第 36 卷 第 3 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.3 Computer Engineering February 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)03—0197—03 文献标识码：A 中图分类号：TN912.34

大间隔高斯混合模型的快速参数更新算法

黄 浩[1]，哈力旦[2]
(1. 新疆大学信息科学与工程学院，乌鲁木齐 830046；2. 新疆大学电气工程学院，乌鲁木齐 830008)

摘 要：针对大间隔高斯混合模型基于 LBFGS 参数更新算法收敛速度慢的不足，提出一种快速参数更新算法。采用构造弱意义辅助函数的方法，得到扩展 Baum-Welch 算法形式的快速参数更新公式。利用大词汇汉语语音库上的声调分类任务来验证训练速度与分类性能。实验结果表明快速参数更新算法只需数次迭代就能收敛至最优结果，较 LBFGS 优化方法在识别性能相当的情况下具有更快的训练速度。

关键词：大间隔；高斯混合模型；声调识别

Rapid Parameter Updating Algorithm for Large Margin Gaussian Mixture Model

HUANG Hao[1], HAlidan[2]
(1. College of Information Science and Engineering, Xinjiang University, Urumuqi 830046;
2. College of Electrical Engineering, Xinjiang University, Urumuqi 830008)

【Abstract】To speed up training of large margin Gaussian mixture model based on LBFGS optimization routing, a rapid model parameter method is proposed. The method formulates extended Baum-Welch algorithm like updating equations by constructing weak-sense auxiliary function. The proposed algorithm is experimented on Mandarin tone recognition tasks. It is shown training can be accomplished within only several iterations, much faster than that of LBFGS optimization routine.

【Key words】 large margin; Gaussian mixture model; tone recognition

1 概述

利用大间隔原理(large margin principle)进行分类器设计是目前模式识别中的重要研究方向之一。大间隔原理指出：分类器对未知数据的分类结果将取决于训练集上的经验风险以及与推广能力紧密相关的间隔。支持向量机(Support Vector Machine, SVM)[1]就是一种基于大间隔原理的机器学习方法，由于其出色的学习性能而成为机器学习界研究的热点，在各种分类任务中获得广泛的应用。大间隔高斯混合模型是文献[2]依照 SVM 原理提出的一种区分性高斯混合模型训练准则，相较 SVM 的一个优点是无需借助核函数来完成线性不可分的样本问题；另一个优点在于其不需要借助其他手段完成多类分类。文献[2]将大间隔高斯混合模型用于音子分类，在标准 TIMIT 语音库上获得了单一分类器条件下音子分类任务的最佳性能。

本文根据现有的大间隔高斯参数的更新方法，给出一种快速大间隔高斯参数的更新方法并利用汉语声调分类任务验证该方法的性能。

2 大间隔高斯混合模型

2.1 目标函数

为简单起见，先讨论每个类使用单个高斯的情况。对于具有 C 个类别的分类器，每个类使用均值 $\boldsymbol{\mu}_i$ 和方差 $\boldsymbol{\Sigma}_i$ 的高斯参数 $\boldsymbol{\theta}_i=(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),i=1,2,\cdots,C$ 来表示。在大间隔高斯模型定义中，将判别函数表示定义为观察样本 $\boldsymbol{x}$ 距离椭球 $(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 中心的马氏距离加一个偏移量常数 ρ_i：

$$g(\boldsymbol{x},\boldsymbol{\theta}_i)=(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Psi}_i(\boldsymbol{x}-\boldsymbol{\mu}_i)+\rho_i \tag{1}$$

其中，$\boldsymbol{\Psi}_i$ 是类 i 的精度矩阵(协方差矩阵 $\boldsymbol{\Sigma}_i$ 的逆)，若定义如下半正定矩阵：

$$\boldsymbol{\Phi}_i=\begin{bmatrix}\boldsymbol{\Psi}_i & -\boldsymbol{\Psi}_i\boldsymbol{\mu}\\ -\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Psi}_i & \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Psi}_i\boldsymbol{\mu}+\rho_i\end{bmatrix} \tag{2}$$

则判别函数可简化为

$$g(z,\boldsymbol{\Phi}_i)=z^{\mathrm{T}}\boldsymbol{\Phi}_i z$$

其中，$z=[\boldsymbol{x}\quad 1]^{\mathrm{T}}$，大间隔高斯混合模型的目标函数定义为[2]

$$\begin{aligned}&\min\sum_{nc}\xi_{nc}+\kappa\sum_i\mathrm{trace}(\boldsymbol{\Psi}_i)\\&\text{s.t. }\ 1+z_n^{\mathrm{T}}(\boldsymbol{\Phi}_{y_n}-\boldsymbol{\Phi}_i)z_n\leqslant\xi_{nc},\ \xi_{nc}\geqslant 0,\forall i\neq y_n,n=1,2,\cdots,N\\&\boldsymbol{\Phi}_i>0,i=1,2,\cdots,C\end{aligned} \tag{3}$$

其中，约束条件 $1+z_n^{\mathrm{T}}(\boldsymbol{\Phi}_{y_n}-\boldsymbol{\Phi}_i)z_n\leqslant\xi_{nc}$ 保证样本 z_n 距离竞争类距离大于 1 个单位距离；非负的松弛因子 ξ_{nc} 表示训练数据中与约束条件冲突的程度；最小化精度矩阵 $\boldsymbol{\Psi}_i$ 的迹用来保证满足约束条件下的参数唯一性；$\kappa>0$ 是交叉验证得到的控制系数，更多关于大间隔高斯混合模型的定义可参见文献[2]。

2.2 参数更新方法

文献[2]指出，式(3)可以转化为半定规划的形式利用内点法求解，但是现有的半定规划优化函数库在处理训练样本数较多的问题时，时间及空间的复杂度将远超过要求。基于上述考虑，文献[2]提出子梯度法优化方法，然而这种方法的收

基金项目：国家自然科学基金资助项目“基于高速、压缩域的数字动态信息检索”(60865001)；新疆高校科研计划基金资助项目“基于音位学属性特征的多语种自动辨识的研究”(XJEDU2008S15)

作者简介：黄 浩(1976－)，男，讲师、博士，主研方向：自动语音识别技术，模式识别，机器学习；哈力旦，教授、硕士

收稿日期：2009-09-17 **E-mail：**hwanghao@gmail.com

敛速度非常慢，因此文献[2]首先利用 LBFGS 无约束优化例程来加快训练速度。本文将简要叙述该优化方法，并继而给出一种在时间上更加有效的参数优化算法。当采用 LBFGS 方法进行参数优化时，需要将式(3)的目标函数转化为单一目标函数的优化问题，文献[2]利用函数 $hinge(u)=\max(0,u)$ 将式(3)转化为

$$R^{lme}=\sum_{n,i\neq y_n} hinge(1+z_n^{\mathrm{T}}(\boldsymbol{\Phi}_{y_n}-\boldsymbol{\Phi}_c)\,z_n)+\kappa\sum_c trace(\boldsymbol{\Psi}_c) \tag{4}$$

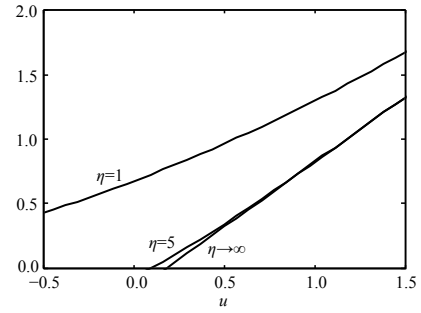
由于 *hinge* 函数不可微，无法直接计算导数进行梯度更新，因此可利用一个平滑的函数 $h(u)$来代替：

$$h(u)=\frac{1}{\eta}\ln\left(1+\exp^{\eta(u)}\right) \tag{5}$$

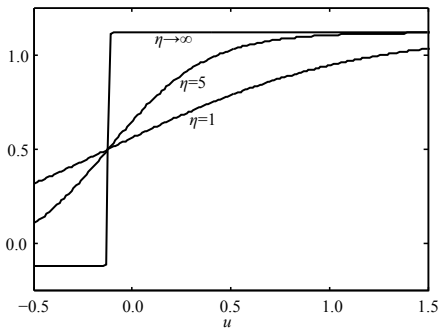
其中，η 为常数，图 1 显示了 $h(u)$函数及其导数，常数 η 越大，$h(u)$将越接近于真实的 $hinge(\eta\to\infty)$函数，在使用 $h(u)$函数时，目标函数对高斯参数的导数可计算为

$$\frac{\partial R^{\mathrm{lme}}}{\partial\boldsymbol{\Phi}_c}=\sum_n\frac{\mathrm{e}^{1+z_n^{\mathrm{T}}(\boldsymbol{\Phi}_{y_n}-\boldsymbol{\Phi}_c)z_n}}{1+\mathrm{e}^{\eta(1+z_n^{\mathrm{T}}(\boldsymbol{\Phi}_{y_n}-\boldsymbol{\Phi}_c)z_n)}}\mathrm{sgn}(c=y_n)z_nz_n^{\mathrm{T}}+\kappa\boldsymbol{A} \tag{6}$$

其中，$\boldsymbol{A}$ 是 $(n+1)\times(n+1)$ 的方阵，其元素 $a_{ii}=1,i=1,2,\cdots,n$ 的值为 1，其余为零。然后将目标函数值以及导数代入 LBFGS 优化例程进行参数的迭代更新。



(a)$h(u)$函数



(b)$h(u)$函数的导数

图 1　平滑 *hinge* 函数 $h(u)$及其导数

3　快速参数更新方法

尽管上述方法能够显著加快训练的速度，但是从实验来看仍需要 80 次~100 次迭代才能收敛至最优，为了进一步加快训练速度，下面将推导利用弱意义辅助函数法得到的基于 EBW 形式的参数更新公式。

考虑多高斯混合的情况，设对每个类别 i 由 J_i 个高斯混合而成，参数表示为 $\boldsymbol{\theta}_{ij},i=1,2,\cdots,C,j=1,2,\cdots,J_i$，令判别函数为高斯产生样本的概率值 $\ln p(\boldsymbol{x}\mid\boldsymbol{q}_{ij})$，类似于式(4)，将目标函数重新定义为

$$R_{\mathrm{ebw}}^{\mathrm{lme}}=-\sum_{n=1}^{N}hinge\left(1+\ln\sum_{i\neq y_n}^{C}\sum_{j=1}^{M_i}p(\boldsymbol{x}_n\mid\boldsymbol{q}_{ij})-\ln p(\boldsymbol{x}_n\mid\boldsymbol{q}_{y_nm_n})\right)+\kappa\sum_i\ln p(\boldsymbol{\theta}_i) \tag{7}$$

其中，高斯先验 $\kappa\sum_c\ln p(\boldsymbol{\theta}_i)$ 用来代替式(4)中最小化精度矩阵的迹，以保证满足约束间隔条件下参数的唯一性。

3.1　弱意义辅助函数

弱意义辅助函数是指辅助函数对参数的导数在当前参数下的取值与目标函数对的参数导数在当前参数值下的取值相等，根据文献[3]，可以构造如下弱意义辅助函数：

$$H(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})=\sum_{n=1}^{N}\frac{\partial R_{\mathrm{ebw}}^{\mathrm{lme}}(\hat{\boldsymbol{\theta}})}{\partial\ln p(\boldsymbol{x}_n\mid\hat{\boldsymbol{\theta}}_{ij})}|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}}\ln p(\boldsymbol{x}_n\mid\hat{\boldsymbol{\theta}}_{ij}) \tag{8}$$

其中，$\boldsymbol{\theta}$ 是原来的模型参数集；$\hat{\boldsymbol{\theta}}$ 是新的模型参数集。

显然式(8)满足弱意义辅助函数的定义。在参数 $\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}$ 时，优化辅助函数按照目标函数优化的方向进行参数的更新。由此，目标函数的优化间接地转化为对弱意义辅助函数的优化。

3.2　参数更新公式

式(8)中求和项是目标函数对模型产生样本 $\boldsymbol{x}_n$ 的对数概率的导数，将其定义为中间变量 $\gamma_{ij}^{\mathrm{lme}}(n)$，则有：

$$\gamma_{ij}^{\mathrm{lme}}(n)=\gamma_{ij}\frac{\mathrm{e}^{\eta[1-g(x_n)]}}{1+\mathrm{e}^{\eta[1-g(x_n)]}}\mathrm{sgn}(i=y_n,j=m_n) \tag{9}$$

其中，$\gamma_{ij}(n)$ 是样本 $\boldsymbol{x}_n$ 属于混合高斯 $\boldsymbol{\theta}_{ij}$ 的后验概率，对于正例样本计算如下 3 个累积量：

$$\begin{aligned}\beta_{ij}^{\mathrm{pos}}&=\sum_{n=1}^{N}\max\left(0,\gamma_{ij}^{\mathrm{lme}}(n)\right)\\ \boldsymbol{\chi}_{ij}^{\mathrm{pos}}&=\sum_{n=1}^{N}\max\left(0,\gamma_{ij}^{\mathrm{lme}}(n)\right)\boldsymbol{x}_n\\ \boldsymbol{Y}_{ij}^{\mathrm{pos}}&=\sum_{n=1}^{N}\max\left(0,\gamma_{ij}^{\mathrm{lme}}(n)\right)\boldsymbol{x}_n\boldsymbol{x}_n^{\mathrm{T}}\end{aligned} \tag{10}$$

对于反例部分的 3 个累积量 $\beta_{ij}^{\mathrm{neg}}$，$\boldsymbol{\chi}_{ij}^{\mathrm{neg}}$，$\boldsymbol{Y}_{ij}^{\mathrm{neg}}$，则可将式(10)中 $\gamma_{ij}^{\mathrm{lme}}(n)$ 替换为 $-\gamma_{ij}^{\mathrm{lme}}(n)$ 计算得到。根据上述累积量，可得到扩展 Baum-Welch 形式[3]的均值和协方差矩阵的参数更新公式为

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{ij}&=\frac{\boldsymbol{\chi}_{ij}^{\mathrm{pos}}-\boldsymbol{\chi}_{ij}^{\mathrm{neg}}+D_{ij}\boldsymbol{\mu}_{ij}}{\beta_{ij}^{\mathrm{pos}}-\beta_{ij}^{\mathrm{neg}}+D_{ij}}\\ \hat{\boldsymbol{\Sigma}}_{ij}&=\frac{\boldsymbol{Y}_{ij}^{\mathrm{pos}}-\boldsymbol{Y}_{ij}^{\mathrm{neg}}+D_{ij}(\boldsymbol{\Sigma}_{ij}+\boldsymbol{\mu}_{ij}\boldsymbol{\mu}_{ij}^{\mathrm{T}})}{\beta_{ij}^{\mathrm{pos}}-\beta_{ij}^{\mathrm{neg}}+D_{ij}}-\boldsymbol{\mu}_{ij}\boldsymbol{\mu}_{ij}^{\mathrm{T}}\end{aligned} \tag{11}$$

其中，D_{ij} 是平滑控制常数，其选取方法在下文讨论。

3.3　平滑控制常数 D_{ij} 的选取

当采用满协方差矩阵时，D_{ij} 的选择必须要保证更新后的协方差矩阵 $\hat{\boldsymbol{\Sigma}}_{ij}$ 为正定，解决的办法是将均值公式带入协方差矩阵更新公式中，可得：

$$\hat{\boldsymbol{\Sigma}}_{ij}=\frac{\boldsymbol{B}_2D_{ij}^2+\boldsymbol{B}_1D_{ij}+\boldsymbol{B}_0}{\beta_{ij}^{\mathrm{c}}+D_{ij}} \tag{12}$$

其中，

$$\begin{aligned}\boldsymbol{B}_2&=\boldsymbol{\Sigma}_{ij}\\ \boldsymbol{B}_1&=\boldsymbol{Y}_{ij}^{\mathrm{com}}+\beta_{ij}^{\mathrm{com}}\left(\boldsymbol{\Sigma}_{ij}+\boldsymbol{\mu}_{ij}\boldsymbol{\mu}_{ij}^{\mathrm{T}}\right)-\left(\boldsymbol{\mu}_{ij}\boldsymbol{x}_{ij}^{\mathrm{comT}}+\boldsymbol{x}_{ij}^{\mathrm{com}}\boldsymbol{\mu}_{ij}^{\mathrm{T}}\right)\\ \boldsymbol{B}_0&=\beta_{ij}^{\mathrm{com}}\boldsymbol{Y}_{ij}^{\mathrm{com}}-\boldsymbol{x}_{ij}^{\mathrm{com}}\boldsymbol{x}_{ij}^{\mathrm{comT}}\end{aligned} \tag{13}$$

由式(12)，为保证协方差矩阵的正定，需要 2 条件：(1)式(12)中分子部分为正定；对于该条件，文献[4]指出可利用求解二次特征值问题来获得满足不等式的最小 D_{ij} 值，设该值为 D_{ij}^{qep}；(2)分母部分为正值，设满足该条件的 D_{ij} 的值为 $E\beta_{ij}^{\mathrm{neg}}$，其中，$E$ 是人工选取的经验值。综合 2 个条件，可

得到满足协方差矩阵正定的选取为 $D_{ij}=\max\left(2D_{ij}^{\text{qep}},E\beta_{ij}^{\text{neg}}\right)$。

4 实验与结果

汉语声调在汉语中承担着重要的构字辨义的作用，因此，声调识别是汉语语音识别系统和计算机辅助汉语语音学习系统的重要组成部分。本文将采用声调识别任务来验证所述方法的训练速度和分类性能。

4.1 系统配置

实验在微软亚洲研究院汉语连续语音库上进行。训练语料包含 454 291 个带调音节，测试语料包含另外 9 570 个带调音节。对于 1 声~5 声类别的高斯混合数目分别为 16 个、16 个、13 个、25 个和 4 个，保证了训练数据的充分性。

4.2 最大似然训练识别结果

首先考察基于采用不同特征条件下的声调识别结果，训练准则全部采用最大似然估计(Maximum Likelihood Estimation, MLE)，使用的特征分别如下：

基音频率($F0$)特征：采用多项式回归将基频序列归一化至 11 维的向量；对数能量特征 ln$E$：长度归一化的对数能量；段动态特征△$F0$：根据文献[5]，将当前段 $F0$ 一阶导数平均。重叠双音调特征(*ditone*)：加入前驱音节的归一化 $F0$ 得到的特征。从表 1 的结果来看，在最大似然训练准则下最佳声调误识率为 31.1%。

表 1　不同特征下的声调识别结果(使用最大似然估计)

训练准则	优化算法	特征	维数	误识率(%)	误识率下降/(%)
MLE	EM	$F0$	11	38.0	-
MLE	EM	ln$E$	22	34.1	10.3
MLE	EM	△$F0$	23	32.8	13.7
MLE	EM	*+ditone*	34	31.1	16.3

4.3 区分性模型训练结果

表 2 给出了区分性模型参数训练的结果，特征全部采用 34 维特征。区分性训练起始参数取自最大似然训练结果。先给出利用大间隔训练参数估计(Large Margin Estimation, LME)得到的声调识别结果，选用重叠双音调特征并利用式(6)进行 LBFGS 参数优化，误识率为 28.3%。可以看出大间隔训练的结果较采用同样特征的最大似然估计得到识别结果(31.1%)性能有明显改善。

表 2　大间隔参数训练的声调识别结果

训练准则	优化算法	特征	维数	误识率/(%)	误识率下降/(%)
LME	LBFGS	*ditone*	34	28.3	25.5
MBR	EBW	*ditone*	34	29.9	21.3
LME	EBW	*ditone*	34	28.0	26.3

实验中还将大间隔参数训练与传统基于贝叶斯风险最小化(Minimum Bayesian Risk, MBR)的区分性准则进行比较。根据文献[3]，一种与 MBR 等价的目标函数可写为

$$R^{\text{MBR}}=\sum_{n}^{N}\sum_{i=1,j=1}^{5,J_i}P(\boldsymbol{\theta}_{ij}\mid\boldsymbol{x}_n)\text{Acc}(\boldsymbol{\theta}_{ij},\boldsymbol{\theta}_r) \tag{14}$$

其中，$P(\boldsymbol{\theta}_{ij}\mid\boldsymbol{x}_n)$ 是给定声调观察值 $\boldsymbol{x}_n$ 下属于高斯 $\boldsymbol{\theta}_{ij}$ 的后验概率；$\text{Acc}(\boldsymbol{\theta}_{ij},\boldsymbol{\theta}_r)$ 是正确率测度，对于正确的类别该值为 1，反之为 0。最大化该目标函数表示提高对训练数据的期望正确度，即降低训练数据中的贝叶斯风险。对于该目标函数下的累积量计算和参数更新与式(10)、式(11)具有相同的形式。但须修改式(9)、式(10)累积量计算的中间变量 γ_{ij}^{lme} 为

$$\gamma_{ij}^{\text{MBR}}(n)=\gamma_{ij}\left(\text{Acc}(\boldsymbol{\theta}_{ij})-c_{avg}\right) \tag{15}$$

其中，γ_{ij} 表示模型的后验概率；c_{avg} 是所有模型的平均正确度。2 种目标函数更新公式形式相似，差别在于累积量中的 2 个权重 $\gamma_{ij}^{\text{MBR}}(n)$ 和 $\gamma_{ij}^{\text{LME}}(n)$：在 MBR 准则下，对于训练样本 $\boldsymbol{x}_n$，对于正确类 $\gamma_{ij}^{\text{MBR}}(n)$ 为正值，对于混淆类则为负值，更新公式对正确类和混淆类参数均会起作用；而在大间隔目标函数下，由于 *hinge* 函数的存在，更新公式只在样本误分时才对参数进行更新，而且这种更新还须满足最小间隔的要求。从识别结果看，大间隔模型准则误识率不仅较最大似然准则有显著下降，较 MBR 准则也有明显下降，这是因为大间隔准则对测试数据有更好的推广能力。

4.4 训练速度的比较

从识别率上采用快速更新算法得到声调误识率为 28.0%，较 LBFGS 优化算法(28.3%)获得相当或者更高的识别结果，但弱意义辅助函数法的优点还在于更快的训练速度。表 3 给出误识率随迭代过程的变化情况，迭代 0 表示两者均从最大似然训练参数开始(误识率为 31.2%)快速参数更新算法需要 5 次左右迭代就能够达到最优结果，而 LBFGS 方法则需要 80 次左右。从存储空间来看，快速算法仅需暂存与高斯参数(均值，方差)相等数量的累积量，这与 LBFGS 方法需要的存储空间相当。从本文任务考察，一次迭代时间为 10 min 左右，在识别结果相当甚至更高的情况下，总的训练时间从 13.3 h 降至 0.8 h，这说明本文提出的快速算法对减少训练的时间复杂度是非常有效的。

表 3　2 种参数更新方法的迭代过程

LBFGS		EBW	
迭代	误识率/(%)	迭代	误识率/(%)
0	31.2	0	31.2
20	28.9	1	30.1
40	28.7	2	29.2
60	28.4	3	28.5
80	28.3	4	28.1
100	28.3	5	28.0

5 结束语

本文提出一种通过构造弱意义辅助函数，推导出扩展 Baum-Welch 形式的大间隔高斯参数优化方法。汉语声调分类实验表明与基于 LBFGS 的优化方法相比，该方法具有收敛速度快、无需调节过多参数的优点，更适合于大规模训练集下的参数估计及对训练速度有快速要求的场合。识别结果表明大间隔目标函数下的模型参数训练较最大似然准则、基于最小贝叶斯风险准则训练的高斯混合声调模型误识率都有明显下降。需要指出，该方法同样适合于其他分类任务条件下的快速参数更新。

参考文献

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York, USA: Springer-Verlag, 1995.

[2] Sha Fei, Saul L K. Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition[C]//Proc. of ICASSP'06. Toulouse, France: IEEE Press, 2006: 265-268.

[3] Povey D. Discriminative Training for Large Vocabulary Speech Recognition[D]. Cambridge, UK: Cambridge University, 2004.

[4] Goel V, Axelrod S, Gopinath R, et al. Discriminative Estimation of Subspace Precision and Mean(SPAM) Models[C]//Proc. of ISCA'03. Geneva, Switzerland: [s. n.], 2003: 2617-2620.

[5] Qian Yao, Lee T, Li Y J. Overlapped Ditone Modeling for Tone Recognition in Continuous Cantonese Speech[C]//Proc. of ISCA'03. Geneva, Switzerland: [s. n.], 2003: 1845-1848.

编辑　任吉慧