区分性模型组合中基于决策树的声学上下文建模方法

黄 浩 [1,2] 李兵虎 [1] 吾守尔 · 斯拉木 [1,2]

摘 要 上下文相关的区分性模型组合的局限在于引入大的模型权重参数集, 在数据有限时容易导致区分性权重训练过拟合. 针对该问题, 本文提出利用决策树进行上下文建模, 采用最小音子错误准则构建决策树以获得最优上下文相关权重参数集. 决策树构造过程中通过评估目标函数的一阶近似增量来加速最优问题集的选择, 并利用精细问题集来获得更好的声学区分能力. 基于多模型组合的语音识别实验表明, 该方法能够增强权重训练对过拟合的鲁棒性, 在大幅减小参数数量的情况下降低误识率, 并优于在特征空间进行组合的方法.

关键词 区分性模型组合, 上下文建模, 声学决策树, 最小音子错误, 语音识别

引用格式 黄浩, 李兵虎, 吾守尔 · 斯拉木. 区分性模型组合中基于决策树的声学上下文建模方法. 自动化学报, 2012, **38**(9): 1449−1458

DOI 10.3724/SP.J.1004.2012.01449

Discriminative Model Combination Using Decision Tree Based Phonetic Context Modeling

HUANG Hao[1,2] LI Bing-Hu[1] SILAMU Wushour[1,2]

Abstract One limitation of context dependent discriminative model combination is that a large number of parameters will be introduced, which is liable to overtraining with limited training data. We propose context modeling using phonetic decision trees in lattice based discriminative model combination. Question in tree node is chosen to optimize the minimum phone error criterion. First order approximation of the objective function increment is used for fast question selection. Results on speech recognition show that the method is capable of improving the robustness to overtraining and obtains error reduction with many fewer parameters. It is also shown that the model combination using tree based context modeling is superior to feature combination approach.

Key words Discriminative model combination, context, decision tree, minimum phone error, speech recognition

Citation Huang Hao, Li Bing-Hu, Silamu Wushour. Discriminative model combination using decision tree based phonetic context modeling. *Acta Automatica Sinica*, 2012, **38**(9): 1449−1458

二次解码是将多种模型信息加入语音识别系统进行融合的后处理过程. 二次解码过程是: 先进行一次解码产生 N 最佳列表 (N-best lists) 或者格 (Lattice) 作为搜索空间, 然后将更为复杂的模型得分加入搜索空间进行组合获得更为准确的识别输出. 在二次解码中, 传统的方法是使用人工经验选取的模型权重值来调节各种模型得分的作用程度. 而在权重参数数量增多时, 依靠人工经验对权重参数取值将变得十分困难. 因此可以利用区分性模型组合[1] 的方法对其进行改进: 将多种模型得分通过区分性训练得到模型权重参数进行加权, 然后再寻找后验概率最大的路径作为最终输出结果. 在文献 [2] 中, 我们提出根据最小音子错误 (Minimum phone error, MPE) 区分性训练准则, 利用扩展 Baum Welch 算法对谱特征模型和声调模型权重进行优化, 来改进汉语语音识别中声调模型集成的性能. 文献 [3−5] 提出了利用区分性训练的模型权重进行声学模型和语言模型的组合. 这些工作都显示出在不同识别任务上、不同模型之间进行区分性模型组合的有效性.

上述工作的另一个共同点就是都采用了上下文相关的权重参数对模型得分进行加权. 本文称之为上下文相关的模型组合. 上下文相关的模型组合的优点在于其能够考虑到当前或者更长范围的语音/语义情景, 从而获得更高精度的识别结果. 但是上下文相关的模型组合由于引入的上下文因素增多, 将会产生大的权重参数集合. 而大参数集合在进行区

收稿日期 2011-07-25 录用日期 2012-03-05
Manuscript received July 25, 2011; accepted March 5, 2012
国家自然科学基金 (60965002, 60865001, 61163026), 新疆高校科研计划培育基金 (XJEDU2008S15), 新疆大学博士科研启动基金 (BS090143) 资助
Supported by National Natural Science Foundation of China (60965002, 60865001, 61163026), Scientific Research Program of the Higher Education Institution of Xinjiang (XJEDU2008S15), and Ph. D. Research Fund in Xinjiang University (BS090143)
本文责任编委 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
1. 新疆大学信息科学与工程学院 乌鲁木齐 830046 2. 新疆大学多语种信息技术实验室 乌鲁木齐 830046
1. Department of Information Science and Engineering, Xinjing University, Urumqi 830046 2. Laboratory of Multi-lingual Information Technology, Xinjiang University, Urumqi 830046

分性权重参数训练时, 由于数据量的不足, 将可能导致训练过拟合, 从而使识别性能下降. 在另一方面, 当上下文可选项增多时, 反复人工试凑选取最有效的声学上下文相关权重参数集将变得十分困难.

针对上述问题, 我们提出在区分性模型组合中进行上下文相关参数集的自动选取方法. 我们采用声学决策树对上下文进行建模, 决策树的叶子节点描述一组特定的声学上下文, 这些声学上下文共享一个模型权重参数. 决策树的构造根据 MPE 目标函数的最优化来进行, 即决策树的节点分裂将最小化训练数据集上的期望错误率. 由于决策树构造过程中需要同时进行参数的优化并评估节点分裂带来目标函数增量, 需要大量格结构的前后向计算, 这将使得决策树的构建十分耗时. 本文提出一种快速的最佳节点问题集选取方法来加快决策树的构建速度. 在备选问题集设计方面, 我们提出细分问题集的设计来加大聚类节点中的区分度. 在微软连续语音数据库上进行的区分性模型组合实验表明了该方法的有效性. 我们将基于隐马尔科夫模型 (Hidden Markov model, HMM) 的谱特征模型, 基于重叠双声调高斯混合声调模型 (Overlapped di-tone Gaussian mixture model, ODGMM), 以及基于多层感知器 (Multi-layer perceptron, MLP) 的神经网络音素分类模型在格上进行模型组合. 实验表明基于决策树的上下文建模能够在大大减小权重参数数量的条件下降低误识率. 这说明该方法能够减小区分性模型权重训练中过拟合的影响. 实验还表明采用区分性模型组合的系统优于采用美尔频率倒谱系数, 平滑基音频率特征与 MLP 后验概率特征合并的 TANDEM[6] 识别系统.

1 上下文相关的区分性模型组合

1.1 基于格的模型组合

在基于格的二次解码中, 格当中每条边的总得分等于几个并行模型得分之和:

$$\psi(a) = \sum_{i}^{I} \lambda_i \psi_i(a) \tag{1}$$

其中, $\psi(a)$ 是 Lattice 中第 a 条边总的模型得分, $\psi_i(a)$ 是该边第 i 个并行模型的得分, λ_i 为各模型的可调节模型权重. 本文将 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_I)$ 称为一个权重参数.

1.2 上下文相关的模型组合

在采用式 (1) 对模型得分进行加权时, 传统方法主要采用人工经验选定全局权重参数. 但是全局权重不能根据不同声学上下文进行模型间插值. 图 1 显示了一个格的简单结构, 格当中每条边显示了当前以及前后音素的类型. 可以看出每条非静音边的声学上下文可以表示为 [c:d/a-b+e], 其中 “a” 表示当前音节的声母类型; “b” 表示当前音节的韵母类型; “c” 表示前一个音节的韵母类型; “d” 表示前一个音节的声调类型; “e” 表示后一个音节的声母类型. 上下文相关的模型权重就是根据前后的发音类型使用不同的权重进行模型得分的调节. 如图 1 中虚线框内的边, 如果权重依赖于当前音节类型, 则格结构中音节类型为 d-ou1 的所有边可赋予一个权重值; 如果权重依赖于韵母三音子类型, 则格结构中所有韵母三音子为 d-ou1+sh 的边可赋予一个权重值.

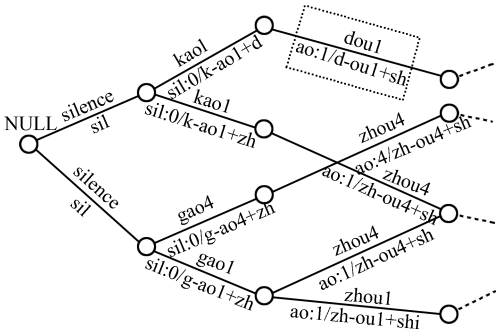


图 1 格及其声学上下文

Fig. 1 Lattice and phonetic contexts

可以看出, 当加入更多的上下文可选项时, 不同的上下文相关权重参数将会急剧增多. 在进行权重参数训练时, 如果有些上下文相关权重没有足够的训练样本, 或者并不具有区分能力的上下文权重参数参与权重训练, 就会导致训练过拟合从而降低识别性能. 因此需要采取合理的上下文建模方法来克服数据稀疏的问题. 我们将在第 2 节讨论该问题.

1.3 区分性模型权重训练目标函数

如果使用全局权重进行模型组合, 最优权重可以通过对校验数据进行性能评估、不断试凑权重值来获得. 而采用上下文相关的权重参数时, 由于可调节参数数量增加, 人工选取这些参数值将变得十分困难. 这时可以采用 MPE 准则[7−8] 对权重参数进行自动调整, 通过优化 MPE 目标函数来降低训练数据的期望误识率. MPE 是目前语音识别中声学建模常用的区分性训练准则, 在国内外已有广泛研究, 在文献 [9] 中则讨论了利用最大似然训练和 MPE 训练得到的声学模型之间进行模型插值的方法. 给定一个具有 U 条语句的练训集 $\mathcal{O} = \{\mathcal{O}_1, \cdots, \mathcal{O}_u, \cdots, \mathcal{O}_U\}$, MPE 准则定义为训练集 Lattice 中所有句子中音子正确度的数学期望[7−8]:

$$\mathcal{F}_{\text{MPE}} = \sum_{u}^{U} \sum_{s}^{S} P^{\kappa}(s|\mathcal{O}_u) A(s, s_r) \tag{2}$$

其中, $\mathcal{O}_u$ 是第 u 条训练语句的观察序列, $P(s|\mathcal{O}_u)$ 为给定观察序列 $\mathcal{O}_u$ 时句子假设 (路径) s 的后验概率. κ 是减少模型概率动态范围的比例系数. $A(s, s_r)$ 是句子假设 s 的正确率测度, 可根据路径中每条边与标注文本 s_r 比对的正确率求和计算. 关于 MPE 方法的更多细节可参见文献 [7−8].

1.4　区分性模型权重训练的优化方法

设 m 是某个声学上下文, $\boldsymbol{\lambda}_m = (\lambda_{m,1}, \cdots, \lambda_{m,I})$ 与其相关的权重参数, $\lambda_{m,i}$ 是其中的第 i 个模型得分的权重, 区分性模型权重训练的目的就是通过调整 $\boldsymbol{\lambda}_m$ 来优化 MPE 目标函数. 设模型权重满足 $\lambda_{m,i} \geq 0$, 以及 $\sum_{i=1}^{I} \lambda_{m,i} = 1$, MPE 意义下的权重优化可按照文献 [2] 中提出的参数更新公式进行:

$$\lambda'_{m,i} = \frac{\kappa \sum\limits_{a \in \mathcal{A}_m} \gamma_a^{\mathrm{MPE}} \lambda_{m,i} \psi_i(a)|_\lambda + C\lambda_{m,i}}{\sum\limits_i \left(\kappa \sum\limits_{a \in \mathcal{A}_m} \gamma_a^{\mathrm{MPE}} \lambda_{m,i} \psi_i(a)|_\lambda + C\lambda_{m,i} \right)} \tag{3}$$

其中, $\lambda_{m,i}$ 是当前使用的权重值, $\lambda'_{m,i}$ 是更新之后的权重值. $\mathcal{A}_m$ 是所有与上下文 m 相关联的边的集合. $\gamma_a^{\mathrm{MPE}} = \gamma_a \left(c(a) - c_{\mathrm{avg}}\right)$ 是 MPE 方法中的重要累积量. γ_a 为格当中通过第 a 条边的后验概率. $c(a)$ 表示包含有弧 a 的所有句子假设的平均正确率. c_{avg} 为 Lattice 中所有句子假设的平均正确率. 这些参量可在格当中进行前–后向计算得到, 具体计算方法可参见文献 [7−8]. 式 (3) 中平滑常数的选取方法为 $C = E \sum_i |\kappa \gamma_a^{\mathrm{MPE}} \lambda_{m,i} \psi_i(a)|_\lambda|$, 其中 E 是正的平滑控制常数. 关于式 (3) 的推导及参数设置可参见文献 [2].

2　基于决策树的声学上下文建模方法

2.1　基于决策树的上下文建模及决策树的建立

在使用上下文相关权重参数时, 简洁的参数集对于保证权重训练的鲁棒性至关重要. 在语音识别中, 声学上下文的建模通常采用决策树进行, 如基于决策树的 HMM 状态绑定方法[10]. 本文将采用声学决策树对上下文相关权重进行建模: 决策树中每个非叶子节点放置一组问题, 一个上下文自根节点开始, 通过回答节点问题到达左/右子节点直至叶子节点. 所有叶子节点聚类的上下文共享一个权重参数, 这些权重参数采用式 (3) 进行训练优化. 识别阶段利用这些权重参数调节模型得分.

决策树的建立从根节点开始, 采用自顶向下的顺序, 按照一定的训练准则进行节点分裂. 在基于决策树的上下文状态绑定中[10], 节点分裂通常根据最大似然准则进行, 决策树的构造是为了最大化训练参考文本 (Reference) 的似然度. 对于二次解码的语音识别任务, 下面给出利用最大似然准则进行模型组合以及权重优化的分析:

最大似然训练常在训练数据的参考文本上进行, 图 2 给出了一条训练语句的参考文本. 由图可以看出, 参考文本的概率为句子中各个边的模型概率得分之和. 设训练集具有 U 条训练语句, 第 u 条语句具有 A_u 条边, 每条边有 I 个并行模型. 在使用全局权重时, 训练集总的似然度可以表示为

$$\mathcal{L} = \sum_{u=1}^{U} \sum_{a=1}^{A_u} \sum_{i=1}^{I} \lambda_i \psi_i(a) \tag{4}$$

调换求和顺序可得:

$$\mathcal{L} = \sum_{i=1}^{I} \lambda_i \left[\sum_{u=1}^{U} \sum_{a=1}^{A_u} \psi_i(a) \right] = \sum_{i}^{I} \lambda_i \Psi_i \tag{5}$$

其中, $\Psi_i = \sum_{u=1}^{U} \sum_{a=1}^{A_u} \psi_i(a)$ 为所有训练语句中第 i 个并行模型得分和. 设 Ψ_i, $i = 1, \cdots, I$ 中的最大值为 Ψ_j. 在满足 $\lambda_i \geq 0$ 以及 $\sum_{i=1}^{I} \lambda_i = 1$ 的条件下, 为最大化 $\mathcal{L}$, 只需将对 Ψ_j 的模型权重 λ_j 设置为 1.0, 其余模型权重设置为零, 目标函数 $\mathcal{L}$ 即可达到最大值. 在本文实验中, 对于每条边, 声调模型得分最大, 其取值范围在 $(-10, 0)$ 区间 (对数概率); 谱特征模型得分最小, 其取值范围在 $(-2\,000, -1\,000)$ 区间. 因此声调模型的概率和 Ψ_T 最大. 在这种情况下, 可将声调模型权重 λ_T 设置为 1.0, 谱特征模型权重 λ_A 及 MLP 模型权重 λ_M 设置为 0.0, 目标函数具有最大值. 其含义为: 解码过程中只考虑声调模型得分而将其他模型得分忽略, 从而失去了模型组合的意义. 上述针对于全局权重的分析同样适用于上下文相关权重. 根据以上分析, 在本文所讨论的模型组合任务中使用最大似然准则进行权重的调整以及决策树的建立是不合适的. 本文将采用 MPE 区分性训练目标函数并在格结构上构建决策树. 在基于决策树的 HMM 状态绑定中, 文献 [11−12] 提出使用区分性准则进行节点分裂构建决策树获得了良好的效果, 然而这些准则都是基于音素分类错误或者基于帧分类错误, 而非更为合理的面向连续语流的 MPE 目标函数, 再者根据这些准则构造决策树并没有在 Lattice 上进行, 从而使得上述方法不适合于在 Lattice 上的二次解码任务.

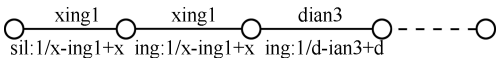


图 2　训练语句的参考文本

Fig.2　Reference of a training utterance

本文将 MPE 作为决策树构造的目标函数, 也就是说寻找训练数据期望错误率最小化意义下的声

学决策树. 在决策树构造中, 对于节点使用不同的问题集将导致不同的上下文参数聚类, 不同的参数聚类又会导致叶子节点更新后的权重参数不同, 进而导致不同的 MPE 目标函数的增量. 因此节点问题集的选择 (决策树的构造) 和模型权重的参数优化需要同步进行. 图 3 显示了结点分裂中最优问题的选择过程. 为不失一般性, 假设根节点 0 中聚集的上下文已经被问题集 QS_21 分裂为左右两个子节点: 节点 1 和节点 2, 接下来需要对节点 2 选择最优问题集. 首先对叶子节点 1 和 2 中的权重参数 $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ 从全局权重 $(\boldsymbol{\lambda}_g, \boldsymbol{\lambda}_g)$ 为起始点进行更新, 得到优化后的权重 $\tilde{\boldsymbol{\lambda}} = \{\tilde{\boldsymbol{\lambda}}_1, \tilde{\boldsymbol{\lambda}}_2\}$, 并在更新参数的基础上计算 MPE 目标函数 $\mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}})$. 第二步利用问题集 q 分裂节点 2, 形成三组可训练权重 $\boldsymbol{\lambda}^q = \{\boldsymbol{\lambda}_1^q, \boldsymbol{\lambda}_3^q, \boldsymbol{\lambda}_4^q\}$, 并从全局权重 $\boldsymbol{\lambda}^q = (\boldsymbol{\lambda}_g, \boldsymbol{\lambda}_g, \boldsymbol{\lambda}_g)$ 进行参数更新至 $\tilde{\boldsymbol{\lambda}}^q = \{\tilde{\boldsymbol{\lambda}}_1^q, \tilde{\boldsymbol{\lambda}}_3^q, \tilde{\boldsymbol{\lambda}}_4^q\}$, 然后计算该参数下的目标函数 $\mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}^q)$. 最后计算利用问题集 q 进行节点分裂带来的目标函数增量:

$$\mathcal{G}_{\rm MPE}(q) = \mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}^q) - \mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}) \tag{6}$$

则 MPE 意义下的决策树构造过程中节点分裂准则就是需要找到问题 q, 使得节点分裂之后带来最大的 MPE 目标函数增量:

$$q_{\rm best} = \arg\max_{q\in\mathcal{Q}} \mathcal{G}_{\rm MPE}(q) \tag{7}$$

其中, $\mathcal{Q}$ 是所有备选问题集. 从式 (6) 和式 (7) 可以看出, 当节点问题集 q 变化时, 需要对节点参数重新进行若干次迭代更新, 而每次更新总累积量 $\sum_{a\in\mathcal{A}_m}\gamma_a^{\rm MPE}$ 的计算需要对所有含有该上下文的格结构进行前后向计算. 当训练语句数量以及问题集数量较大时, 最优问题集的选择将十分耗时, 因此决策树构建中的快速问题集选择是十分重要的.

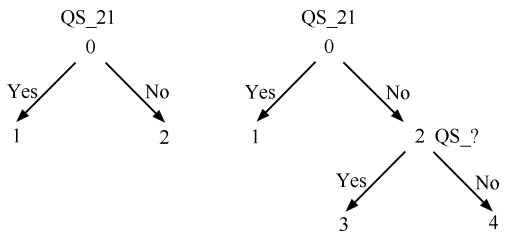


图 3 决策树节点分裂过程

Fig.3 Tree node splitting

2.2 问题集的快速选择

基于式 (3) 的权重参数优化过程一般在数次迭代后收敛至最优, 而第一次迭代更新带来的目标函数增量将占整个迭代过程中目标函数增量的绝大部分 (参见实验中目标函数随迭代的变化过程), 因此最优问题集的选取只需要评估第一次训练迭代后的目标函数增量. 因第一次迭代中目标函数对各边对数概率的导数 $\gamma_a^{\rm MPE}$ 可以预先计算并存储, 所以叶子节点的权重参数只需要根据叶子节点中聚类的训练样本分别进行累积然后进行参数更新, 而无需再进行前后向计算. 更新参数之后, 需要计算该参数下的目标函数增量. 在式 (6) 中第二项与问题 q 无关, 将其去除并不影响最优问题集的选取, 所以最优问题集选取准则可简化为

$$q_{\rm best} = \arg\max_q \mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}^q) \tag{8}$$

利用一阶泰勒级数将 $\mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}^q)$ 展开得到:

$$\mathcal{F}_{\rm MPE}(\tilde{\boldsymbol{\lambda}}^q) \approx \mathcal{F}_{\rm MPE}(\boldsymbol{\lambda}^q) + \left(\frac{\partial\mathcal{F}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right)^{\rm T}\Big|_{\boldsymbol{\lambda}^q}(\tilde{\boldsymbol{\lambda}}^q - \boldsymbol{\lambda}^q) \tag{9}$$

其中第一项为全局权重下的 MPE 目标函数值且为常量, 将上式等号右边第二项定义为 MPE 目标函数的近似增量, 其标量形式为

$$\mathcal{G}_{\rm MPE}^p(q) = \sum_m\sum_i \frac{\partial\mathcal{F}(\boldsymbol{\lambda})}{\partial\lambda_{m,i}}\Big|_{\boldsymbol{\lambda}_g}(\tilde{\lambda}_{m,i}^q - \lambda_{g,i}) \tag{10}$$

其中, $\lambda_{g,i}$ 是第 i 个模型得分的全局权重. 当式 (9) 的近似条件满足时, 最优问题集的选择可按照近似增量最大化进行:

$$q_{\rm best} = \arg\max_q \mathcal{G}_{\rm MPE}^p(q) \tag{11}$$

在式 (9) 中, 目标函数对节点权重参数的导数计算公式为

$$\frac{\partial\mathcal{F}(\boldsymbol{\lambda})}{\lambda_{m,i}}\Big|_{\boldsymbol{\lambda}_g} = \kappa\sum_{a\in\mathcal{A}_m}\gamma_a^{\rm MPE}\psi_i(a) \tag{12}$$

式中, $\sum_{a\in\mathcal{A}_m}$ 为训练数据 Lattice 中所有与该节点上下文 m 相关联的边 $a\in\mathcal{A}_m$ 的求和. 由于 $\gamma_a^{\rm MPE}$ 可以预先计算存储后直接使用, 因此可以无需再进行 Lattice 的前后向计算. 需要注意的是, 在更新后的参数与全局参数偏离不十分大的情况下, 式 (9) 中的近似条件才能满足. 本文把采用直接评估目标函数增长获得最佳节点分裂问题集的方法简称为 "精确分裂方法", 把采用评估近似增量获得最佳问题集的方法称为 "近似分裂方法".

3 实验与结果

本文采用微软亚洲研究院语音识别工具箱[13] 提供的带调音节输出实验来验证上下文建模方法在区分性模型组合中的有效性. 在识别过程中, 我们没有使用语言模型从而更好地讨论所提方法对声学性能的影响. 语音数据采样率为 16 bit/16 kHz. 训

练语料包含 100 个说话人发声的 19 688 条语句, 测试语料包含另外 25 个说话人发声的 500 条语句, 共计 9 570 个带调音节. 另外选取独立于上述数据的 2 000 句朗读语音作为开发集进行数据校验.

Lattice 数据的准备、决策树的构建、模型权重的训练以及输出测试结果的过程总结如下: 1) 利用 HMM 谱特征模型对训练数据和测试数据产生格结构, 通过强制对齐来获得格当中每条边的声母与韵母的起止时间; 2) 在韵母部分提取基频 F_0 值然后计算声调得分, 利用 MLP 音素分类器计算每条边声韵母音素后验得分; 3) 在标记好各模型得分的训练集的格结构上进行声学决策树构建; 4) 由于在树的生长过程中权重参数没有完全优化 (分裂过程中只进行了 1 次迭代更新), 因此在决策树建好以后, 需要利用式 (3) 在叶节点参数上进行 10 次迭代的区分性权重训练直至 MPE 目标函数收敛; 5) 使用优化过的权重参数调节测试数据格结构中的三种模型得分进行二次解码得到最终的输出序列, 即

$$\psi(a) = \lambda_{m,A}\psi_A(a) + \lambda_{m,T}\psi_T(a) + \lambda_{m,M}\psi_M(a) + \psi_{WP} \tag{13}$$

其中, $\psi_A(a)$, $\psi_T(a)$ 和 $\psi_M(a)$ 分别是谱特征模型得分, 声调模型得分和 MLP 音素后验概率得分. ψ_{WP} 是词惩罚值. m 表示决策树中的叶子节点, $\boldsymbol{\lambda}_m = (\lambda_{m,A}, \lambda_{m,T}, \lambda_{m,M})$ 分别为该节点三个模型得分的可调节权重. 节点 m 可以根据边 a 的上下文根据决策树得到. ψ_{WP} 是词惩罚值. 下面给出文中采用的局部模型/分类器的实验配置.

3.1　局部分类模型

3.1.1　谱特征模型

谱特征模型采用状态绑定的上下文相关三音子 HMM . 观察向量采用 39 维, 包括倒谱均值归一化的 12 阶美尔频率倒谱系数 (Mel frequency cepstral coefficient, MFCC)、归一化对数能量及其一阶、二阶导数. 先利用最大似然训练方法得到单音子模型, 再利用决策树的状态绑定将单音子模型集扩展为上下文相关三音子集. 谱特征模型共计 2 392 个绑定状态, 每状态 8 个混合高斯. 在最大似然训练的基础上, 再进行 MPE 区分性 HMM 参数训练获得谱特征模型的最佳性能. 然后对数据集进行解码产生格结构. 最后对格结构的每条边, 利用该边的谱特征模型和谱特征进行前后向计算得到该边的谱特征模型得分.

3.1.2　高斯混合声调模型

声调模型采用基于段特征的重叠双音调高斯混合模型[14]. 先利用谱特征模型对训练数据进行强制对齐获得声韵母分割, 然后对韵母部分 (浊音段) 提取基频值和对数能量值. 声调特征包含四个部分: 1) 当前音节浊音段时长归一化的基频特征; 2) 当前音节浊音段的对数能量; 3) 当前音节浊音段基频序列的一阶导数的平均值; 4) 前一个音节浊音段的时长归一化基频特征. 模型训练利用 k 均值聚类初始化高斯参数并进行期望最大算法更新高斯参数, 对于一声至轻声的 5 个声调模型, 混合高斯数分别为 10, 10, 9, 16, 3 个. 在期望最大化训练的基础上再利用大间隔[15] 训练获得更好的声调分类性能. 声调分类器在测试数据上的声调误识率为 28.5 %. 声调模型建立之后, 对 Lattice 每条边的浊音段 (韵母) 提取声调特征, 并计算该边声调的后验概率作为声调模型得分.

3.1.3　多层感知器音素分类模型

在 MLP 音素分类模型中, 对于每一时刻 t, 将当前帧和前后各 4 帧共 9 帧的 MFCC 特征连接作为输入. 在 39 维 MFCC 的情况下, 输入层共有 351 个输入节点, 隐层数目设为 5 000 个, 输出层的数目与单音素数目相同. 在采用声母、无调韵母作为建模单元的情况下共有 66 个音素后验概率输出 (包括 sil 和 sp 模型). MLP 音素分类模型训练自相同的训练数据, 利用 Quicknet 神经网络工具箱[16] 中的 qnstrn 进行 MLP 模型参数训练. 我们对训练数据和测试数据分别进行了基于帧的音素分类测试, 帧分类错误率分别为 21.4 % 和 23.3 %. 获得音素分类模型之后, 使用 qnsfwd 工具根据格的每条边的起止时刻计算边内各帧的音素后验概率. 每条边的 MLP 模型得分为该边内各帧音素后验概率之和. 表 1 总结了三种模型/分类器的识别结果.

表 1　局部模型/分类器性能

Table 1　Performance of the models/classifier

模型	测试名称	错误率 (%)
谱特征模型 (MLE)	带调音节识别	48.7
谱特征模型 (MPE)	带调音节识别	40.9
声调模型	声调分类	28.5
神经网络	音素分类	23.3

3.2　决策树构造与备选问题集设计

实验中将每个带调音节 (声 – 韵母组合) 作为决策树的根节点进行初始化, 因此共有 1 497 棵声学决策树. 由于式 (9) 中的近似条件并不总能满足, 因此采用两种节点分裂方法 (精确分裂和近似分裂) 得到的决策树并不完全相同, 但是从实验发现两种分裂方法得到的决策树节点问题集重叠度达到 60 %, 而且使用精确分裂方法得到的最佳问题集 q_{best} 总在

近似分裂得到的近似增量最大的前 N_q 个问题集中. 因此可采取问题集剪枝的办法加快效率: 先通过近似分裂方法保留近似增量最大的 N_q 个问题集, 然后再利用精确分裂方法从这 N_q 个问题集中选取目标函数增量最大的问题集. 实验表明当 $N_q \geq 3$ 时,采用问题集剪枝方法构造出的决策树与精确分裂方法获得的决策树完全等同. 另一种加速问题集选择的方法是直接使用近似分裂获得最优问题集, 虽然这种方法构造出的决策树与精确分裂得到的决策树有不同, 但从识别结果看, 两者并无明显差别. 而近似分裂决策树构造速度远远快于上述精确分裂和问题集剪枝的方法. 使用近似分裂准则的决策树构造过程在四核心 Intel Q9400 CPU 上进行并行计算, 整个决策树构造时间将小于 1 倍实时时间. 后续实验将直接给出利用近似分裂获得的决策树的结果.

备选问题集 $\mathcal{Q}$ 的设计也是获得最优决策树及最佳识别性能的关键. 实验中决策树问题集修改自微软语音工具箱[11] 原用于构造上下文相关三音子模型的问题集, 该问题集以汉语普通话的发音属性为设计依据. 由于决策树根节点已经表示为边的带调音节类型, 因此决策树问题将集中考虑如下声学情景: 该边前驱音节的韵母类型; 该边前驱音节的声调类型; 该边后续音节的声母类型; 该边是不是具有静音段的前驱或者后继. 该问题集中共有 99 组问题集, 称为备选问题集 1. 如下是备选问题集 1 的举例:

```
QS_0 {*:*/*-*+b,*:*/*-*+p,*:*/*-*+m}
QS_1 {*:*/*-*+b,*:*/*-*+d,*:*/*-*+g}
    ...
QS_35 {a:*/*-*+*,an:*/*-*+*,ao:*/*-*+*,
       ang:*/*-*+*,ai:*/*-*+*}
QS_92 {*:1/*-*+*}
QS_97 {sil:0/*-*+*}
    ...
QS_98 {*:*/*-*+sil}
```

3.3 结果与分析

首先给出基于全局权重的识别结果, 模型得分利用全局权重参数 $\boldsymbol{\lambda} = (\lambda_A, \lambda_T, \lambda_M)$ 加权. 在 $\lambda_A + \lambda_T + \lambda_M = 1$ 的条件下, 全局权重参数 $\boldsymbol{\lambda}$ 可以利用全局搜索 (Grid search) 的方法通过评估训练集上的 MPE 目标函数获得. 本文实验采用简化的办法: 先确定未归一化的全局权重 $\boldsymbol{\lambda}^* = (\lambda_A^*, \lambda_T^*, \lambda_M^*)$. 其方法为: 固定谱特征模型权重 $\lambda_A^* = 1$ 不变, 不断调整声调模型权重 λ_T^* 直至训练集上的 MPE 目标函数达到最优; 然后保持 λ_A^*, λ_T^* 固定不变, 不断调整 MLP 音素分类模型权重 λ_M^* 直至训练集上的 MPE 目标函数达到最优.

归一化的全局权重 $\boldsymbol{\lambda}$ 可使用 $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*/(\lambda_A^* + \lambda_T^* + \lambda_M^*)$ 获得. 实验中从上述过程中获得的未归一化权重为 $\boldsymbol{\lambda}^* = (1.0, 15.0, 0.8)$, 归一化之后得到 $\boldsymbol{\lambda} =$ (0.0595, 0.8929, 0.0476), 可以看到各模型权重处于不同的数量级. 为保证权重训练中数值计算的稳定性以及便于观察权重训练后权重值的变化程度, 我们使用固定调节因子 α, β, γ 获得归一化全局权重: $\lambda_A^* = \alpha\lambda_A$, $\lambda_T^* = \beta\lambda_T$, $\lambda_M^* = \gamma\lambda_M$, 实验中选取 $\alpha =$ 3.0, $\beta = 45.0$, $\gamma = 2.4$, 使得 $\boldsymbol{\lambda} = (\lambda_A, \lambda_T, \lambda_M) =$ (1/3, 1/3, 1/3), 此时每条边的得分计算式为

$$\psi(a) = \alpha\lambda_A\psi_A(a) + \beta\lambda_T\psi_T(a) + \lambda_M\gamma\psi_M(a) + \psi_{WP} \tag{14}$$

在后续实验中, 调节因子 α, β, γ 保持不变. 上下文相关的权重训练初始权重值均从全局权重 $\boldsymbol{\lambda} = (\lambda_A, \lambda_T, \lambda_M) = (1/3, 1/3, 1/3)$ 初始化.

表 2 列出了全局权重条件下不同模型互相组合的测试结果. 谱特征模型与声调模型组合时, 误识率从仅使用谱特征模型的 40.9 % 降为 34.8 %. 当谱特征模型与音素分类器组合时, 误识率为 37.1 %. 三种模型得分共同作用时, 误识率降为 32.7 %. 结果都表明了模型组合对识别性能的贡献.

表 2 利用全局权重组合结果

Table 2 Model combination using global weighting

谱特征模型	声调模型	音素分类	误识率 (%)	$-\Delta$ (%)
MPE	无	无	40.9	0
MPE	有	无	34.8	14.9
MPE	无	有	37.1	9.3
MPE	有	有	**32.7**	20.1

表 3 给出采用上下文相关模型权重进行模型组合的结果. 其中前半部分给出人工选取上下文的结果. 从 [c:d/a-b+e] 五种上下文可以任选其中的若干种组成参数集合, 实验中我们测试了如下的上下文组合: 1) 当前带调音节类型 (Center syllable, CT), 其形式为 [*:*/a-b+*], 用来考虑不同的带调音节类型; 2) 当前音节 + 左上下文 (Center syllable + Left context, CL), 用来考虑当前音节和前一个发音韵母类型, 其形式为 [c:d/a-b+*]; 3) 带调音节 + 右上下文 (Center syllable + Right context, CR), 用来考虑当前音节和下一个声母类型. 其形式为 [*:*/a-b+e]; 4) 当前音节 + 左右上下文 (Center syllable + Left and Right context, CLR), 可以同时考虑到当前音节类型, 前一个韵母类型和后一个声母类型, 形式为 [c:d/a-b+e]. 表 3 中同时给出了上

表 3　上下文相关模型组合结果

Table 3　Model combination using context dependent weighting

上下文	N_w	误识率 (%)	$-\Delta$ (%)
全局	1	32.7	0
当前音节 (CT)	1.5 k	32.0	2.1
+ 左上下文 (CL)	231 k	29.2	10.7
+ 右上下文 (CR)	42 k	30.9	5.5
+ 左右上下文 (CLR)	4.7 M	29.8	8.9
决策树 (问题集 1)	7.7 k	28.9	11.6
决策树 (问题集 2)	9.3 k	**27.6**	15.6

下文相关权重的数目 N_w, 对于四种权重分配策略权 N_w 分别为 1 497 个、231 k、42 k 以及 4.7 M, 可见考虑到的上下文类型增多, 参数数量将会急剧增大. 上下文参数集选定之后, 将所有训练权重从全局权重初始化, 然后利用式 (3) 对权重参数进行更新, 并利用更新后的权重进行测试输出. 表 4 给出了训练集和开发集上四种权重参数集上的权重训练期望误识率随迭代变化过程 ($1 - \frac{1}{N}\mathcal{F}_{\mathrm{MPE}}$, N 为训练数据中总的音素数目). 表 5 给出了测试集上误识率随权重训练的迭代变化过程. 从表 4 上半部分来看, 随着可训练权重数量的增多, 期望误识率总体上都呈单调下降趋势且逐步收敛至最优, 而且权重数目越多单调性越明显. 对于 CLR 参数集上进行权重训练得到的期望误识率下降最大 (在第 10 次迭代后期望误识率下降为 $0.377 - 0.0700 = 0.307$).

从表 5 中测试集的识别结果随迭代的变化过程来看, 在权重参数数量较小时, 如 CT 参数集 (参数数量 $N_w = 1.5\,\mathrm{k}$) 和 CR 参数集 ($N_w = 42\,\mathrm{k}$), 测试集的误识率在几次迭代降至最低之后基本保持不变. 而对于权重参数数量较多的 CL 参数集和 CLR 参数集, 测试集误识率先下降之后又逐步升高. 因此需要先确定最佳的迭代次数, 并将最佳迭代次数下的识别结果作为最终识别结果. 从表 4 上半部分 CL 和 CLR 参数集在训练集上的期望误识率来看, 在第 10 次迭代之后期望误识率降至最低, 但此时测试集误识率并非最低. 因此在权重数量增多时, 按照训练数据上的期望误识率来选择最佳迭代次数是无效的. 一种选取最佳迭代次数的方法是根据校验集的性能来确定最佳迭代次数. 表 4 的下半部分给出了开发集上的期望误识率随迭代的变化过程, 我们根据开发集上的期望误识率最低值确定测试集的最佳迭代次数, 并且将最佳迭代次数时测试集的识别结果列于

表 4　迭代过程中的期望误识率 (人工选择上下文)

Table 4　Expected error rate of weight training iterations (manually selected contexts)

数据集	上下文	Iter = 0	1	2	3	4	5	6	7	8	9	10
训练集	当前音节	0.377	0.365	0.362	0.360	0.359	0.359	0.359	0.358	0.358	0.358	0.358
	+ 左上下文	0.377	0.304	0.265	0.243	0.230	0.222	0.217	0.214	0.212	0.210	0.209
	+ 右上下文	0.377	0.344	0.329	0.322	0.318	0.316	0.315	0.314	0.314	0.314	0.314
	+ 左右上下文	0.377	0.249	0.179	0.139	0.114	0.0991	0.0889	0.0819	0.0772	0.0731	0.0700
开发集	当前音节	0.421	0.412	0.411	0.411	0.409	0.409	0.407	0.409	0.408	0.409	0.407
	+ 左上下文	0.421	0.393	0.382	0.375	0.372	0.374	0.380	0.382	0.389	0.391	0.394
	+ 右上下文	0.421	0.407	0.403	0.405	0.401	0.399	0.401	0.399	0.403	0.400	0.402
	+ 左右上下文	0.421	0.383	0.377	0.391	0.401	0.415	0.430	0.442	0.453	0.465	0.477

表 5　迭代过程中的误识率 (人工选择上下文) (%)

Table 5　Recognition error rate of weight training iterations (manually selected contexts) (%)

数据集	上下文	Iter = 0	1	2	3	4	5	6	7	8	9	10
测试集	当前音节	32.7	32.1	31.9	32.1	32.1	32.2	32.0	32.1	32.1	32.1	32.0
	+ 左上下文	32.7	30.1	29.8	28.6	29.2	29.1	29.7	29.6	29.8	30.3	30.8
	+ 右上下文	32.7	31.2	31.0	30.9	30.8	30.9	30.8	31.0	31.1	31.0	31.1
	+ 左右上下文	32.7	29.8	29.8	30.8	31.6	33.2	34.3	35.4	36.8	37.5	38.3

表 3. 从结果来看, 对于上述四种权重集的带调音节误识率分别为从全局权重的 32.7 % 降为 32.0 %、29.2 %、30.9 % 和 29.8 %.

从四种参数集的结果比较来看, CL 权重参数集误识率下降最为明显, 这说明一个音节的发音受到上一个音节声学上下文影响最大. 对于 CLR 参数集, 误识率虽然比 CR 参数集略低, 但仍然比最优的 CL 参数集高. 尽管其训练集上的期望误识率最低, 但测试集上的误识率并非最低. 这是由于可训练权重参数过多, 一些并不能带来区分性性能的上下文相关参数参与优化, 导致训练过拟合从而影响识别性能. 训练过拟合另一方面体现在: 在迭代过程中, 期望误识率随着迭代次数单调下降, 而测试集的误识率在降到最低之后又开始回升, 而且参数数量越多误识率回升情况越严重. 这是由于对提高性能没有帮助的上下文相关参数继续参与优化反而降低了识别性能.

接下来给出利用决策树进行上下文建模的结果. 第一组实验备选问题集采用第 3.2 节列出的 99 条的问题集, 我们称之为问题集 1. 决策树构造重复节点分裂过程直至式 (10) 和式 (11) 中近似增量 $\mathcal{G}^{p}_{\mathrm{MPE}}$ 小于某一预先设定的门限值 τ. 表 6 给出了不同门限值下的权重参数个数 (叶子节点数目) N_w、校验集上的期望误识率和测试集上的误识率. 较小的门限 τ 导致更多的叶节点. 叶子节点较少时, 参数过度聚集不能带来良好的区分能力; 叶子节点过多时, 无用的上下文又会影响系统性能. 因此 τ 的设定是获取更大区分能力和引入无用上下文之间的折中. 实验中, 我们根据开发集的期望误识率选择门限 $\tau = 2.0$.

表 6 不同门限 τ 的结果

Table 6 Results of different threshold τ

τ	N_w	开发集期望误识率	测试集误识率 (%)
1.0	12.5 k	0.385	29.9
1.5	9.4 k	0.377	29.2
2.0	7.7 k	0.373	28.9
2.5	6.7 k	0.380	29.5
3.0	5.9 k	0.388	30.1

表 7 和表 8 给出了采用决策树上下文建模参数集上的 10 次训练迭代结果, 包括训练集、校验集上的期望误识率和测试集的误识率. 最佳迭代次数仍然使用开发集上期望误识率最低时的迭代次数来确定, 并将此时测试集的识别结果列于表 3. 从决策树建模结果来看, 采用备选问题集 1 获得上下文参数集的误识率为 28.9 %, 优于人工选取上下文的结果. 通过观察问题集 1 以及从构造出的决策树叶子节点中聚类的声学上下文, 我们发现一些叶子节点仍然聚集了较多数量的上下文训练样本 (Lattice 的边), 粗略的问题集设计使得一些具有较好区分能力的上下文仍然与区分能力较弱的上下文聚集在一起, 因此可以在问题集设计时加入一些更为精细的问题集来考虑单一音素上下文:

QS_a_1 {a:1/*-*+*} ···
QS_b {*:*/*-*+b} ···

这批问题集称为问题集 2. 与问题集 1 的区别在于, 问题集 2 只表示了前驱带调韵母或者后继声母一种

表 7 迭代过程中的期望误识率 (决策树上下文建模)

Table 7 Expected error rate of weight training iterations (decision tree based context modeling)

数据集	问题集	Iter = 0	1	2	3	4	5	6	7	8	9	10
训练集	问题集 1	0.377	0.322	0.304	0.296	0.294	0.291	0.291	0.290	0.290	0.289	0.289
	问题集 2	0.377	0.310	0.292	0.285	0.282	0.280	0.279	0.278	0.277	0.276	0.276
开发集	问题集 1	0.421	0.380	0.375	0.380	0.377	<u>0.373</u>	0.375	0.374	0.376	0.373	0.375
	问题集 2	0.421	0.369	0.360	0.364	0.361	<u>0.359</u>	0.361	0.362	0.360	0.363	0.361

表 8 迭代过程中的误识率 (决策树上下文建模) (%)

Table 8 Recognition error rate of weight training iterations (decision tree based context modeling) (%)

数据集	上下文	Iter = 0	1	2	3	4	5	6	7	8	9	10
测试集	问题集 1	32.7	29.3	29.3	29.0	29.2	<u>28.9</u>	29.1	28.9	29.1	28.8	29.0
	问题集 2	32.7	28.4	28.3	27.6	27.7	<u>27.6</u>	27.5	27.7	27.5	27.6	27.6

音素类型, 而不是像问题集 1 那样把若干种音素类型聚合在一起. 问题集 2 中的条目共 543 条. 通过加入精细问题集进行决策树构造, 可训练权重数量(叶子节点数目) 从 7.7 k 增至 9.3 k, 误识率进一步从 28.9 % 降至 27.6 %. 在只使用 9.3 k 个权重组的情况下, 较人工选取上下文的性能最优的 CL 参数集 (权重参数数目为 231 k) 误识率低 1.6 %. 这说明在人工选取的声学上下文中绝大多数不具备对识别性能有贡献的区分能力, 这些参数的引入使得权重训练发生过拟合, 反而降低了识别性能. 而采用决策树的上下文建模方法可以将这些无用上下文聚集起来, 消除其带来的过拟合的影响. 从每次训练迭代的结果来看, 采用人工选取上下文在参数数量较大时训练多次后误识率又会逐步回升, 因此需要加入校验集来确定最佳迭代次数. 而采用决策树进行上下文建模时, 误识率收敛到最优之后基本没有出现随后续迭代又回升的现象, 这使得我们无需再通过校验集确定最佳迭代次数. 这也从另一个角度说明了利用决策树自动选取上下文参数集对过训练的鲁棒性, 也说明了基于格的二次解码过程中, 上下文建模对于区分性模型组合的重要性.

3.4 与基于特征组合方法识别结果的比较

语音识别中对于基频特征和 MLP 音素后验概率特征的另外一种建模方法就是将这些特征与谱特征流合并在特征空间进行组合, 并在组合特征上进行声学模型训练, 然后进行一次解码得到输出结果, 这种方法称为 TANDEM 方法[5]. 为了比较模型组合与特征组合的结果, 本文进行了基于组合特征的语音识别实验. 组合特征采用传统 39 维 MFCC 特征、MLP 音素分类器后验概率输出通过主分量分析降维得到的 25 维特征, 以及清音段插值平滑的基频序列及其一阶、二阶差分 (3 维), 在共计 67 维特征基础上进行 HMM 建模. 参数训练采用 MPE 区分性训练, 带调音节输出误识率为 29.7 %, 而基于决策树上下文建模的模型组合误识率为 27.6 %. 由此可见, 基于决策树自动选取上下文的模型组合方法要优于特征组合方法, 我们认为误识率的下降的原因是长范围的上下文相关的模型权重参数对模型得分进行了更加精细的调整.

4 结论

本文对二次解码过程中上下文相关的区分性模型组合方法进行了研究. 针对上下文相关模型组合带来的参数集过大的问题, 提出了使用决策树对声学上下文进行建模的方法. 并以最小音子错误准则为目标函数进行决策树构造. 我们还讨论了该方法在实际应用中重要的快速问题集选择方法, 并就备选问题集的设计进行了改进. 连续语音识别实验表明: 与人工选取上下文的方法相比, 决策树上下文建模基础上的区分性模型组合能够在大大减少参数数量的情况下降低系统误识率, 从而提高模型权重训练的鲁棒性. 实验结果还表明, 基于决策树上下文建模的区分性模型组合的结果要优于特征组合的结果. 该方法最突出的优点在于无需反复人工试凑选择上下文, 而自动获得最优的上下文相关参数集.

References

1 Beyerlein P. Discriminative model combination. In: Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Santa Barbara, USA: IEEE, 1997. 238−245

2 Huang H, Zhu J. Discriminative incorporation of explicitly trained tone models into lattice based rescoring for Mandarin speech recognition. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vagas, USA: IEEE, 2008. 1541−1544

3 Hoffmeister B, Liang R Y, Schlüter R, Ney H. Log-linear model combination with word-dependent scaling factors. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech). Brighton, UK: Interspeech, 2009. 248−541

4 Liu X Y, Gales M J F, Woodland P C. Use of contexts in language model interpolation and adaptation. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech). Brighton, UK: Interspeech, 2009. 360−363

5 Liu X Y, Gales M J F, Hieronymus J L, Woodland P C. Language model combination and adaptation using weighted finite state transducers. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, USA: IEEE, 2010. 5390−5393

6 Ellis D P W, Singh R, Sivadas S. Tandem acoustic modeling in large-vocabulary recognition. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Salt Lake City, USA: IEEE, 2001. 1201−1204

7 Povey D, Woodland P C. Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Orlando, USA: IEEE, 2002. 105−108

8 Povey D. Discriminative Training for Large Vocabulary Speech Recognition [Ph. D. dissertation], Cambridge University, UK, 2004

9 Wu Ya-Hui, Liu Gang, Guo Jun. Research on model combination based on model confusion. *Acta Automatica Sinica*, 2009, **35**(5): 551−555
(吴娅辉, 刘刚, 郭军. 基于模型混淆度的模型组合算法研究. 自动化学报, 2009, **35**(5): 551−555)

10 Young S J, Odell J J, Woodland P C. Tree-based state tying for high accuracy acoustic modelling. In: Proceedings of the

1994 Workshop on Human Language Technology. Stroudsburg, USA: ACL, 1994. 307−312

11 Gao S, Lee C H. A discriminative decision tree learning approach to acoustic modeling. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech). Geneva, Switzerland: ISCA, 2003. 1833−1836

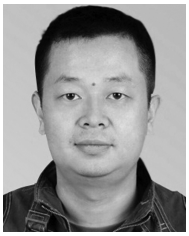
12 Wiesler S, Heigold G, Nuβbaum-Thom M, Schlüter R, Ney H. A discriminative splitting criterion for phonetic decision trees. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech). Makuhari, Japan: ISCA, 2010. 54−57

13 Chang E, Shi Y, Zhou J L, Huang C. Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech). Aalborg, Denmark: ISCA, 2001. 2779−2782

14 Qian Y, Lee T, Li Y J. Overlapped ditone modeling for tone recognition in continuous Cantonese speech. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech). Geneva, Switzerland: ISCA, 2003. 1845−1848

15 Huang Hao, Abudureyimu Halidan. Rapid parameter updating algorithm for large margin Gaussian mixture model. *Computer Engineering*, 2010, **36**(3): 197−199
(黄浩, 哈力旦 · 阿布都热依木. 大间隔高斯混合模型的快速参数更新算法. 计算机工程, 2010, **36**(3): 197−199)

16 The ICSI Quicknet Tools [Online], available: http://www.icsi.berkeley.edu/Speech/qn.html, March 15, 2012



黄 浩 新疆大学信息科学与工程学院副教授. 2008 年在上海交通大学电子工程系获博士学位. 主要研究方向语音识别, 多媒体人机交互技术. 本文通信作者. E-mail: huanghao@xju.edu.cn
(**HUANG Hao** Associate professor in the Department of Information Science and Engineering, Xinjiang University. He received his Ph. D. degree from Shanghai Jiao Tong University in 2008. His research interest covers speech recognition and multi-media human-machine interaction. Corresponding author of this paper.)

李兵虎 新疆大学信息科学与工程学院硕士研究生. 主要研究方向为语音识别, 语音信号处理.
E-mail: binghulee@gmail.com
(**LI Bing-Hu** Master student in the Department of Information Science and Engineering, Xinjiang University. His research interest covers speech recognition and speech signal processing.)

吾守尔 · 斯拉木 新疆大学信息科学与工程学院教授. 主要研究方向为语音识别, 语音合成, 多语种信息处理.
E-mail: wushour@xju.edu.cn
(**SILAMU Wushour** Professor in the Department of Information Science and Engineering, Xinjiang University. His research interest covers speech recognition, speech synthesis, and multi-lingual information processing.)