研究领域为信号和信息处理. 近年来主要从事语音信号处理研究, 主持和参加国家 863 高科技项目语音识别的研究.