鉴别性最大后验概率线性回归说话人自适应研究

齐耀辉[1,2,3]， 潘复平[2]， 葛凤培[2]， 颜永红[1,2]
(1.北京理工大学 信息与电子学院，北京 100081；2.中国科学院声学研究所
中国科学院语言声学与内容理解重点实验室，北京 100190；3.河北师范大学
物理科学与信息工程学院，河北，石家庄 050024)

摘 要：为增强自适应后的声学模型的鉴别能力，提出了一种基于最大互信息(MMI)的鉴别性最大后验概率线性回归(MMI-DMAPLR)说话人自适应方法. 将最大互信息准则和最大后验概率(MAP)准则相结合，设计了一个新的目标函数来估计基于线性变换的自适应方法中的变换参数，在最大后验概率估计中加入了鉴别性. 大词汇量连续语音识别的实验结果表明，新方法在增强声学模型与测试数据的匹配性的同时，可以有效提高声学模型的鉴别能力，在少量自适应数据的情况下，其性能比最大后验概率线性回归(MAPLR)相对提高 4.8%.

关键词：最大似然线性回归；最大后验概率线性回归；最大互信息；说话人自适应
中图分类号：TN 912.3 **文献标志码**：A **文章编号**：1001-0645(2015)09-0946-05
DOI：10.15918/j.tbit1001-0645.2015.09.013

Investigation on Discriminative Maximum a Posteriori Linear Regression for Speaker Adaptation

QI Yao-hui[1,2,3], PAN Fu-ping[2], GE Feng-pei[2], YAN Yong-hong[1,2]
(1.School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;
2.Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustic,
Chinese Academy of Science, Beijing 100190, China; 3.College of Physics and Information
Engineering, Hebei Normal University, Shijiazhuang, Hebei 050024, China)

Abstract: In order to increase the discriminative capability of the adapted acoustic model, the maximum mutual information based discriminative maximum a posteriori linear regression (MMI-DMAPLR) adaptation method was proposed. Combining the maximum mutual information criterion with maximum a posteriori (MAP) criterion, a new objective function was designed to estimate the transform parameters of adaptation method based on the linear transformation, to increase the discriminative capability in maximum a posteriori estimation. The experimental results in large vocabulary continuous recognition show that the proposed method can both enhance the match degree between the acoustic model and the test data and the discriminative power of acoustic model. Compared with maximum a posteriori linear regression (MAPLR), the proposed method can obtain 4.8% relative reduction in word error rate when the amount of data is limited.

Key words: maximum likelihood linear regression; maximum a posteriori linear regression;

收稿日期：2013-05-24
基金项目：国家自然科学基金资助项目(10925419,90920302,61072124,11074275,11161140319,91120001,61271426)；中国科学院战略性先导科技专项(XDA06030100,XDA06030500)；国家"八六三"计划项目(2012AA012503)；中科院重点部署项目(KGZD-EW-103-2)
作者简介：齐耀辉(1978—)，女，博士，讲师，E-mail：qiyaohui@hccl.ioa.ac.cn.

maximum mutual information; speaker adaptation

声学模型与测试数据的不匹配是引起自动语音识别系统性能下降的重要原因之一. 引起声学模型与测试数据失配的因素有很多,其中说话人是一个重要因素. 在实际应用中,从每个测试说话人获得的数据有限,因此直接训练说话人相关模型是不现实的,为此研究者提出了说话人自适应技术,利用少量的说话人相关数据对说话人无关模型进行自适应,从而提高系统的识别性能. 目前说话人自适应已经成为自动语音识别系统中的重要组成部分.

基于线性变换的声学模型自适应是目前比较流行的说话人自适应方法[1],该方法假设声学模型在自适应前与后存在线性变换关系,并且不同模型参数间可以共享同一变换. 最大似然线性回归(maximum likelihood linear regression,MLLR)[2]是一种比较常用的基于线性变换的声学模型自适应方法,其采用仿射变换,并通过最大化线性变换后的声学模型在自适应数据上的似然值来估计线性变换的参数. 虽然在自适应数据充足的情况下,MLLR 可以获得很好的性能,但当自适应数据有限时,由于不能对变换参数进行有效估计,反而会使系统的性能下降. 为了避免过拟合问题,有研究者在估计仿射变换的参数时融入变换矩阵的先验信息,利用最大后验概率(maximum a posteriori,MAP)准则估计变换矩阵,提出了最大后验概率线性回归(maximum a posteriori linear regression,MAPLR)[3]. 还有研究者利用变换后的均值参数的先验信息来实现最大后验概率线性回归[4-6]. 随着鉴别性声学模型训练所表现出的优异性能,出现了一些用鉴别性准则来估计线性回归参数的自适应方法,例如最小分类错误线性回归(minimum classification error linear regression,MCELR)[7]、最小词分类错误线性回归(minimum word classification error linear regression,MWCELR)[8]、最小音素错误线性回归(minimum phone error linear regression,MPELR)[9-10]、软分类边缘估计线性回归(soft margin estimation linear regression,SMELR)[11]. 采用鉴别性准则估计线性回归参数能够在一定程度上弥补最大似然估计的不足,从而得到更优的模型参数. 近来,Tsao Y 等[12]在目标函数中引入似然率(likelihood ratio,LR),提出了鉴别性最大后验概率线性回归(discriminative maximum a posteriori linear regression,DMAPLR),利用识别的 n-best 结果进行线性回归参数估计,增加采用最大后验准则估计的变换矩阵的鉴别能力.

对于以隐马尔科夫模型为基础的声学模型,本文研究其在少量自适应数据下的说话人自适应方法. 为了结合最大后验概率估计与鉴别性训练的优点,改善语音识别系统的性能,设计了一个新的目标函数,将 MAP 准则与 MMI 准则进行线性组合,来估计线性回归自适应方法中的变换矩阵,提出了基于最大互信息的鉴别性最大后验概率线性回归(MMI-DMAPLR). 在实现时,采用 word-lattice 来代表竞争词序列. 对于大词汇量语音识别而言,word-lattice 能比 n-best 结果更有效地表示竞争词序列. 因此 MMI-DMAPLR 在估计变换矩阵时可以融入更多的竞争词序列的信息,有利于提高自适应后的模型的鉴别性. 构建了大词汇量连续语音识别系统,用 863 数据库进行有监督自适应,评估 MMI-DMAPLR 的自适应性能,结果表明该方法优于 MLLR、MAPLR 和 DMAPLR.

1　最大后验概率与最大互信息相结合的目标函数

在基于最大似然与最大后验概率估计的线性回归自适应算法中,用线性回归函数 F_φ 变换原始 HMM 集合 λ 中的参数,使 λ 与测试数据相匹配. φ 是要估计的参数. 给定 R 个观察序列 $\{O_1,O_2,\cdots,O_r,\cdots,O_R\}$ 及其相应的标注文本 $\{w_1,w_2,\cdots,w_r,\cdots w_R\}$,最大似然估计的目标是使语音特征在正确标注文本上的似然值最大化,其目标函数为

$$F_{\mathrm{ML}}(\varphi,\lambda)=\sum_{r=1}^{R}\lg P(O_r\mid\varphi,\lambda,w_r). \quad (1)$$

基于最大后验概率估计的线性回归模型自适应算法的目标函数为

$$F_{\mathrm{MAP}}(\varphi,\lambda)=\sum_{r=1}^{R}\lg[P(O_r\mid\varphi,\lambda,w_r)P(\varphi,\lambda)]. \quad (2)$$

式中 $P(\varphi,\lambda)$ 为 φ 和 λ 的联合概率分布.

MMI 准则最大化正确文本的后验概率,其形式为

$$F_{\mathrm{MMI}}(\varphi,\lambda)=\sum_{r=1}^{R}\lg\frac{P(O_r\mid\varphi,\lambda,w_r)P(w_r)}{\sum_i P(O_r\mid\varphi,\lambda,w_i)P(w_i)}.\tag{3}$$

其中分母项对代表解码空间的所有可能的词序列求和. 分子部分实际上就是最大似然准则,因此最大化 MMI 准则的过程中不仅让特征在正确文本上的似然值增大,同时让特征在其他干扰文本上的似然值减小. 对模型参数的调整不仅针对正确的模型,同时也兼顾到竞争模型的影响,使得到的模型参数具有更大的散度,增加了模型间的区分度.

本文将 MAP 准则和 MMI 准则进行线性组合,来估计线性回归函数中的参数 φ,目标函数的形式为

$$F(\varphi,\lambda)=(1-\alpha)F_{\mathrm{MAP}}(\varphi,\lambda)+\alpha F_{\mathrm{MMI}}(\varphi,\lambda).\tag{4}$$

其中的权重系数 $0<\alpha<1$. 用该目标函数估计变换矩阵可以增加 MAPLR 的鉴别能力,从而改善系统的识别性能.

2 基于最大互信息的鉴别性最大后验概率线性回归

在基于线性变换的声学模型自适应方法中,假设声学模型参数 λ 在自适应前后存在线性变换关系. 线性回归自适应算法采用的变换形式是仿射变换. 目前的自动语音识别系统中大多都使用连续隐马尔科夫模型(hidden Markov model,HMM)进行声学建模,其最重要的待自适应的参数是各高斯分量的均值矢量,因为说话人之间的差别主要由均值描述. 只对均值矢量作变换时,声学模型中第 m 个高斯自适应前后的关系如下:

$$\hat{\boldsymbol{\mu}}_m=\boldsymbol{A}\boldsymbol{\mu}_m+\boldsymbol{b}=\boldsymbol{W}\boldsymbol{\xi}_m.\tag{5}$$

式中:$\boldsymbol{\xi}_m$ 为扩展均值矢量$[1\quad\boldsymbol{\mu}_m^{\mathrm{T}}]^{\mathrm{T}}$;$\boldsymbol{W}$ 为一个 $D\times(D+1)$维的扩展变换矩阵$[\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}]^{\mathrm{T}}$,$\boldsymbol{\mu}_m$ 和$\hat{\boldsymbol{\mu}}_m$ 分别为自适应前后的声学模型参数集中的第 m 个均值矢量. 需要估计的参数为扩展变换矩阵 $\boldsymbol{W}$.

采用如式(1)所示的最大似然准则估计变换矩阵时,$\boldsymbol{W}$ 的第 i 行用下式计算[2]:

$$\boldsymbol{W}_{(i)}^{\mathrm{T}}=\boldsymbol{G}_{(i)}^{-1}\boldsymbol{K}_{(i)},\tag{6}$$

$\boldsymbol{G}_{(i)}$和 $\boldsymbol{K}_{(i)}$的计算方法如下[2]:

$$\boldsymbol{G}_{(i)}=\sum_m\frac{1}{\sigma_{m(i)}^2}\boldsymbol{\xi}_m\boldsymbol{\xi}_m^{\mathrm{T}}\sum_t\gamma_m(t).\tag{7}$$

$$\boldsymbol{K}_{(i)}=\sum_m\frac{1}{\sigma_{m(i)}^2}\boldsymbol{\xi}_m\sum_t\gamma_m(t)O(t)_{(i)}.\tag{8}$$

式中:$O(t)_{(i)}$为第 t 帧观察矢量的第 i 维;$\gamma_m(t)$为 t 时刻处于第 m 个高斯的后验概率;$\sigma_{m(i)}^2$为第 m 个高斯的对角协方差矩阵的第 i 个元素.

最大后验概率准则加入了变换后的模型参数的先验分布 $P(\varphi,\lambda)$. 本文假定第 m 个高斯的均值的先验分布 $P(\boldsymbol{W},\boldsymbol{\xi}_m)$是多变量正态分布:

$$P(\boldsymbol{W},\boldsymbol{\xi}_m)=\frac{\exp[-\frac{1}{2}(\boldsymbol{W}\boldsymbol{\xi}_m-\boldsymbol{\eta}_m)^{\mathrm{T}}(\beta\boldsymbol{V}_m)^{-1}(\boldsymbol{W}\boldsymbol{\xi}_m-\boldsymbol{\eta}_m)]}{(2\pi)^{D/2}\mid\beta\boldsymbol{V}_m\mid^{1/2}}.\tag{9}$$

式中:$\boldsymbol{\eta}_m$ 和 $\boldsymbol{V}_m$ 为超参数;β 为控制最大后验概率准则中先验密度权重的比例因子.

采用如式(4)所示的 MAP 准则与 MMI 准则线性组合的目标函数估计变换矩阵,得到基于最大互信息的鉴别性最大后验概率线性回归自适应算法. 在 MMI-DMAPLR 中,$\boldsymbol{W}$ 还是用式(6)计算,但其中 $\boldsymbol{G}_{(i)}$和 $\boldsymbol{K}_{(i)}$的计算方法与上面不同:

$$\boldsymbol{G}_{(i)}=\sum_m\left[\frac{\sum_t\gamma_m^{\mathrm{num}}(t)-\alpha\sum_t\gamma_m^{\mathrm{den}}(t)+\alpha D_m}{\sigma_{m(i)}^2}+\frac{\varepsilon}{v_{m(i)}^2}\right]\boldsymbol{\xi}_m\boldsymbol{\xi}_m^{\mathrm{T}}.\tag{10}$$

$$\boldsymbol{K}_{(i)}=\sum_m\left[\frac{\sum_t\gamma_m^{\mathrm{num}}(t)O(t)_{(i)}-\alpha\sum_t\gamma_m^{\mathrm{den}}(t)O(t)_{(i)}+\alpha D_m\tilde{\mu}_{mi}}{\sigma_{m(i)}^2}+\frac{\varepsilon\eta_{m(i)}}{v_{m(i)}^2}\right]\boldsymbol{\xi}_m.\tag{11}$$

式中:$\varepsilon=(1-\alpha)/\beta$;$v_{m(i)}^2$和 $\eta_{m(i)}$分别为超参数 V_m 和 $\boldsymbol{\eta}_m$ 的第 i 个元素;$\tilde{\mu}_{mi}$为用初始变换矩阵计算得到的均值矢量 $\tilde{\boldsymbol{\mu}}_m$ 的第 i 个元素;D_m 为平滑因子;$\gamma_m^{\mathrm{num}}(t)$和 $\gamma_m^{\mathrm{den}}(t)$分别为在正确路径和整个搜索空间上计算的高斯后验概率.

3 实验结果和分析

本文搭建了一个汉语大词汇量连续语音识别系统对提出的基于最大互信息的鉴别性最大后验概率线性回归自适应方法的性能进行测试. 训练集约 65 h. 测试数据来自 6 个说话人,3 男 3 女,每人 260 句共1 560句. 训练数据和测试数据均来自国家 863 语音库. 在下面的实验中,均采用有监督的批处理自适应模式,识别结果是 6 个测试说话人的识别结果的平均值.

实验中所使用的特征为感知线性预测系数(PLP),包括 13 维静态特征及对应的一阶、二阶、三阶差分,经过异方差线性鉴别分析(HLDA)之后特征维数从 52 维降为 39 维. 声学模型采用三音子建模方式,声学建模单元采用带四声调的声韵母,共 179 个,每个声韵母用 5 状态从左到右连续隐马尔科夫模型(HMM)来描述,其中第一个状态和最后一个状态为虚状态,声学模型经过基于决策树的状态聚类之后最终的状态数为 5 955,每个状态输出 8 个高斯分量,每个高斯分布的协方差矩阵为对角化的.

基线系统的声学模型为性别无关、说话人无关的声学模型. 为了将 MMI-DMAPLR 方法的性能与其他自适应方法做比较,分别采用 MLLR、MAPLR、DMAPLR 及本文提出的 MMI-DMAPLR 自适应方法对基线声学模型进行自适应,然后在相同的识别环境下对测试数据进行识别. 实验中采用全局变换矩阵,即声学模型集合中的所有高斯分量共享一个变换矩阵.

对于 MMI-DMAPLR 方法而言,在做自适应之前,首先要生成分子 lattice 和分母 lattice. 分子 lattice 用标注文本进行强制对齐,然后在每个词节点上加上语言模型得分得到的;分母 lattice 是用 Unigram 语言模型对自适应集进行识别得到. 在 lattice 上计算统计量时,本文采用了"Exact-match[13]"方式,即直接利用每个音素的起始和结束时间,先在这个时间段上,计算出每个音素的似然值,然后在整个 lattice 上,采用前向后向算法计算每个音素的前向和后向概率,以及整个 lattice 的输出概率. 第 m 个高斯的先验分布中的超参数 $\boldsymbol{\eta}_m$ 和 $\boldsymbol{V}_m$ 的值,分别取基线声学模型的均值和方差. 式(10)和(11)中的 ε 和 α 取经验值,在 MAPLR 中分别为 0.2 和 0,在 MMI-DMAPLR 中分别为 0.2 和 0.4. 在 DMAPLR 中,$\boldsymbol{G}_{(i)}$ 和 $\boldsymbol{K}_{(i)}$ 计算公式中各个参数的取值如下:$\{h_1=1, h_2=0.2, h_3=0.4\}$,$\lambda_n=1/N$,$N=8$. 其计算公式可参考文献[12].

首先研究自适应数据量很少时,MMI-DMAPLR 算法的自适应效果. 对基线系统的声学模型分别采用 MLLR、MAPLR、DMAPLR 和 MMI-DMAPLR 算法进行了自适应. 表 1 给出了自适应数据为 4,5,6 句(约 6.8,8.5,10.2 s)时,几种自适应算法在测试集上的字错误率. 基线系统的字错误率是 11%. 由表 1 的实验结果可见:① 当自适应数据少时(只有大约 6.8 s),MLLR 算法无法进行有效的自适应,而 MAPLR、DMAPLR 和 MMI-DMAPLR 可以有效提高系统的识别性能,字错误率分别从 11%下降到 10.4%,10.0%和 9.9%. 当自适应语句增多时,MMI-DMAPLR 算法的性能也明显优于 MLLR 和 MAPLR 算法的性能;② 在各种数量的自适应语句情况下,MMI-DMAPLR 算法的性能均优于 DMAPLR.

表 1　MMI-DMAPLR 的自适应性能

Tab.1　The performance of MMI-DMAPLR

自适应方法	字错误率/%			
	0 句	4 句	5 句	6 句
MLLR	11.0	11.1	9.9	9.6
MAPLR	11.0	10.4	9.7	9.3
DMAPLR	11.0	10.0	9.6	9.2
MMI-DMAPLR	11.0	9.9	9.4	9.1

第 2 个实验中,研究 MMI-DMAPLR 自适应算法的渐进性. 图 1 分别对应自适应句子数为 3,4,5,…,14 句时几种自适应方法的自适应效果. 图 1 中横坐标 N 表示自适应句子数,纵坐标 E 表示系统的字错误率. 从图 1 中可以看出,当自适应语句大于 7 时,各种自适应方法的性能提高都趋于饱和. 显然,MMI-DMAPLR 与其他自适应方法的渐进性相同,但其收敛性好.

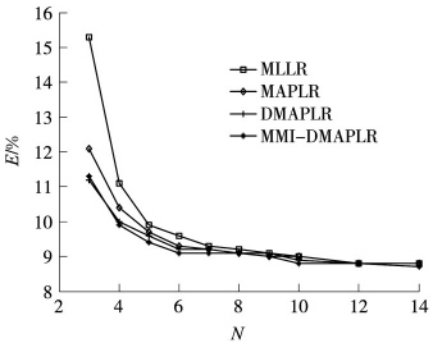

图 1　MMI-DMAPLR 的渐进性

Fig. 1　The asymptotic property of MMI-DMAPLR

4　结束语

本文将 MAP 准则与 MMI 准则进行线性组合,设计了一个新的目标函数,并用此目标函数来估计线性回归自适应方法中的变换矩阵,提出了基于最大互信息的鉴别性最大后验概率线性回归自适应算法. 实验结果表明,将 MMI 准则和 MAP 准则相结合可以有效提高自适应之后声学模型的鉴别能力. MMI-DMAPLR 自适应算法可以在极少量的说话

人数据情况下达到更优的性能提升,收敛速度明显快于传统算法.

参考文献:

[1] Shinoda K. Speaker adaptation techniques for automatic speech recognition[C]// Proceedings of APSIPA ASC. Xi'an, China: [s.n.], 2011.

[2] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition [J]. Computer Speech and Language, 1998,12(2):75-98.

[3] Chesta C, Siohan O, Lee C H. Maximum a posteriori linear regression for hidden Markov model adaptation [C]// Proceedings of Eurospeech. Budapest, Hungary: [s.n.], 1999:211-214.

[4] Lin C H, Wang W J. Maximum a posteriori linear regression for speaker adaptation with the prior of mean [C]// Proceedings of EUPSICO. [S.l.]: IEEE, 2000-01-04.

[5] Tsao Y, Isotani R, Kawai H, et al. An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition[C]// Proceedings of ISCSLP. Tainan: [s.n.], 2010:29-32.

[6] Hu Tingyao, Tsao Y, Lee Lin-shan. Discriminative fuzzy clustering maximum a posteriori linear regression for speaker adaptation[C]// Proceedings of Interspeech. Portland, USA: [s.n.], 2012.

[7] Wu J, Huo Q. A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models [J]. IEEE Trans on Audio, Speech and Language Processing, 2007,15(2):478-488.

[8] Zhu B, Yan Z J, Hu Y, et al. Investigation on adaptation using different discriminative training criteria based linear regression and MAP[C]// Proceedings of ISCSLP. Kunming, China: [s.n.], 2008:93-96.

[9] Wang L, Woodland P C. MPE-based discriminative linear transform for speaker adaptation[J]. Computer Speech and Language, 2008,22(3):256-272.

[10] Pirhosseinloo Sh, Javadi Sh. A combination of maximum likelihood Bayesian framework and discriminative linear transforms for speaker adaptation[J]. International Journal of Information and Electronics Engineering, 2012,2(4):552-555.

[11] Matsuda S, Tsao Y, Li J, et al. A study on soft margin estimation of linear regression parameters for speaker adaptation[C]// Proceedings of Interspeech. Brighton, UK:[s.n.], 2009:1603-1606.

[12] Tsao Y, Isotani R, Kawai H, et al. Increasing discriminative capability on MAP-based mapping function estimation for acoustic model adaptation[C]// Proceedings of ICASSP. Prague, Czech:[s.n.], 2011:5320-5323.

[13] Povey D. Discriminative training for large vocabulary speech recognition[D]. Cambridge: Cambridge University, 2004.

(责任编辑:刘芳)