一种高效的语音关键词检索系统

罗骏, 欧智坚
(清华大学 电子工程系，北京 100084)

摘 要：针对音频检索任务中的关键词检索提出一种新的基于拼音图的两阶段检索系统，可以高效地从大量语音数据中检索出感兴趣的文本信息，从而达到为国家安全服务的目的。该系统分为预处理阶段和检索阶段。预处理阶段将语音数据识别成具有高覆盖率的拼音图，在这一过程中通过若干次的无监督最大似然线性回归自适应算法渐次提高拼音图的质量。检索阶段响应用户的频繁查询，只需在拼音图中查找出与关键词拼音匹配的拼音串，并采用基于 N 元拼音文法的前后向算法计算置信度以实现对检索结果的筛选。实验表明：系统具有较高的召回率和正确率，且检索阶段仅需 0.01 倍实时，可以满足快速检索的需要。

关键词：信息检索；关键词检索；拼音图；置信度
中图分类号：TP391 **文献标识码**：B **文章编号**：1000-436X(2006)02-0113-06

Efficient keyword spotting system for information retrieval

LUO Jun, OU Zhi-jian
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: A new two-stage keyword spotting system was proposed based on syllable graph for audio information retrieval task, which could efficiently spot the interesting words in mass speech data, thus serve for the national security. It comprised two stages – preprocessing and searching. In the preprocessing stage, the audio data was recognized into syllable graph which included high accuracy syllable candidates, and unsupervised MLLR (maximum likelihood linear regression) adaptation was carried out iteratively to further improve the accuracy of the syllable graph. In the searching stage, to answer the frequent queries from users, searching for matched keywords was only scanned in the graph for likely syllable strings. A forward-backward algorithm based on syllable *N*-grammar was used to calculate confidence measures for further filtering of the searching result. Experimental results show the system achieved good performances in both recall rate and accuracy rate, and in the searching stage only 0.01 times of real time is needed, which can meet the demand for fast retrieval.

Key words: information retrieval; keyword spotting; syllable graph; confidence measure

1 关键词检索系统设计方案

近年来，随着多媒体信息的日益增长，多媒体（视频、音频）数据库越来越成为广泛关注的热点。基于文本检索的常规信息检索技术无法满足对大量多媒体数据的检索需要。本文主要针对音频检索[1]中的关键词检索（keyword spotting）任务，提出一种基于拼音图的两阶段检索系统以实现快速、有效的信息获取。

针对音频的关键词检索指用户按照一定方式

收稿日期：2005-11-20；修回日期：2005-12-20
基金项目：国家自然科学基金资助项目（60402029）
Foundation Items: The National Natural Science Foundation of China (60402029)

（语音或者文本）输入关键词并从音频数据中返回关键词所在的特定语音段的检索过程。关键词检索系统通常可以分为如下两类。

一类是单阶段系统[2]，识别即在由关键词模型和非关键词模型（或者称为废料模型）并联的网络上进行。当关键词表改变时，系统必须再次进行识别，因此检索速度慢，不适合在用户需要反复修改查询条件的场合下使用。

另一类是两阶段系统[3,4]。预处理阶段通过一般意义的连续语音识别将语音数据转化为文本，只要运行一次。以后为响应用户的检索只需在文本中查找匹配的关键词。

本文采用两阶段的系统设计方案，该系统与一般的两阶段系统有所不同，主要区别在于：

(1) 预处理阶段只进行声学层的识别，即采用零语言模型，输出为拼音结果（拼音图[5]），这样做的好处在于拼音图往往比文本结果有高得多的覆盖率[6]，从而保证系统有较高的召回率。拼音图由利用拼音文法的 Token Passing[7]算法生成。由于预处理阶段在后台执行，且只需要进行一次，对处理速度没有特别的要求，因此在识别过程中可以反复利用无监督自适应调整码本以逐次提高拼音图的质量。

(2) 检索阶段在拼音图上查找拼音匹配结果，利用拼音文法模型而不是语言模型计算拼音串的置信概率完成关键词的筛选。

两阶段系统的框图如图 1 所示。

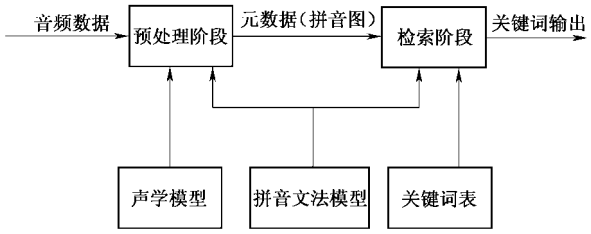

图 1 两阶段系统的整体框图

1.1 预处理（识别）阶段

预处理阶段即通常的声学识别过程。考虑到训练得到的码本与测试数据之间可能存在失配，在此过程需要利用识别结果指导进行无监督的最大似然线性回归（MLLR, maximum likelihood linear regression）自适应[8]，然后再识别。这样的过程可以重复若干次，直至识别性能趋近最优。一般而言，自适应的循环重复两到三次能对识别性能带来较大的改善，此后则性能的变化不再显著。

完整的预处理过程可以用图 2 表示。

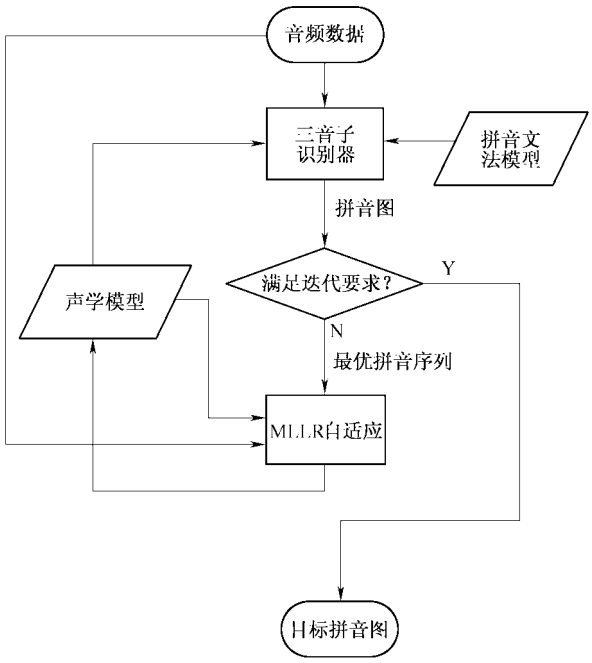

图 2 预处理阶段流程图

1.1.1 声学识别结果的拼音图结构表示

识别器采用基于语音学分类的三音子（triphone）识别单元和三元拼音文法，声学模型为基于段长分布的隐含马尔可夫模型（DDBHMM[9], duration distribution based hidden Markov model）。根据 Token Passing 识别算法在音节边界保留多个 Token 得到覆盖占优路径的拼音图。一个典型的拼音图结构如图 3 所示。

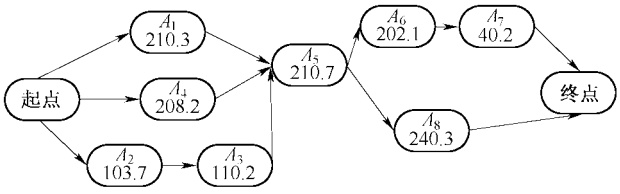

图 3 拼音图结构

拼音图是一个完整的有向图。图上从起点到终点的完整路径构成一个可能的识别结果候选。每个除起点和终点外的图上节点代表一个音节，保存音节标注、前后连接关系以及该音节的声学匹配得分（亦即部分似然得分）。为了便于区分，每个节点的序号 i 是惟一的，其音节标注用 A_i

表示。

1.1.2 无监督 MLLR 自适应

由于环境噪声、说话人性别、口音等各种因素的影响，语音数据呈现很大的可变性，在一定条件下训练的码本并不能普适地应用于所有场合，无监督自适应的目的在于根据测试数据调整码本以减少失配带来的影响。

MLLR 是一种行之有效的自适应方法[8]，其基本思想是：针对特定的说话人，其码本可以用 SI 码本经过线性变换后的说话人自适应（SA，speaker adaptive）码本来表示。即 SA 码本中的均值矢量 $\hat{\boldsymbol{u}}_i$ 可以表示为

$$\hat{\boldsymbol{u}}_i = \boldsymbol{U}\boldsymbol{u}_i + \boldsymbol{b} \tag{1}$$

其中，$\boldsymbol{u}_i = [\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \cdots \boldsymbol{u}_{iN}]^{\mathrm{T}}$ 表示 N 维的 SI 码本均值矢量，$\boldsymbol{U}$ 为 $N \times N$ 维的线性变换矩阵，$\boldsymbol{b}$ 表示偏移量。该式也可以方便地表示为

$$\hat{\boldsymbol{u}}_i = \hat{\boldsymbol{W}}\boldsymbol{\xi}_i \tag{2}$$

其中，$\hat{\boldsymbol{W}} = [\boldsymbol{b}, \boldsymbol{U}]$，$\boldsymbol{\xi}_i = [1, \boldsymbol{u}_i^{\mathrm{T}}]^{\mathrm{T}}$。对于不同的均值矢量可以有不同的变换矩阵 $\hat{\boldsymbol{W}}$，为了叙述简便起见，这里不加区分。MLLR 码本自适应的目的即估算变换矩阵 $\hat{\boldsymbol{W}}$ 从而更新得到 SA 码本。

由于在执行检索任务之前无法获得目标数据的相关信息，本系统采用批处理方式的无监督自适应，即：利用上一轮循环的最优识别结果作为标注，对整批数据统一完成统计后再进行码本的更新。

1.2 检索阶段

检索阶段需要完成两项任务：匹配关键词的查找和关键词的置信度计算。

第一步在拼音图的所有可能路径中把匹配的所有关键词都找出来，第二步通过置信度门限删除不可信的结果，以防止虚警过多。在短关键词的检索过程中置信度的计算尤其显得重要。

1.2.1 动态匹配的关键词查找过程

假设关键词可以用音节序列 $K = k_1 k_2 \cdots k_P$ 表示，P 为关键词长度，令 $\Lambda(i, j)$ 表示在节点 i 匹配到第 j 个音节的匹配记录，即

$$\Lambda(i, j) = \{(m_1, m_2, \cdots m_j) \mid A_{m_1} = k_1, A_{m_2} = k_2, \cdots A_{m_j} = k_j, m_j = i\} \tag{3}$$

令 $\Omega(i)$ 表示第 i 个节点相连的父节点编号集合，则

$$\Lambda(i, j) = \begin{cases} \bigcup\limits_{s \in \Omega(i)} \{(T, i) \mid T \in \Lambda(s, j-1)\}, A_i = k_j \\ \phi, A_i \neq k_j \end{cases} \tag{4}$$

当 $j = P$ 时即完成关键词的匹配，$\Lambda(i, P)$ 表示在节点 i 完成匹配的记录。

1.2.2 置信度计算

置信度得分以拼音串后验概率为依据[10]。以 $\boldsymbol{x}_1^{\mathrm{T}} = \boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_T$ 表示完整的观测特征序列，在拼音图中查找到匹配的节点编号序列为 $m_1^P = m_1 m_2 \cdots m_p$（$A_{m_1} = k_1, A_{m_2} = k_2, \cdots A_{m_P} = k_P$），则置信度得分可以表示为

$$P_{\mathrm{score}} = P(m_1^P \mid \boldsymbol{x}_1^{\mathrm{T}}) \tag{5}$$

可以利用前后向算法计算后验概率得分。采用 N 元拼音文法，只考虑前 $N-1$ 个音节对当前音节的影响。引入前向概率 $\boldsymbol{\Phi}(h_1^{N-1})$ 和后向概率 $\boldsymbol{\Psi}(h_1^{N-1})$ 两个记号（h_1^{N-1} 表示长度为 $N-1$ 的在图上顺序相连的任意合理拼音串的编号序列），并用 t_s^i，t_e^i 分别表示节点 i 的开始和终止时间。前向概率的定义式如下

$$\boldsymbol{\Phi}(h_1^{N-1}) \triangleq \sum_{\substack{w_1^n \\ w_{n-N+2}^n = h_1^{N-1}}} \prod_{i=1}^{n} p(\boldsymbol{x}_{t_s^{w_i}}^{t_e^{w_i}} \mid A_{w_i}) \cdot p(w_i \mid w_{\max\{1, i-N+1\}}^{i-1}) \tag{6}$$

其中，w_1^n 表示从整个拼音图的起点开始，最后 $N-1$ 个节点恰好是 h_1^{N-1} 的任意合理拼音串。后向概率的定义式如下

$$\boldsymbol{\Psi}(h_1^{N-1}) \triangleq \sum_{\substack{w_1^n \\ w_1^{N-1} = h_1^{N-1}}} \prod_{i=1}^{n} p(\boldsymbol{x}_{t_s^{w_i}}^{t_e^{w_i}} \mid A_{w_i}) \cdot \prod_{j=N}^{n} p(w_j \mid w_{j-N+1}^{j-1}) \tag{7}$$

其中，w_1^n 表示一直延续到整个拼音图的终点为止，前 $N-1$ 个节点恰好是 h_1^{N-1} 的任意合理拼音串。

可知有如下的递推式成立

$$\boldsymbol{\Phi}(h_2^N) = p(\boldsymbol{x}_{t_s^{h_N}}^{t_e^{h_N}} \mid A_{h_N}) \cdot \sum_{h_1} \boldsymbol{\Phi}(h_1^{N-1}) \cdot p(h_N \mid h_1^{N-1}) \tag{8}$$

$$\boldsymbol{\Psi}(h_1^{N-1}) = \sum_{h_N} \boldsymbol{\Psi}(h_2^N) \cdot p(\boldsymbol{x}_{t_s^{h_1}}^{t_e^{h_1}} \mid h_1) \cdot p(h_N \mid h_1^{N-1}) \tag{9}$$

引入两个记号后，后验概率得分可以表示为

$$P(m_1^P \mid \boldsymbol{x}_1^{\mathrm{T}}) = \frac{P(m_1^P, \boldsymbol{x}_1^{\mathrm{T}})}{P(\boldsymbol{x}_1^{\mathrm{T}})} \tag{10}$$

$$P(\boldsymbol{x}_1^{\mathrm{T}}) = \sum_{\substack{w_{N-1} \in \Omega^* \\ w_1^{N-1}}} \boldsymbol{\Phi}(w_1^{N-1}) \tag{11}$$

$$P(m_1^P, \boldsymbol{x}_1^T) = \sum_{w_1^{N-2}} \sum_{v_1^{N-2}} \Phi(w_1^{N-2} m_1) \cdot \Psi(m_p v_1^{N-2}) \cdot \prod_{j=2}^{p+N-2} p(m_j \mid m_{j-N+1}^{j-1}) \cdot \prod_{i=2}^{p-1} p(\boldsymbol{x}_{t_s^i}^{t_e^i} \mid A_{m_i}) \quad (12)$$

其中，$\Omega*$ 表示与终点相连的节点集合，w_1^{N-2} 和 v_1^{N-2} 分别表示拼音串 m_1^P 的合理前接和后续拼音串（长度为 $N-2$），$m_i = \begin{cases} w_{N-2+i}, & i \leqslant 0 \\ v_{i-P}, & i > P \end{cases}$。

1.2.3 Token Passing 的概念模型表示

如果从更一般的形式来理解关键词查找与置信度计算过程，这两个过程实际上可以用一个模型——Token Passing 概念模型来表示，示意图如图 4 所示。

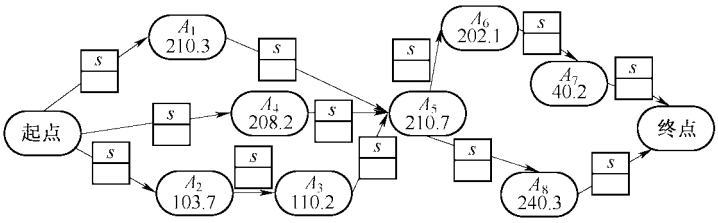

图 4 Token Passing 概念模型

图 4 中的 S 表示与 Token 相关的匹配得分，除此之外 Token 还需要保留匹配相关的信息（实际上传递的 Token 往往不止一个，为简便起见就不在图中标出了）。对于关键词查找而言，匹配得分非零即一（能够匹配或者不能够匹配），匹配信息为当前匹配的关键词位置。对于置信度计算的前向算法而言，匹配得分为前向概率，匹配信息为前 $N-1$ 状态序列（后向算法只需要简单地将图反向，过程是完全一致的）。每次 Token 传递只需要根据本地得分更新 Token 即可，可以根据这一模型构建统一的递推算法。

2 关键词检索的实验结果

对两阶段系统而言，第一阶段的目标是提供高声学覆盖率、高质量的拼音图以供检索，而第二阶段需要可靠的置信度依据以完成检索结果的筛选。两部分在整体目标一致的前提下相对独立，实验将分阶段考察每步过程的影响。

2.1 实验系统介绍

第一阶段的识别使用基于 DDBHMM 的 Triphone 识别器，输出概率分布形式采用混合高斯分布（32 个分量）。自适应的循环次数为 2，最终识别得到候选宽度为 20 的拼音图。

第二阶段检索过程中声学得分和拼音模型得分采取不同的加权，声学距离权重为 0.06，拼音文法模型权重为 0.85。

对比实验采用基于大词汇量连续语音识别(LVCSR, large vocabulary continuous speech recognition)的两阶段系统。LVCSR 系统首先将待检索的语料识别成文本，检索阶段在文本中查找关键词。文本结果只给出最优候选，因此不需要计算置信度，完成匹配的文本串即为检索输出。

数据选用广播新闻语料，共 8 个文件，包含多个说话人，语音长度共计约 6h，语音特征（14 维 MFCC、1 维归一化能量以及它们的一阶和二阶差分，共 45 维特征）的规模为 362MB。

2.2 实验结果分析讨论

2.2.1 识别性能

第一阶段考察声学识别率以及无监督自适应的效果。分别比较有调一候选、无调一候选和有调 20 候选的误识率，比较结果如表 1 所示。

表 1 声学识别率比较

	无自适应	一次自适应	两次自适应
一候选误识率(有调)	57.23%	49.47%	46.28%
一候选误识率(无调)	43.58%	37.08%	34.90%
20 候选误识率(有调)	17.87%	16.16%	15.99%

这里的一候选、20 候选指每个音节的候选宽度，即：平均到每个音节的候选数。

实验结果表明，无监督的 MLLR 自适应能够对声学层结果带来较大改善。尽管相应的处理时间较长（无自适应大约为 2 倍实时，一次自适应系统为 4 倍实时，两次自适应系统为 6 倍实时），但由于预处理只需要且可以后台执行，不占用检索时间，因此多次的自适应可以接受。在实用系统的搭建中，用户也可根据需求在性能和处理时间上取适当的折衷。

2.2.2 检索性能

(1) 拼音文法对检索性能的影响

为了提高检索的准确性，系统采用 N 元拼音文法模型。N 元文法模型假设当前词与前 $N-1$ 个词相关，N 越大，引入的历史信息越多，置信度得分越准确；N 越小，置信得分中声学匹配起到的作用也就越大。选取 500 个常用关键词，对广播新闻语料中的 8 个文件进行实验。系统的准确率及召回率曲线如图 5 所示。

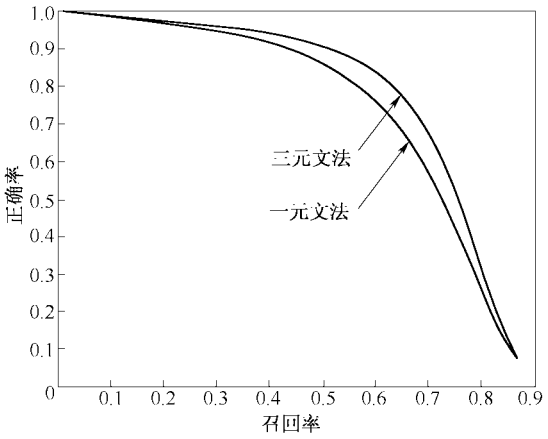

图 5　一元文法与三元文法的比较

可见三元文法的性能要比一元文法更好，在同样的召回率下其准确率更高。

(2) 拼音图系统与 LVCSR 系统的性能比较

分别选择二字词、三字词各 500 个，以置信度 0.1 作为门限进行筛选，低于门限的检出关键词被剔除。检索结果与基于 LVCSR 的两阶段系统进行比较，结果如表 2 所示（在表中两次实验的结果分别用“拼音图”和“LVCSR”表示）。

表 2　关键词检测系统性能（正确率＝正确数/检出总数，召回率＝正确数/关键词总数）

		检出总数	关键词总数	正确数	正确率	召回率
二字词	拼音图	12 844	12 932	9 335	72.68%	72.19%
	LVCSR	5 338	12 932	3 353	62.81%	25.93%
三字词	拼音图	1 187	1 340	985	82.98%	73.51%
	LVCSR	1 539	1 340	697	45.29%	52.01%

考察置信度门限和召回率、正确率之间的关系，可以得到如图 6、图 7 的性能曲线。

从实验结果可以看出，加入置信度判断之后，系统在基本不降低召回率的情况下大大提高了正确率，避免了过高的虚警率。与 LVCSR 系统相比，拼音图的覆盖率更高，检索结果也更稳健。检索阶段的实时率（这里的实时率指完成一次关键词检索所需要的时间）低于 0.01 倍实时，适合用户反复修改检索输入的情况。

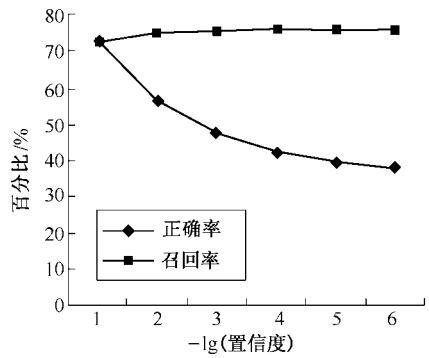

图 6　二字词正确率、召回率与门限的关系

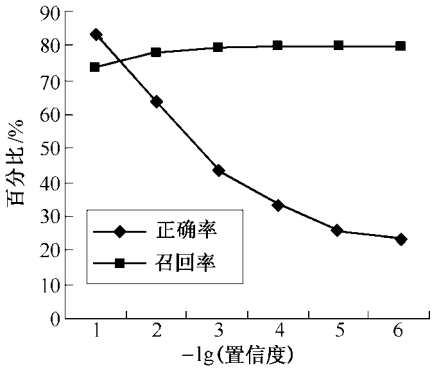

图 7　三字词正确率、召回率与门限的关系

3　结论

本文介绍了一个完整的基于拼音图的两阶段关键词检索系统。预处理阶段后台执行，不占用用户时间，尽可能地为检索提供高质量的元数据；检索阶段不但能够提供较高召回率和准确率，而且可以保证检索速度，满足快速检索的需要。事实上，这种两阶段的设计与互联网文本搜索引擎（例如 Yahoo，Google）的设计思想是一致的，此类音频检索系统可以为国家信息安全管理提供更好的帮助。

参考文献：

[1] 欧智坚, 罗骏, 谢达东, 赵贤宇, 林晖, 王作英. 多功能语音/音频信息检索系统的研究与实现[A]. 全国网络与信息安全技术研讨会[C]. 北京，2004.106-112.

OU Z J, LUO J, XIE D D, ZHAO X Y, LIN H,WANG Z Y. A multifunctional speech/audio information retrieval system[A]. Proc China Network & Information Security Technology Conference (NetSec'2004)[C]. Beijing,2004.106-112.

[2] WILPON J, RABINER L, LEE L, *et al*. Automatic recognition of

keywords in unconstrained speech using hidden markov models[J]. IEEE Trans on Acoustics, Speech and Signal Processing, 1990,38(11): 1870-1878.

[3] GAROFOLO J, AUZANNE C, VOORHEES E. The TREC spoken document retrieval track: a success story[A]. Proc of TREC-8[C].1999. 107-116.

[4] JOHNSON S E, JOURLIN P, MOORE G L, *et al*.The cambridge university spoken document retrieval system[A]. Proc of the IEEE International Conference on Acoustics, Speech, and Signal Processing [C].1999. 49-52.

[5] LIU Y, HARPER M P, JOHNSON M T, *et al*. The effect of pruning and compression on graphical representations of the output of a speech recognizer[J]. Computer Speech and Language, 2003, 7(4):329-356.

[6] CARDILLO P S, CLEMENTS M, MILLER M S. Phonetic searching vs LVCSR: how to find what you really want in audio archives[J].International Journal of Speech Technology, 2002, 5(1): 9-22.

[7] YOUNG S J, RUSSEL N H, THORNTON J H S. Token passing: a simple conceptual model for connected speech recognition systems[EB/OL].http:// svr-www.eng.cam.ac.uk,1989.

[8] LEGGETTER C J, WOODLAND P C. Maximum likelihood linear regression for speaker adaptation of continuous density HMM[J].Computer Speech and Language, 1995,9(1):171-186.

[9] ZHAO Q W, WANG Z Y, LU D J. A study of duration in continuous speech recognition based on DDBHMM[A]. Proc 6rd European Conf on Speech Communication and Technology (Eurospeech'99)[C]. Budapest, 1999.1511-1514.

[10] WESSEL F,SCHLÜTER R, MACHERTY K, *et al*. Confidence measures for large vocabulary continuous speech recognition[J].IEEE Trans on Speech and Audio Processing,2001,9(3):288-298.

作者简介：

罗骏（1978-），男，浙江台州人，清华大学博士，主要研究方向为语音识别。

欧智坚（1975-），男，湖南衡阳人，博士，清华大学讲师，主要研究方向为信号与信息处理。

（上接第 112 页）

作者简介：

李丽萍（1976-），女，河南济源人，中国科学院软件研究所博士生、助理研究员，主要研究方向为信息系统安全和网络安全。

贺也平（1962-），男，甘肃兰州人，中国科学院软件研究所研究员、博士生导师，主要研究方向为密码学、安全协议和信息安全。

卿斯汉（1939-），男，湖南邵阳人，中国科学院软件研究所研究员、博士生导师，主要研究方向为信息系统安全理论和技术以及可信计算技术。

沈晴霓（1970-），女，江西宜春人，讲师，中国科学院软件研究所博士生，主要研究方向为信息系统安全和网络安全。