基于音素解码的语种识别系统联合自适应算法研究

邓 妍 [1] 张卫强 [1] 刘 加 [1]

摘 要 针对真实环境下的语种识别, 信道类型和通话内容等非语种方面因素的不同都会造成测试和训练条件的不匹配, 从而影响系统的识别性能. 本文以音素识别器后接向量空间模型 (Phone recognizer followed by vector space model, PRVSM) 为语种识别系统, 引入联合自适应算法来解决系统中测试和训练条件的失配问题. 研究了三种自适应方法用于系统的不同阶段: 1) 基于受约束的最大似然线性回归 (Constrained maximum likelihood linear regression, CMLLR) 的声学模型自适应; 2) 基于全局 N 元文法的音位特征向量自适应; 3) VSM 模型中的支持向量机 (Support vector machines, SVM) 自适应. 在综合采用多种自适应技术后, PRVSM 系统的性能有了较大的提高, 在 NIST LRE 2009 测试库上对于 30 s、10 s 和 3 s 的测试段, 基于不同音素识别器的 PRVSM 系统的等错误率 (Equal error rate, EER) 分别相对降低了 18 %～23 %、12 %～20 % 以及 5 %～9 %.

关键词 语种识别, 音素识别器后接向量空间模型, 联合自适应, 受约束的最大似然线性回归, 支持向量机自适应

引用格式 邓妍, 张卫强, 刘加. 基于音素解码的语种识别系统联合自适应算法研究. 自动化学报, 2012, **38**(4): 652−658

DOI 10.3724/SP.J.1004.2012.00652

Research on Joint Adaptation for Phonotactic Language Recognition

DENG Yan[1] ZHANG Wei-Qiang[1] LIU Jia[1]

Abstract For language recognition in real application, a variety of non-language sources (i.e., channel, content, etc.) will induce mismatch between training and test utterances, which affects the recognition accuracy. This paper introduces joint adaptation to deal with the mismatch problem for the phone recognition followed by vector space model (PRVSM) system. We investigate three adaptation methods in different stage of the system: 1) acoustic model adaptation using constrained maximum likelihood linear regression (CMLLR); 2) phonotactic feature adaptation using the universal N-grams; 3) adapt-SVM for the vector space model(VSM).The joint adaptation is carried out by combining these methods and significant improvements can be obtained. Experiments on the NIST LRE 2009 evaluation corpus show that there are relative decreases of 18 %～23 %, 12 %～20 % and 5 %～9 % in EER for the 30 s, 10 s and 3 s test conditions, respectively.

Key words Language recognition, phone recognizer followed by vector space model (PRVSM), joint adaptation, constrained maximum likelihood linear regression (CMLLR), adapt-support vector machines (SVM)

Citation Deng Yan, Zhang Wei-Qiang, Liu Jia. Research on joint adaptation for phonotactic language recognition. *Acta Automatica Sinica*, 2010, **36**(5): 652−658

收稿日期 2011-01-17 录用日期 2011-07-07
Manuscript received January 17, 2011; accepted July 7, 2011
国家自然科学基金 (60931160443, 61005019) 资助
Supported by National Natural Science Foundation of China (60931160443, 61005019)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 清华大学电子工程系清华信息科学与技术国家实验室 (筹) 北京 100084
1. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084

语种识别主要研究如何从语音中获取语种相关信息. 随着国际交流的不断发展, 语种识别在多语言交流领域以及国家信息安全相关领域都发挥了越来越重要的作用[1−2].

目前与说话人无关的语种识别系统主要分为两大类: 基于声学特征的系统以及基于音素解码的系统. 其中, 基于声学特征的系统主要采用高斯混合模型 (Gaussian mixture model, GMM) 或者支持向量机 (Support vector machines, SVM) 对偏移差分倒谱特征进行建模, 近年来也有采用 SVM 对 GMM 的均值和方差超矢量进行建模的. 基于音素解码的系统则主要采用二叉树或者 SVM 等对音素的 N 元文法进行建模, 音素识别器后接向量空间模型 (Phone recognizer followed by vector space model, PRVSM) 是这类系统中广泛使用而且最有效的方法之一[2]. 在 PRVSM 系统中, 首先利用音素识别器对语音进行识别得到对应的音素序列或者音素格[2−3], 然后统计音素的 N 元文法, 将统计值拼接成特征向量, 最后利用 SVM 进行建模得到向量空间模型 (Vector space model, VSM), 基于 VSM 进行语种识别. 在近几次由美国国家标准技术局 (National Institute of Standards and Technology, NIST) 举办的语种识别评测中, PRVSM 系统都获得了不错的识别性能, 但是该系统还存在以下几点不足. 首先, 语音中信道等非语种方面的差异会对音素识别器的性能带来一定的影响, 造成音素识别结果不够稳健[4−5], 从而影响整个系统的识别性能; 其次, 对于一段语音, 它不可能包含所有的 N 元文法, 利用 N 元文法的统计值进行拼接将得到一个稀疏的音位特征向量, 仅利用一段语音对高阶 N 元文法进行统计是不够准确的; 此外, 不同语音段的通话内容千差万别, 因此, 测试语音中的某些 N 元文法可能在训练语料中没有出现过, 这将导致训练和测试样本的音位特征在分布上存在一定的偏移, 从而影响 VSM 模型的稳健性.

为了弥补 PRVSM 系统的上述不足, 研究人员提出了多种方法进行改进. 如针对音素识别器引入说话人自适应训练和声道长度归一化来消除说话人的影响[6], 在音素解码的过程中采用受约束的最大似然线性回归 (Constrained maximum likelihood linear regression, CMLLR) 自适应方法来去除信道的差异[7], 以及利用基于全局或者低阶 N 元文法的语言模型自适应方法来获得对高阶 N 元文法的稳健估值[8]. 本文对系统中的上述问题进行深入研究, 提出对系统进行联合自适应, 通过综合利用声学模型和语言模型的自适应方法来解决训练和测试条件的不匹配以及高维特征向量的稳健估值问题. 首先针对音素识别采用 CMLLR 进行声学模型的自适应, 解决训练和测试语音的信道不匹配问题, 以获得稳健的音素识别结果; 然后利用基于全局 N 元文法的自适应方法来实现对高维特征向量的稳健估值; 最后, 针对 VSM 模型, 提出 SVM 自适应的方法来消除训练和测试样本由于通话内容不同而产生的分布上的差异, 进一步解决由于训练和测试条件的不匹配对 VSM 模型的影响. 为了充分利用不同的自适应技术来尽量提高 PRVSM 系统的稳健性, 还对不同自适应技术之间的互补性进行了研究.

本文的组织结构如下: 首先简介所用的 PRVSM 语种识别系统, 接下来详细阐述提出的自适应算法, 然后给出 PRVSM 系统的自适应框架, 最后是实验结果与分析.

1　PRVSM 系统

在 PRVSM 系统中, 通常采用一个或多个音素识别器 (基于多个音素识别器的系统称之为 PPRVSM (Parallel phone recognizers followed by vector space model), 即并行音素识别器后接向量空间模型) 对语音进行识别, 基于识别结果对音素之间的连接关系 (N 元文法) 进行统计建模和识别, 其系统框架如图 1 所示. 由图 1 可知, PRVSM 系统分为三个部分: 音素识别、N 元文法统计以及 VSM 建模和识别. 下面将对这三个部分进行简单的介绍.

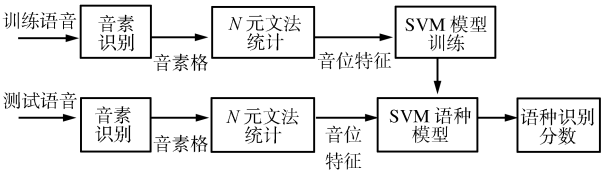

图 1　PRVSM 系统结构

Fig. 1　Framework of the PRVSM system

1.1　音素识别

在 PRVSM 系统中, 一般采用音素识别器对输入的语音进行解码. 由于解码只是将语音转换成便于分析的音素序列, 因此, 音素识别器可以和待识别的目标语种无关. 目前, 音素识别器主要采用语音识别中常用的隐马尔科夫模型 (Hidden Markov model, HMM), 表示如下:

$$P(\boldsymbol{O}|\lambda)=\sum_{\forall s}\prod_{t=1}^{T}P(\boldsymbol{o}_t|s_t,\lambda)*P(s_t|s_{t-1},\lambda)$$

其中, λ 表示 HMM 的模型参数, $\boldsymbol{O}$ 表示长为 T 的声学特征向量 (如梅尔域倒谱系数等) 序列, s 是状态序列, $P(\boldsymbol{o}_t|s_t,\lambda)$ 表示已知模型参数 λ、由状态 s_t 生成观测向量 $\boldsymbol{o}_t$ 的概率, 这种概率函数主要有两种建模方式: GMM[1−3] 和人工神经网络 (Artificial neural network, ANN)[9−10]. 鉴于后端的分类是根据不同语种中音素之间的连接关系不同来进行语种识别, 故该解码过程不需要利用任何的先验语法知识进行约束. 一般来说, 音素识别器的最优识别结果并不够稳健, 因此有学者提出采用音素格来代替最优音素序列, 从多候选的识别结果中获取音素之间的统计关系[3].

1.2　N 元文法统计

对于从音素识别器得到的某个音素序列: $S=s_1\cdots s_i\cdots s_n$, N 元文法的统计量 $c(d_i)$ 是 d_i 在音素序列 S 中出现的次数 (d_i 表示单个音素或者多个音素的组合), 这是仅考虑音素识别器的最优识别结果的情况. 而对于音素格, 由于它是一个由边和节点所组成的图结构 (节点表示时间, 边表示识别得到的假设音素), 涵盖了多候选的识别结果, 因此 $c(d_i)$ 是该音素格中所有可能序列 S 中 $c(d_i)$ 出现次数的期望值, 计算公式如下[3]:

$$c(d_i)=\sum_{S\in\ell}p(S|\ell)c(d_i|S)=\sum_{s_j\cdots s_{j+N-1}\in\ell}\left[\alpha(s_j)\beta(s_{j+N-1})\prod_{k=j}^{j+N-1}\xi(s_k)\right]$$

其中, $p(S|\ell)$ 表示音素序列 S 在音素格 ℓ 中出现的概率, $\alpha(s_j)$ 表示 N 元文法 $d_i=s_j\cdots s_{j+N-1}$ 中起始节点 s_j 的前向概率, $\beta(s_{j+N-1})$ 表示其终止节点 s_{j+N-1} 的后向概率, $\xi(s_k)$ 表示音素 s_k 的声学似然概率. 利用统计量 $c(d_i)$, 可以根据下式来计算 d_i 在该音素格中出现的概率: $p(d_i|\ell)=c(d_i)/\sum_i c(d_i)$. 接下来, 将这些概率值拼接成高维特征向量来得到音位特征. 对于音素序列 $S=s_1\cdots s_i\cdots s_n$, 如果仅考虑二元文法, 则其对应的音位特征向量可以表示为

$$\boldsymbol{X}=[p(d_1|\ell),p(d_2|\ell),\cdots,p(d_F|\ell)]^{\mathrm{T}}$$

其中, $F = f + f^2$, f 表示音素识别器中音素的个数, $p(d_i|\ell)$ 表示对音素格进行统计得到的 N 元文法 d_i 的概率值.

1.3 VSM 建模和识别

目前, 学者们一般采用 SVM 对音位特征向量进行建模和识别[2, 9, 11]. SVM 首先将输入向量映射到高维空间, 然后在高维拓展空间中构造最优超平面分界面来解决分类问题.

SVM 是通过对多个核函数 $K(\boldsymbol{X}_1, \boldsymbol{X}_2)$ 进行求和来构建一个两类分类器, 因此, SVM 分类器的识别性能取决于核函数的构造. 对于 VSM 模型, 主要采用文献 [12] 中所提出的基于词频对数似然比 (Term frequency log likelihood ratio, TFLLR) 的核函数, 表示如下:

$$K(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sum_i p_n(d_i|\ell_1) p_n(d_i|\ell_2) = \sum_i \frac{p(d_i|\ell_1)}{\sqrt{p(d_i|\text{all})}} \frac{p(d_i|\ell_2)}{\sqrt{p(d_i|\text{all})}} \tag{1}$$

其中, ℓ_1, ℓ_2 分别表示任意两段语音所对应的音素格, $p(d_i|\text{all})$ 表示 N 元文法在所有的用于训练的音素格中出现的概率. 由式 (1) 可知, TFLLR 核函数实际上就是在对特征向量进行归一化之后采用线性内积核函数来度量不同样本之间的相似性. 在利用 SVM 对音位特征向量建立语种模型时, 一般采用 "one-versus-rest" 的方法进行训练, 将不属于目标语种的所有其他特征向量看作非目标训练样本. 最终的语种识别主要基于 SVM 分类器的输出, 对于任意的一个特征向量 $\boldsymbol{X}$, SVM 的输出 $f(\boldsymbol{X})$ 为

$$f(\boldsymbol{X}) = \sum_i \alpha_i y_i K(\boldsymbol{X}, \boldsymbol{X}_i) + d \tag{2}$$

其中, y_i 为理想的输出, α_i 为加权系数且满足 $\sum_i \alpha_i y_i = 0$, $\boldsymbol{X}_i$ 为支持向量, d 为偏移量.

实验证明, 高阶 N 元文法比低阶 N 元文法具有更强的区分性[2]. 然而特征向量的维数一般随着 N 的增加呈指数增长, 这给 SVM 的训练带来一定的困难, 但是并非所有的高阶 N 元文法都具有良好的区分性[9]. 根据这一特点, 学者们提出根据模型参数建立目标语种相关的音素子集或者区分性关键词集合的方法进行语种识别[9, 11], 这样可以在系统中引入高阶 N 元文法, 加强 VSM 模型的区分能力, 同时又不会显著地增加计算资源.

2 自适应

针对真实环境下的语种识别, 信道类型、环境噪音、说话人以及通话内容等因素的不同都会造成测试和训练条件的失配. 对于 PRVSM 系统, 信道的变化会在一定程度上影响音素识别器的性能和稳健性, 从而导致系统的语种识别性能下降. 同时, 由于不同语音段的通话内容千差万别, 而其包含的有效语音的持续时间较短, 一般不超过 30 s, 使得单个的语音段中出现的各种不同发音的数目有限, 从而导致高阶 N 元文法对应的音位特征向量是极度稀疏的, 这样会导致某些 N 元文法在测试集中存在而在训练集中不存在的现象, 从而影响 VSM 模型的稳健性. 因此, 在各种失配条件中, 信道的不匹配和通话内容的不同是 PRVSM 系统性能下降的最主要原因. 本文将采用联合自适应的方法来解决 PRVSM 系统中的上述问题.

2.1 基于 CMLLR 的声学模型自适应

对于 PRVSM 系统来说, 理想的音素识别器应该是: 对任何人在任何情况下所说的同一句话进行解码, 都可以得到相同的识别结果. 但是由于信道等干扰因素会在一定程度上影响音素识别器的识别性能, 导致在不同条件下得到的识别结果不同, 从而影响系统的语种识别性能. 对此, 学者们提出在音素识别器的训练或者识别过程中对声学模型进行自适应来消除这些干扰因素的影响. 目前, 已有研究人员对不同的声学模型自适应方法进行了综合对比, 从对比的结果来看, 最好的方法就是对声学模型采用说话人自适应训练, 然后在识别的过程中采取 CMLLR 方法进行自适应[13]. 但是, 与只在识别的过程中利用 CMLLR 进行自适应相比, 这种最好的方法对识别性能的改善并不十分明显, 而且该方法需要额外对语音再进行一次识别. 对于 PRVSM 系统来说, 识别是最耗时的一个阶段, 因此, 本文将只考虑在识别的过程中利用 CMLLR 对声学模型进行自适应.

CMLLR 算法是最大似然线性回归 (Maximum likelihood linear regression, MLLR) 算法的进一步推广[14], 它采用了同一个变换矩阵对 GMM 模型的均值和方差进行更新. 设第 c 个类的 GMM 模型对应的 CMLLR 拓展变换矩阵 W_c 为 $[\boldsymbol{b}_c \ A_c]$, 则通过自适应变换后的均值向量 $\hat{\boldsymbol{\mu}}$ 和方差 $\hat{\sum}$ 可以表示为

$$\hat{\boldsymbol{\mu}} = A'_c \boldsymbol{\mu} + \boldsymbol{b}'_c, \qquad \hat{\sum} = A'_c \sum A_c'^{\mathrm{T}}$$

其中, $A'_c = A_c^{-1}$, $\boldsymbol{b}'_c = -A_c \boldsymbol{b}_c$ 分别为根据自适应训练语音, 用最大似然准则统计出的变换参数和偏移量, 这与只自适应均值的 MLLR 参数估计过程类似.

由此可见, 与 MLLR 被用来描述模型均值的变换关系不同, CMLLR 矩阵还包含了模型方差间的变换关系. 将上述变换式与 GMM 模型相结合, 可得如下表达式:

$$\hat{\boldsymbol{o}}(t) = A_c \boldsymbol{o}(t) + \boldsymbol{b}_c = W_c \boldsymbol{\zeta}_t \tag{3}$$

其中, $\hat{\boldsymbol{o}}(t), \boldsymbol{o}(t)$ 分别表示变换前后的语音特征向量, $\boldsymbol{\zeta}_t = [1, \boldsymbol{o}(t)]^{\mathrm{T}}$ 表示拓展的语音特征向量. 在 CMLLR 自适应中, 为了减少计算量, 通常采用式 (3) 对特征进行变换来实现声学模型的自适应.

本文将采用 CMLLR 方法对输入的语音进行无监督的自适应, 对训练和测试的语音段中由于非语种因素而引起的不匹配进行补偿. 具体实现步骤如下:

步骤 1. 利用音素识别器对输入的语音进行维特比解码, 得到最优的音素序列;

步骤 2. 将识别得到的最优音素序列作为语音的真实标注, 采取最大似然准则对 CMLLR 的变换矩阵进行迭代估计;

步骤 3. 根据估计得到的变换矩阵, 利用式 (3) 对输入语音的特征向量进行变换;

步骤 4. 利用音素识别器对变换后的特征向量进行解码, 生成音素格, 基于音素格对音素的 N 元文法进行统计.

2.2 基于全局 N 元文法的音位特征自适应

目前, 解决 N 元文法的稳健估值问题主要有两种途径[8]: 一是利用低阶的 N 元文法进行插值; 二是利用在整个训练集上统计得到的全局 N 元文法进行自适应. 由于对于 10 s 以下的短时语音段, 低阶 N 元文法的估值不够稳健, 因此, 本文将利用全局 N 元文法进行自适应.

与基于最大后验概率的语言模型自适应相似, 利用全局 N 元文法对单个训练样本的特征向量进行自适应可以通过插值来实现, 表达式如下所示:

$$p'(d_i|\ell) = \alpha p(d_i|\ell) + (1 - \alpha) p(d_i|\text{all})$$

其中, $p'(d_i|\ell)$, $p(d_i|\ell)$ 分别表示自适应前后所对应的 N 元文法 d_i 的概率值, $p(d_i|\text{all})$ 表示对所有训练样本进行统计

得到的 N 元文法 d_i 的概率值, α 为插值系数, 一般通过实验来确定. 接下来, 将利用自适应得到的特征向量 $\boldsymbol{X}' = [p'(d_1|\ell), p'(d_2|\ell), \cdots, p'(d_F|\ell)]^{\mathrm{T}}$ 来建立 VSM 语种模型或者对模型进行匹配打分.

2.3 SVM 自适应

对于 SVM 等有监督的分类器而言, 一般假设训练和测试数据都服从相同的分布. 然而, 由于一些干扰因素的存在, 导致测试和训练条件不匹配, 从而使得测试和训练数据在分布上产生一定的偏移, 这会降低 SVM 的分类性能. 为了解决该问题, 有研究人员提出通过利用少量与测试数据相匹配的带标注的数据, 对现有的分类器进行自适应, 从而在一定程度上消除由干扰因素所造成的概率分布上的差异[15]. 该方法主要是通过对 SVM 目标函数中的正则化因子进行修正来实现, 修正的准则就是不仅要最小化 SVM 分类器在训练集上的分类错误, 同时还要尽量减少自适应后的 SVM 分类器和原始 SVM 之间的不一致性. 本文将在 VSM 模型中借鉴该方法.

令 $\{\boldsymbol{X}_i, y_i\}$ $(i = 1, \cdots, M)$ 表示与测试数据相匹配的带标注的样本集, 有 M 个样本, $y_i \in \{-1, 1\}$ 为二进制标注. 对于线性 SVM 分类器, 由式 (1) 和式 (2) 可知其输出表达式为

$$f(\boldsymbol{X}) = \sum_i \alpha_i y_i b(\boldsymbol{X})^{\mathrm{T}} b(\boldsymbol{X}_i) + d = b(\boldsymbol{X})^{\mathrm{T}} \boldsymbol{w}$$

其中, $\boldsymbol{w} = \left[\sum_i \alpha_i y_i b(\boldsymbol{X}_i) \;\; d\right]^{\mathrm{T}}$. $f(\boldsymbol{X})$ 的目标优化函数可表示如下:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N} \xi_i \tag{4}$$

其中, $\xi_i \geq 0$; $y_i x_i^{\mathrm{T}} \boldsymbol{w} \geq 1 - \xi_i, \forall i$; C 为惩罚因子. 由于信道等干扰因素的存在, 导致训练得到的 SVM 模型参数和测试集对应的模型参数之间会存在一定的偏移. 假设训练得到的 SVM 分类器的函数表达式为

$$f'(\boldsymbol{X}) = b^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{w}', \qquad \boldsymbol{w}' = \left[\sum_i \alpha_i' y_i' b(\boldsymbol{X}_i') \;\; d'\right]^{\mathrm{T}}$$

其中, $\{\boldsymbol{X}_i', y_i'\}$ $(i = 1, \cdots, M')$ 表示大小为 M' 的带标注的训练样本集. 一般来说, 很难获取大量和测试数据相匹配的训练样本, 因此并不能利用这一部分数据单独进行训练来得到一个稳健的分类模型. 尽管样本集 $\{\boldsymbol{X}_i', y_i'\}$ 和 $\{\boldsymbol{X}_i, y_i\}$ 对应的模型参数不一致, 但是两个样本集是密切相关的. 因此, 可将 $f'(\boldsymbol{X})$ 作为初始化模型, 利用少量和测试条件相匹配的数据进行自适应来得到新的 SVM 模型参数. 由于自适应的目的就是尽量消除训练和测试条件的不匹配, 在目标优化函数中就表现为最小化两个模型参数 $\boldsymbol{w}$ 和 $\boldsymbol{w}'$ 之间的不一致性, 因此, 可通过对式 (4) 中的正则化因子 $||\boldsymbol{w}||^2$ 进行修正来实现[15]:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w}'||^2 + C\sum_{i=1}^{N} \xi_i \tag{5}$$

对式 (5) 采用拉格朗日乘子法, 将与 $f'(\boldsymbol{X})$ 有关的模型参数视为常量, 则式 (5) 的最小化问题可转为二次规划问题, 一般采用序贯最小化算法对其进行求解. 最终可得自适应后的 SVM 模型参数 $\boldsymbol{w}$ 的表达式:

$$\boldsymbol{w} = \boldsymbol{w}' + \sum_i \alpha_i y_i b(\boldsymbol{X}_i) \tag{6}$$

由式 (6) 可见, 自适应后的模型可以看作在初始化模型中增加了新的支持向量, 这些支持向量都是通过利用自适应数据 $\{\boldsymbol{X}_i, y_i\}$ 进行 SVM 训练而得到的.

直接应用式 (6) 对 VSM 模型进行自适应意味着原始的训练数据和新的自适应数据在自适应过程中起着同样重要的作用. 但是这种前提条件在实际应用中较难满足, 主要有以下几个原因: 1) 一般用于自适应的数据比较少, 因此很难对测试数据的概率分布准确地进行估计; 2) 训练和测试数据的分布偏差太大, 以至于在自适应时原始模型 $\boldsymbol{w}'$ 基本不起作用; 3) 在很难获得与测试数据相匹配的带标注的自适应数据时, 需要首先利用现有的模型对测试数据进行自动标注, 一般这种自动标注可信度有限, 因此, 并不能利用这些数据进行可靠的自适应. 在上述情况下, 需要在自适应的同时根据不同模型的性能来进行调整. 为此, 本文提出在式 (5) 中加入加权因子进行进一步的修正:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w} - \beta\boldsymbol{w}'||^2 + C\sum_{i=1}^{N} \xi_i \tag{7}$$

其中, 加权因子 β 可以通过实验来确定. 易得式 (7) 的最优解为

$$\boldsymbol{w} = \beta\boldsymbol{w}' + \sum_i \alpha_i y_i b(\boldsymbol{X}_i) \tag{8}$$

3 PRVSM 系统的联合自适应

在第 2 节介绍的各种自适应方法中, 基于 CMLLR 的声学模型自适应是为了消除训练和测试语音在信道方面的差异, 而基于全局 N 元文法的自适应是为了获得高阶 N 元文法的稳健估值, SVM 自适应则是为了去除由于通话内容的不同所带来的影响. 这三种自适应方法都是应用在系统的不同阶段, 从不同的角度来获得稳健的音位特征以及 VSM 模型. 因此, 本文将综合利用各种自适应技术来对系统进行联合自适应. 具体实现如图 2 所示.

首先利用音素识别器对输入语音的 CMLLR 变换矩阵进行估计, 然后利用 CMLLR 矩阵对输入语音的特征进行变换, 利用变换后的特征进行音素识别得到音素格; 接下来首先利用所有的训练样本对全局 N 元文法进行统计, 然后利用统计得到的全局 N 元文法对每个训练样本的 N 元文法进行自适应, 将自适应后的 N 元文法进行拼接, 得到音位特征向量; 最后利用自适应后的音位特征向量进行 SVM 训练得到初始的 VSM 模型, 基于该模型对测试语音进行语种识别, 将识别结果作为标注来进行 SVM 自适应, 对初始 VSM 模型进行更新, 得到最终的自适应后的 VSM 模型, 基于该模型对测试语音的音位特征进行匹配打分, 得到最终的语种识别分数.

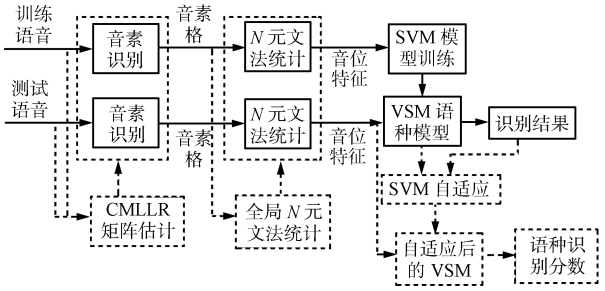

图 2　PRVSM 系统的自适应框图

Fig. 2　Adaptation to the PRVSM system

在 PRVSM 系统中, 这三种自适应算法的基本原理以及具体实现都不尽相同. CMLLR 自适应可以对测试和训练数据信道条件的不匹配进行补偿, 获得稳健的音素识别结果, 从而有利于提取稳健的音位特征并对其进行建模. 音位特征向量的自适应可以较好地解决基于单样本进行统计所导致的数据稀疏问题, 从而便于更好地对 N 元文法进行估计来提取音位特征. 最后, SVM 自适应可以在一定程度上解决训练和测试数据的音位特征分布不一致的问题, 从而有利于建立更好的 VSM 模型. 但是 CMLLR 自适应并不能完全保证可以消除信道条件的不匹配所带来的影响, 而且, 这种不匹配对音素识别结果所造成的影响也可以通过音位特征的自适应或是 SVM 自适应来进行补偿. 与此类似, 音位特征向量自适应也不能保证对所有的 N 元文法都能稳健地进行估值, 由此引发的音位特征在分布上的差异也可以利用 SVM 自适应来进行修正. 综上所述, 联合采用各种自适应算法应该可以获得比单个的自适应算法更好的识别性能. 但是, 根据上面的分析可知, 对于测试和训练数据在某些方面 (如信道) 的不匹配, 可以利用不同的自适应方法从不同的角度来进行补偿. 因此, 这三种不同的自适应方法在某些层面也存在着一定程度的重叠, 从而导致不同自适应方法之间的互补性较弱. 也就是说, 并非所有的自适应方法都可以有效地加入系统中进行联合自适应, 而是需要对不同的组合加以考虑, 通过对比来选择一种最优的组合方案, 以尽量提高系统的语种识别性能.

4 实验设置及结果

4.1 数据库

本文的实验在 2009 年度 NIST 语种识别评测的测试库上进行. 测试任务是要区分 23 个不同的语种: 阿姆哈拉语、波斯尼亚语、汉语广东话、海地克里奥尔语、克罗地亚语、达里语、美国英语、印度英语、波斯语、法语、格鲁吉亚语、豪撒语、印地语、韩语、汉语普通话、普什图语、葡萄牙语、俄语、西班牙语、土耳其语、乌克兰语、乌尔都语、越南语[16]. 总共有 41 793 段测试语音, 其中, 30 s、10 s 和 3 s 三种名义长度的测试段分别有 14 166、13 847 和 13 780 段. NIST LRE 2009 的测试数据的信道条件主要有两种: 会议电话和广播. 对于会议电话语音, 主要采用 CallFriend 数据库、OGI 数据库和 NIST 提供的历年的开发集和测试集; 而对于广播语音, 则主要采用 NIST 提供的美国之音 (Voice of American, VOA) 中的 VOA3 数据集. 开发集数据主要来自 LRE03、LRE05 以及没有参与训练的部分 VOA3 数据集. 由于数据库中的语音较长, 因此先利用语音端点检测将语音切分成小的语音段, 每一段中有效语音的持续时间约为 30 s.

4.2 参数设置和系统描述

本文利用中文音素识别器来对语音进行解码得到音素格. 该中文识别器的训练数据库为约 30 h 的会议电话数据. 采用基于 HMM 的声学模型, 其中每个 HMM 为 3 状态的上下文相关的 HMM, 每个状态采用 GMM 进行建模, 每个 GMM 模型由 32 个分量构成.

与以往的音素识别器不同, 本文将采用 52 维的时频倒谱 (Time-frequency cepstral, TFC) 特征取代传统的梅尔域倒谱系数和感知域线性预测系数特征. 对于特征提取, 语音信号先去直流和预加重 (因子为 0.97), 然后经过帧长为 25 ms、帧移为 10 ms 的汉明窗, 最后将连续 9 帧的基本特征向量拼接成时频倒谱矩阵, 对该矩阵进行离散余弦变换 (Discrete cosine transform, DCT), 具体的 TFC 特征的提取过程可参考文献 [17]. 为了去除信道卷积噪声, 本文对 TFC 特征进行了倒谱均值减和方差归一化. 同时, 为了提高音素识别器的性能, 在最大似然训练的基础上, 还采用最小音素错误特征训练法对音素模型进行区分性训练.

为了提高 VSM 模型的区分性, 采用与文献 [9] 类似的方法来选择高阶 N 元文法的区分性子集. 对于 CMLLR 自适应技术, 采取对语音和非语音类分别估计变换矩阵, 并且只对大于 6 s 的语音段进行自适应[7]. 对于基于全局 N 元文法的音位特征自适应和 SVM 自适应中的可变参数, 均利用开发集上的实验结果进行确定.

4.3 实验结果

本文将采用 NIST 的标准指标: 等错误率 (Equal error rate, EER) 和平均检测代价函数 (Decision cost function, 表示为 Cavg), 在 NIST LRE 2009 测试库上 (包括三种测试条件: 30 s、10 s 和 3 s) 对算法的有效性进行验证. 各实验详细设计和结果如下.

4.3.1 自适应技术实验结果

在本文的系统中, 用到的自适应技术较多. 因此将首先对各种自适应技术的性能进行对比, 然后再研究它们之间的互补性, 并在最后综合运用不同的自适应技术进行联合自适应. 实验结果如表 1 所示. 由表 1 可见, 每种自适应技术对系统性能的提升都是有益的. 总的来说, 声学模型自适应不如 SVM 自适应和音位特征自适应, 这表明信道条件的不匹配对 PRVSM 系统没有特别显著的影响, 而基于声学特征的语种识别系统则对信道条件比较敏感, 这也是基于音素解码的语种识别系统的一大优势. 当综合运用各种自适应技术时, 与单个的自适应技术相比, 系统的性能均有所提升, 这是因为这三种自适应技术分别运用在系统的不同阶段: 音素识别、音位特征提取和建模, 使得各种技术之间能够有效互补. 但是, 有一点比较特殊的是, 对于 10 s 和 3 s 的测试段, 在进行 CMLLR 自适应后性能均有所下降. 这是因为限定对长度大于 6 s 的语音段进行 CMLLR 自适应, 训练集的语音段一般都在 30 s 左右, 大部分都能进行 CMLLR 自适应, 而对于 10 s 和 3 s 的测试段, 大部分都没有进行 CMLLR 自适应, 导致训练和测试的不匹配, 从而影响系统的最终性能. 在综合运用各种补偿技术之后, 系统的性能反而不如不用 CMLLR, 即表 1 中的 (b + c). 这主要是由于联合采用 b 与 c 这两种自适应技术已经可以使系统的性能变得比较稳健, 若使系统的性能得到进一步的提高, 应采用更可靠的自适应方法; 而 CMLLR 自适应依赖于音素识别结果, 通常音素识别的错误率较高 (50 % 左右), 因此, CMLLR 自适应的效果一般不太理想, 若在 b + c 的基础上再加上 CMLLR 自适应, 反而会使系统的识别性能有所下降. 因此, 本文将综合运用基于全局 N 元文法的音位特征自适应和 SVM 自适应技术对系统进行联合自适应.

4.3.2 系统融合性能

这部分实验将对基于不同音素识别器的 PRVSM 系统自适应前后的性能进行比较, 系统的性能以及融合后的结果如表 2 所示. 表 2 中的捷克语、匈牙利语以及俄语音素识别器由布尔诺理工大学开发[11]. 从表 2 可知, 在自适应后, 基于单个音素识别器的 PRVSM 系统的性能均有明显的提高, 对于 30 s、10 s 和 3 s 的测试段, EER 分别相对降低了 18 %～23 %、12 %～20 % 以及 5 %～9 %. 与此同时, 本文也对多个 PRVSM 系统进行融合组成 PPRVSM 系统来验证提出的自适应算法的有效性. 从实验结果来看, 尽管融合多个音素识别器的 PPRVSM 系统的性能一般来说比较稳健, 但是在自适应后, 系统的性能仍有一定幅度的提升, 在 30 s、10 s 和 3 s 的测试库上, EER 分别相对降低了 3.2 %、15.2 % 以及 6.9 %. 不过在 30 s 的测试条件下, PPRVSM 系统性能的改

表 1 自适应技术实验结果 (%)
Table 1 Experimental results for adaptation (%)

自适应技术	30 s		10 s		3 s	
	EER	Cavg	EER	Cavg	EER	Cavg
PRVSM 基线系统	2.92	2.95	7.17	7.40	19.05	19.27
基于 CMLLR 的声学模型自适应 (a)	2.77	2.79	7.65	7.91	19.11	19.45
基于全局 N 元文法的音位特征自适应 (b)	2.65	2.67	6.85	7.08	18.63	18.97
SVM 自适应 (c)	2.39	2.38	5.80	6.00	17.46	17.69
a + b	2.54	2.53	7.38	7.53	19.00	19.36
a + c	2.29	2.31	6.53	6.66	17.75	18.00
b + c	2.23	2.23	5.71	5.87	17.33	17.54
a + b + c	2.30	2.28	6.31	6.41	17.65	18.00

表 2 系统融合实验结果 (%)
Table 2 Experimental results for fusion (%)

测试库	方法	捷克语		匈牙利语		俄语		汉语		融合	
		EER	Cavg	EER	Cavg	EER	Cavg	EER	Cavg	EER	Cavg
30 s	联合自适应前	3.87	3.87	3.09	3.10	2.86	2.84	2.92	2.95	1.56	1.52
	联合自适应后	2.96	3.02	2.53	2.54	2.18	2.12	2.23	2.23	1.51	1.47
10 s	联合自适应前	9.28	9.41	7.72	7.88	7.22	7.23	7.17	7.40	4.06	4.09
	联合自适应后	8.11	8.17	6.68	6.75	5.93	5.92	5.71	5.87	3.44	3.53
3 s	联合自适应前	22.92	22.93	20.94	20.93	19.38	19.29	19.05	19.27	14.05	14.21
	联合自适应后	21.77	21.75	19.50	19.63	18.01	18.07	17.33	17.54	13.08	13.11

善并不十分明显, 有两个主要原因: 1) 一般来说, 不同音素识别器的识别结果具有一定的互补性, 因此通话内容的不同对 PPRVSM 系统的性能并没有太大的影响; 2) 与 10 s 以及 3 s 的短时语音段相比, 基于 30 s 的长时语音对 N 元文法进行统计可以得到更准确的统计值. 因此, 在 30 s 的测试库上对 PPRVSM 系统进行自适应并不会给系统带来很大的增益.

5 结论

针对 PRVSM 系统的训练和测试条件不匹配以及 N 元文法的稳健估值问题, 本文提出了联合自适应算法. 为了对系统进行联合自适应, 首先研究了三种不同的自适应技术: 基于 CMLLR 的声学模型自适应、基于全局 N 元文法的音位特征向量自适应以及 SVM 自适应技术, 这三种技术被应用于系统的不同阶段来解决不同的问题. 与此同时, 还研究了不同自适应技术之间的互补性来有效地对系统进行联合自适应.

实验结果表明, 在采用联合自适应技术之后, PRVSM 系统的性能可以得到显著的提高. 尤其是在综合采用基于全局 N 元文法的自适应和 SVM 自适应技术对系统进行自适应后, PRVSM 系统能够得到最佳性能, 最终融合系统的 EER 在 NIST LRE 2009 的三种测试条件: 30 s、10 s 以及 3 s 测试库上分别达到了 1.51 %、3.44 % 和 13.08 %. 在下一阶段的工作中, 将研究如何提高基于 CMLLR 的声学模型自适应算法在 10 s 和 3 s 的测试库上的有效性, 并与其他自适应技术进行有效的互补.

References

1 Zissman M A. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 1996, **4**(1): 31−44

2 Li H, Ma B, Lee C H. A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(1): 271−284

3 Gauvain J L, Messaoudi A, Schwenk H. Language recognition using phone lattices. In: Proceedings of the 8th International Conference on Spoken Language Processing. Jeju Island, Korea: ISCA, 2004. 1283−1286

4 Lippmann R P, Carlson B A. Speech recognition by humans and machines under conditions with severe channel variability and noise. In: Proceedings of the Applications and Sciences of Artificial Neural Networks. Orlando, USA: SPIE, 1997. 46−57

5 Huang X D, Lee K F. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1993, **1**(2): 150−157

6 Shen W, Reynolds D. Improved phonotactic language recognition with acoustic adaptation. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium: ISCA, 2007. 358−361

7 BenZeghiba M F, Gauvain J L, Lamel L. Context-dependent phone models and models adaptation for phonotactic language recognition. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia: ISCA, 2008. 313−316

8 Xu B, Song Y, Dai L. The adaptation schemes in PR-SVM based language recognition. In: Proceedings of the 6th International Symposium on Chinese Spoken Language Processing. Kunming, China: IEEE, 2008. 1−4

9 Richardson F S, Campbell W M. Language recognition with discriminative keyword selection. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, USA: IEEE, 2008. 4145−4148

10 Matejka P, Schwarz P, Cernocky J, Chytil P. Phonotactic language identification using high quality phoneme recognition. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: ISCA, 2005. 2237−2240

11 Tong R, Ma B, Li H, Chng E S. Target-oriented phone tokenizers for spoken language recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, USA: IEEE, 2008. 4221−4224

12 Campbell W M, Campbell J P, Reynolds D A, Jones D A, Leek T R. Phonetic speaker recognition with support vector machines. In: Proceedings of the Advances in Neural Information Processing System. Vancouver, Canada: MIT Press, 2003. 1377−1384

13 Leung C C, Ma B, Li H. Parallel acoustic model adaptation for improving phonotactic language recognition. In: Proceedings of the Speaker and Language Recognition Workshop. Brno, Czech Republic: ISCA, 2010. 246−250

14 Digalakis V V, Rtischev D, Neumeyer L G. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 1995, **3**(5): 357−366

15 Yang J, Yan R, Hauptmann A G. Adapting SVM classifiers to data with shifted distributions. In: Proceedings of the 7th IEEE International Conference on Data Mining Workshops. Omaha, USA: IEEE, 2007. 69−76

16 National Institute of Standards and Technology. The 2009 NIST Language Recognition Evaluation Plan (LRE09) [Online], available: http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf, December 4, 2011

17 Zhang W Q, He L, Deng Y, Liu J, Johnson M T. Time-frequency cepstral features and heteroscedastic linear discriminant analysis for language recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(2): 266−276

邓 妍 清华大学电子工程系博士研究生. 主要研究方向为语种识别与说话人识别. 本文通信作者. E-mail: dengyan05@gmail.com
(**DENG Yan** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. Her research interest covers language recognition and speaker recognition. Corresponding author of this paper.)

张卫强 清华大学电子工程系助理研究员. 主要研究方向为时频分析, 高阶统计量, 音频检索, 说话人识别, 语种识别.
E-mail: wqzhang@tsinghua.edu.cn
(**ZHANG Wei-Qiang** Assistant professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers time-frequency analysis, higher-order statistics, audio retrieval, speaker recognition, and language recognition.)

刘 加 清华大学电子工程系教授. 主要研究方向为语音识别, 说话人识别, 语种识别, 表现力语音合成, 语音编码, 语言理解.
E-mail: liuj@tsinghua.edu.cn
(**LIU Jia** Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition, speaker recognition, language recognition, expressive speech synthesis, speech coding, and spoken language understanding.)